USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Una medida de influencia en la clasificación a través de la teoría de juegos cooperativos

Roi Gómez Salvador

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Una medida de influencia en la clasificación a través de la teoría de juegos cooperativos

Roi Gómez Salvador

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística e Investigación Operativa.

Título: Una medida de influencia en la clasificación a través de la teoría de juegos cooperativos.

Breve descripción do contido

En este trabajo se pretende revisar la referencia [1]. En ella se parte de que si un conjunto de datos ha sido analizado por medio de un clasificador binario [2], se plantea la cuestión de cuáles fueron los factores más importantes para determinar el resultado de la clasificación. Para dar respuesta, en [1] se emplea un enfoque axiomático para caracterizar de forma única una medida de influencia: una función que, dado un conjunto de individuos clasificados, genera un valor para cada característica correspondiente a su influencia en la determinación del resultado de la clasificación. Además se estudia la relación de la medida de influencia con el valor de Banzhaf de la teoría de juegos cooperativos clásica [3]. Para finalizar el trabajo, se ilustrará la medida de influencia a través de ejemplos y conjuntos de datos.

Bibliografía

[1] Datta, A., Datta, A., Procaccia, A. D., and Zick, Y. (2015), "Influence in classification via cooperative game theory." Twenty-Fourth International Joint Conference on Artificial Intelligence, 511–517.
[2] Hastie, T., Tibshirani, R., and Friedman, J. (2008), The Elements of Statistical Learning. Springer Series in Statistics.
[3] Sánchez Rodríguez, E. and Vidal Puga, J. (2014), Juegos Coalicionales. Universidade de Vigo. Servizo de Publicacións, ed.

Índice general

Resumen

Partiendo de un conjunto de datos que han sido clasificados en dos clases, queremos saber cuáles han sido los factores más importantes para determinar la clase que les ha sido asignada. Para ello se usará una aproximación axiomática con la finalidad de definir una medida de influencia única. Esta medida será una función que, a partir de un conjunto de datos clasificados, devolverá un valor para cada característica correspondiente a la influencia que ha tenido a la hora de determinar la clasificación. Primero se comenzará con una breve introducción a los juegos cooperativos con utilidad transferible, ya que axiomatizar una medida de influencia y una solución de teoría de juegos tienen características comunes, y varios modelos de clasificación. Después de definir la medida de influencia, se muestra que tiene una forma intuitiva cuando el clasificador es lineal. Y por último, aplicamos la medida en un conjunto de datos relacionados con pacientes afectados por el COVID-19 dependiendo de si han tenido algún incidente o no.

Abstract

Starting from a collection of data that have been classified into two classes, we want to know which factors have been the most important in determining the classification outcome. For this purpose, an axiomatic approach will be used in order to define a unique measure of influence. This measure will be a function that, given a set of classified data, outputs a value for each feature corresponding to its influence in determining the classification outcome. We will first begin with a brief introduction to cooperative games with transferable utility, since axiomatizing an influence measure and a game theory solution have common aspects, and several classification models. After defining the influence measure, we show that it has an intuitive form when the classifier is linear. Finally, we apply the measure on a dataset related to patients affected by COVID-19 depending on whether they have had any incident or not.

Introducción

En este último año y medio hemos pasado una pandemia mundial a causa del coronavirus, con un gran número de contagios y, en consecuencia, de fallecimientos. No hay una definición clara y matemática de ola pandémica, pero se puede diferenciar fácilmente la primera ola. Centrándonos en esta primera ola, y a partir de los datos del estudio de Davila-Pena et al. (2021) que analiza los pacientes de COVID-19 de Galicia entre marzo y abril de 2020, queremos ver cuáles han sido las características más influyentes a la hora de que estos individuos hayan tenido alguna incidencia. Se entiende que ha tenido una incidencia si el paciente ha estado hospitalizado, si ha necesitado UCI o si ha fallecido. Para ello usaremos y analizaremos la medida definida en Datta et al. (2015), la cual permite ver la influencia de cada característica.

Esta medida tiene propiedades que se pueden relacionar con la Teoría de Juegos Cooperativos, por lo que, en el primer capítulo comenzamos con un breve resumen, en concreto sobre los juegos cooperativos con utilidad transferible (véase Sánchez Rodríguez and Vidal Puga, J., 2014) centrándonos sobre todo en definir ciertas propiedades que caracterizan a dos valores importantes: el valor de Shapley (1953) y el valor de Banzhaf (1964).

Para aplicar la medida de influencia partimos de un conjunto de datos clasificados, por lo que, aunque nuestro objetivo no es la clasificación, para enmarcar nuestro problema estudiaremos los problemas de clasificación (véase Hastie et al., 2008). Partimos del aprendizaje estadístico para hacer una aproximación de la predicción utilizando el modelo lineal y el del vecino más cercano. A continuación, se mencionan diferentes métodos para clasificar datos como el análisis cluster y los árboles de clasificación, un método más visual en el cual hemos añadido un claro ejemplo.

Ya en la parte principal de nuestro trabajo, suponemos que tenemos un conjunto de datos de individuos, B, donde cada individuo $\mathbf{a} \in B$ puede ser pensado como un vector de n características, donde la i-ésima coordenada de $\mathbf{a}$ corresponde con el estado de la

i-ésima característica. Cada $\mathbf{a}$ tiene un valor $v(\mathbf{a})$, llamado puntuación de $\mathbf{a}$. A partir del conjunto de datos B, se calcula una medida $\phi_i(N, B, v)$ que cuantifica la importancia de la característica i-ésima en la determinación de la clase (o etiqueta) de los puntos en B. Esta medida de influencia se caracteriza axiomáticamente y se discuten brevemente posibles extensiones.

Finalmente, aplicaremos esta medida en el conjunto de datos relacionados con pacientes afectados por el COVID-19 mencionados en el primer párrafo y las conclusiones sobre la medida de influencia.

Capítulo 1

Teoría de juegos cooperativos

La teoría de juegos es la teoría matemática que se ocupa de resolver los problemas de decisión interactivos donde intervienen varios agentes, llamados jugadores, que toman distintas decisiones, llamadas estrategias. La teoría de juegos puede dividirse en dos partes: los juegos estratégicos o no cooperativos y los juegos coalicionales o cooperativos.

En este capítulo nos centraremos en los problemas de negociación coalicional. En uno de tales problemas se consideran los resultados posibles de una hipotética cooperación, por lo tanto, se deben especificar también los beneficios de la cooperación, la de todos los jugadores y la de cada posible coalición de ellos.

En estos problemas no es necesaria la unanimidad para que un reparto sea aprobado, sino que puede haber algunos grupos de jugadores capaces de forzar determinados repartos. También supondremos que los beneficios generados por cada coalición pueden repartirse libremente entre ellos. Por eso, trataremos con problemas de negociación coalicional con utilidad transferible.

Para caracterizar estos problemas usaremos los llamados juegos cooperativos con utilidad transferible (juegos TU) que definimos formalmente a continuación.

Definición 1.1. Un juego TU es un par (N,v) donde N es el conjunto finito de jugadores y $v\colon 2^N \to \mathbb{R}$ es una función que cumple $v(\emptyset) = 0$.

El conjunto 2^N está formado por todos los subconjuntos de N y la función v se denomina función característica del juego. Dada una coalición $S \subset N$, $v(S)$ representa los beneficios que obtendrían los jugadores de S, independientemente de cómo actúe el resto de los jugadores. Denotamos por G^N al conjunto de todos los juegos TU con conjunto de

jugadores N.

En los juegos coalicionales se pueden llevar a cabo distintas operaciones. Dados $v_1, v_2 \in G^N$, $S \subset N$ y $\lambda \in \mathbb{R}$, se define una operación interna en el conjunto G^N como la suma de juegos:

$$(v_1 + v_2)(S) = v_1(S) + v_2(S),$$

y una operación externa como el producto de un juego por un escalar:

$$(\lambda v_1)(S) = \lambda v_1(S).$$

Definición 1.2. Sea $(N,\, v) \in G^N$. Diremos que $(N,\, v)$ es:

- Superaditivo si $v(S \cup T) \geq v(S) + v(T)$ para todo $S,\, T \subset N,\, S \cap T = \emptyset$.
- Subaditivo si $v(S \cup T) \leq v(S) + v(T)$ para todo $S,\, T \subset N,\, S \cap T = \emptyset$.
- Aditivo si $v(S \cup T) = v(S) + v(T)$ para todo $S,\, T \subset N,\, S \cap T = \emptyset$.

Podemos observar que en los juegos superaditivos se priorizará la cooperación, ya que la unión de dos grupos disjuntos cualesquiera no provoca una disminución de los beneficios. Por lo tanto, el objetivo implícito de estos juegos es que se forme la gran coalición N y repartir entre ellos el beneficio.

Definición 1.3. Un reparto es un vector $x = (x_1, ..., x_n) \in \mathbb{R}^N$ donde cada coordenada x_i representa la cantidad asignada al jugador i. Dado un reparto x, la suma de las cantidades asignadas a los miembros de una coalición $S \subset N$ se denota por $x(S) = \sum_{i \in S} x_i$.

En teoría de juegos este reparto se rige por dos requisitos comunes:

- Racionalidad individual. Dado $v \in G^N$, un reparto $x = (x_1, ..., x_n) \in \mathbb{R}^N$ verifica la propiedad de racionalidad individual si cada jugador recibe un pago que no es inferior a lo que puede garantizarse por sí mismo, es decir, $x_i \geq v(\{i\})$ para todo $i \in N$.
- Eficiencia. Dado $v \in G^N$, un reparto $x = (x_1, ..., x_n) \in \mathbb{R}^N$ será eficiente si distribuye el valor de la gran coalición $v(N)$ entre los jugadores, esto es, si $x(N) = x_1 + \cdots + x_n = v(N)$.

El primer requisito establece que cada jugador debe obtener, al menos, lo que puede garantizarse por sí mismo. En cuanto al segundo, es razonable que si los jugadores van a

formar la coalición N, entonces se reparta el valor total disponible.

El concepto de solución de un juego se puede estudiar desde dos puntos de vista: indicando qué repartos son mejores que otros siguiendo un cierto criterio, solución de tipo conjunto, o dejando que un árbitro recomiende a los jugadores un reparto del valor, solución puntual. En primer lugar vamos a formalizar de forma matemática el concepto de solución de un juego coalicional, que recoge ambos casos.

Definición 1.4. Una solución definida en un dominio $\Omega \subset G^N$, es una correspondencia $\Psi \colon \Omega \to \mathbb{R}^N$ que asocia a cada juego (N, v) en Ω un subconjunto $\Psi(N, v) \subset \mathbb{R}^N$.

En el caso de que la solución sea siempre unitaria hablamos de solución puntual, o también llamada valor. En otro caso, se trata de una solución tipo conjunto. Nosotros nos centraremos en las soluciones puntuales, basadas en la idea de ecuanimidad, es decir, tratan de proponer para cada juego TU un reparto ecuánime que sea aceptable para todos los jugadores. Estas soluciones puntuales o valores son funciones de la forma $f \colon G^N \to \mathbb{R}^N$. En concreto, estudiaremos el valor de Shapley (1953) y el valor de Banzhaf (1964).

Para ello, comenzaremos definiendo algunos conceptos y estableciendo ciertas propiedades que caracterizan el valor de Shapley.

Definición 1.5. Un juego $v \in G^N$ se denomina simple si cumple que para toda $S \subset N$, $v(S) = 0$ o $v(S) = 1$, $v(N) = 1$ y que v es un juego monótono, esto es, $v(S) \leq v(T)$ para cualesquiera S, $T \subset N$ con $S \subset T$.

Se denotará por S^N el conjunto de los juegos simples con conjunto de jugadores N. Para un juego simple (N, v) y una coalición $S \subset N$, se dice que S es ganadora si $v(S) = 1$; en caso contrario, se dice que S es perdedora.

Definición 1.6. Sea $v \in G^N$ un juego TU:

1. Decimos que $i \in N$ es un jugador nulo de v si, para cada $S \subset N$, $v(S \cup \{i\}) = v(S)$.
2. Dos jugadores $i, j \in N$ se denominan simétricos en v si, para cada coalición $S \subset N \setminus \{i, j\}$, se tiene que $v(S \cup \{i\}) = v(S \cup \{j\})$.

Propiedades. Sea f un valor y considérense las siguientes propiedades:

- Un valor f satisface la eficiencia si, para todo $v \in G^N$,

$$\sum_{i \in N} f_i(v) = v(N).$$

- Un valor f satisface la propiedad de jugador nulo si, para todo $v \in G^N$ y para todo $i \in N$ jugador nulo de v, se tiene que $f_i(v) = 0$.
- Un valor f satisface la simetría si, para todo $v \in G^N$ y para todo par de jugadores $i, j \in N$ simétricos en v, se tiene que $f_i(v) = f_j(v)$.
- Un valor f satisface aditividad si, para todo par de juegos $v, w \in G^N$, $f(v + w) = f(v) + f(w)$.
- Un valor f satisface poder total si, para todo $v \in G^N$ se tiene que:

$$\sum_{i=1}^{n} f_i(v) = \frac{1}{2^{n-1}} \sum_{i=1}^{n} \sum_{S \subset N \setminus \{i\}} |v(S \cup i) - v(S)|.$$

Es fácil ver que la eficiencia indica que f debe repartir $v(N)$ entre todos los jugadores. La propiedad de jugador nulo quiere decir que los jugadores que no generan beneficios no deben recibir nada. La simetría significa que debemos tratar por igual a jugadores que aportan lo mismo. La aditividad es básicamente un requerimiento técnico que, aunque no está conectado con la idea de ecuanimidad, tampoco parece ir en contra de tal idea. La propiedad de poder total establece que el pago total obtenido por los jugadores es la suma de las medias de las contribuciones marginales de todos los jugadores. Por lo tanto, si un valor es eficiente no puede satisfacer la propiedad de poder total.

A continuación, definimos el valor de Shapley y enunciamos un teorema que prueba que las cuatro primeras propiedades caracterizan el valor de Shapley.

Definición 1.7. El valor de Shapley está dado por:

$$\phi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} (v(S \cup \{i\}) - v(S)),$$

para todo $v \in G^N$ y todo $i \in N$, donde s y n denotan los cardinales de S y N, respectivamente.

Teorema 1.8. *Existe un único valor en G^N que verifica las propiedades de eficiencia, jugador nulo, simetría y aditividad. Este valor es el valor de Shapley.*

Definición 1.9. El valor de Banzhaf de un jugador $i \in N$ está dado por:

$$\beta_i(v) = \frac{1}{2^{n-1}} \sum_{S \subset N \setminus \{i\}} (v(S \cup \{i\}) - v(S)),$$

para todo $v \in G^N$ y todo $i \in N$.

Teorema 1.10. *Existe un único valor en G^N que verifica las propiedades de jugador nulo, simetría, aditividad y poder total. Este valor es el valor de Banzhaf.*

Este valor tiene una interpretación simple en términos probabilísticos. Supongamos que se elige al azar una coalición S y al unirse el jugador i recibe su contribución marginal. Si todas las coaliciones tienen la misma probabilidad de ser elegidas, el valor de Banzhaf del jugador i es su pago esperado.

En el caso del valor de Shapley, se supone que se elige un orden o permutación al azar de los jugadores y que cada jugador recibe su contribución a la coalición de jugadores que se forma en el orden dado antes de su llegada. Si todos los órdenes tienen la misma probabilidad, es fácil ver que el valor de Shapley de cada jugador es su pago esperado de acuerdo con el mecanismo anteriormente descrito.

Nótese que sólo una propiedad diferencia estas dos caracterizaciones: el valor de Shapley verifica eficiencia mientras que el valor de Banzhaf verifica poder total.

Capítulo 2

El problema de clasificación

El aprendizaje estadístico es una disciplina que tiene aplicaciones en muchas áreas de la ciencia, las finanzas o la industria. Incluye un gran número de métodos y técnicas pertenecientes a diversos campos como la estadística, la gestión de datos y la inteligencia artificial.

2.1. Dos aproximaciones sencillas a la predicción

Cuando se trata de aprendizaje estadístico, normalmente se tiene una medida de salida que será nuestra variable respuesta, casi siempre cualitativa o categórica, que se pretende predecir a partir de un conjunto de características, las variables independientes. Partimos de un conjunto de datos de orientación, donde observamos las salidas y las características para un conjunto de objetos (o personas). A partir de estos datos construimos un modelo de predicción que nos permitirá predecir la salida de un nuevo objeto.

Denotamos por X la variable de entrada. Cuando X sea un vector escribiremos cada componente como X_i. A la variable de salida la denotaremos por Y, y cuando hagamos una predicción pasaremos a llamarla $\hat{Y}$.

2.1.1. Modelos lineales

En el método lineal simple, tenemos una variable salida Y y una colección de variables entrada $X_1, X_2, ..., X_p$. Predecimos la salida Y como $\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$, donde el término $\hat{\beta}_0$ es el intercepto. Suele ser común incluirlo en $\hat{\beta}$ añadiendo la constante 1 en X, de modo que la predicción resulte en $\hat{Y} = X^T \hat{\beta}$. Así, podemos tomar una función en un espacio de entradas p-dimensional de la forma $f(X) = X^T \hat{\beta}$, la cual es lineal y su

gradiente $f'(X) = \hat{\beta}$ es un vector en el espacio de entradas que señala la dirección con la inclinación más alta.

Por lo tanto, si tenemos N individuos, obtenemos el modelo de regresión:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i,$$

donde Y_i es la variable salida del i-ésimo individuo, $x_{i,1}, ..., x_{i,p}$ las variables entradas del mismo y ϵ_i el error asociado a dicho individuo. En forma matricial se puede expresar así:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

Ahora anticipamos una observación sobre el problema de clasificación. Nuestro espacio de datos es, en ese caso, divisible en regiones de acuerdo con las características de la clasificación. Estas regiones pueden ser más o menos suaves, dependiendo del enfoque usado para el método de predicción, por ejemplo mínimos cuadrados o el vecino más cercano.

Por un lado, la aproximación de mínimos cuadrados toma los coeficientes β para minimizar el residuo de sumas al cuadrado:

$$RSS(\beta) = \sum_{i=1}^{N} (Y_i - X_i^T \beta)^2.$$

$RSS(\beta)$ es una función cuadrática de los parámetros y, por lo tanto, el mínimo siempre existe, pero puede no ser único. Para resolverlo, pasaremos a notación matricial:

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta),$$

donde $\mathbf{X}$ es una matriz $N \times (p+1)$ siendo cada fila un vector de entrada, e $\mathbf{Y}$ un N-vector de salidas del conjunto de partida. Se llega así a la ecuación normal:

$$\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\beta) = 0.$$

Si $\mathbf{X}^T\mathbf{X}$ es no singular, la solución única viene dada por:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

y el valor ajustado a la entrada i-ésima X_i es $\hat{Y}_i = \hat{Y}(X_i) = X_i^T\hat{\beta}$.

Es decir, una vez que sabemos los estimadores de los parámetros, $\hat{\beta}$, podemos calcular los ajustes o predicciones para la muestra de unos individuos X_i con $i \in \{1, ..., N\}$, de la forma $\hat{Y}(X_i) = X_i^T\hat{\beta}$ con $i \in \{1, ..., n\}$, o bien, en notación matricial, $\hat{\mathbf{Y}}(X) = X^T\hat{\beta}$.

2.1.2. Modelos de vecinos más cercanos

En cambio, la aproximación del vecino más cercano usa las observaciones del conjunto de entrenamiento, τ, más cercanas a X en el espacio de entradas para conseguir $\hat{Y}$. En concreto, el vecino k-cercano ajustado por $\hat{Y}$ es definido por:

$$\hat{Y}(X) = \frac{1}{k} \sum_{X_i \in N_k(X)} Y_i,$$

donde $N_k(X)$ denota la vecindad de X definida por los k puntos más cercanos a X en el conjunto de partida. La cercanía implica una métrica, la cual asumiremos como la distancia euclidiana por el momento. Por lo tanto, encontramos las k observaciones con los X_i más cercanos a X en el espacio de entrada, y el promedio de sus respuestas.

2.2. Modelos lineales para clasificación

Volviendo al problema de clasificación usando regresión, si tenemos K clases, el modelo de regresión ajustado para la k-ésima clase sería: $\hat{f}_k(X) = \hat{\beta}_{k0} + \hat{\beta}_k^T X$. La decisión de frontera entre dos clases k y l es un conjunto de puntos que cumplen $\hat{f}_k(X) = \hat{f}_l(X)$, es decir, el conjunto $\left\{X : (\hat{\beta}_{k0} - \hat{\beta}_{l0}) + (\hat{\beta}_k - \hat{\beta}_l)^T X = 0\right\}$, que puede ser un conjunto afín o un hiperplano. Este tipo de regresión modela funciones discriminantes $\delta_k(X)$ para cada clase, y después clasifica a X en la clase con mayor valor. Por lo tanto, si $\delta_k(X)$ es lineal en X, los límites de decisión serán lineales.

Consideramos la regresión lineal de una matriz de indicadores. Existe un clasificador G que toma valores en el conjunto $\{1, ..., K\}$ de grupos y las respuestas se codifican mediante variables indicadoras $Y_1, ..., Y_K$ donde $Y_k = 1$ en el grupo k y 0 en otro caso. De esta forma tenemos el vector $Y = (Y_1, ..., Y_K)$ y las N entradas diferentes. Así, tenemos una matriz $\mathbf{Y}$ de tamaño $N \times K$ que solamente contiene 0's y 1's, de forma que cada fila tiene un único 1. Ajustamos un modelo de regresión lineal a cada columna de $\mathbf{Y}$ simultáneamente y el ajuste está dado por:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

donde $\mathbf{X}$ es la matriz de diseño, con $p + 1$ columnas correspondientes a las p entradas y una columna de 1's para el intercepto. Cabe destacar que, para cada columna de vectores respuestas $\mathbf{Y}_k$, tenemos un vector de coeficientes. Por lo tanto, obtenemos una matriz $(p + 1) \times K$ de coeficientes:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Ahora, para obtener la clasificación de una nueva observación x primero calculamos el valor ajustado de salida $\hat{f}(x)^T = (1, x^T)\hat{\beta}$, un vector de tamaño K. Después identificamos la componente más grande y clasificamos mediante: $\hat{G}(x) = \text{argmax}_{k \in G} \hat{f}_k(x)$. Se obtiene de esta forma el grupo al que pertenece.

2.3. Análisis clúster

Por otra parte, en el análisis de clústeres también tenemos la finalidad de agrupar las observaciones o variables en grupos con características semejantes. Aquí, nos centraremos en el problema de partición de datos, donde queremos dividir los datos en un número de grupos prefijados. Supongamos una muestra de n elementos con p variables y los queremos dividir en K grupos. Para ello, disponemos del algoritmo de k-medias descrito a continuación.

Partimos de K puntos como centros de los grupos iniciales que pueden ser escogidos aleatoriamente o arbitrariamente. En segundo lugar, calculamos las distancias euclidianas de cada elemento al centro de los K grupos, y asignamos cada elemento al grupo más próximo. Esta asignación se realiza secuencialmente: con cada nuevo punto incluido en un grupo se vuelven a calcular las coordenadas de la nueva media de dicho grupo. Comprobamos mediante un criterio si es posible mejorar la asignación de los elemento y en caso contrario tendríamos terminado el método.

El criterio mencionado en el párrafo anterior es la *suma de cuadrados dentro de los grupos*, SSIG en adelante, para todas las variables, es decir, la suma ponderada de las varianzas de las variables en los grupos:

$$\text{SSIG} = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2,$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k, y $\bar{x}_{jk}$ la media de la variable j en el grupo k. Por lo tanto, el criterio viene dado por:

$$\text{mín SSIG} = \text{mín} \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2,$$

donde n_k es el número de elementos del grupo k y s_{jk}^2 es la varianza de la variable j en dicho grupo. De esta forma, al minimizar las varianzas de todas las variables en los grupos,

obtenemos grupos con las características más parecidas.

Por último, existen varios métodos para calcular el número de grupos óptimo en el algoritmo de k-medias. Uno de los más usados consiste en realizar un test F aproximado de reducción de variabilidad, comparando la SSIG con K grupos con la de $K+1$, y calculando la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional. El test viene dado por:

$$F = \frac{\text{SSIG}(K) - \text{SSIG}(K+1)}{\text{SSIG}(K+1)/(n-K-1)}.$$

Este valor F se compara con una distribución F con p y $p(n-K-1)$ grados de libertad, a pesar de que no siempre se puede aplicar, ya que los datos pueden no verificar las hipótesis de la distribución F.

2.4. Validación cruzada

La validación cruzada es una técnica para evaluar modelos de *machine learning* mediante el entrenamiento de varios modelos en subconjuntos de los datos de entrada disponibles y evaluarlos con el subconjunto complementario de los datos. Es decir, en la validación cruzada de k iteraciones o *k-fold* se dividen los datos en k subconjuntos (*folds*). Uno de los subconjuntos se utiliza como datos de prueba y el resto, $k-1$, como datos de entrenamiento.

El proceso de validación cruzada se repite durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. El error se calcula como la media aritmética de los errores de cada iteración para obtener un único resultado. Si denotamos a MSE_i por el error en la iteración i-ésima, entonces el error de la validación cruzada se estima por:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i.$$

2.5. Métodos basados en árboles

Los árboles de decisión tienen como base la segmentación del espacio de predicción en un número de regiones simples. Los árboles de decisión pueden ser aplicados a los problemas de regresión y a los problemas de clasificación. Se les llama árboles de regresión, si la variable salida es cuantitativa, y árboles de clasificación si la variable es cualitativa.

2.5.1. Árboles de regresión

Partimos de un conjunto de N observaciones, cada una con p variables de entrada y una variable respuesta. Esto es, (x_i, y_i) para cada $i = 1, 2, ..., N$, con $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$. Para ambos modelos tenemos una partición del espacio de predicción, donde cada región se representa por R_m con $m \in \{1, ..., M\}$, y creamos un modelo para la variable respuesta con una constante c_m para cada región:

$$f(x) = \sum_{m=1}^{M} c_m \mathcal{I}(x \in R_m).$$

Si se escoge como criterio el mínimo de la suma de cuadrados $\sum(y_i - f(x_i))^2$, es fácil ver que el mejor $\hat{c}_m$ es el promedio de los y_i en la región R_m:

$$\hat{c}_m = \text{prom}(y_i : x_i \in R_m).$$

Buscar la mejor partición binaria en términos del mínimo de suma de cuadrados, en general, es muy costoso computacionalmente. Por lo tanto, se usa el siguiente algoritmo. Comenzando con todo el conjunto de datos, se considera una variable de separación j y un punto de separación s, y definimos los siguientes semiplanos:

$$R_1(j, s) = \{x : x_j \leq s\} \text{ y } R_2(j, s) = \{x : x_j > s\}.$$

Después se busca la variable j y el punto s de forma que resuelvan:

$$\arg\min_{j,\, s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right].$$

Ahora, para un j y un s dados, el mínimo interno se resuelve de la forma:

$$\hat{c}_1 = \text{prom}(y_i : x_i \in R_1(j, s)) \text{ y } \hat{c}_2 = \text{prom}(y_i : x_i \in R_2(j, s)).$$

Para cada variable de división, la determinación del punto de división s puede hacerse muy rápidamente y, por tanto, recorriendo todas las entradas la determinación del mejor par $(j,\, s)$ es viable. Una vez encontrada la mejor división, separamos los datos en las dos regiones resultantes y repetimos el proceso de separación en cada una de las dos regiones. Luego, este proceso se repite en todas las regiones resultantes hasta tener un árbol con las dimensiones deseadas. Un árbol demasiado grande podría sobreajustar los datos, mientras que un árbol pequeño podría no capturar una estructura importante.

Una regla común a seguir es crear un primer árbol T_0 grande, parando cuando los nodos tengan un tamaño pequeño. Después este árbol se poda mediante el algoritmo que

precisaremos a continuación. Definimos un subárbol $T \subset T_0$ como cualquier árbol que se puede obtener podando el árbol T_0, esto es, colapsando cualquier número de sus nodos internos (no terminales). Se usará m como índice de cada nodo terminal, representando cada región R_m. Sea $|T|$ el número de nodos terminales en T y N_m el número de observaciones en R_m. Tomando:

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i,$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2, \tag{2.1}$$

definimos el criterio de complejidad de los costes para un $\alpha \in \mathbb{R}^+$:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|.$$

La idea es encontrar, para cada α, un subárbol $T_\alpha \subseteq T_0$ que minimice $C_\alpha(T)$. La afinación del parámetro $\alpha \geq 0$ actúa sobre la compensación entre el tamaño del árbol y su buen ajuste a los datos. Cuanto mayor valor tenga α más pequeño será el árbol T_α y, al revés, cuanto menor sea α más grande será T_α. Es fácil observar que para $\alpha = 0$ la solución es el árbol T_0.

Para escoger T_α se usa la *poda del eslabón más débil*: sucesivamente se colapsan los nodos internos que producen el menor incremento por nodo en $\sum_m N_m Q_m(T)$, y continuamos hasta que producimos un único nodo. Esto da una secuencia finita de subárboles, y se puede ver que esta secuencia debe contener a T_α. Para más detalles ver Breiman et al. (1984) o Ripley (1996). La estimación de α es lograda con la validación cruzada 5- o 10-*fold*: escogemos el valor $\hat{\alpha}$ para minimizar la validación cruzada de la suma de cuadrados. El árbol final es $T_{\hat{\alpha}}$.

2.5.2. Árboles de clasificación

Centrándonos en los árboles de clasificación donde nuestra variable respuesta toma valores de la forma $1, 2, ..., K$, el único cambio necesario en el algoritmo del árbol es sobre el criterio de división de los nodos y de poda del árbol. Para regresión se usó la medida de impureza de nodo del error cuadrático $Q_m(T)$ definido en (2.1), pero no es idóneo para clasificación. Ahora, en un nodo m, definimos:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} \mathcal{I}(y_i = k),$$

como la proporción de observaciones de la clase k en el nodo m, donde N_m son las observaciones de la región R_m. Se clasifican las observaciones del nodo m en la clase $k(m) = \arg\max_k \hat{p}_{mk}$, la clase mayoritaria en el nodo m. Para calcular la impureza del nodo utilizamos:

- Error de clasificación: $\frac{1}{N_m} \sum_{i \in R_m} \mathcal{I}(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$.
- Índice de Gini: $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$.
- Entropía cruzada o desviación: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

Para dos clases, si tomamos p como la proporción de la segunda clase tenemos para las tres medidas anteriores los valores $1 - \max(p, 1-p)$, $2p(1-p)$ y $-p\log p - (1-p)\log(1-p)$, respectivamente. Cualquiera de las tres medidas son utilizadas para la depuración del árbol, pero el error de clasificación es el preferible si el objetivo es la precisión de la predicción del árbol podado final. Llamaremos $Q_m(T)$ al criterio escogido y continuaremos como en los árboles de regresión.

2.5.3. Ejemplo *spam*

El conjunto de datos *Spambase* (véase Dua and Graff, G., 2017) consiste en un total de 4601 mensajes de correo electrónico. La variable respuesta es binaria, toma los valores de `email` o `spam`, y hay 57 predictores: 47 predictores cuantitativos, el porcentaje de palabras en el correo que coinciden con unas palabras dadas (`business`, `address`, `internet`, `free`, ...); 6 predictores cuantitativos, el porcentaje de caracteres en el correo que coinciden con unos dados (`ch;`, `ch(`, `ch[`, `ch!`, `ch$` y `ch#`); `CAPAVE`, la longitud media de la secuencia ininterrumpida de letras mayúsculas; `CAPMAX`, la longitud de la secuencia ininterrumpida de letras mayúsculas más larga; y `CAPTOT`, la suma de la longitud de las secuencias ininterrumpidas de letras mayúsculas.

En el código tomaremos `spam` como 1 y `email` como 0. Se selecciona un conjunto de testeo de 1536 mensajes escogidos al azar, dejando el resto de 3065 observaciones en el conjunto de prueba.

Se aplica el método del árbol de clasificación usando la medida de la desviación para hacer crecer el árbol (Figura 2.1) y el criterio del error de clasificación para podarlo. La Figura 2.2 muestra el error de la validación cruzada frente al tamaño del árbol podado.

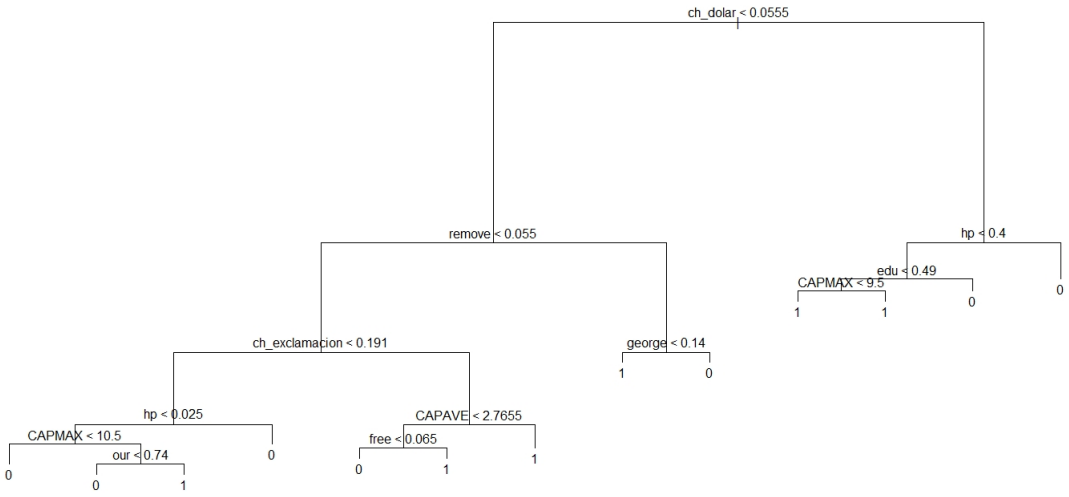

Figura 2.1: Árbol sin podar.

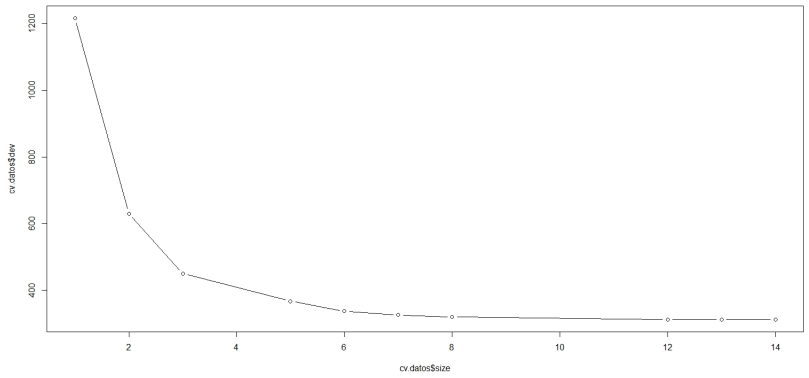

Figura 2.2: Resultados para el ejemplo de `spam`. Número de nodos terminales frente al error de la validación cruzada.

Obtenemos que el error se aplana alrededor de los 12 nodos terminales, dando lugar así al árbol podado de la Figura 2.3.

Si consideramos el primer nodo, representa que se toma la rama derecha si el porcentaje de los caracteres que son $ supera el 5,55 %. De la misma forma, si es frecuente la palabra `hp`, se puede suponer que el correo es del trabajo y por lo tanto se clasifica como `email`.

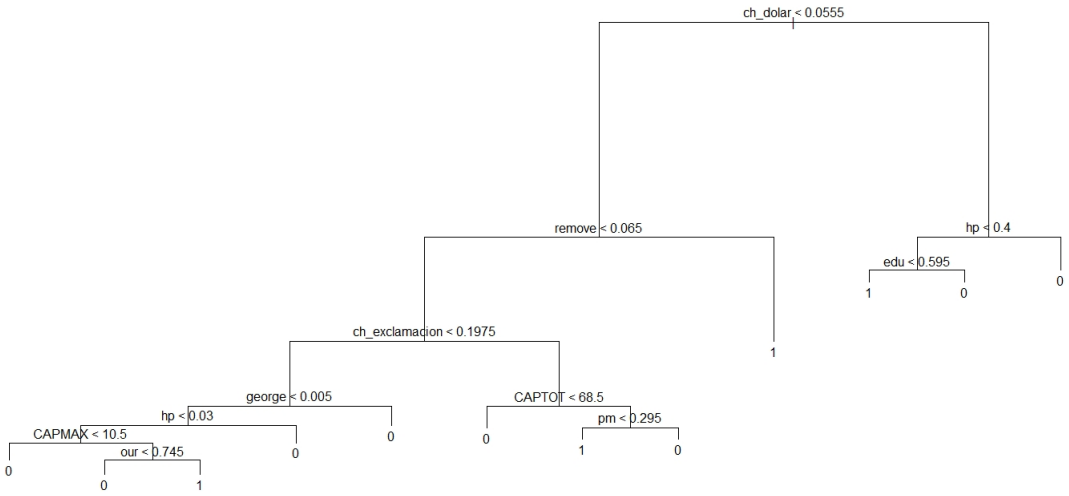

Figura 2.3: Árbol podado.

Capítulo 3

Influencia en la clasificación

Partiendo de un conjunto de datos clasificados en dos grupos distintos queremos averiguar cuáles fueron las características más importantes para determinar el resultado de la clasificación. Para ello, caracterizaremos de forma única una medida de influencia, una función que, dado un conjunto de puntos clasificados, genera un valor para cada característica correspondiente a su influencia en la determinación del resultado de la clasificación. En caso de que el clasificador sea lineal, la medida de influencia adquiere una forma geométrica intuitiva.

Suponiendo que tenemos un conjunto de datos de individuos, B, donde cada individuo $\mathbf{a} \in B$ puede ser pensado como un vector de características (por ejemplo, $\mathbf{a}$ = (edad, género, ...)), donde la i-ésima coordenada de $\mathbf{a}$ corresponde con el valor de la i-ésima característica. Cada $\mathbf{a}$ tiene un valor $v(\mathbf{a})$, llamado puntuación de $\mathbf{a}$. Así, nos centraremos en ver, a partir de un conjunto de datos B de varios vectores de características y sus valores, la influencia que ha tenido cada característica al determinar ese valor.

Formalmente, dado un conjunto $N = \{1, ..., n\}$ de características o atributos y un conjunto de datos B de características de los individuos, donde cada perfil $\mathbf{a}$ tiene un valor $v(\mathbf{a})$, nos gustaría calcular una medida $\phi_i(N, B, v)$ que corresponde con la importancia de la característica i-ésima para determinar la etiqueta de los puntos en B.

Seguiremos un enfoque axiomático, el cual presenta similitudes con la teoría de juegos cooperativos, para definir una medida de influencia. Veremos que esta medida de influencia es la única medida que satisface algunas propiedades naturales.

3.1. Caracterización axiomática

Dado un conjunto de características $N = \{1, ..., n\}$, sea A_i el conjunto de posibles valores o estados que la característica i puede tomar; por ejemplo, la i-ésima característica puede ser el género, en ese caso $A_i = \{\text{hombre, mujer, otro}\}$. Se nos dan las salidas parciales de una función sobre un conjunto de datos que contiene perfiles de características. Es decir, tenemos un subconjunto B de $A = \prod_{i \in N} A_i$, y los valores, $v(\mathbf{a})$, para cada $\mathbf{a} \in B$. Por lo tanto, nos referimos a que no sabemos la verdadera estructura de v, pero sabemos cuáles son los valores que toma en el conjunto de datos B. Formalmente, nuestra entrada es una tripla $G = \langle N, B, v \rangle$, donde $v \colon A \longrightarrow \mathbb{Q}$ es una función que asigna un valor $v(\mathbf{a})$ para cada $\mathbf{a} \in B$. Nos referimos a G como el conjunto de datos. Cuando $v(\mathbf{a}) \in \{0, 1\}$ para todo $\mathbf{a} \in B$, entonces v es una clasificación binaria. Cuando $B = A$ y $|A_i| = 2$ para todo $i \in N$, el conjunto de datos se puede ver como un juego cooperativo estándar TU, y es un juego simple si $v(\mathbf{a}) \in \{0, 1\}$.

Ahora, nos centraremos en ver cuán influyente es la característica i. Por ello, nuestra salida deseada es la medida $\phi_i(G)$ que está asociada con cada característica i. La medida $\phi_i(G)$ debería ser una buena métrica de la importancia de i determinando los valores de v sobre B. Antes de ver que la medida de influencia existe y es la única que satisface ciertos axiomas naturales, describiremos varios de los axiomas. Comenzaremos por el axioma de simetría.

Dado un conjunto de datos $G = \langle N,\ B,\ v \rangle$ y una aplicación biyectiva $\sigma \colon N \longrightarrow N$, definimos $\sigma G = \langle \sigma N,\ \sigma B,\ \sigma v \rangle$ de forma natural. σN tiene todas las características reetiquetadas de acuerdo con σ (i.e., el índice i ahora es $\sigma(i)$); σB es $\{\sigma \mathbf{a} : \mathbf{a} \in B\}$, y $\sigma v(\sigma \mathbf{a}) = v(\mathbf{a})$ para todo $\sigma \mathbf{a} \in \sigma B$. Dada una aplicación biyectiva $\tau \colon A_i \longrightarrow A_i$ sobre el conjunto de posibles estados de una característica $i \in N$, definimos $\tau G = \langle N,\ \tau B,\ \tau v \rangle$ de forma similar.

Definición 3.1. Una medida de influencia ϕ satisface la propiedad de la simetría de las características si es invariante ante la reetiquetación de las características: dado un conjunto de datos $G = \langle N,\ B,\ v \rangle$ y una biyección $\sigma \colon N \longrightarrow N$, $\phi_i(G) = \phi_{\sigma(i)}(\sigma G)$ para todo $i \in N$. Una medida de influencia ϕ satisface la propiedad de la simetría de los estados si es invariante ante la reetiquetación de los estados de una característica: dado un conjunto de datos $G = \langle N,\ B,\ v \rangle$, $i \in N$ y una biyección $\tau \colon A_i \longrightarrow A_i$, $\phi_j(G) = \phi_j(\tau G)$ para todo $j \in N$. Nótese que es posible que $i \neq j$. Una medida que satisface simetría de las características y de los estado se dice que satisface el axioma de simetría (Sim).

La propiedad de simetría de las características es una extensión natural del axioma

de simetría definido para juegos cooperativos. Sin embargo, la simetría de los estados no tiene mucho sentido en los juegos cooperativos clásicos; se podría traducir diciendo que para cualquier conjunto de jugadores $S \subset N$ y cualquier $j \in N$, el valor de i es el mismo si tratamos a S como $S \setminus \{j\}$ y a $S \setminus \{j\}$ como S. Mientras en el contexto de juegos cooperativos esto es más bien poco informativo, hemos hecho un uso no trivial de ello en el análisis posterior.

En segundo lugar describiremos una condición suficiente para que una característica sea no influyente, es decir, una característica que no afecte en los resultados de ninguna forma. Formalmente, una característica $i \in N$ es *dummy* si $v(\mathbf{a}) = v(\mathbf{a}_{-i}, b)$[1] para todo $\mathbf{a} \in B$, y para todo $b \in A_i$ tal que $(\mathbf{a}_{-i}, b) \in B$.

Definición 3.2. Una medida de influencia ϕ satisface la propiedad de característica *dummy* (D) si $\phi_i(G) = 0$ siempre que i sea una característica *dummy* en el conjunto de datos G.

La propiedad de característica *dummy* es una extensión estándar de la propiedad de jugador *dummy* (o nulo) usada en la caracterización del valor en juegos cooperativos. Sin embargo, cuando utilizamos conjuntos de datos reales, es posible que no exista $\mathbf{a} \in B$ tal que $(\mathbf{a}_{-i}, b) \in B$.

En teoría de juegos cooperativos se emplea una noción de aditividad en la caracterización del valor de Shapley y el valor de Banzhaf. Dados dos conjuntos de datos $G_1 = \langle N, B, v_1 \rangle$ y $G_2 = \langle N, A, v_2 \rangle$, definimos $G = \langle N, A, v \rangle = G_1 + G_2$ tomando $v(\mathbf{a}) = v_1(\mathbf{a}) + v_2(\mathbf{a})$ para todo $\mathbf{a} \in B$.

Definición 3.3. Una medida de influencia ϕ satisface la aditividad (AD) si $\phi_i(G_1 + G_2) = \phi_i(G_1) + \phi_i(G_2)$ para todo par de conjuntos de datos $G_1 = \langle N, B, v_1 \rangle$ y $G_2 = \langle N, B, v_2 \rangle$ y para todo $i \in N$.

El axioma de aditividad es comúnmente usado en el análisis axiomático de reglas de división de beneficios en juegos cooperativos; sin embargo, falla a la hora de conseguir una noción satisfactoria de influencia en nuestro contexto.

Ahora mostraremos que cualquiera medida que satisfaga la aditividad, adicionalmente con la simetría y la propiedad de característica *dummy*, debe evaluar a cero a todas las características. Para demostrar esto, primero definiremos la clase de conjuntos simples de datos.

[1] Nótese que $(a_{-i}, b) = (a_1, ..., a_{i-1}, b, a_{i+1}, ..., a_n)$.

Definición 3.4. Sea $U_a = \langle N, A, u_a \rangle$ un conjunto de datos definido por el clasificador u_a, donde $u_a(\mathbf{a}') = 1$ si $\mathbf{a}' = \mathbf{a}$, y $u_a(\mathbf{a}') = 0$ si $\mathbf{a}' \neq \mathbf{a}$. El conjunto de datos U_a es conocido como conjunto simple de datos sobre $\mathbf{a}$.

Además, se puede comprobar que la aditividad implica que para cualquier escalar $\alpha \in \mathbb{Q}$, $\phi_i(\alpha G) = \alpha\phi_i(G)$, donde el conjunto de datos αG tiene el valor de cada punto multiplicado por un factor α para cualquier conjunto de datos G.

Haciendo uso de las propiedades anteriores, se puede probar el siguiente resultado.

Proposición 3.5. *Cualquiera medida de influencia que satisfaga los axiomas de simetría, dummy y aditividad, evalúa en cero todas las características.*

Demostración. Primero, se demostrará que para cualquier $\mathbf{a}$, $\mathbf{a}' \in A$ y cualquier $b \in A_i$, se da que $\phi_i(U_{(\mathbf{a}_{-i},b)}) = \phi_i(U_{(\mathbf{a}'_{-i},b)})$. Esto es así, porque se puede definir una aplicación biyectiva de $U_{(\mathbf{a}_{-i},b)}$ a $U_{(\mathbf{a}'_{-i},b)}$ tal que para todo $j \in N \setminus \{i\}$, se intercambian a_j y a'_j. Por la propiedad de simetría de los estados, $\phi_i(U_{(\mathbf{a}_{-i},b)}) = \phi_i(U_{(\mathbf{a}'_{-i},b)})$.

Ahora, si ϕ es aditiva, entonces tenemos que para cualquier conjunto de datos $G = \langle N, B, v \rangle$, $\phi_i(G) = \sum_{\mathbf{a}\in B} v(\mathbf{a})\phi_i(U_{\mathbf{a}})$. Esto es, la influencia de una característica debe ser la suma de sus influencias sobre los conjuntos simples de datos sobre $\mathbf{a}$ ponderadas por el valor $v(\mathbf{a})$.

Por último, supongamos por contradicción que existe algún conjunto simple $U_{\bar{\mathbf{a}}}$ ($\bar{\mathbf{a}} \in B$) para el cual alguna característica $i \in N$ no tiene una influencia igual a cero. Esto es, se asume que $\phi_i(U_{\bar{\mathbf{a}}}) \neq 0$. Se define un conjunto de datos $G = \langle N, A, v \rangle$ como sigue. Para todo $\mathbf{a} \in A$ tal que $\mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}$, se toma $v(\mathbf{a}) = 1$, y $v(\mathbf{a}) = 0$ si $\mathbf{a}_{-i} \neq \bar{\mathbf{a}}_{-i}$. En el conjunto de datos resultante, $v(\mathbf{a})$ está determinado solamente por los valores de las características en $N \setminus \{i\}$; en otras palabras, $v(\mathbf{a}) = v(\mathbf{a}_{-i}, b)$ para todo $b \in A_i$, por lo que la característica i es *dummy*. De acuerdo con el axioma de característica *dummy*, se tiene que $\phi_i(G) = 0$; sin embargo,

$$\begin{aligned} 0 = \phi_i(G) &= \sum_{\mathbf{a}:v(\mathbf{a})=1} \phi_i(U_{\mathbf{a}}) = \sum_{b\in A_i} \phi_i(U_{(\bar{\mathbf{a}}_{-i},b)}) \\ &= \sum_{b\in A_i} \phi_i(U_{\bar{\mathbf{a}}}) = |A_i|\phi_i(U_{\bar{\mathbf{a}}}) > 0, \end{aligned}$$

donde la segunda igualdad viene dada por la descomposición de G en los conjuntos simples, y en la cuarta igualdad se aplica la simetría. Por lo tanto, se llega a una contradicción y

se puede afirmar que cualquier medida de influencia que satisfaga los axiomas de simetría, *dummy* y aditividad, evalúa en cero la influencia de todas las características.

□

Como muestra la Proposición 3.5, la aditividad, simetría y la propiedad *dummy* no consiguen describir una influencia de modo satisfactorio. Un lector familiarizado con la caracterización axiomática del valor de Shapley encontrará este resultado decepcionante: la caracterización clásica de los valores de Shapley y Banzhaf asume la aditividad.

En lo que sigue, mostraremos la caracterización axiomática de una medida de influencia, por medio de la definición de un axioma alternativo, que evoca una propiedad descrita por Lehrer (1988) que hace uso de ciertas uniones e intersecciones, en el contexto de la teoría de juegos. A partir de aquí, asumiremos que en todos los conjuntos de datos, estos están clasificados por un clasificador binario. Escribimos $W(B)$ para referirnos al conjunto de los individuos de B tales que $v(\mathbf{a}) = 1$ y los llamaremos perfiles ganadores. Por otra parte, escribiremos $L(B)$ para referirnos al conjunto de los individuos de B tales que $v(\mathbf{a}) = 0$ y los llamaremos perfiles perdedores. Por lo tanto, podemos escribir $\phi_i(W(B), L(B))$, en vez de $\phi_i(G)$. Así que, dados dos conjuntos disjuntos $W, L \subset A$ podemos definir el conjunto de datos como $G = \langle W, L\rangle$ y la influencia de i como $\phi_i(W(B), L(B))$, sin tener que poner explícitamente N, B y v.

Definición 3.6. Una medida de influencia ϕ satisface la propiedad de la unión disjunta (DU) si para cualquier $Q \subset A$, y cualesquiera $R, R' \subset A\backslash Q$, entonces $\phi_i(Q, R)+\phi_i(Q, R') = \phi_i(Q, R \cup R')$, y $\phi_i(R, Q) + \phi_i(R', Q) = \phi_i(R \cup R', Q)$.

Esto es, dado el resultado de una clasificación binaria de dos conjuntos de datos $G_1 = \langle W, L_1\rangle$ y $G_2 = \langle W, L_2\rangle$, el axioma de la unión disjunta afirma que la capacidad de una característica de afectar a los elementos de salida de G_1 es independientes de la capacidad que tiene sobre los de G_2, si los perfiles ganadores son los mismos en ambos conjuntos de datos.

Reemplazando el axioma de aditividad por el de unión disjunta obtenemos una medida de influencia única, de la siguiente forma:

$$\chi_i(G) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(a_{-i},b)\in B} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|, \quad \forall i \in N. \tag{3.1}$$

Esta medida contabiliza el número de veces que un cambio en el estado de la característica i afecta en la clasificación de los datos de salida. Si normalizamos χ y lo dividimos

entre $|B|$, la medida resultante tiene la siguiente interpretación: se toma un vector $\mathbf{a} \in B$ al azar, y cuenta el número de puntos en A_i para los cuales $(a_{-i}, b) \in B$ e i cambia el valor de $\mathbf{a}$. Se puede observar que, cuando todas las características tienen dos estados y $B = A$, entonces χ coincide con el valor de Banzhaf (1964).

A continuación se presenta un lema que caracteriza una medida de influencia que satisface (D), (Sim) y (DU) cuando el conjunto de datos contiene solo una única característica.

Lema 3.7. *Sea ϕ una medida de influencia que satisface la simetría de los estados, y sean $G_1 = \langle \{i\}, A_i, v_1 \rangle$ y $G_2 = \langle \{i\}, A_i, v_2 \rangle$ dos conjuntos de datos con una única característica i. Si el número de estados ganadores de G_1 y G_2 son idénticos, entonces $\phi_i(G_1) = \phi_i(G_2)$.*

El lema anterior implica que para juegos con una única característica, el valor de la característica solo depende del número de estados ganadores, en lugar de su identidad.

Ahora se enunciará un teorema que demuestra que ϕ es la única medida de influencia que satisface los tres axiomas anteriores, salvo el producto por una constante.

Teorema 3.8. *Una medida de influencia ϕ satisface (D), (Sim) y (DU) si y solo si existe una constante C tal que para todo conjunto de datos $G = \langle N, B, v \rangle$*

$$\phi_i(G) = C \cdot \chi_i(G), \quad \forall i \in N.$$

Demostración. Es fácil probar que χ satisface los tres axiomas, por lo tanto nos centraremos en la otra implicación del "*si y solo si*".

Asumimos que tenemos el conjunto A como datos; la demostración es válida incluso si se asume que tenemos un subconjunto arbitrario $B \subset A$. Denotaremos $W = W(A)$ y $L = L(A)$ como los conjuntos de los perfiles ganadores y perdedores respectivamente. Dado un $\mathbf{a}_{-i} \in A_{-i}$, definimos $L_{\mathbf{a}_{-i}} = \{\bar{\mathbf{a}} \in L : \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$ y $W_{\mathbf{a}_{-i}} = \{\bar{\mathbf{a}} \in W : \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$.

Utilizando la propiedad de la unión disjunta, se puede descomponer $\phi_i(W, L)$ de la forma:

$$\phi_i(W, L) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \sum_{\bar{\mathbf{a}}_{-i} \in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\bar{\mathbf{a}}_{-i}}). \tag{3.2}$$

Si en todos los sumandos de (3.2) resulta $\bar{\mathbf{a}}_{-i} \neq \mathbf{a}_{-i}$, entonces la característica i es *dummy* dado el conjunto de datos proporcionado. En efecto, los perfiles de estados están

en $W_{\mathbf{a}_{-i}}$ o en $L_{\bar{\mathbf{a}}_{-i}}$; esto es, si $v(\mathbf{a}_{-i}, b) = 0$ entonces $(\mathbf{a}_{-i}, b)$ no se observa, y si $v(\bar{\mathbf{a}}_{-i}, b) = 1$, entonces $(\bar{\mathbf{a}}_{-i}, b)$ no se observa. Por lo tanto, se concluye que:

$$\phi_i(W, L) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}). \tag{3.3}$$

Ahora consideremos $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}})$. Como ϕ satisface la propiedad de simetría de los estados, el Lema 3.7 implica que ϕ_i solo tiene la posibilidad de depender de $\mathbf{a}_{-i}$, $|W_{\mathbf{a}_{-i}}|$ y $|L_{\mathbf{a}_{-i}}|$. Luego, para cualesquiera $\mathbf{a}_{-i}$ y $\mathbf{a}'_{-i}$ tal que $|L_{\mathbf{a}_{-i}}| = |L_{\mathbf{a}'_{-i}}|$ y $|W_{\mathbf{a}_{-i}}| = |W_{\mathbf{a}'_{-i}}|$, por el Lema 3.7, $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}) = \phi_i(W_{\mathbf{a}'_{-i}}, L_{\mathbf{a}'_{-i}})$. En otras palabras, ϕ_i solo depende de $|W_{\mathbf{a}_{-i}}|$, $|L_{\mathbf{a}_{-i}}|$, y no de la identidad de $\mathbf{a}_{-i}$.

Por lo tanto, se puede ver a ϕ_i para una única característica como función de dos parámetros, w y l en $\mathbb{N}$, donde w es el número de los estados ganadores y l el número de los perdedores. De acuerdo con la propiedad de característica *dummy*, se sabe que $\phi_i(w, 0) = \phi_i(0, l) = 0$; es más, la propiedad de la unión disjunta dice que $\phi_i(x, l) + \phi_i(y, l) = \phi_i(x + y, l)$, y que $\phi_i(w, x) + \phi(w, y) = \phi_i(w, x + y)$. Ahora se probará que $\phi_i(w, l) = \phi_i(1, 1)wl$.

Se probará por inducción en $w + l$. Para $w + l = 2$ la afirmación es clara. Asumimos sin pérdida de generalidad que $w > 1$ y $l \geq 1$; entonces podemos escribir $w = x + y$ con $x, y \in \mathbb{N}$ tal que $1 \leq x, y < w$. Por nuestra observación anterior,

$$\begin{aligned} \phi_i(w, l) &= \phi_i(x, l) + \phi_i(y, l) \\ &= \phi_i(1, 1)xl + \phi_i(1, 1)yl = \phi_i(1, 1)wl, \end{aligned}$$

donde la segunda igualdad se tiene por la hipótesis de inducción. Entonces, $\phi_i(1, 1)$ es la influencia de la característica i cuando hay exactamente un estado ganador y uno perdedor. Denotaremos $\phi_i(1, 1) = c_i$. Definimos:

$$W_i(\mathbf{a}_{-i}) = \{b \in A_i : v(\mathbf{a}_{-i}, b) = 1\} \text{ y } L_i(\mathbf{a}_{-i}) = A_i \setminus W_i(\mathbf{a}_{-i}).$$

Por lo tanto, $|W_{\mathbf{a}_{-i}}| = |W_i(\mathbf{a}_{-i})|$ y $|L_{\mathbf{a}_{-i}}| = |L_i(\mathbf{a}_{-i})|$. Juntando todo lo anterior, se obtiene que:

$$\phi_i(G) = c_i \sum_{\mathbf{a}_{-i} \in A_i} |W_i(\mathbf{a}_{-i})| \cdot |L_i(\mathbf{a}_{-i})|. \tag{3.4}$$

Ahora solo queda ver que la medida dada en (3.4) es igual (salvo el producto por una constante) a χ_i. En efecto, (3.4) es igual a $\sum_{\mathbf{a} \in A : v(\mathbf{a}) = 0} |W_i(\mathbf{a}_{-i})|$, lo que a su vez es igual

a $\sum_{\mathbf{a}\in A:v(\mathbf{a})=0}\sum_{b\in A_i}|v(\mathbf{a}_{-i},b)-v(\mathbf{a})|$. De forma análoga, tenemos que (3.4) es igual a:

$$\sum_{\mathbf{a}\in A:v(\mathbf{a})=1}\sum_{b\in A_i}|v(\mathbf{a}_{-i},b)-v(\mathbf{a})|.$$

Por lo tanto, se tiene la siguiente igualdad:

$$\sum_{\mathbf{a}_{-i}\in A_i}|W_i(\mathbf{a}_{-i})|\cdot|L_i(\mathbf{a}_{-i})|=\frac{1}{2}\sum_{\mathbf{a}\in A}\sum_{b\in A_i}|v(\mathbf{a}_{-i},b)-v(\mathbf{a})|.$$

En particular, para cada conjunto de datos $G=\langle N,A,v\rangle$ y para todo $i\in N$, existe alguna constante C_i tal que $\phi_i(G)=C_i\chi_i(G)$. Para concluir la demostración, se mostrará que $C_i=C_j$ para todo $i,j\in N$. Sea $\sigma\colon N\longrightarrow N$ una biyección que intercambia i con j; entonces $\phi_i(G)=\phi_{\sigma(i)}(\sigma G)$. Por la propiedad de simetría de las características,

$$\begin{aligned}C_i\chi_i(G)=\phi_i(G)=\phi_{\sigma(i)}(\sigma G)=\phi_j(\sigma G)\\ =C_j\chi_j(\sigma G)=C_j\chi_i(G).\end{aligned}$$

Por lo tanto, hemos llegado a que $C_i=C_j$ para todo i, $j\in N$. □

3.1.1. Ejemplo 1

Ahora realizaremos un ejemplo sencillo donde aplicar la medida de influencia presentada y ver una clara relación con el valor de Banzhaf. Por lo tanto, vamos a considerar un clasificador binario que clasifica en 0 o 1. Cada vector $\mathbf{a}\in B$ informa sobre un total de tres atributos que toman valores de 0 o 1, es decir, es un vector de la forma $\mathbf{a}=(a_1,a_2,a_3)$ donde $a_i\in\{0,1\}$ para todo $i\in\{1,2,3\}$. Hay un total de $2\times2\times2=8$ combinaciones de los estados de estos atributos. Supongamos que tenemos un vector de cada tipo y que se han clasificado de la forma que aparece en la Tabla 3.1.

a	$v(\mathbf{a})$	**a**	$v(\mathbf{a})$
$(0,0,0)$	0	$(0,1,1)$	1
$(0,0,1)$	1	$(1,0,1)$	0
$(0,1,0)$	0	$(1,1,0)$	0
$(1,0,0)$	0	$(1,1,1)$	1

Tabla 3.1: Datos del ejemplo 1.

Ahora procederemos a calcular la medida de influencia, primero para la característica $i=1$. Fijamos un vector $\mathbf{a}$, por ejemplo $(0,0,1)$, y se calcula la diferencia en la clasificación

cuando cambia el valor del primer atributo y pasa a ser 1 (véase la ecuación (3.1)). En este caso, el cambio de estado ha producido un cambio en la clasificación, de 1 a 0, por lo que la diferencia de los valores en valor absoluto será 1. De forma análoga ocurrirá al considerar el vector $(1, 0, 1)$, por lo tanto, la influencia aumenta en 2, uno por cada vector. Otro posible caso sería el del vector $(0, 0, 0)$, cuyo valor es 0, al igual que el de $(1, 0, 0)$, por lo tanto no hay cambio en la clasificación y entonces no aumenta la influencia. De la misma forma se procede con el resto de vectores, obteniéndose el siguiente resultado:

$$\begin{aligned}\chi_1(G) =& 2 \cdot |v(1,0,0) - v(0,0,0)| + 2 \cdot |v(1,0,1) - v(0,0,1)| \\ &+ 2 \cdot |v(1,1,0) - v(0,1,0)| + 2 \cdot |v(1,1,1) - v(0,1,1)| \\ =& 2 \cdot 0 + 2 \cdot 1 + 2 \cdot 0 + 2 \cdot 0 = 2.\end{aligned}$$

De forma análoga se calcula la influencia de las otras dos características:

$$\begin{aligned}\chi_2(G) =& 2 \cdot |v(0,1,0) - v(0,0,0)| + 2 \cdot |v(0,1,1) - v(0,0,1)| \\ &+ 2 \cdot |v(1,1,0) - v(1,0,0)| + 2 \cdot |v(1,1,1) - v(1,0,1)| \\ =& 2 \cdot 0 + 2 \cdot 0 + 2 \cdot 0 + 2 \cdot 1 = 2,\end{aligned}$$

$$\begin{aligned}\chi_3(G) =& 2 \cdot |v(0,0,1) - v(0,0,0)| + 2 \cdot |v(0,1,1) - v(0,1,0)| \\ &+ 2 \cdot |v(1,0,1) - v(1,0,0)| + 2 \cdot |v(1,1,1) - v(1,1,0)| \\ =& 2 \cdot 1 + 2 \cdot 1 + 2 \cdot 0 + 2 \cdot 1 = 6.\end{aligned}$$

Tenemos, para cada una de las tres características, una influencia de 2, 2 y 6, respectivamente. Por lo tanto, se puede decir que la tercera característica es la más influyente.

Ahora veamos la relación con el valor de Banzhaf. En este ejemplo se cumple que todas las características tienen dos estados y $B = A$, por lo tanto χ coincide con el valor de Banzhaf. Identificamos las distintas combinaciones de estados de las características como las posibles coaliciones de características, S, tal que el estado 1 se interpreta como que la característica correspondiente, i, "participa en la coalición" y el estado 0 como que "no participa". En la Tabla 3.2, así dadas las coaliciones, recogemos los valores de los juegos asociados, v^i, uno para cada característica. Si para $S \subset N$, a_S es el elemento de B correspondiente a S, e $i \in \{1, 2, 3\}$, se define el juego v^i:

$$v^i(\mathrm{S}) = \begin{cases} \text{mín}\left\{v(a_S), v(a_{S \cup i})\right\} & \text{si } i \notin S, \\ \text{máx}\left\{v(a_S), v(a_{S \setminus \{i\}})\right\} & \text{si } i \in S. \end{cases} \tag{3.5}$$

a	S	$v^1(S)$	$v^2(S)$	$v^3(S)$
$(0,0,0)$	$\{\emptyset\}$	0	0	0
$(0,0,1)$	$\{3\}$	0	1	1
$(0,1,0)$	$\{2\}$	0	0	0
$(1,0,0)$	$\{1\}$	0	0	0
$(0,1,1)$	$\{2,3\}$	1	1	1
$(1,0,1)$	$\{1,3\}$	1	0	0
$(1,1,0)$	$\{1,2\}$	0	0	0
$(1,1,1)$	$\{1,2,3\}$	1	1	1

Tabla 3.2: Juegos TU asociados al conjunto de datos.

Ahora, para cada i, se calcula el valor de Banzhaf del jugador i en el juego v^i:

$$\begin{aligned}\beta_1(v^1) =& \frac{1}{2^{3-1}}\left[(v(1,0,0)-v(0,0,0))+(v(1,0,1)-v(0,0,1))\right.\\ &\left.+(v(1,1,0)-v(0,1,0))+(v(1,1,1)-v(0,1,1))\right]\\ =& \frac{1}{2^{3-1}}\cdot(0+1+0+0)=\frac{1}{4},\end{aligned}$$

$$\begin{aligned}\beta_2(v^2) =& \frac{1}{2^{3-1}}\left[(v(0,1,0)-v(0,0,0))+(v(0,1,1)-v(0,0,1))\right)\\ &+(v(1,1,0)-v(1,0,0))+(v(1,1,1)-v(1,0,1))]\\ =& \frac{1}{2^{3-1}}\cdot(0+0+0+1)=\frac{1}{4},\end{aligned}$$

$$\begin{aligned}\beta_3(v^3) =& \frac{1}{2^{3-1}}\left[(v(0,0,1)-v(0,0,0))+(v(0,1,1)-v(0,1,0))\right)\\ &+(v(1,0,1)-v(1,0,0))+(v(1,1,1)-v(1,1,0))]\\ =& \frac{1}{2^{3-1}}\cdot(1+1+0+1)=\frac{3}{4}.\end{aligned}$$

Los valores de Banzhaf obtenidos coinciden con las medidas de influencia calculadas anteriormente si normalizamos sus valores dividiendo por $|B|=8$.

3.1.2. Ejemplo 2

Vamos a considerar de nuevo el ejemplo anterior, con un clasificador binario y cada vector $\mathbf{a}\in B$ con tres atributos que toman valores de 0 o 1, pero en este caso vamos a tener en cuenta que en nuestra muestra puede haber vectores repetidos o que no son observados en ningún momento. Por lo tanto, supongamos que tenemos los resultados de la Tabla 3.3,

donde la primera columna son las posibles valores del vector **a**, la segunda las veces que ese vector ha salido repetido en nuestra muestra y la tercera cómo lo ha clasificado v.

a	frecuencia(**a**)	v(**a**)
$(0,0,0)$	3	0
$(0,0,1)$	2	1
$(0,1,0)$	0	-
$(1,0,0)$	2	0
$(0,1,1)$	4	1
$(1,0,1)$	1	0
$(1,1,0)$	1	0
$(1,1,1)$	3	1

Tabla 3.3: Datos del ejemplo 2.

Como se puede observar en la segunda columna, B tiene 16 elementos. Otro dato a tener en cuenta sería la combinación $(0,1,0)$ de atributos, que no ha sido observada en ningún momento, por lo tanto, a la hora de calcular la influencia no podemos tener en cuenta cómo ha sido clasificado ni cómo cambia dicha clasificación si se modifica el valor de alguno de los atributos.

Procedamos a calcular la medida de influencia, primero para la característica $i = 1$. Como hemos hecho en el ejemplo anterior, fijamos un vector **a**, por ejemplo $(0,0,1)$, y se calcula la diferencia en la clasificación cuando cambia el valor del primer atributo y pasa a ser 1 (véase la ecuación (3.1)). Se produce un cambio en la clasificación, de 1 a 0, por lo que la diferencia de los valores va a ser 1. Además, como el vector $(0,0,1)$ se repite 2 veces y, de forma análoga, el vector $(1,0,1)$ se repite solamente 1 vez, se puede decir que la influencia aumenta en $2 + 1 = 3$. Otro posible caso sería el del vector $(0,0,0)$, cuyo valor es 0, al igual que el de $(1,0,0)$, por lo tanto no hay cambio en la clasificación y entonces no aumenta la influencia. De la misma forma se procede con el resto de vectores, obteniendo ahora el siguiente resultado:

$$\begin{aligned}\chi_1(G) =& (3+2)\cdot|v(1,0,0)-v(0,0,0)| + (2+1)\cdot|v(1,0,1)-v(0,0,1)| \\ &+ 0\cdot|v(1,1,0)-v(0,1,0)| + (4+3)\cdot|v(1,1,1)-v(0,1,1)| \\ =& 5\cdot 0 + 3\cdot 1 + 0 + 7\cdot 0 = 3.\end{aligned}$$

De forma análoga se calcula la influencia de las otras dos características:

$$\begin{aligned}\chi_2(G) =& 0 \cdot |v(0,1,0) - v(0,0,0)| + (2+4) \cdot |v(0,1,1) - v(0,0,1)| \\ &+ (2+1) \cdot |v(1,1,0) - v(1,0,0)| + (1+3) \cdot |v(1,1,1) - v(1,0,1)| \\ =& 0 + 6 \cdot 0 + 3 \cdot 0 + 4 \cdot 1 = 4,\end{aligned}$$

$$\begin{aligned}\chi_3(G) =& (3+2) \cdot |v(0,0,1) - v(0,0,0)| + 0 \cdot |v(0,1,1) - v(0,1,0)| \\ &+ (2+1) \cdot |v(1,0,1) - v(1,0,0)| + (1+3) \cdot |v(1,1,1) - v(1,1,0)| \\ =& 5 \cdot 1 + 0 + 3 \cdot 0 + 4 \cdot 1 = 9.\end{aligned}$$

Tenemos, para cada una de las tres características, una influencia de 3, 4 y 9, respectivamente. De este modo, se puede decir que la tercera característica es la más influyente. En este caso $B \neq A$, por lo tanto, no se puede relacionar con el valor de Banzhaf.

3.2. Influencia en clasificadores lineales

En esta sección se presentará la aplicación de los resultados anteriores a la clase de clasificadores lineales. Para esta clase de funciones, la medida de influencia tiene un interpretación intuitiva.

Un clasificador lineal está definido por un hiperplano en $\mathbb{R}^n$, de forma que todos los puntos que están a un lado del hiperplano toman el valor 1, y todos los puntos del otro lado toman el valor 0. Formalmente, se asocia un peso $\mathrm{w}_i \in \mathbb{R}$ a todas las características de N (se asume que $\mathrm{w}_i \neq 0$ para todo $i \in N$); un punto $\mathbf{x} \in \mathbb{R}^n$ vale 1 si $\mathbf{x} \cdot \mathbf{w} \geq q$, donde $q \in \mathbb{R}$ es un parámetro dado. La función de clasificación $v \colon \mathbb{R}^n \to \{0,1\}$ viene dada por:

$$v(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} \cdot \mathbf{w} \geq q, \\ 0 & \text{en otro caso.} \end{cases} \tag{3.6}$$

Fijado el valor de x_i a algún $b \in \mathbb{R}$, se considera el conjunto:

$$W_i(b) = \left\{ \mathbf{x}_{-i} \in \mathbb{R}^{n-1} : v(\mathbf{x}_{-i}, b) = 1 \right\};$$

se observa que si $b < b'$ y $\mathrm{w}_i > 0$, entonces $W_i(b) \subset W_i(b')$ (si $\mathrm{w}_i < 0$ entonces $W_i(b') \subset W_i(b)$). Dados dos valores $b, b' \in \mathbb{R}$, se denota por:

$$D_i(b,b') = \left\{ \mathbf{x}_{-i} \in \mathbb{R}^{n-1} : v(\mathbf{x}_{-i}, b) \neq v(\mathbf{x}_{-i}, b') \right\}.$$

Por la anterior observación, si $b < b'$ entonces $D_i(b, b') = W_i(b') \setminus W_i(b)$, y si $b > b'$ entonces $D_i(b, b') = W_i(b) \setminus W_i(b')$.

Supongamos que en vez de tomar valores en $\mathbb{R}^n$, se toman en $[0, 1]^n$, entonces podemos definir $|D_i(b, b')| = Vol(D_i(b, b'))$, donde:

$$Vol(D_i(b, b')) = \int_{\mathbf{x}_{-i} \in [0,1]^{n-1}} |v(\mathbf{x}_{-i}, b') - v(\mathbf{x}_{-i}, b)| \partial \mathbf{x}_{-i}.$$

En otras palabras, para medir la influencia total de fijar el estado de la característica i a b, se calcula el volumen total de $D_i(b, b')$ para todo $b' \in [0, 1]$, esto es $\int_{b'=0}^{1} Vol(D_i(b, b')) \partial b$. Por lo tanto, la influencia total de fijar el estado i a b es $\int_{\mathbf{x} \in [0,1]^n} |v(\mathbf{x}_{-i}, b) - v(\mathbf{x})| \partial \mathbf{x}$. La influencia total de i será entonces la influencia total de sus estados, i. e.:

$$\int_{b=0}^{1} \int_{\mathbf{x} \in [0,1]^n} |v(\mathbf{x}_{-i}, b) - v(\mathbf{x})| \partial \mathbf{x} \partial b. \tag{3.7}$$

La fórmula en la ecuación (3.7) se denota por $\chi_i(\mathbf{w}; q)$. La ecuación (3.1) es una versión discretizada de la ecuación (3.7); el resultado de la sección anterior se puede extender a un conjunto continuo, con solo unos pocos cambios en la demostración.

Ahora se mostrará en el siguiente teorema que la medida dada en (3.7) concuerda con los pesos de una manera natural.

Teorema 3.9. *Sea v un clasificador lineal definido por $\mathbf{w}$ y q; entonces $\chi_i(G) \geq \chi_j(G)$ si y solo si $|\mathrm{w}_i| \geq |\mathrm{w}_j|$.*

3.3. Extensiones de la medida de influencia de las características

Anteriormente se presentó una caracterización axiomática de la influencia de la característica, donde el valor de cada vector de características es 0 o 1. En esta sección se presentarán algunas posibles extensiones más de la medida, y las variaciones en los axiomas que sean necesarias.

3.3.1. Influencia de los estados

Al igual que se ha visto la influencia que tiene cierta característica, uno también se puede preguntar por la influencia de cierto estado de una característica. En otras palabras, en vez de medir la influencia de una característica, medir la influencia de una característica

i en cierto estado. El resultado descrito en la caracterización axiomática se puede extender fácilmente a este caso. Es más, el resultado de imposibilidad descrito en la Proposición 3.5 ya no se cumple cuando medimos la influencia del estado, se puede reemplazar la propiedad de la unión disjunta por la de aditividad para obtener una clasificación alternativa de la influencia de estado.

3.3.2. Influencia de los pesos

Supongamos ahora que además del conjunto de datos B, se tiene una función de pesos $w\colon B \to \mathbb{R}$. $w(\mathbf{a})$ puede ser pensado como el número de apariciones del vector $\mathbf{a}$ en el conjunto de datos o la probabilidad de que $\mathbf{a}$ aparezca. En la Sección 3.1 se asumió implícitamente que todos los puntos sucedían con la misma frecuencia e igual importancia. Una extensión simple de la unión disjunta y del axioma de la simetría para la variante con pesos permitiría demostrar que la única medida de influencia con pesos que satisface estos axiomas es:

$$\chi_i^w(B) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(\mathbf{a}_{-i},b)\in B} w(\mathbf{a})|v(\mathbf{a}_{-i},b) - v(\mathbf{a})|.$$

3.3.3. Medida de distancia general

Supongamos que en vez de un clasificador $v\colon A \to \{0,1\}$ se tiene una pseudo-distancia, es decir, una función $d\colon A \times A \to \mathbb{R}$ que satisface $d(\mathbf{a},\mathbf{a}') = d(\mathbf{a}',\mathbf{a})$, $d(\mathbf{a},\mathbf{a}) = 0$ y la desigualdad triangular. Nótese que es posible que $d(\mathbf{a},\mathbf{a}') = 0$ con $\mathbf{a} \neq \mathbf{a}'$. Es posible realizar un análisis axiomático en este entorno general, pero requiere más suposiciones sobre el comportamiento de la medida de influencia. Tal enfoque axiomático conduce a mostrar que la medida de influencia:

$$\chi_i^d(B) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(\mathbf{a}_{-i},b)\in B} d((\mathbf{a}_{-i},b),\mathbf{a}),$$

está definida únicamente a través de varios axiomas naturales. Los axiomas adicionales son extensiones simples de la propiedad de la unión disjunta, y un requisito mínimo indicando que cuando $B = \{\mathbf{a},(\mathbf{a}_{-i},b)\}$, entonces la influencia de una característica es $\alpha d((\mathbf{a}_{-i},b),\mathbf{a})$ para una constante α independiente de i.

Capítulo 4

Ejemplo COVID-19

En este capítulo aplicaremos a un ejemplo práctico la medida de influencia dada anteriormente. Se han obtenido datos de los pacientes afectados por el COVID-19 correspondientes a la primera ola de Galicia, en marzo y abril de 2020, estudiados en Davila-Pena et al. (2021). Nos centraremos en el área sanitaria de Santiago-Barbanza, con un total de 1631 pacientes. De cada paciente se conoce la edad, el sexo y todos sus antecedentes médicos. A la hora de realizar el estudio se ha agrupado la edad en dos grupos distintos, los pacientes con menos de 60 años (representado por 0) y los de 60 o más años (representado por 1) y en el sexo puede tomar solamente los estados de *hombre* y *mujer*. Por otra parte, en los antecedentes médicos se han seleccionado 5 de los considerados influyentes: bronquitis, neumonía, demencia, diabetes y obesidad; marcados con 1 si la han padecido y 0 en caso contrario. Además, se sabe de cada paciente si ha tenido alguna incidencia o no, siendo esta nuestra clase a la hora de ser clasificados. Se entiende que ha tenido una incidencia si el paciente ha estado hospitalizado, si ha necesitado UCI o si ha fallecido.

Primero se hará un breve análisis descriptivo de los datos recolectados. El número total de pacientes que han tenido alguna incidencia en Santiago-Barbanza es de 169, lo que implica un 10,36 % de los pacientes. Luego, se pueden observar en la Figura 4.1 las distintas frecuencias de las edades, haciendo distinción entre los que han tenido alguna incidencia y los que no. En el intervalo de edad inferior a 60 años, con un total de 910 pacientes, se han contabilizado solamente 11 pacientes con alguna incidencia, y el resto, con edad igual o superior a 60 años, ha habido un total de 721 paciente de los cuales 158 han tenido alguna incidencia. Además, en la Figura 4.2 se puede ver representado el número de pacientes diferenciados por su sexo, con 644 hombres y 987 mujeres en total, y a su vez, divididos entre los que han tenido alguna incidencia o no, con 107 hombre y

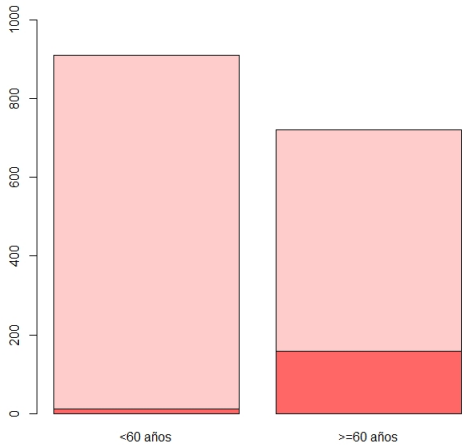

Figura 4.1: Frecuencia de cada intervalo de edad. El primer color desde la base indica los pacientes con alguna incidencia y, el segundo, los que no han tenido ninguna incidencia.

62 mujeres. Vamos a prescindir de esta característica a la hora de calcular la influencia. Además, podemos ver en la Figura 4.3 el número de pacientes que han tenido cada uno de los antecedentes escogidos: bronquitis, neumonía, demencia, diabetes y obesidad. En cada caso tenemos un total de 36, 20, 56, 178 y 221 pacientes. Por otra parte, se puede

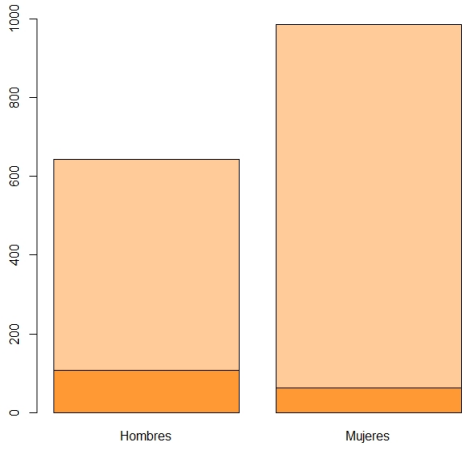

Figura 4.2: Frecuencia de los sexos. El primer color desde la base indica los pacientes con alguna incidencia y, el segundo, los que no han tenido ninguna incidencia.

diferenciar en cada grupo el número de pacientes que han tenido alguna incidencia, siendo 5, 5, 23, 57 y 39, respectivamente. Para calcular la influencia vamos a escoger tres de estos antecedentes para simplificar, los cuales son bronquitis, demencia y diabetes.

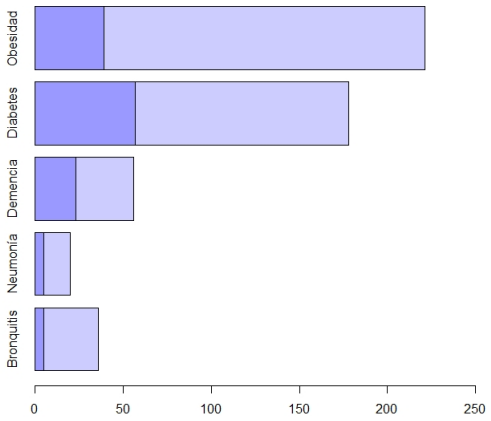

Figura 4.3: Frecuencia de cada uno de los distintos tipos de antecedentes. El primer color desde la izquierda indica los pacientes con alguna incidencia y, el segundo, los que no han tenido ninguna incidencia.

Ahora procederemos a calcular la medida de influencia de cada característica. Por lo tanto, como al final hemos escogido cuatro características: edad y los antecedentes de bronquitis, demencia y diabetes, tenemos un total de $2 \times 2 \times 2 \times 2 = 16$ combinaciones y, a su vez, cada una de ellas puede tomar el valor de 0 o 1 según cómo se haya clasificado.

Comenzamos con la influencia de la edad. Entonces, para poder contabilizar el número de veces que un cambio en el estado de la edad afecte en la clasificación de cada paciente, tenemos que tener en cuenta dos cosas. Primero, hay que tomar un vector $\mathbf{a}$ que exista en el conjunto de datos B. Y segundo, a la hora de hacer la diferencia de los valores, $v(\mathbf{a}_{-1}, b) - v(\mathbf{a})$, se contabilizará si, y solo si, existe un vector $(\mathbf{a}_{-1}, b)$ con valor contrario a $\mathbf{a}$, ya que v toma valores de 0 o 1. Por lo tanto, este vector $\mathbf{a}$ con un valor concreto, $v(\mathbf{a})$, hará aumentar la influencia las veces que se repitan esas mismas características y, de la misma forma, con el vector $(\mathbf{a}_{-1}, b)$. Por ejemplo, si tomamos el vector $\mathbf{a} = (0, 0, 0, 0)$ con valor $v(\mathbf{a}) = 0$, se contabilizará el número de veces que se haya repetido de esa misma forma, en el caso de que exista algún vector $(1, 0, 0, 0)$ con valor 1.

Por lo tanto, para simplificar los pasos, primero calcularemos el número de veces que se repite cada vector con ciertas características y cierto valor. Se obtiene así la Tabla 4.1.

a	$v(\mathbf{a})$	frecuencia($\mathbf{a}$)	**a**	$v(\mathbf{a})$	frecuencia($\mathbf{a}$)
$(0,0,0,0)$	0	865	$(0,0,0,1)$	0	20
$(0,0,0,0)$	1	8	$(0,0,0,1)$	1	3
$(1,0,0,0)$	0	420	$(1,0,0,1)$	0	94
$(1,0,0,0)$	1	87	$(1,0,0,1)$	1	44
$(0,1,0,0)$	0	12	$(0,1,0,1)$	0	1
$(0,1,0,0)$	1	0	$(0,1,0,1)$	1	0
$(1,1,0,0)$	0	12	$(1,1,0,1)$	0	5
$(1,1,0,0)$	1	2	$(1,1,0,1)$	1	2
$(0,0,1,0)$	0	1	$(0,0,1,1)$	0	0
$(0,0,1,0)$	1	0	$(0,0,1,1)$	1	0
$(1,0,1,0)$	0	30	$(1,0,1,1)$	0	1
$(1,0,1,0)$	1	14	$(1,0,1,1)$	1	8
$(0,1,1,0)$	0	0	$(0,1,1,1)$	0	0
$(0,1,1,0)$	1	0	$(0,1,1,1)$	1	0
$(1,1,1,0)$	0	1	$(1,1,1,1)$	0	0
$(1,1,1,0)$	1	1	$(1,1,1,1)$	1	0

Tabla 4.1: Recuento de datos.

Repitiendo el mismo procedimiento con las otras tres características obtenemos las influencias para la edad, bronquitis, demencia y diabetes siguientes: 1573, 736, 723 y 1615, respectivamente. Por lo tanto, se puede observar que la diabetes es la más influyente seguida de la edad. Para una mejor visualización, en la Figura 4.4 se representa un diagrama de barras donde la altura es la influencia de cada característica. Se puede observar que la edad y la diabetes son más influyentes que las dos restantes. En el Apéndice A se presenta el código de `R` necesario para el cálculo de las influencias.

Después de ver la influencia de distintos atributos a nivel del área sanitaria de Santiago-Barbanza, procedemos a realizar el análisis a nivel de toda Galicia. En este caso, hay un total de 10454 pacientes de los cuales un 8,4 % han tenido alguna incidencia. De la misma forma se vuelve a calcular la influencia de las mismas características, dando en este caso:

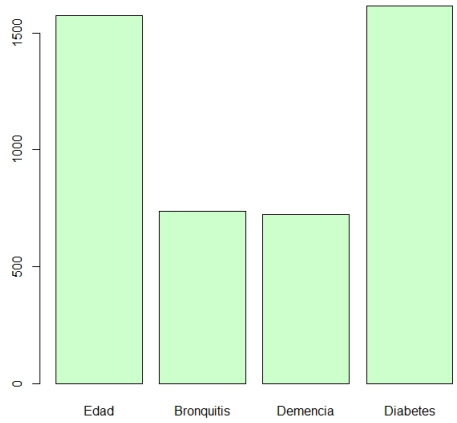

Figura 4.4: Influencia de cada una de las características en el área de Santiago-Barbanza.

10183, 10406, 4964 y 10444 para las características de edad, bronquitis, demencia y diabetes, respectivamente. Por lo tanto, tenemos de nuevo que la diabetes es la característica con mayor influencia a nivel de Galicia seguido de la bronquitis y la edad.

Por otra parte, nos hemos centrado también en la incidencia concreta de fallecimiento a nivel de Galicia, y tenemos que del total de pacientes hubo 544 fallecidos. Calculando la influencia de las cuatro características obtenemos: 10008, 5006, 4920 y 10341, respectivamente. De esta forma, volvemos a tener con mayor influencia la diabetes seguida de cerca por la edad.

Podemos concluir a partir de estos dos últimos casos que, la característica con mayor influencia a la hora de que un paciente haya tenido alguna incidencia es la diabetes seguida de la edad, al igual que ocurría si clasificábamos según el fallecimiento. En cambio, centrándonos en la influencia de la bronquitis, esta era mayor en el caso de contar todas las incidencias que si solamente clasificábamos según si fallecían o no. Por lo tanto, se puede pensar que la gente que tiene como antecedente médico la bronquitis esté más predispuesta a tener alguna incidencia del tipo de hospitalización o UCI antes que el fallecimiento.

Capítulo 5

Conclusiones

En este trabajo, principalmente se ha estudiado y analizado una medida de influencia en la clasificación. Esta medida de influencia ha sido definida a partir únicamente de una serie de axiomas naturales, y con la posibilidad de ser extendida fácilmente a otros contextos. La principal ventaja de esta aproximación es la posibilidad de usarla con un conocimiento mínimo sobre el algoritmo o prioridades del clasificador. En el ejemplo del COVID-19, se puede observar que se parte simplemente del conjunto de todos los pacientes, sin conocer previamente en qué tipo de personas es más agresivo el virus.

La medida de influencia puede ser extendida de varias formas. La medida χ presentada contabiliza el número de veces que un cambio en el estado de cierta característica produce un cambio en el clasificador. Sin embargo, se puede dar el caso de que el conjunto de datos no contenga alguno de los vectores $\mathbf{a}$, $\mathbf{a}' \in B$, tal que $\mathbf{a}' = (\mathbf{a}_{-i}, b)$, como es en el caso de nuestro ejemplo. Se puede observar en la Tabla 4.1 que no tenemos ningún paciente con las características de tener menos de 60 años y con los antecedentes de bronquitis, demencia y con o sin diabetes, por ejemplo. En nuestro caso, hemos prescindido simplemente de esos perfiles a la hora de calcular la influencia, pero un problema abierto sería extender la medida de manera que no deseche posibles perfiles. Esta forma de proceder se ha visto reflejada al estudiar la influencia de las características en la clasificación de una incidencia, pues si bien al considerar el área sanitaria de Santiago-Barbanza la patología previa de la bronquitis no tenía gran influencia, esta se incrementó al realizar el análisis a nivel de toda Galicia, cuando se observaron perfiles de pacientes no existentes en el primer análisis. Esto parece aconsejar la búsqueda de una medida que podría equilibrar los resultados.

Por último, los resultados de la parte experimental son más bien ilustrativos que informativos, ya que simplemente hemos analizado los pacientes a partir de cuatro caracterísi-

cas. Y en función de esas cuatro características, hemos visto que las más influyentes son la diabetes y la edad, pero tampoco se podría concluir que son las más influyentes en general, ya que, escogiendo más o unas distintas podrían variar los resultados. Por lo que, un paso a mayores sería la de ir escogiendo más características para obtener unos resultados más realistas y en consecuencia útiles para servir de apoyo a la toma de decisiones.

Apéndice A

Programación con R

A.1. Preparación de los datos

```
Pacientes=read.csv2("Pacientes_01.csv", sep=";", header=FALSE,
stringsAsFactors=TRUE)
Antecedentes=read.csv2("Antecedentes_01.csv", sep=";", header=FALSE,
stringsAsFactors=TRUE)

m=length(Pacientes[,1])

datos=Pacientes[,c(4,2)]

#Creamos los dos grupos de edad <60 (0) / >=60 (1)
n=which(datos[,2]<60)
v=rep(0,length(n))

datos[,2]=replace(datos[,2],1:m, 1)
datos[,2]=replace(datos[,2],n, v)

#Columna Bronquitis SI(1)/NO(0)
datos=cbind(datos,Br=rep(0,m))
n=which(Antecedentes[,2]=='R78' | Antecedentes[,2]=='R79')
ln=length(n)
for (i in 1:ln){
for (j in 1:m) {
if(Antecedentes[n[i],1]==Pacientes[j,1]) {
```

```
datos[j,3]=1}
}}

#Columna Demencia SI(1)/NO(0)
datos=cbind(datos,Dem=rep(0,m))
n=which(Antecedentes[,2]=='P70')
ln=length(n)
for (i in 1:ln){
for (j in 1:m) {
if(Antecedentes[n[i],1]==Pacientes[j,1]) {
datos[j,4]=1}
}}

#Columna Diabetes SI(1)/NO(0)
datos=cbind(datos,Diab=rep(0,m))
n=which(Antecedentes[,2]=='T89' | Antecedentes[,2]=='T90'
| Antecedentes[,2]=='W85')
ln=length(n)
for (i in 1:ln){
for (j in 1:m) {
if(Antecedentes[n[i],1]==Pacientes[j,1]) {
datos[j,5]=1}
}}
```

A.2. Medida de influencia

```
a=datos
va=rep(0,32)

for (w in 0:1){
#diabetes
for (l in 0:1){
#demencia
for (i in 0:1){
#bronquitis
for (j in 0:1){
```

```
#edad
for (k in 0:1){
#incidencia
index=which(a[,1]==k & a[,2]==j & a[,3]==i & a[,4]==l & a[,5]==w)
va[16*w+8*l+4*i+2*j+k+1]=length(a[index,1])
}}}}}

chi_1=0
for (w in 0:1){
#diabetes
for (l in 0:1){
#demencia
for (i in 0:1){
#bronquitis
for (k in 0:1){
#incidencia
if (va[16*w+8*l+4*i+2*0+k+1]>0 & va[16*w+8*l+4*i+2*1+(1-k)+1]>0){
chi_1=chi_1 + va[16*w+8*l+4*i+2*0+k+1]+va[16*w+8*l+4*i+2*1+1-k+1]
}}}}}

chi_2=0
for (w in 0:1){
#diabetes
for (l in 0:1){
#demencia
for (j in 0:1){
#edad
for (k in 0:1){
#incidencia
if (va[16*w+8*l+4*0+2*j+k+1]>0 & va[16*w+8*l+4*1+2*j+1-k+1]>0){
chi_2=chi_2 + va[16*w+8*l+4*0+2*j+k+1]+va[16*w+8*l+4*1+2*j+1-k+1]
}}}}}

chi_3=0
for (w in 0:1){
#diabetes
```

```
for (i in 0:1){
#bronquitis
for (j in 0:1){
#edad
for (k in 0:1){
#incidencia
if (va[16*w+8*0+4*i+2*j+k+1]>0 & va[16*w+8*1+4*i+2*j+1-k+1]>0){
chi_3=chi_3 + va[16*w+8*0+4*i+2*j+k+1]+va[16*w+8*1+4*i+2*j+1-k+1]
}}}}}

chi_4=0
for (l in 0:1){
#demencia
for (i in 0:1){
#bronquitis
for (j in 0:1){
#edad
for (k in 0:1){
#incidencia
if (va[16*0+8*l+4*i+2*j+k+1]>0 & va[16*1+8*l+4*i+2*j+1-k+1]>0){
chi_4=chi_4 + va[16*0+8*l+4*i+2*j+k+1]+va[16*1+8*l+4*i+2*j+1-k+1]
}}}}}

#influencia edad
chi_1
#influencia bronquitis
chi_2
#influencia demencia
chi_3
#influencia diabetes
chi_4
```

Bibliografía

Banzhaf, J. F. (1964), "Weighted voting doesn't work: A mathematical analysis." *Rutgers Law Review*, 19, 317–343.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*. CRC press.

Datta, A., Datta, A., Procaccia, A. D., and Zick, Y. (2015), "Influence in classification via cooperative game theory." *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 511–517.

Davila-Pena, L., García-Jurado, I., and Casas-Méndez, B. (2021), "Assessment of the influence of features on a classification problem: an application to COVID-19 patients." *arXiv preprint. arXiv:2104.14958*.

Dua, D. and Graff, G. (2017), "UCI machine learning repository." URL `http://archive.ics.uci.edu/ml`.

Hastie, T., Tibshirani, R., and Friedman, J. (2008), *The Elements of Statistical Learning*. Springer Series in Statistics.

Lehrer, E. (1988), "An axiomatization of the Banzhaf value." *International Journal of Game Theory*, 17, 89–99.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*. Cambridge University Press.

Sánchez Rodríguez, E. and Vidal Puga, J. (2014), *Juegos Coalicionales*. Universidade de Vigo. Servizo de Publicacións, ed.

Shapley, L. S. (1953), "A value for n-person games." *Contributions to the Theory of Games*, 2, 307–317.