UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Julio 2025, Vol. 18, No. 1, pp. 1–27

Comparación de cinco modelos de *machine learning* para la predicción de las elecciones presidenciales en Colombia: una perspectiva con datos composicionales

Comparison of five machine learning models for the prediction of presidential elections in Colombia: a compositional data perspective

Paula Andrea Leal Varón.[a]
paulalealv@usantotomas.edu.co

Germán Andrés Galeano Ortiz.[b]
germangaleano@usantotomas.edu.co

Wilmer Pineda-Rios.[c]
wilmerpineda@usta.edu.co

Resumen

En los últimos años, numerosas investigaciones han empleado técnicas de *machine learning* y análisis de datos composicionales en distintos campos de estudio. Sin embargo, su integración en el análisis electoral sigue siendo escasa. Por tal razón, este trabajo integra ambos enfoques aplicando cinco modelos de machine learning: *random forest*, *gradient boosting*, *support vector machines*, *k-nearest neighbors*, y *feedforward neural networks*, para predecir los resultados de las elecciones presidenciales en Colombia a nivel municipal, considerando los datos como composicionales. Específicamente, se pronostica la distribución de votos de cada municipio en el espectro ideológico unidimensional Izquierda-Derecha. De esta forma, se busca no solo mejorar la precisión de las predicciones, sino también generar un avance importante en las metodologías aplicadas al análisis electoral. Los modelos se entrenaron con el 70 % de los datos de las elecciones presidenciales entre 2002 y 2022, y se evaluó su rendimiento en el 30 % restante. Los algoritmos mostraron desempeños similares entre las transformaciones de cada espectro ideológico con porcentajes de variabilidad entre el 56 % y 94 % en la predicción de la proporción de votos, destacándose el modelo de *feedforward neural networks* con la transformación log-cociente centrada, que alcanzó los mejores resultados.

–

***Palabras clave*:** elecciones precidenciales, *machine-learning*, *random forest*, *gradient boosting*, *support vector machines*, *k-nearest neighbors*, *feedforward neural networks*, datos composicionales, logaritmos de cocientes.

Abstract

In recent years, numerous studies have employed machine learning techniques and compositional data analysis in various fields of study. However, their integration into electoral analysis remains limited. For this reason, this work combines both approaches by applying five machine learning models: random forest, gradient boosting, support vector machines, k-nearest neighbors, and feedforward neural networks, to predict the results of the presidential elections in Colombia at the municipal level, onsidering the data as compositional. Specifically, it forecasts the vote distribution in each municipality along a unidimensional

[a]Estudiante, Maestría en Estadística Aplicada, Universidad Santo Tomás, Bogotá
[b]Estudiante, Maestría en Estadística Aplicada, Universidad Santo Tomás, Bogotá
[c]Docente, Facultad de Estadística, Universidad Santo Tomás, Bogotá

Left-Right ideological spectrum. This approach aims not only to improve prediction accuracy but also to comtribute a significant advancement in methodologies applied to electoral analysis. The models were trained on 70 % of the presidential election data from 2002 to 2022 and evaluated on the remaining 30 %. The algorithms demonstrated similar performance across transformations of each ideological spectrum, with variability percentages between 56 % and 94 % in predicting vote proportions, with the feedforward neural networks model using the centered log-ratio transformation achieving the best results. –

***Keywords*:** *presidential elections*, *machine-learning*, *random forest*, *gradient boosting*, *support vector machines*, *k-nearest neighbors*, *feedforward neural networks*, *compositional data*, *log-ratios*.

1. Introducción

En la actualidad, los algoritmos de *machine-learning* han tomado mayor visibilidad por sus múltiples resultados en la optimización de procesos y la toma decisiones confiables en diferentes áreas del conocimiento. Esto, debido a que pueden aprender de los datos sin depender de la programación basada en reglas. Su origen surge en el año 1950 cuando el gran Alan Turing creó el "Test de Turing" para determinar si una máquina era realmente inteligente. Para pasar el test, la máquina tenía que ser capaz de engañar a un humano haciéndole creer que era humana en lugar de un computador. Posteriormente, en el año 1952, Arthur Samuel escribió el primer programa de ordenador capaz de aprender; el software era un programa que jugaba a las damas y que mejoraba su juego partida tras partida (González 2019). Estos hechos, son el punto de partida de la continua evolución de una de las ramas de la inteligencia artificial.

En la sociedad, estudiar el comportamiento de la población de votantes, en cualquier proceso eleccionario, para predecir resultados futuros es un objetivo estudiado hace varios años por medios de comunicación y encuestas de opinión pública. Sin embargo, existen destacados hechos mundiales que han afectado la confiabilidad de estas entidades como: el triunfo de Donald Trump en Estados Unidos, la victoria del Brexit en Gran Bretaña, el alto abstencionismo en las elecciones municipales de Chile y el triunfo del No en el plebiscito de Colombia. En la literatura, se registran numerosos estudios que analizan el comportamiento electoral para predecir elecciones usando diferentes metodologías estadísticas entre las que sobresalen las redes neuronales, los modelos híbridos, el análisis de sentimientos y los algoritmos de *machine learning*.

En el mundo, muchas investigaciones han analizado y pronosticado los resultados electorales de algunos países, tomando como fuente de información los datos proporcionados por plataformas de redes sociales como Facebook y Twitter, y usando el aprendizaje de máquinas orientado en el análisis de sentimientos (Khan et al. 2021). Particularmente en Chile, (Santander et al. 2017) utilizó información de la red social Twitter para predecir con la mayor precisión posible los resultados electorales de las primarias, empleando técnicas de inteligencia computacional. En México, (Borges et al. 2016) analizó y comparó los algoritmos de K-vecinos, criterios de convergencia y la metodología de clasificación LAMDA para predecir las elecciones en el estado de Quintana Roo. De igual forma, (Aguilar López & Aquino López 2015) usaron la regresión de dirichlet para predecir las elecciones municipales de León de los Aldama.

En el contexto Colombiano, se han reportado algunos estudios para predecir el comportamiento electoral. Por ejemplo, (Cerón-Guzmán & León-Guzmán 2016) predijeron los resultados electorales del año 2014 a partir del análisis de sentimientos con la información proporcionada por las redes sociales. Igualmente, (Cuervo & Guerrero 2019) propusieron un modelo híbrido para predecir el desenlace de la primera vuelta en las elecciones presidenciales en 2018 cuyo objetivo era minimizar el error absoluto y mejorar la calidad de la predicción. Finalmente, (Baquero & Rosero 2019) estimaron los resultados de las elecciones presidenciales de segunda vuelta a nivel municipal utilizando los algoritmos de redes neuronales, *gradient boosting* y *random forest*; siendo *gradient boosting* el que presentó predicciones más cercanas a la realidad.

El análisis de datos composicionales es de especial interés debido a que permite estudiar la información en conjunto y no por separado, logrando una visión no sesgada de la realidad. En la literatura

colombiana, se evidencian algunos estudios en ciencias políticas para analizar el comportamiento de los resultados electorales tomando la fuente de información como un conjunto de datos composicionales. Entre estos, se destaca el análisis de los resultados del plebiscito por la paz en Colombia desarrollado por (Plata Rincón 2017) y el análisis de los procesos eleccionarios en Colombia usando la regresión dirichlet, un modelo lineal multivariado y un modelo mixto multivariado (Liscano Fierro 2017). Aunque es creciente el número de estudios que usan técnicas de *machine learning* y datos composicionales en diversos campos, su combinación en el sector político sigue siendo poco explorada. Este trabajo, contribuye de manera significativa al integrar ambos enfoques, no solo para mejorar la precisión de las predicciones, sino también como un avance importante en las metodologías aplicadas al análisis electoral. En particular, se evaluó el rendimiento cinco algoritmos de *machine learning*: *random forest*, *gradient boosting*, *support vector machines*, *k-nearest neighbors* y *feedforward neural networks*, junto a tres transformaciones log-cociente para predecir la distribución de votos de las elecciones presidenciales en el espectro ideológico unidimensional Izquierda-Derecha de cada municipio en Colombia.

2. Sistema político colombiano

Colombia es un Estado social de derecho, definido como una república unitaria y descentralizada con autonomía territorial, basado en principios democráticos, participativos y pluralistas. Estos valores, establecidos en el artículo 1 de la Constitución, reflejan una nación donde el poder político emana de los ciudadanos a través de su participación en elecciones. El voto es fundamental para la participación ciudadana, permitiendo a los colombianos influir en las decisiones que afectan a la sociedad y fortaleciendo el sistema democrático (De Colombia et al. 1991).

La legislación colombiana establece que en las elecciones generales se eligen mediante voto popular los siguientes cargos: presidente y vicepresidente de la república, senadores y representantes a la cámara. Estas elecciones se celebran cada cuatro años en años pares (por ejemplo, 2010, 2014, 2018, etc.). Se distinguen de las elecciones locales, que se llevan a cabo en años impares (por ejemplo, 2011, 2015, 2019, etc.). Las elecciones para el Congreso se convocan el segundo domingo de marzo, mientras que las de presidente y vicepresidente se realizan el último domingo de mayo (Barrios et al. 2018).

2.1. Presidente de la República

Las responsabilidades de cada uno de los cargos de elección popular están claramente definidas por la ley. La Constitución establece que todo funcionario público es responsable de cumplir sus funciones sin omitirlas ni excederse en ellas. El artículo 189 de la Constitución Política de Colombia detalla las funciones del Presidente de la República, quien actúa como Jefe de Estado, Jefe de Gobierno y Máxima Autoridad Administrativa.

Los criterios que deben satisfacer todos aquellos que aspiren a ocupar los cargos de Presidente y Vicepresidente están definidos en los artículos 191 y 197 de la Constitución Política de Colombia (De Colombia et al. 1991). Estas disposiciones se centran en establecer requisitos mínimos que garantizan la competencia y la capacidad de quienes aspiran a liderar la nación.

El sistema de elección prevé dos vueltas para las elecciones presidenciales. La primera vuelta se celebra el último domingo del mes de mayo y en ella podrán participar todos los candidatos que se hayan inscrito. Si en estas votaciones ninguno de los candidatos obtiene, por lo menos, la mitad más uno del total de votos depositados, se celebrará una nueva votación (segunda vuelta) en la que sólo participarán los dos candidatos que obtuvieron las votaciones más altas.

2.2. Partidos políticos

Los partidos políticos en Colombia cumplen un rol fundamental en el sistema democrático, actuando como entidades de interés público que representan diversos sectores de la sociedad y participan en el proceso electoral bajo principios e ideales específicos. Desde el siglo XIX, los partidos han sido vistos como organizaciones estructuradas con ideologías definidas. En 1848, Ezequiel Rojas defendió el liberalismo en su artículo "La Razón de mi Voto", apoyando a José Hilario López (Isaza 2009). Al año siguiente, Mariano Ospina Rodríguez y José Eusebio Caro publicaron el "Programa Conservador", estableciendo los principios del conservatismo (Partido Conservador 2021).

En otro sentido, no existe unanimidad sobre los orígenes de los partidos políticos en Colombia. Algunos textos como "Los partidos políticos en Colombia: entre la realidad y la ficción" de (Gechem Sarmiento 2009) señalan que los partidos aparecen como reflejo de las divisiones sociales que llegan al campo de lo político; es decir, para poder hablar de partidos políticos se requiere, por lo menos, dos organizaciones opuestas que trasladen a la escena políticas los grandes conflictos de la sociedad civil.

En la Ley 130 de 1994 se establece el estatuto básico de los partidos y movimientos políticos en Colombia. En el segundo artículo, se definen los partidos políticos como instituciones permanentes que reflejan el pluralismo político, promueven y canalizan la participación ciudadana, y contribuyen a la formación y expresión de la voluntad popular, con el objetivo de acceder al poder, a los cargos de elección popular y de influir en las decisiones políticas y democráticas de la Nación. Los movimientos políticos, por su parte, son asociaciones de ciudadanos que se constituyen libremente para influir en la formación de la voluntad política o para participar en las elecciones. Tanto los partidos como los movimientos políticos que cumplan con todos los requisitos constitucionales y legales obtendrán personería jurídica.

La Constitución Política de Colombia establece en al artículo 107 que los partidos y movimientos políticos deben organizarse democráticamente y regirse por principios fundamentales como la transparencia, objetividad, moralidad, equidad de género, y la obligación de presentar y divulgar sus programas políticos. Además, garantiza a todos los ciudadanos el derecho a fundar, organizar y desarrollar partidos y movimientos políticos, así como la libertad de afiliarse a ellos o retirarse. Se prohíbe a los ciudadanos pertenecer simultáneamente a más de un partido o movimiento político con personería jurídica. Por otra parte, los partidos y movimientos políticos son responsables de cualquier violación o incumplimiento de las normas que regulan su organización, funcionamiento o financiación. También, se establece que los dirigentes de los partidos y movimientos políticos deben fomentar procesos de democratización interna y fortalecer el régimen de bancadas (De Colombia et al. 1991).

2.3. Espectro ideológico

El espectro ideológico en política es una manera de categorizar y entender las diferentes posiciones y creencias políticas que pueden tener individuos, grupos, o partidos políticos. Estas posiciones se organizan en un continuo que va desde la extrema izquierda hasta la extrema derecha, y permite una clasificación más detallada de las diversas ideologías y orientaciones políticas.

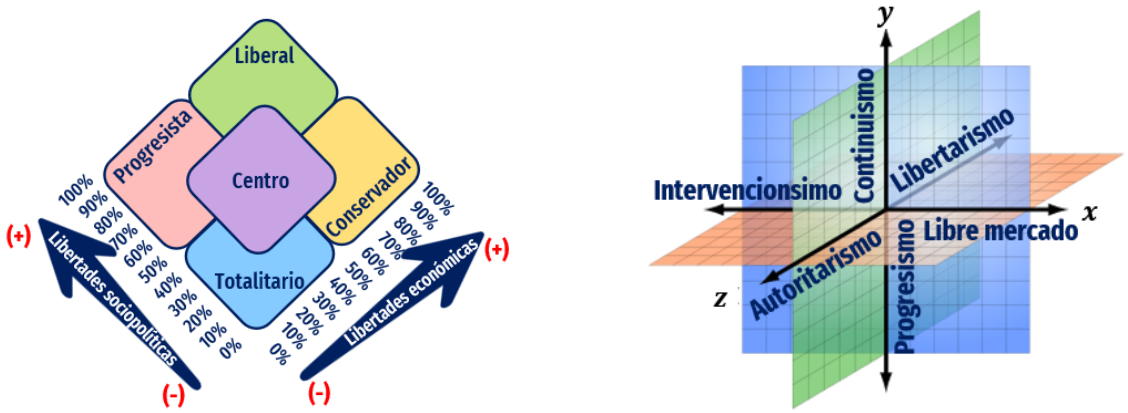
En la actualidad, la definición del espectro ideológico de un partido o movimiento político puede variar según el contexto político, histórico y social de cada país. Este espectro puede ser más complejo debido a la diversidad de estructuras organizativas y las circunstancias particulares de cada nación. La forma más común de definir el espectro ideológico es a través del eje unidimensional de Izquierda - Derecha (Ver Figura 1), este clasifica las posiciones ideológicas y políticas en función de su proximidad a dos extremos opuestos: la izquierda y la derecha. Este concepto surge durante la Revolución Francesa, donde los diputados se ubicaban en el lado izquierdo o derecho del parlamento según su postura respecto al rey y al cambio social, (Triglia 2015).

En este eje, las posiciones políticas de izquierda suelen asociarse con valores como la igualdad social, la justicia económica, la intervención estatal en la economía, y la defensa de los derechos sociales y civiles. Por otro lado, las posiciones de derecha suelen asociarse con valores como el libre mercado, el

Figura 1: Representación del eje unidimensional de Izquierda - Derecha. Elaboración propia.

individualismo, la menor intervención estatal en la economía y un mayor énfasis en la tradición y el orden social. Este enfoque tiende a simplificar las opciones para los electores y facilita la comunicación entre estos y los partidos políticos, (Triglia 2015).

El espectro ideológico también se puede analizar a través de modelos bidimensionales y tridimensionales que consideran varios ejes relevantes; El gráfico de Nolan, es una representación visual para clasificar las posiciones políticas en dos dimensiones: económica y social. El Modelo de Friesian. busca clasificar las ideologías políticas en tres ejes principales: preferencias políticas, económicas y sociales.



(a) Gráfico de Nolan

(b) Modelo de Friesian

Figura 2: Representación gráfica de los modelos bidimensional y tridimensional del espectro ideológico, respectivamente. Elaboración propia.

En Colombia, varios candidatos presidenciales han afirmado que no es posible analizar la política colombiana en términos de derecha e izquierda. Sin embargo, Luis Javier Orjuela en su artículo "Quién es quién en el espectro político colombiano" sostiene que esta afirmación es falsa, ya que la política siempre estará asociada a posiciones de izquierda y derecha, (Orjuela Escobar 2022).

Por lo tanto, considerando hechos históricos como las guerras de partidos, el reparto del poder durante el bipartidismo del Frente Nacional, la creación de grupos guerrilleros de extrema izquierda, los procesos de paz que condujeron a la Constitución de 1991, la reelección presidencial en dos periodos y el proceso de paz con la guerrilla de las FARC, se ha decidido adoptar para este trabajo un espectro ideológico unidimensional de Izquierda-Derecha.

3. De la parte al todo: datos composicionales

Los datos composicionales (DaCo) hacen referencia a cualquier vector $\mathbf{x}$ para el cual sus componentes pueden estar expresados como partes de un total. Un dato composicional $\mathbf{x} = (x_1, x_2, ..., x_D)^{'}$ con D partes, es un vector con componentes estrictamente positivas, tal que la suma de todas ellas es igual a una constante k. Su espacio muestral es el símplex S^D, definido por

$$S^D = \{(x_1, x_2, ..., x_D)^{'} : x_i > 0; \sum_{i=1}^{D} x_i = k\} \tag{1}$$

Para el caso $D = 3$, el símplex S^3 suele representarse mediante el diagrama ternario, triángulo equilátero de altura k (véase la Figura 5). Existe una correspondencia biunívoca entre los datos composicionales con 3 partes y los puntos del diagrama ternario. Un dato composicional $\mathbf{x} = (x_1, x_2, x_3)^{'}$ se corresponde con el punto que dista x_1, x_2 y x_3, respectivamente, de los lados opuestos a los vértices 1, 2 y 3.

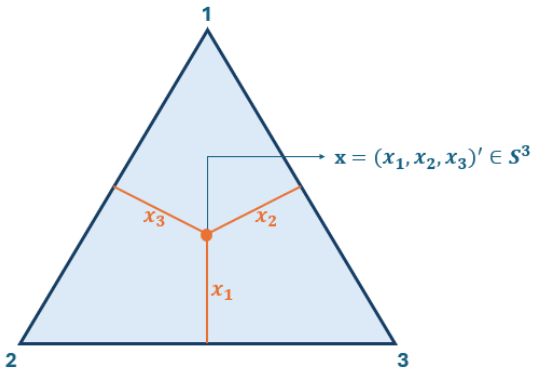


Figura 3: Representación de un dato composicional $(x_1, x_2, x_3)^{'}$ en el símplex S^3. Elaboración propia.

La esencia de los datos composicionales radica en que la geometría del espacio muestral, donde se definen los vectores de componentes, difiere de la clásica geometría euclidiana de $\mathbb{R}^D$. Esta diferencia fundamental impide la aplicación directa de técnicas multivariantes comúnmente utilizadas de tal forma que requiere el desarrollo de métodos estadísticos y transformaciones específicas que respeten la estructura relativa de los datos composicionales.

3.1. Geometría Aitchison

Los problemas del análisis estadístico de los datos composicionales fueron resueltos por primera vez en 1982 por Aitchison; quien argumentó que todas las dificultades de interpretación vienen motivadas por centrar la atención en las magnitudes absolutas de las partes $x_1, x_2, ..., x_D$ de una composición. La atención debe centrarse en la magnitud relativa de las partes; es decir, en los cocientes x_i/x_j con $i, j = 1, 2, ..., D$ y $i \neq j$. También, es importante mencionar que la geometría del espacio muestral sobre el que se define un vector de proporciones difiere de la geometría euclidiana clásica de $\mathbb{R}^D$. Por tal razón, las técnicas multivariantes convencionales que se fundamentan en dicha geometría, no son directamente aplicables a los datos composicionales.

A partir de los conceptos propuestos por (Aitchison 1982) se desarrolla la geometría de Aitchison que proporciona una base sólida para el análisis de datos composicionales, permitiendo el uso de técnicas estadísticas que respetan la naturaleza intrínseca de estos datos y evitando interpretaciones erróneas y las correlaciones espurias que pueden surgir al aplicar técnicas multivariantes tradicionales a datos composicionales.

3.2. Transformaciones log-cociente

La metodología de Aitchison se fundamenta en la transformación de datos composicionales al espacio real multivariante. Esto implica que el espacio muestral para los cocientes entre las partes se sitúa en el octante positivo de $\mathbb{R}^{D-1}$ y al calcular los logaritmos de estos cocientes, el espacio resultante también corresponde a $\mathbb{R}^{D-1}$. De este modo, se pueden aplicar las técnicas estadísticas convencionales de manera más adecuada.

Existen diversas posibilidades de transformación de los datos, todas ellas basadas en los logaritmos de cocientes entre las partes de un dato composicional. A continuación, se describen las tres transformaciones más comunes.

3.2.1. Transformación log-cociente aditiva

La transformación log-cociente aditiva (alr) de $\mathbf{x} \in S^D$ sobre $\mathbb{R}^{D-1}$ se define como:

$$\begin{aligned} \mathbf{w} = alr(\mathbf{x}) &= \left(ln\left(\frac{x_1}{x_D}\right), ln\left(\frac{x_2}{x_D}\right), ..., ln\left(\frac{x_{D-1}}{x_D}\right)\right) \\ &= (ln(x_1) - ln(x_D), ln(x_2) - ln(x_D), ..., ln(x_{D-1}) - ln(x_D)) \end{aligned} \tag{2}$$

La transformación inversa alr^{-1} está definida de $\mathbb{R}^{D-1}$ sobre $\mathbf{x} \in S^D$ de la siguiente forma:

$$\begin{aligned} \mathbf{x} = alr^{-1}(\mathbf{w}) &= \mathbf{C}(exp(w_1), exp(w_2), ..., exp(w_D), 1) \\ &= \mathbf{C}(\mathbf{x}) \end{aligned} \tag{3}$$

Esta trasformación es biyectiva pero no es simétrica en las partes de $\mathbf{x}$ ya que la parte del denominador adquiere un protagonismo especial respecto al resto. Es decir, un cambio en el orden de las componentes produce a su vez un cambio en el denominador de cada cociente. Este hecho condujo a (Aitchison 1982) a introducir la transformación log-cociente centrada.

3.2.2. Transformación log-cociente centrada

La transformación log-cociente centrada (clr) de $\mathbf{x} \in S^D$ sobre $\mathbb{R}^D$ se define como:

$$\begin{aligned} \mathbf{z} = clr(\mathbf{x}) &= \left(ln\left(\frac{x_1}{g(\mathbf{x})}\right), ln\left(\frac{x_2}{g(\mathbf{x})}\right), ..., ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right) \\ &= (ln(x_1) - ln(g(\mathbf{x})), ln(x_2) - ln(g(\mathbf{x})), ..., ln(x_D) - ln(g(\mathbf{x}))) \end{aligned} \tag{4}$$

donde

$$g(\mathbf{x}) = \left[\prod_{j=1}^{D} x_j\right]^{1/D} \tag{5}$$

es la media geométrica de las D partes de $\mathbf{x}$.

La transformación inversa clr^{-1} está definida de $\mathbb{R}^D$ sobre $\mathbf{x} \in S^D$ de la siguiente forma:

$$\begin{aligned} \mathbf{x} = clr^{-1}(\mathbf{z}) &= \mathbf{C}(exp(z_1), exp(z_2), ..., exp(z_D)) \\ &= \mathbf{C}(\mathbf{x}) \end{aligned} \tag{6}$$

Esta trasformación es biyectiva y simétrica entre las partes. Su imagen es el hiperplano de $\mathbb{R}^D$ que pasa por el origen y es ortogonal al vector de unidades; encontrándose aquí una nueva dificultad ya que la suma de los componentes del vector transformado es igual a cero. A raíz de esta dificultad (Egozcue et al. 2003) define una isometría entre los espacios S^D y $\mathbb{R}^{D-1}$ con el propósito de resolver los inconvenientes de las dos transformaciones anteriores.

3.2.3. Transformación log-cociente isométrica

Dada una base ortonormal $e_1, e_2, ..., e_{D-1}$ del símplex S^D, se define la transformación log-cociente isométrica (ilr) de una composición $\mathbf{x} \in S^D$ sobre $\mathbb{R}^{D-1}$ como:

$$\mathbf{y} = ilr(\mathbf{x}) = (\langle \mathbf{x}, e_1 \rangle_a, \langle \mathbf{x}, e_2 \rangle_a, ..., \langle \mathbf{x}, e_{D-1} \rangle_a) \tag{7}$$

Los componentes del vector ilr transformado son las coordenadas de la composición $\mathbf{x}$ respecto a la base $e_1, e_2, ..., e_{D-1}$. Esta transformación, como su nombre lo indica, es isométrica, y los datos ilr transformados se representan en los habituales ejes ortogonales. La transformación inversa ilr^{-1} está definida de $\mathbb{R}^{D-1}$ sobre $\mathbf{x} \in S^D$ de la siguiente forma:

$$\mathbf{x} = ilr^{-1}(\mathbf{y}) = \bigoplus_{j=1}^{D-1} y \otimes e_j \tag{8}$$

Está transformación presenta dificultades en determinar cuál es la base ortonormal más adecuada para un problema específico, aquella que proporcione expresiones que faciliten la interpretación de los resultados. Las coordenadas ilr no suelen ser fáciles de interpretar, por lo que una opción es utilizar las coordenadas y expresar los resultados en la base canónica de $\mathbb{R}^D$ sin abandonar el símplex.

4. *Machine learning (ML)*

El *machine learning* o aprendizaje automático es un conjunto de métodos que pueden detectar automáticamente patrones en los datos y luego usarlos para predecir o clasificar; es decir, se trata de hacer que las computadoras modifiquen o adapten sus acciones para que estas acciones sean más exactas, donde la exactitud se mide por qué tan bien las acciones elegidas reflejan la correcta (ver, por ejemplo (Marsland 2011)).

Cualquier tarea de *machine learning* se puede dividir en una serie de tareas más sencillas. (Marsland 2011) describe 5 pasos para aplicar el aprendizaje automático; *recolectar los datos*, *explorar y preparar los datos*, *entrenar el modelo*, *evaluar el rendimiento del modelo* y *mejorar el rendimiento del modelo.*

En *machine learning* son múltiples los tipos de problemas que se pueden resolver. Sin embargo, es necesario distinguir la clasificación de los algoritmos para abordar de forma correcta el problema. Existen muchas formas de clasificar los modelos, pero principalmente existen dos tipos: paramétricos y no paramétricos. Los modelos paramétricos tienen la ventaja de ser a menudo más rápidos de usar, pero la desventaja de hacer suposiciones más sólidas sobre la naturaleza de las distribuciones de los datos. Los modelos no paramétricos son más flexibles, pero pueden ser computacionalmente costosos con grandes volúmenes de datos. Otro enfoque de clasificación de algoritmos, los divide en tres categorías: supervisados, no supervisados y por refuerzo. Este trabajo se enfoca en los algoritmos supervisados, cuyo objetivo según (Murphy 2012) en su libro "Machine Learning A Probabilistic Perspective" es aprender una relación entre entradas y salidas a partir de un conjunto de entrenamiento. En problemas supervisados, la salida puede ser categórica, en cuyo caso se trata de un problema de clasificación, o continua, lo cual corresponde a un problema de regresión.

4.1. Modelos propuestos

En el presente estudio, se exploran cinco modelos de *machine learning* para la predicción de las elecciones presidenciales en Colombia. Los modelos seleccionados son:

1. ***Support Vector Regression* (SVR)** se destaca por su capacidad de ajustar los datos dentro de un margen de tolerancia especifico, permitiendo un enfoque en maximizar el margen de error para predecir valores continuos.
2. ***Random Forest Regression* (RFR)** permite al combinar múltiples árboles de decisión manejar datos complejos y ruidos, además, genera predicciones robustas con una interpretación clara de la importancia de las variables.
3. ***Gradient Boosting Regression* (GBR)** sobresale por su capacidad de corregir errores sucesivos mediante la optimización de las predicciones con una secuencia de árboles de decisión, especialmente en escenarios donde son relevantes las pequeñas variaciones en los datos.
4. ***K-Nearest Neighbors Regression* (k-NN)** se destaca por su simplicidad y efectividad en capturar patrones dado que esta basado en la proximidad de puntos similares.
5. ***Feedforward Neural Networks* (FNN)** es reconocido por su capacidad para captar relaciones complejas mediante la combinación de múltiples capas y neuronas, permitiendo capturar dinámicas ocultas en los datos.

5. Predicción de las elecciones presidenciales en Colombia

Este proceso abarca la recolección, exploración y preparación de los datos, así como el entrenamiento, evaluación e interpretación de los modelos. El objetivo de analizar el comportamiento electoral aplicando y evaluando de manera efectiva los cinco modelos (SVR, RFR, GBR, k-NN y FNN) junto con las tres transformaciones log-cociente (*alr*, *clr*, *ilr*) para predecir la distribución de votos de las elecciones presidenciales en el espectro ideológico unidimensional Izquierda-Derecha de cada municipio en Colombia.

5.1. Recolección de los datos: desde el voto hasta el dato

Para predecir la distribución de votos dentro del espectro ideológico unidimensional de Izquierda-Derecha de las elecciones presidenciales de Colombia a nivel municipal, utilizando técnicas de *machine learning* con datos composicionales, se analizaron 139,235 observaciones correspondientes a los datos electorales municipales desde 1994 hasta 2022, proporcionados por la Registraduría Nacional del Estado Civil. Estos datos incluyen el total de votos en cada municipio y su distribución entre votos en blanco, nulos, no marcados, y los votos para cada candidato, tanto en primera como en segunda vuelta, según corresponda.

Además, se tienen 38 variables auxiliares disponibles, que abarcan indicadores relacionados con finanzas públicas, educación, conflicto armado y seguridad, salud, así como demografía y población. Estos datos, están desagregados por municipios desde el año 2000 y fueron obtenidos de la plataforma TerriData, lo que permitió un análisis detallado y robusto de las dinámicas territoriales y su impacto en los resultados electorales.

La clasificación ideológica de los candidatos presidenciales en cada año electoral, se fundamento en la consulta de diversas fuentes bibliográficas correspondientes a cada candidato en el año respectivo. En los casos donde no se encontró información clara o suficiente sobre la orientación política del candidato, se optó por asignar una posición neutral en el espectro ideológico, colocándolo en Centro. Esta metodología busca ofrecer una clasificación equilibrada y fundamentada, evitando suposiciones sin respaldo documental cuando los datos disponibles son limitados.

5.2. Exploración y preparación de los datos: radiografía del voto

Inicialmente, se analizó la distribución de observaciones (municipios) con registro de votos por año electoral, se observa que entre 1994 y 2002 hubo un aumento notable en el número de municipios, pasando de 999 a 1,096, lo cual puede atribuirse a la creación de nuevos municipios y la expansión de la cobertura electoral en zonas previamente afectadas por problemas de infraestructura o conflicto armado. A partir de 2002, el número de municipios se estabiliza en torno a 1,103, manteniéndose constante hasta 2022. Por esta razón, el análisis se centrará en los datos de 2002 a 2022, ya que este período ofrece mayor estabilidad en la cantidad de municipios; además, las variables auxiliares disponibles cubren desde el año 2000.

Para el año 2002, faltaban datos electorales de cinco municipios (Norosí, Guachené, Medio Atrato, San José de Uré y Tuchín) y para 2006, de cuatro municipios (Norosí, Guachené, San José de Uré y Tuchín) ya que todos fueron creados después de estos años. Para completar esta información, se identificó el municipio "padre" al que pertenecían antes de su constitución como municipios independientes. Posteriormente, se calculó la media de la proporción de votos entre el nuevo municipio y su municipio "padre", tomando como referencia los resultados electorales de los años posteriores a su creación. Esto permitió generar una distribución promedio de votos que fue aplicada a las elecciones sin datos. Finalmente, el total de votos para estos municipios se estimó en función de la distribución obtenida y la proporción de votos del municipio padre. Esta estrategia implementada en *Python*, se utilizó porque los municipios nuevos heredan las características demográficas y electorales de sus territorios de origen, lo que permitió aprovechar la relación histórica de votos entre los municipios recién creados y sus municipios "padre" para generar estimaciones precisas y coherentes.

También, en las elecciones de 2002, debido a las condiciones del conflicto armado, la guerrilla impidió el desarrollo de los comicios en 14 municipios (Murindó, Alto Baudó, Pisba, La Salina, Recetor, Sácama, Santa Rosa, Pulí, Calamar, Miraflores, El Calvario, San Juanito, Magüí y Carurú) mediante la destrucción del material de votación. Además, se reportaron combates, hostigamientos contra la fuerza pública, intentos de atentados terroristas y bloqueos de vías. Entonces, para completar la información de estos municipios, en *Python* se realizó una estimación del número de votos utilizando el método de *Holt-Winters*, un algoritmo iterativo que pronostica el comportamiento de la serie en base a promedios ponderados de los datos anteriores. De tal forma, que se tomó los resultados de las elecciones posteriores y se reorganizaron los años de forma descendente para estimar la cantidad de votos y se determinó la proporción de votos por espectro ideológico empleando el promedio de la proporción de votos de los otros años.

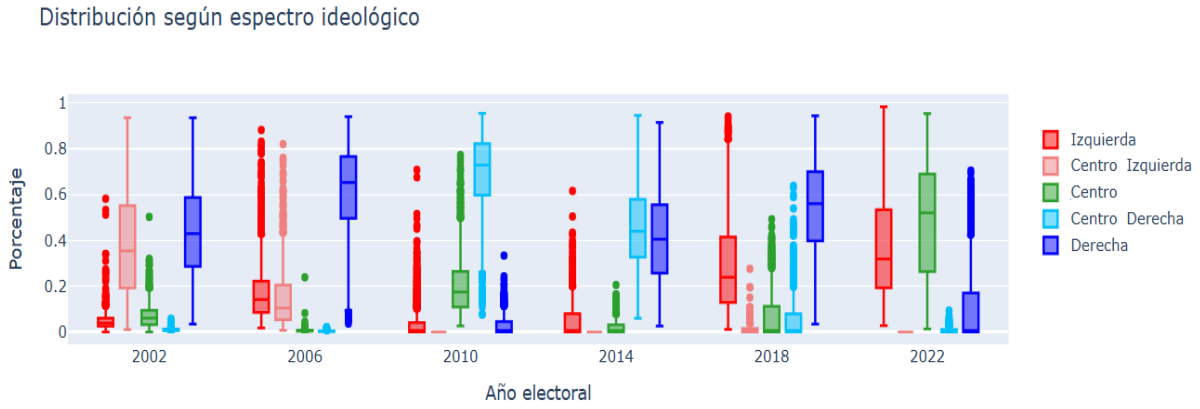


Figura 4: Distribución de votos de los espectros ideológicos en las elecciones presidenciales (2002-2022). Elaboración propia.

La Figura 4, muestra la distribución de votos entre los diferentes espectros ideológicos, los cuales han

variado significativamente con el tiempo. En 2002 y 2010, el apoyo a la Izquierda fue limitado debido a la política de seguridad democrática bajo el presidente Álvaro Uribe y Juan Manuel Santos, que consolidó una tendencia de apoyo a propuestas de Derecha y Centro-Derecha, respectivamente. En 2006, la dispersión de votos aumentó debido al liderazgo de Carlos Gaviria, reflejando un creciente cuestionamiento sobre la estrategia militar en la población civil en algunos sectores. En 2018 y 2022, el apoyo creció considerablemente, alcanzando su máximo en 2022 con el liderazgo de Gustavo Petro, quien promovió temas sociales y políticas progresistas.

En el pensamiento de Centro-Izquierda, en 2002 hubo un apoyo notable impulsado por figuras como Ingrid Betancourt y Horacio Serpa, quienes consolidaron una coalición alternativa con propuestas de cambio sin una inclinación radical. Sin embargo, este apoyo disminuyó en 2006 debido al crecimiento de la Izquierda. A partir de 2010, el apoyo casi desapareció, posiblemente por la polarización política y el fortalecimiento de la Derecha y la Izquierda.

Asimismo, para el Centro, en 2002 y 2006 los votos fueron bajos y presentaron poca dispersión; en cambio, en 2010 hubo un aumento notable con la candidatura de Antanas Mockus. Sin embargo, el apoyo disminuyó nuevamente en 2014 y 2018, con un aumento en 2022 debido a la polarización política, que impulsó el interés en una opción equilibrada. Por otro lado, Centro-Derecha tuvo un apoyo bajo en 2002 y 2006, pero aumentó significativamente en 2010 con Juan Manuel Santos, quien continuó las políticas de Uribe con un enfoque moderado. Este apoyo disminuyó en 2018 y 2022 debido a la preferencia por opciones más polarizadas.

Finalmente, la Derecha, en 2002 y 2006 contaba con un apoyo sólido, pero cayó en 2010 con la transición de liderazgo. Desde 2014, el apoyo se recuperó parcialmente, impulsado por discursos de seguridad, aunque en 2022 el respaldo disminuyó nuevamente, mostrando una fragmentación en el electorado.

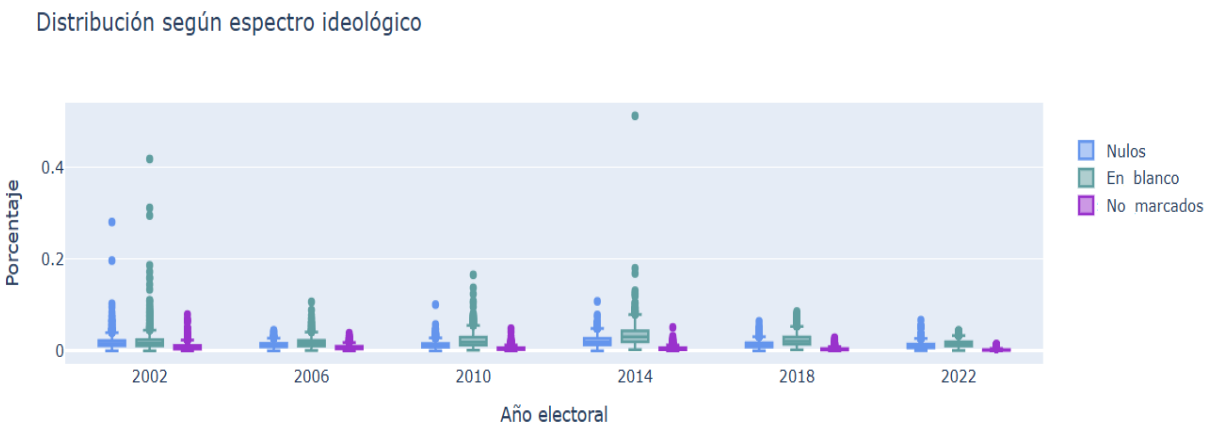


Figura 5: Distribución de votos de en blanco, nulos y no marcados en las elecciones presidenciales (2002-2022). Elaboración propia.

La Figura 5, muestra que los votos nulos, en blanco y no marcados presentan patrones específicos. En 2002, muestran una alta dispersión con algunos municipios registrando altos porcentajes, reflejando descontento y falta de opciones representativas. A partir de 2006, los porcentajes se estabilizaron en niveles bajos, probablemente gracias a mejoras en la educación electoral y la consolidación de opciones políticas claras. Sin embargo, mantienen una proporción baja en las elecciones analizadas, de tal forma que se decidió excluirlos del estudio, enfocándose exclusivamente en los votos del espectro ideológico Izquierda-Derecha para resaltar las preferencias ideológicas de los votantes.

Las 38 variables auxiliares disponibles, que abarcan indicadores de finanzas públicas, educación, conflicto armado y seguridad, salud, demografía y población, fueron analizadas preliminarmente calculando las correlaciones. En este proceso, se identificaron 15 variables con una correlación superior al 96 % y posteriormente se analizó la varianza, encontrando 3 variables con varianza cercana a cero. Estas variables

fueron eliminadas antes de proceder con la imputación de datos faltantes.

Los valores faltantes fueron imputados utilizando el método *mice (Multiple Imputation by Chained Equations)* en *R-Studio*, empleando el enfoque "*norm.predict*" que ajusta modelos de regresión lineal para predecir estos valores a partir de las otras variables en el conjunto de datos. Sin embargo, algunas variables seguían presentando valores vacíos, por lo que se aplicó el método de *Holt-Winters* para estimar aquellos valores. Es importante mencionar que 8 municipios de los 1,103 no presentan indicadores, expluyéndolos del estudio. Finalmente, los indicadores seleccionados para el estudio se presentan en la siguiente tabla.

Tabla 1: Clasificación de los indicadores según su dimensión.

Dimensión	**Indicador**
Educación	- Cobertura neta en educación - Total
Finanzas públicas	- % de ingresos corrientes destinados a funcionamiento. - % de ingresos corrientes que corresponden a recursos propios. - % de ingresos que corresponden a transferencias. - % del gasto total destinado a inversión. - Déficit o superávit total. - Gastos corrientes per cápita. - Gastos totales per cápita. - Ingresos corrientes per cápita. - Gastos de funcionamiento per cápita. - Ingresos corrientes per cápita. - Ingresos no tributarios per cápita. - Ingresos per cápita por impuesto a la Industria y al comercio. - Ingresos per cápita por impuesto predial. - Ingresos tributarios per cápita. - Transferencias de los ingresos corrientes. - Transferencias per cápita de los ingresos corrientes.
Conflicto armado y seguridad	- Eventos de minas antipersonal.
Demografía y población	- Hombres mayores de edad.
Salud	- Tasa de mortalidad infantil en menores de 5 años. - Tasa de mortalidad (x cada 1.000 habitantes).

Nota. Esta tabla presenta la clasificación de los indicadores según una temática específica (Tomado de `https://terridata.dnp.gov.co/`).

A continuación, se presentan algunos gráficos que muestran el comportamiento de indicadores de interés, contrastados con diferentes espectros ideológicos con el objetivo de analizar cómo estos varían en función de las posiciones ideológicas, permitiendo identificar algunas tendencias.

En la Figura 6, se representa la variabilidad del promedio de votos en los espectros ideológicos según los deciles de la tasa de mortalidad por cada 1,000 habitantes. El espectro de Izquierda muestra una caída inicial en áreas de menor mortalidad, pero lo incrementa en los deciles altos, atrayendo a votantes en contextos de mayor vulnerabilidad que buscan reformas sociales significativas. El Centro Izquierda mantiene una tendencia descendente uniforme con una leve recuperación al final captando a perfiles moderados en áreas más afectadas. El Centro, por otro lado, gana apoyo constantemente, especialmente en los deciles altos, probablemente por su enfoque pragmático y en la preferencia de propuestas equilibradas y no polarizadas. La Centro Derecha tiene un pico temprano y luego disminuye, reflejando un apoyo en áreas de mortalidad intermedia en donde prevalece la preocupación por el orden y la estabilidad. La Derecha crece al inicio, alcanza su máximo cerca del decil 7 y desciende después, mostrando menor apoyo en zonas de alta mortalidad donde aumenta la demanda de cambios estructurales.

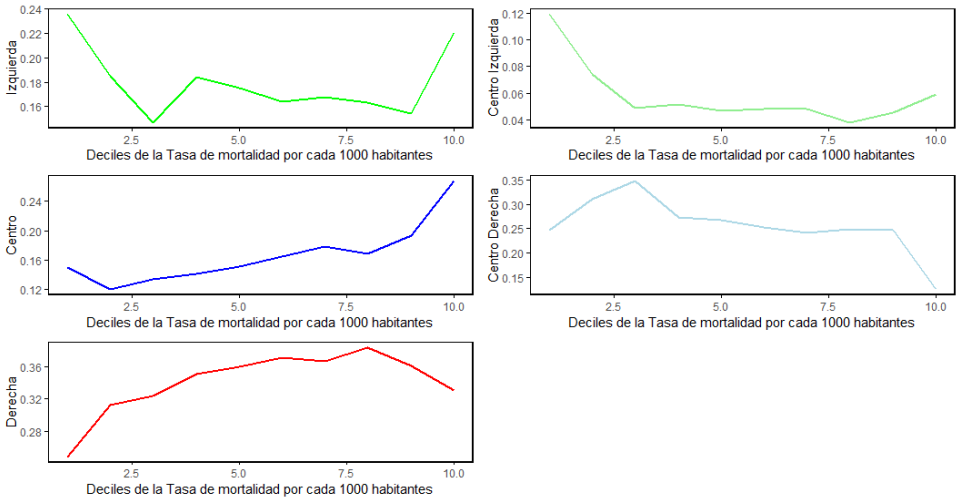


Figura 6: Relación entre la tasa de mortalidad y el promedio de votos de cada espectro ideológico. Elaboración propia.

En la Figura 7, se observa la relación entre el número de hombres mayores de edad y el promedio de votos de cada espectro ideológico. La Izquierda gana apoyo constante en los deciles superiores, lo que indica una mayor identificación de los hombres mayores de edad con cambios sociales y redistribución económica. En contraste, el Centro Izquierda tiene un patrón irregular con una tendencia descendente en los deciles superiores, indicando una menor conexión de esta idologia con las prioridades de los hombres mayores de edad. El Centro, se recupera gradualmente después de una caída inicial, sugiriendo que su enfoque equilibrado es atractivo en los deciles más altos. La Centro Derecha alcanza un pico en los primeros deciles y luego disminuye, indicando menor atractivo para hombres mayores de edad. La Derecha, por su parte, pierde apoyo constantemente en los deciles superiores, reflejando una desconexión progresiva de los hombres mayores hacia sus políticas, prefiriendo enfoques que atienden mejor las necesidades de estabilidad social.

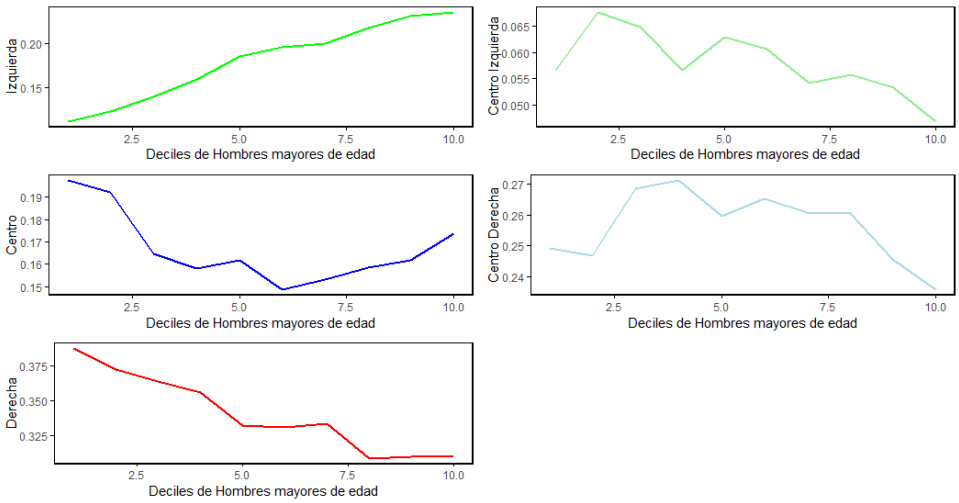


Figura 7: Relación entre el número de hombres mayores de edad y el promedio de votos de cada espectro ideológico. Elaboración propia.

La Figura 8, muestra cómo varía el promedio de votos en los espectros ideológicos según los deciles de gasto en inversión. El espectro de Izquierda duplica su apoyo, de 0.15 en los primeros deciles a 0.30 en

el decil más alto, sugiriendo que un mayor gasto en inversión pública favorece el respaldo a políticas de corte izquierdistas. El Centro Izquierda experimenta una caída constante, disminuye de 0.11 a casi 0, reflejando una pérdida de apoyo ante políticas de inversión más fuertes, posiblemente reflejando una percepción de insuficiencia en sus propuestas ante políticas de inversión más robustas. El Centro aumenta de 0.14 a 0.18 en los primeros deciles y luego se estabiliza, mostrando atractivo en contextos de inversión media. La Centro Derecha alcanza su punto máximo en el decil 3 (0.27) y luego desciende, indicando menor atractivo con alta inversión. La Derecha muestra una tendencia de caída constante, desciende de 0.36 a 0.25, lo que sugiere una disminución en su atractivo en escenarios de mayor inversión pública.

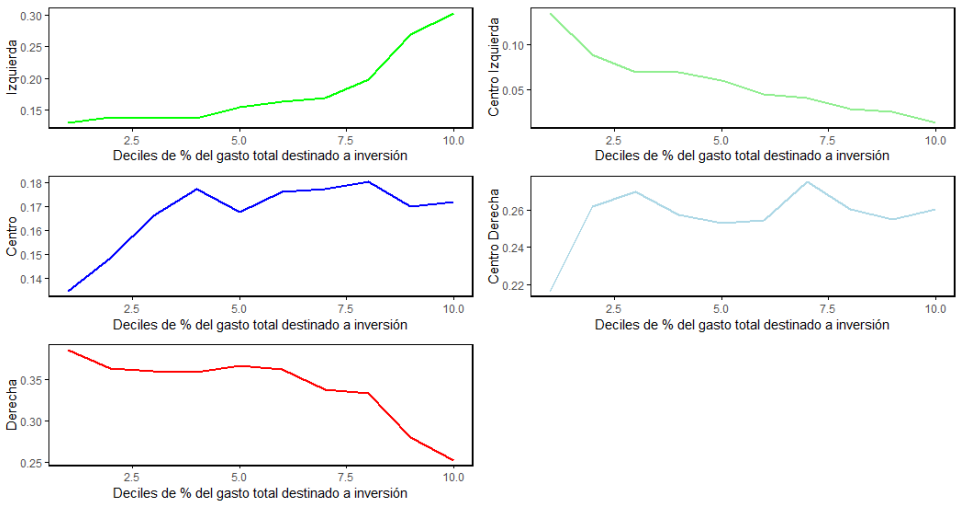


Figura 8: Relación entre el porcentaje del gasto total destinado a inversión y el promedio de votos de cada espectro ideológico. Elaboración propia.

Ahora, se presenta el análisis de las elecciones presidenciales desde una perspectiva composicional permitiendo explorar cómo se distribuyen las preferencias electorales entre los distintos espectros ideológicos.

(a) Elecciones 2018-2022. (b) Elecciones totales.

Figura 9: Diagramas ternarios de distribución ideológica (Centro, Derecha y Centro Derecha) para las elecciones 2018-2022 y totales, tomando el espectro de Centro como el punto fijo. Elaboración propia.

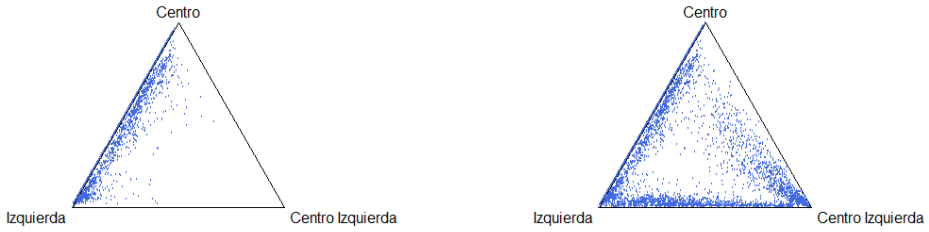
La Figura 9, presenta los diagramas ternarios que muestran las dinámicas ideológicas de Centro, Centro Derecha y Derecha en las elecciones de 2018 y 2022 (a) frente al total histórico de elecciones (b). En el gráfico de 2018 y 2022, se observa una dispersión significativa de los votos hacia las posiciones de Derecha y Centro, lo que sugiere una fragmentación del electorado hacia posturas más conservadoras en

estos años específicos. En contraste, el gráfico que resume el total de las elecciones muestra una marcada concentración de los votos en el Centro Derecha y Derecha, lo que indica que históricamente las posiciones con inclinaciones derechistas han sido predominantes en el panorama electoral. Este análisis, revela una evolución en las preferencias ideológicas de los votantes, con un electorado recientemente más dividido y menos cohesionado, mostrando una mayor diversidad en sus inclinaciones conservadoras, con preferencia por posturas moderadas.

(a) Elecciones 2018-2022. (b) Elecciones totales.

Figura 10: Diagramas ternarios de distribución ideológica (Centro, Derecha, Izquierda) para las elecciones 2018-2022 y totales, tomando el espectro de Centro como el punto fijo. Elaboración propia.

La Figura 10, presenta en el primer diagrama una mayor dispersión hacia el Centro y la Izquierda, lo cual sugiere que en este periodo reciente hubo una inclinación hacia posiciones moderadas y de Izquierda. Por el contrario, el segundo diagrama que abarca las elecciones históricas, muestra una distribución más equilibrada entre los tres espectros ideológicos: Centro, Derecha e Izquierda. Esta mayor uniformidad en el segundo gráfico indica que a lo largo del tiempo los votantes han estado repartidos de forma más diversa entre las distintas ideologías, sin una preferencia tan marcada hacia el Centro o la Izquierda como en el periodo 2018-2022.



(a) Elecciones 2018-2022. (b) Elecciones totales.

Figura 11: Diagramas ternarios de distribución ideológica (Centro, Izquierda, Centro Izquierda) para las elecciones 2018-2022 y totales, tomando el espectro de Centro como el punto fijo. Elaboración propia.

El primer diagrama ternario (a) que se presenta en la Figura 11, muestra que los votos están concentrados hacia el vértice del Centro e Izquierda, indicando una inclinación mayoritaria del electorado hacia posiciones de Izquierda en este periodo reciente. Por otro lado, el segundo diagrama (b) que representa

las elecciones históricas, la dispersión es más amplia extendiéndose de manera más uniforme entre los tres espectros ideológicos. Esto refleja que a lo largo del tiempo, los votantes han mostrado una mayor diversidad en sus preferencias ideológicas, distribuyéndose no solo en el Centro sino también en la Izquierda y el Centro Izquierda. En términos generales, los diagramas reflejan que en elecciones recientes el electorado ha mostrado una preferencia más marcada por posiciones centradas y de Izquierda, mientras que en el contexto histórico ha existido una mayor pluralidad ideológica.

5.3. Entrenamiento y evaluación: estimando la predicción electoral

A continuación, se presentan los análisis de los modelos aplicados a 1,095 municipios de Colombia para predecir la distribución de votos en el espectro ideológico unidimensional de Izquierda-Derecha a partir de 22 variables independientes (20 indicadores, el periodo de elección (primera/segunda vuelta) y el departamento). Se emplean cinco modelos de *machine learning* en combinación con tres transformaciones log-cociente, evaluando cuál es más adecuado para representar la distribución ideológica. Los modelos se entrenaron con datos de las elecciones presidenciales entre 2002 y 2018, y se evaluaron con los resultados de 2022.

5.3.1. *Support Vector Regression*

En el caso del *support vector regression*, se entrenaron 364 modelos cada uno con una configuración distinta de parámetros. Los parámetros iterados incluyeron la función *kernel*, el parámetro *gamma* (que define el alcance de influencia de cada punto, aplicable solo a *kernels* no lineales) y el parámetro de costo. Se evaluaron tres tipos de *kernels*: radial, sigmoidal y polinomial, con el objetivo de determinar cuál proporcionaba el mejor ajuste para cada espectro ideológico, se optimizó el parámetro *gamma*, testeando con valores de 0.001, 0.01, 0.1, 1 y 10, y se evaluaron valores de 1000, 100, 10, 1 y 0.1 para el costo. A continuación, se presentan las configuraciones óptimas seleccionadas para cada espectro ideológico con cada una de las transformaciones.

Tabla 2: Configuraciones óptimas para cada espectro ideológico en cada transformación.

Modelo	Espectro ideológico	Transformación	*kernel*	Gamma	Costo
SVR	Izquierda	*arl*	Radial	0,01	100
	Centro Izquierda			0,001	1000
	Centro Derecha			0,001	1000
	Derecha			0,01	100
	Izquierda	*irl*		0,001	1000
	Centro Izquierda			0,001	100
	Centro Derecha			0,001	1000
	Derecha			0,001	1000
	Izquierda	*clr*		0,01	100
	Centro Izquierda			0,001	1000
	Centro Derecha			0,001	1000
	Derecha			0,001	100
	Centro			0,01	100

Nota. Las transformaciones *alr* e *ilr* presentan configuración para cuatro espectros ideológicos debido a que al aplicarlas, se reduce la dimensionalidad de la composición. Elaboración propia.

En la Figura 12, se presenta el rendimiento del modelo en términos de R^2 y $RMSE$ para diferentes transformaciones log-cociente (*alr*, *clr* e *ilr*) en los diferentes espectros ideológicos. En general, se observa que la transformación *clr* tiende a mejorar el R^2 en comparación con las otras transformaciones en la mayoría de los espectros, indicando que esta transformación permite al modelo capturar mejor la variabilidad de los datos. Por ejemplo, en el espectro de Centro la transformación *clr* tiene un R^2 de 0.898 superior al de *alr* e *ilr*, lo que sugiere un ajuste superior en este caso. En el espectro de Centro Derecha,

la transformación *clr* también muestra el mejor R^2 (0.923) junto con el menor $RMSE$ (0.088) indicando tanto una buena capacidad explicativa como una menor magnitud de error de predicción. Sin embargo, en el espectro de Derecha aunque *alr* y *ilr* tienen valores de R^2 similares (0.833 y 0.832 respectivamente), el $RMSE$ de *alr* es ligeramente menor, lo que muestra que esta transformación podría ser preferible en este caso. En resumen, la transformación *clr* parece ser la más adecuada en general para maximizar el ajuste del modelo, aunque en ciertos espectros ideológicos específicos, otras transformaciones también ofrecen resultados competitivos en términos de error y variabilidad explicada.

Izquierda		
Transformación	RMSE	R2
alr	0,084	0,838
clr	0,085	0,834
ilr	0,089	0,818

Centro - Izquierda		
Transformación	RMSE	R2
alr	0,078	0,740
clr	0,075	0,762
ilr	0,075	0,763

Centro		
Transformación	RMSE	R2
alr	0,074	0,893
clr	0,072	0,898
ilr	0,079	0,879

Centro - Derecha		
Transformación	RMSE	R2
alr	0,092	0,918
clr	0,088	0,923
ilr	0,093	0,914

Derecha		
Transformación	RMSE	R2
alr	0,118	0,833
clr	0,120	0,828
ilr	0,119	0,832

Figura 12: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. Elaboración propia.

5.3.2. *Random Forest Regression*

Para este algoritmo se evaluaron 234 modelos, probando diferentes combinaciones de hiperparámetros para optimizar su rendimiento. En cuanto al número de árboles (*ntree*) se configuraron 500 y 1000, para el número de variables a considerar en cada división (*mtry*) se probaron valores de 6, 8 y 10, y para el número mínimo de observaciones en cada nodo terminal (*nodesize*) se configuraron valores de 10, 15 y 20. También, se utilizó el criterio de importancia *IncNodePurity* (incremento en la pureza del nodo) para evaluar la relevancia de las variables en cada una de las configuraciones; este criterio permitió identificar las variables que más contribuyen a mejorar la pureza en las divisiones de los nodos en el modelo reflejando así su importancia en la predicción.

Tabla 3: Top cinco de las variables más importantes.

Variable	Top
Periodo	1
% de ingresos corrientes que corresponden a recursos propios	2
Gastos totales per cápita	3
Transferencias de los ingresos corrientes	4
Ingresos tributarios per cápita	5

Nota. Estas son las variables que mostraron una mayor frecuencia de aparición como las más importantes en los modelos para cada transformación en cada espectro ideológico. Elaboración propia.

La Tabla 3, presenta el top 5 de las variables que mostraron una mayor frecuencia de aparición como las más importantes en los modelos diseñados para cada transformación en cada espectro ideológico. Las variables identificadas están relacionadas con el contexto de los modelos ya que la mayoría están ligadas a la capacidad económica y fiscal; indicadores clave para entender la dinámica en los diferentes espectros ideológicos. La variable "Periodo" es importante porque refleja los cambios temporales que pueden influir en el comportamiento de otros indicadores económicos. El " % de ingresos corrientes que corresponden a recursos propios" y los "ingresos tributarios per cápita" son indicadores de la autonomía fiscal, lo cual es relevante para evaluar la sostenibilidad financiera y la capacidad de generación de ingresos de cada sector. Los "gastos totales per cápita" ofrecen una medida del gasto público y la distribución de

recursos impactando directamente en el bienestar social y, por ende, en la percepción ideológica de los ciudadanos. Finalmente, las "transferencias de los ingresos corrientes" representan fondos recibidos, que pueden depender de políticas centralizadas o descentralizadas; un aspecto importante que también puede variar según el contexto ideológico. Estas variables son esenciales para comprender el impacto financiero y económico en cada espectro ideológico, lo que justifica su alta frecuencia en los modelos.

Las configuraciones óptimas seleccionadas para cada espectro ideológico con cada una de las transformaciones se observan en la Tabla 4.

Tabla 4: Configuraciones óptimas para cada espectro ideológico en cada transformación.

Modelo	**Espectro ideológico**	**Transformación**	***ntree***	***mtry***	***nodesize***
RFR	Izquierda	*arl*	500	10	10
	Centro Izquierda		1000	10	10
	Centro Derecha		1000	10	10
	Derecha		500	10	10
	Izquierda	*irl*	500	10	10
	Centro Izquierda		1000	10	10
	Centro Derecha		1000	10	10
	Derecha		1000	10	10
	Izquierda	*clr*	500	6	10
	Centro Izquierda		1000	10	10
	Centro Derecha		500	10	10
	Derecha		500	10	10
	Centro		500	10	10

Nota. Las transformaciones *alr* e *ilr* presentan configuración para cuatro espectros ideológicos debido a que al aplicarlas, se reduce la dimensionalidad de la composición. Elaboración propia.

El rendimiento del modelo en términos de R^2 y $RMSE$ para las tres transformaciones log-cocientes en los cinco espectros ideológicos se muestran en la Figura 13. En el espectro de Centro Derecha, la transformación *alr* logra el mejor ajuste, con un R^2 de 0.920 y un $RMSE$ bajo de 0.090 indicando tanto una alta capacidad explicativa como un error de predicción reducido. En el espectro de Centro, *alr* también proporciona buenos resultados con un R^2 de 0.906 y un $RMSE$ más bajo (0.070) sugiriendo que esta transformación captura adecuadamente las relaciones en este grupo. Para los otros espectros, las transformaciones *clr* e *ilr* presentan un rendimiento más competitivo. En particular, para el espectro de Derecha, la transformación *ilr* obtiene un R^2 de 0.792 y un $RMSE$ de 0.132 mostrando un equilibrio entre precisión y ajuste. En términos generales, la transformación *alr* parece ser la más efectiva para maximizar el ajuste en varios espectros, aunque *clr* e *ilr* también ofrecen resultados satisfactorios dependiendo del espectro ideológico específico.

Izquierda		
Transformación	RMSE	R2
alr	0,098	0,778
clr	0,110	0,723
ilr	0,105	0,747

Centro - Izquierda		
Transformación	RMSE	R2
alr	0,080	0,727
clr	0,078	0,742
ilr	0,078	0,737

Centro		
Transformación	RMSE	R2
alr	0,070	0,906
clr	0,087	0,854
ilr	0,083	0,866

Centro - Derecha		
Transformación	RMSE	R2
alr	0,090	0,920
clr	0,096	0,909
ilr	0,096	0,910

Derecha		
Transformación	RMSE	R2
alr	0,134	0,785
clr	0,134	0,789
ilr	0,132	0,792

Figura 13: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. Elaboración propia.

5.3.3. *Gradient Boosting Regression*

En el caso de *gradient boosting regression*, se entrenaron 1,053 modelos cada uno con una combinación distinta de parámetros para optimizar el rendimiento. Los rangos de los parámetros incluyeron valores de tasa de aprendizaje (*shrinkage*) de 0.01, 0.1 y 0.3, permitiendo ajustar la velocidad con la que el modelo aprende, la profundidad de interacción (*interaction.depth*) de 1, 5 y 10 para controlar la complejidad de las interacciones entre variables en cada árbol, el número mínimo de observaciones por nodo (*n.minobsinnode*) de 5, 10 y 15, los valores de fracción de muestreo (*bag.fraction*) de 0.65, 0.8 y finalmente el número total de árboles (*ntree*). A continuación, se presentan las configuraciones óptimas seleccionadas para cada espectro ideológico con cada una de las transformaciones.

Tabla 5: Configuraciones óptimas para cada espectro ideológico en cada transformación.

Modelo	**Espectro ideológico**	**Transformación**	***n.t***	***i.d***	***n.m***	***s-k***	***b.f***
GBR	Izquierda	*arl*	499	10	15	0,1	0,8
	Centro Izquierda		492	5	5	0,1	0,8
	Centro Derecha		435	10	5	0,1	0,8
	Derecha		499	10	15	0,1	0,8
	Izquierda	*irl*	488	10	10	0,1	0,8
	Centro Izquierda		341	10	5	0,1	0,8
	Centro Derecha		493	10	15	0,1	0,8
	Derecha		482	10	5	0,1	0,8
	Izquierda	*clr*	495	10	5	0,1	0,8
	Centro Izquierda		246	10	5	0,1	0,8
	Centro Derecha		379	10	5	0,1	0,8
	Derecha		498	10	5	0,1	0,8
	Centro		481	10	5	0,1	0,8

Nota. Las transformaciones *alr* e *ilr* presentan configuración para cuatro espectros ideológicos debido a que al aplicarlas, se reduce la dimensionalidad de la composición. (*n.trees (n.t), interaction.depth (i.d), n.minobsinnode (n.m), shrin- kage (s-k), bag.fraction (b.f)*. Elaboración propia.

El rendimiento del modelo en términos de R^2 y $RMSE$ para las transformaciones log-cocientes en los cinco espectros ideológicos se muestran en la Figura 14. En general, la transformación clr ofrece el mejor rendimiento en la mayoría de los grupos, logrando los valores más bajos de $RMSE$ y los más altos de R^2 en Izquierda, Centro Izquierda, Centro Derecha y Derecha, lo que indica un mejor ajuste en esos casos. En particular, cabe destacar el rendimiento para el espectro Centro Derecha, la transformación *clr* obtiene un R^2 de 0.939 y un $RMSE$ de 0.079, lo que muestra un equilibrio entre precisión y ajuste. En el espectro de Centro, la transformación *alr* logra el mejor ajuste con un R^2 de 0.908 y un $RMSE$ bajo de 0.068, indicando tanto una alta capacidad explicativa como un error de predicción reducido.

Izquierda

Transformación	RMSE	R2
alr	0,087	0,827
clr	0,084	0,830
ilr	0,086	0,828

Centro - Izquierda

Transformación	RMSE	R2
alr	0,077	0,746
clr	0,075	0,759
ilr	0,075	0,756

Centro

Transformación	RMSE	R2
alr	0,068	0,908
clr	0,071	0,903
ilr	0,072	0,900

Centro - Derecha

Transformación	RMSE	R2
alr	0,079	0,938
clr	0,079	0,939
ilr	0,080	0,937

Derecha

Transformación	RMSE	R2
alr	0,117	0,838
clr	0,112	0,849
ilr	0,114	0,845

Figura 14: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. Elaboración propia.

5.3.4. *K-Nearest Neighbors Regression*

En este algoritmo se entrenaron 351 modelos con distintas combinaciones de parámetros. En primer lugar, se evaluaron diferentes tipos de *kernel* (rectangular, triangular y *epanechnikov*) para analizar cómo influye la ponderación de las observaciones cercanas en la predicción y finalmente se especificó el tipo de distancia que el modelo utilizaría para calcular la proximidad entre observaciones: Euclidiana, *Manhattan* y *Minkowski*. A continuación, se presentan las configuraciones óptimas seleccionadas para cada espectro ideológico con cada una de las transformaciones.

Tabla 6: Configuraciones óptimas para cada espectro ideológico en cada transformación.

Modelo	Espectro ideológico	Transformación	Vecinos	Distancia	*Kernel*
KNN	Izquierda	*arl*	15	Minkowski	Triangular
	Centro Izquierda		10	Minkowski	Triangular
	Centro Derecha		10	Minkowski	Triangular
	Derecha		10	Euclidiana	Rectangular
	Izquierda	*irl*	15	Minkowski	Epanechnikov
	Centro Izquierda		10	Minkowski	Triangular
	Centro Derecha		10	Minkowski	Triangular
	Derecha		10	Minkowski	Triangular
	Izquierda	*clr*	15	Minkowski	Epanechnikov
	Centro Izquierda		10	Euclidiana	Rectangular
	Centro Derecha		10	Minkowski	Triangular
	Derecha		15	Minkowski	Triangular
	Centro		10	Minkowski	Triangular

Nota. Las transformaciones *alr* e *ilr* presentan configuración para cuatro espectros ideológicos debido a que al aplicarlas, se reduce la dimensionalidad de la composición. Elaboración propia.

La Figura 15, presenta el rendimiento del modelo en términos de R^2 y $RMSE$ para las transformaciones log-cociente en los cinco espectros ideológicos. En Centro Derecha, la transformación *clr* destaca con un R^2 de 0.886 y un $RMSE$ más bajo de 0.108 indicando un buen ajuste y una baja magnitud de error, haciendo que esta transformación sea la más efectiva para este grupo. En el espectro de Centro, la transformación *ilr* alcanza el R^2 más alto (0.818) y un $RMSE$ relativamente bajo (0.097), sugiriendo que esta transformación permite capturar adecuadamente las relaciones en este espectro. En el espectro de Derecha, *clr* también muestra un buen rendimiento con un R^2 de 0.776 y un $RMSE$ de 0.140 superando en rendimiento a las otras transformaciones en este grupo. En los espectros de Izquierda y Centro Izquierda ninguna transformación destaca claramente en cuanto a ajuste y error, aunque *clr* e *ilr* presentan R^2 y $RMSE$ similares, lo que muestra que ambas pueden capturar información relevante en estos espectros. En general, *clr* parece ofrecer los mejores resultados en términos de equilibrio entre precisión y capacidad explicativa en la mayoría de los espectros ideológicos.

Izquierda		
Transformación	RMSE	R2
alr	0,133	0,648
clr	0,104	0,755
ilr	0,104	0,756

Centro - Izquierda		
Transformación	RMSE	R2
alr	0,086	0,686
clr	0,085	0,696
ilr	0,085	0,689

Centro		
Transformación	RMSE	R2
alr	0,104	0,795
clr	0,097	0,817
ilr	0,097	0,818

Centro - Derecha		
Transformación	RMSE	R2
alr	0,135	0,832
clr	0,108	0,886
ilr	0,108	0,885

Derecha		
Transformación	RMSE	R2
alr	0,203	0,562
clr	0,140	0,776
ilr	0,141	0,772

Figura 15: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. Elaboración propia.

5.3.5. *Feedforward Neural Networks*

En el caso de las *feedforward neural networks*, se entrenaron 195 redes diferentes variando los parámetros para optimizar su rendimiento. Los parámetros iterados son el número de épocas (*epochs*) con los siguientes valores: 50, 100 y 500, el número de capas ocultas (*hidden layers*) con el número de neuronas que se probaron con las siguientes configuraciones : (10, 10), (20, 10), (50, 25), (100, 50, 25) y (150, 100, 50) y la función de activación ReLU y ReLU *with dropou*. A continuación, se presentan las configuraciones óptimas seleccionadas para cada espectro ideológico con cada una de las transformaciones.

Tabla 7: Configuraciones óptimas para cada espectro ideológico en cada transformación.

Modelo	**Espectro ideológico**	**Transformación**	**Función de activación**	***Epochs***	***Hidden***
FNN	Izquierda	*arl*	ReLU	50	(150,100,50)
	Centro Izquierda			100	(10,10)
	Centro Derecha			500	(10,10)
	Derecha			50	(100,50,25)
	Izquierda	*irl*		50	(150,100,50)
	Centro Izquierda			50	(150,100,50)
	Centro Derecha			500	(10,10)
	Derecha			50	(150,100,50)
	Izquierda	*clr*		50	(150,100,50)
	Centro Izquierda			100	(10,10)
	Centro Derecha			500	(10,10)
	Derecha			50	(150,100,50)
	Centro			50	(100,50,25)

Nota. Las transformaciones *alr* e *ilr* presentan configuración para cuatro espectros ideológicos debido a que al aplicarlas, se reduce la dimensionalidad de la composición. Elaboración propia.

Las tablas en la Figura 16, comparan las transformaciones en cuanto a su desempeño en los diferentes espectros ideológicos. Se observa que en Centro la transformación *clr* sobresale con un R^2 de 0.907 y un $RMSE$ bajo de 0.069, indicando una excelente capacidad del modelo para explicar la variabilidad de los datos y mantener un bajo error de predicción. En el espectro de Centro Derecha, nuevamente *clr* se destaca alcanzando un R^2 de 0.940 y un $RMSE$ de 0.078 confirmando su efectividad en capturar las características de este grupo ideológico. Por otro lado, para Derecha aunque *clr* muestra un rendimiento competitivo con un R^2 de 0.838 el $RMSE$ es ligeramente superior (0.117) comparado con los otros espectros, sugiriendo una mayor complejidad en los datos de este espectro. En los espectros de Izquierda y Centro Izquierda, *clr* también obtiene los valores más altos de R^2 en comparación con las otras transformaciones, destacándose especialmente en Centro Izquierda con un R^2 de 0.750.

Izquierda		
Transformación	RMSE	R2
alr	0,092	0,810
clr	0,087	0,831
ilr	0,086	0,831

Centro - Izquierda		
Transformación	RMSE	R2
alr	0,085	0,694
clr	0,078	0,750
ilr	0,084	0,702

Centro		
Transformación	RMSE	R2
alr	0,068	0,910
clr	0,069	0,907
ilr	0,074	0,893

Centro - Derecha		
Transformación	RMSE	R2
alr	0,085	0,930
clr	0,078	0,940
ilr	0,085	0,929

Derecha		
Transformación	RMSE	R2
alr	0,130	0,812
clr	0,117	0,838
ilr	0,122	0,825

Figura 16: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. Elaboración propia.

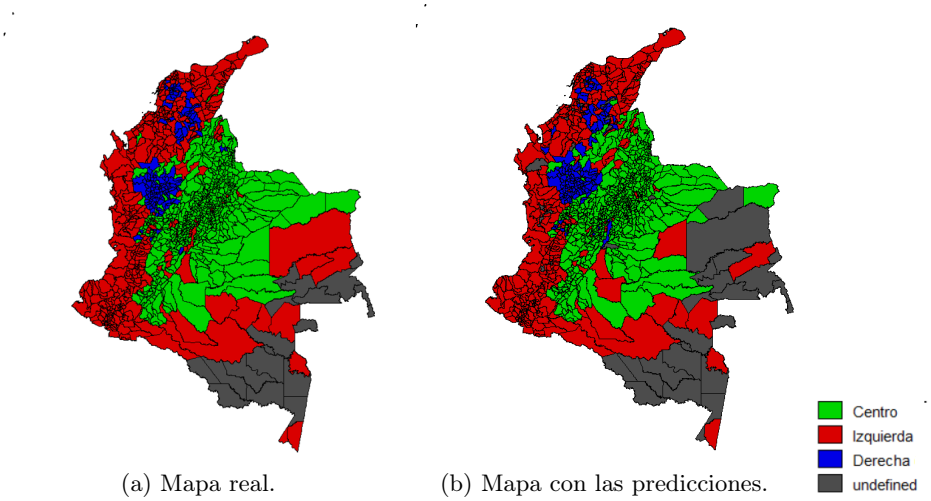
La Figura 17, presenta el consolidado del rendimiento de los modelos de *machine learning*: *support vector regression*, *random forest regression*, *gradient boosting regression*, *K-nearest neighbors regression* y *feedforward neural network* en términos de R^2 y $RMSE$ aplicados a las tres transformaciones log-cociente (*alr*, *clr* y *ilr*) para los cinco espectros ideológicos. Inicialmente, se puede observar que el *feedforward neural network* con transformación *clr* es el modelo más efectivo en los espectros de Izquierda y Centro Derecha, y en Centro con la transformación *alr* logrando un alto R^2 y un $RMSE$ bajo, mientras que en Centro Izquierda y Derecha el *gradient boosting regression* con transformación *clr* es el más preciso. En términos generales, el *feedforward neural network* con transformación *clr* es una buena elección para varios espectros con ajustes específicos en Centro Izquierda, Centro Derecha y Derecha. También, se observa que la transformación *clr* obtuvo los mejores resultados en varios modelos y espectros, destacándose en Centro Izquierda, Centro Derecha y Derecha, especialmente con el modelo *gradient boosting regression*, *K-nearest neighbors regression* y *feedforward neural network*.

Modelo	Izquierda			Centro - Izquierda			Centro			Centro - Derecha			Derecha		
	Transformación	RMSE	R2	Transformación	RMSE	R2	Transformación	RMSE	R2	Transformación	RMSE	R2	Transformación	RMSE	R2
Support Vector Regression	alr	0,084	0,838	alr	0,078	0,740	alr	0,074	0,893	alr	0,092	0,918	alr	0,118	0,833
	clr	0,085	0,834	clr	0,075	0,762	clr	0,072	0,898	clr	0,088	0,923	clr	0,120	0,828
	ilr	0,089	0,818	ilr	0,075	0,763	ilr	0,079	0,879	ilr	0,093	0,914	ilr	0,119	0,832
Random Forest Regression	alr	0,098	0,778	alr	0,080	0,727	alr	0,070	0,906	alr	0,090	0,920	alr	0,134	0,785
	clr	0,110	0,723	clr	0,078	0,742	clr	0,087	0,854	clr	0,096	0,909	clr	0,134	0,789
	ilr	0,105	0,747	ilr	0,078	0,737	ilr	0,083	0,866	ilr	0,096	0,910	ilr	0,132	0,792
Gradient Boosting Regression	alr	0,087	0,827	alr	0,077	0,746	alr	0,068	0,908	alr	0,079	0,938	alr	0,117	0,838
	clr	0,084	0,830	clr	0,075	0,759	clr	0,071	0,903	clr	0,079	0,939	clr	0,112	0,849
	ilr	0,086	0,828	ilr	0,075	0,756	ilr	0,072	0,900	ilr	0,080	0,937	ilr	0,114	0,845
K-Nearest Neighbors Regression	alr	0,133	0,648	alr	0,086	0,686	alr	0,104	0,795	alr	0,135	0,832	alr	0,203	0,562
	clr	0,104	0,755	clr	0,085	0,696	clr	0,097	0,817	clr	0,108	0,886	clr	0,140	0,776
	ilr	0,104	0,756	ilr	0,085	0,689	ilr	0,097	0,818	ilr	0,108	0,885	ilr	0,141	0,772
Feedforward Neural Network	alr	0,092	0,810	alr	0,085	0,694	alr	0,068	0,910	alr	0,085	0,930	alr	0,130	0,812
	clr	0,087	0,831	clr	0,078	0,750	clr	0,069	0,907	clr	0,078	0,940	clr	0,117	0,838
	ilr	0,086	0,831	ilr	0,084	0,702	ilr	0,074	0,893	ilr	0,085	0,929	ilr	0,122	0,825

Figura 17: Consolidado de las métricas R^2 y $RMSE$ de los modelos para cada transformación en cada espectro ideológico. Elaboración propia.

A continuación, se presenta el mapa de Colombia que muestra los espectros ideológicos ganadores por municipio en la primera vuelta de las elecciones presidenciales de 2022, junto con el mapa de la predicción usando el modelo *feedforward neural network* con la transformación *clr*.

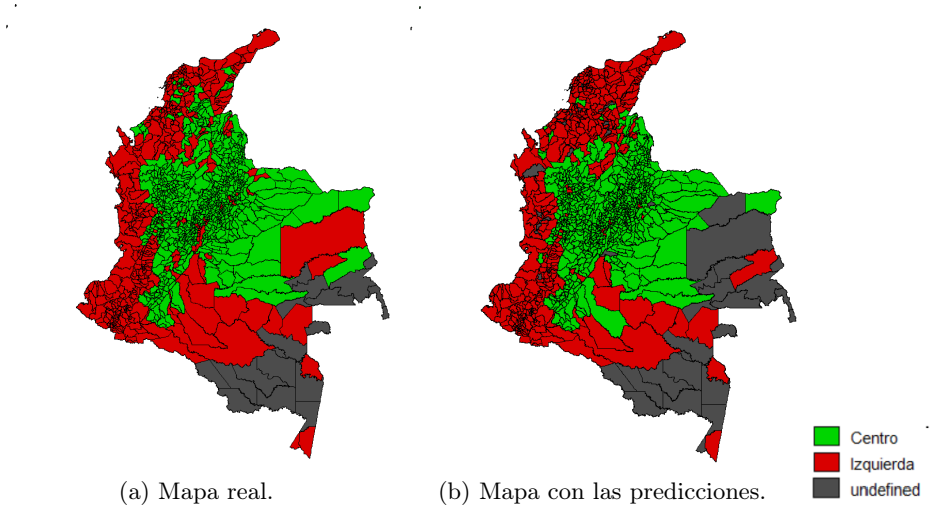
Es importante señalar para el análisis que los municipios en color gris en el mapa (a) representan las zonas no municipalizadas, mientras que en el mapa (b) corresponden tanto a las zonas no municipalizadas como a los municipios que no fueron estimados debido a la falta de reportes de indicadores. La Figura 18, muestra que el modelo logra predecir correctamente el espectro ideológico ganador para la primera vuelta en aproximadamente el 90 % de los municipios. Por ejemplo, en ciudades capitales como: Bogotá, Tunja y Armenia, las predicciones no coinciden con los resultados reales de las elecciones de 2022. En Bogotá, donde ganó la Izquierda, el modelo predijo Derecha; en Tunja, donde también ganó la Izquierda, el modelo predijo Centro; y en Armenia, donde ganó el Centro, el modelo predijo Izquierda. Estas discrepancias podrían deberse a factores específicos de cada capital, como la complejidad de sus dinámicas socioeconómicas. La ciudad de Bogotá frecuentemente es epicentro de polarización política dadas las diferentes posiciones ideológicas lo cual genera patrones de voto que varían significativamente entre localidades. Este comportamiento se puede observar al analizar las estimaciones; la estimación de votos para el espectro de Izquierda de 40.8 % y para Derecha de 45.5 %. De igual forma, en Tunja se observa polarización con una estimación de votos del 39 % para el espectro de Izquierda y un 45.7 % para el espectro de Centro. En cambio, para Armenia se observa una distribución de los votos entre los espectros Izquierda, Centro y Derecha, con una estimación de votos del 36 % para el espectro de Izquierda, 31.1 % para el espectro Centro y un 31.2 % para el espectro de Derecha.



(a) Mapa real. (b) Mapa con las predicciones.

Figura 18: Mapas del espectro ideológico ganador en cada municipio de Colombia para la primera vuelta de las elecciones presidenciales 2022. Elaboración propia.

El análisis anterior también se lleva a cabo para la segunda vuelta de las elecciones presidenciales de 2022. La Figura 19, muestra cómo el modelo logra predecir correctamente el espectro ideológico ganador en aproximadamente el 89 % de los municipios. Por ejemplo, Bogotá sigue siendo una excepción notable, ya que la predicción del modelo para la capital no coincide con el resultado real de la elección. Esta discrepancia resalta los desafíos que enfrenta el modelo para capturar la complejidad y las particularidades de la ciudad. Así, a pesar del buen desempeño del modelo a nivel general, su capacidad predictiva en Bogotá se ve limitada debido a la propia naturaleza de la ciudad y al gran número de votantes que representa a nivel nacional.



(a) Mapa real. (b) Mapa con las predicciones.

Figura 19: Mapas del espectro ideológico ganador en cada municipio de Colombia para la segunda vuelta de las elecciones presidenciales 2022. Elaboración propia.

Posteriormente, se presenta el mapa de Colombia que muestra la predicción usando el modelo *feedforward neural network* con la transformación *clr* de los espectros ideológicos ganadores por municipio para las elecciones presidenciales de 2026. Para realizar el pronóstico, se estimaron los indicadores por medio del método de *Holt Winter*, obteniendo un valor estimado de lo que podría presentar para este proceso electoral.

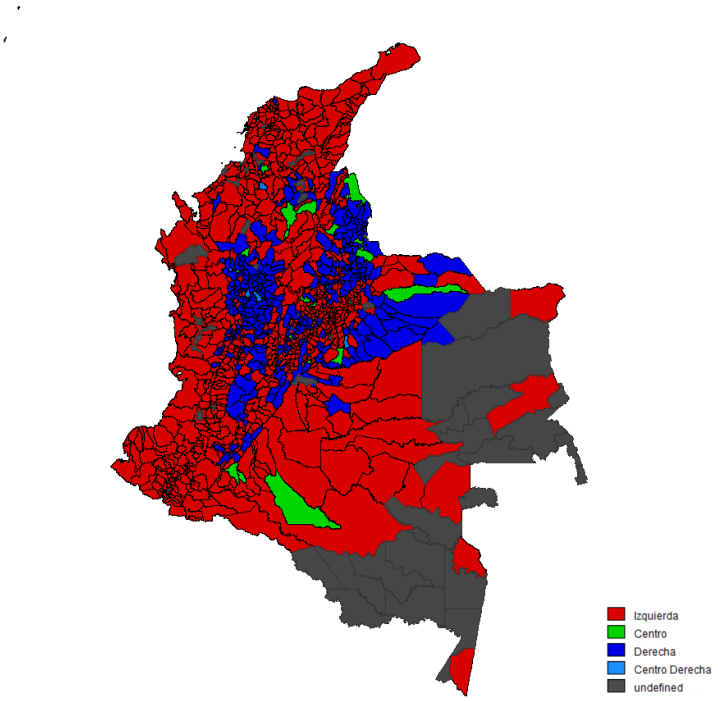


Figura 20: Mapas del espectro ideológico ganador en cada municipio de Colombia para la primera vuelta de las elecciones presidenciales 2026. Elaboración propia.

La Figura 20, muestra un dominio significativo de la Izquierda en las elecciones presidenciales de 2026, alcanzando el 66.2 % de los municipios de Colombia, lo cual refleja en gran medida la continuidad de la preferencia que obtuvo en las elecciones de 2022. Este dominio no solo se mantiene, sino que también se amplía en regiones como la Costa Atlántica, el Magdalena Medio y la Orinoquía, áreas donde la presencia del conflicto armado ha sido históricamente más evidente y puede influir en las tendencias de voto hacia posturas de cambio o justicia social que suelen asociarse con la Izquierda. En los municipios donde la Izquierda predomina, la proporción promedio de votos para este espectro alcanza el 59 %, seguido de la Derecha con un 24.2 %. Por otro lado, en el 31 % de los municipios donde la Derecha obtiene la mayoría, la proporción promedio de votos es del 48.9 %, con un 28.9 % de apoyo para la Izquierda. Este análisis muestra una posible estructura de preferencia polarizada en el país, con áreas específicas donde las inclinaciones ideológicas están marcadas, influenciadas tanto por factores históricos y sociales como por las dinámicas propias de cada región.

6. Conclusiones

Este trabajo contribuye al entendimiento de las elecciones presidenciales en Colombia al explorar el comportamiento de los votantes y su impacto en la predicción de resultados electorales. Durante años, predecir estos resultados ha sido un objetivo de interés para medios de comunicación y estudios de opinión pública. Hoy en día, la literatura científica ha dado pasos significativos al integrar técnicas avanzadas, como la minería de datos y el análisis de datos composicionales, para profundizar en los patrones de comportamiento electoral. Sin embargo, hasta la fecha son escasos los estudios que combinan de manera integral los enfoques de *machine learning* y datos composicionales para analizar y predecir tendencias de voto. De tal forma que este trabajo no solo llena ese vacío en la investigación, sino que también ofrece un modelo robusto para capturar y entender la complejidad del comportamiento electoral en un contexto cambiante, donde las dinámicas sociopolíticas pueden influir significativamente en las preferencias de los

votantes. Al incorporar técnicas de *machine learning* con datos composicionales, se abre una nueva vía de análisis que permite observar matices en la distribución del voto, proporcionando un marco innovador para futuras investigaciones en el ámbito de la elecciones políticas.

Las transformaciones log-cociente para datos composicionales aplicadas a los resultados de las elecciones presidenciales fueron esenciales para resolver los problemas de dependencia y reducir la colinealidad, mejorando así la interpretabilidad de los análisis. En el ámbito electoral, esto resulta particularmente importante, ya que las proporciones de votos distribuidas en el espectro ideológico unidimensional Izquierda-Derecha están restringidas a una suma constante, lo cual complicaba la aplicación de métodos tradicionales. En este estudio, se concluye que aunque los resultados entre las transformaciones son similares, la transformación *clr* es la que ofrece los mejores resultados, destacándose en los cinco modelos para los distintos espectros ideológicos. Esta ventaja se debe a que la transformación *alr* y la *ilr*, requieren la elección de un componente de referencia o una base ortonormal, siendo el espectro de Centro la opción utilizada en este caso, mientras que la *clr* no depende de estos elementos, evitando sesgos y proporcionando resultados más robustos y coherentes. Además, la *clr* ofrece una representación simétrica y equilibrada de las proporciones, lo que facilita la interpretación directa de las relaciones entre las categorías ideológicas en el contexto electoral. A su vez, elimina eficazmente la colinealidad inherente en los datos composicionales, mejorando la estabilidad y precisión del modelo sin la necesidad de ajustes adicionales, lo que no siempre es posible con las otras transformaciones.

La imputación de datos faltantes para las variables que describen características generales de los municipios representa un desafío significativo para la precisión y confiabilidad de las estimaciones en el análisis electoral. La falta de información en algunos municipios de Colombia, donde no se obtuvo ningún registro en los procesos electorales estudiados, introduce un alto grado de incertidumbre en las proyecciones y modelos. La ausencia de datos esenciales, como los datos de finanzas públicas, educación, conflicto armado y seguridad, salud, así como demografía y población de estos municipios, limita la capacidad del modelo para reflejar con precisión las tendencias electorales y puede influir en el sesgo de las predicciones. Además, una mayor tasa de datos faltantes no solo eleva la incertidumbre en los resultados, sino que también complica la imputación misma, ya que se reduce la cantidad de datos disponibles para extrapolar valores realistas en los municipios menos documentados.

La evaluación de los cinco modelos de *machine learning* para predecir los resultados de las elecciones presidenciales en Colombia, aplicando un enfoque composicional, permitió identificar diferencias fundamentales en su rendimiento. Entre ellos, el modelo *feedforward neural network* destacó por ofrecer los mejores resultados en las métricas de $RMSE$ y R^2, lo que sugiere una capacidad superior para captar patrones complejos en los datos electorales composicionales y adaptarse a las particularidades de las proporciones de voto. En segundo lugar, el *gradient boosting regression* y el *support vector regression* demostraron un rendimiento robusto, aunque algo menor en comparación con el modelo de red neuronal. En contraste, los modelos *K-nearest neighbors regression* y *random forest regression* presentaron desempeños inferiores en las métricas utilizadas, probablemente estos modelos no atraen correctamente la dependencia inherente entre los espectros ideológicos. Por lo tanto, se destaca el *feedforward neural network* como el más adecuado para capturar la complejidad de los patrones de voto en el contexto colombiano.

Aunque no se encontraron estudios que combinen modelos de *machine learning* y datos composicionales en el ámbito electoral, es relevante mencionar los resultados obtenidos en investigaciones previas como la de (Baquero & Rosero 2019) y (Castaño-Gómez et al. 2019), donde los modelos de *gradient boosting* y *decision tree* respectivamente presentaron los mejores resultados para predecir las elecciones presidenciales en Colombia, en nuestro caso, el modelo *feedforward neural network* se destacó, particularmente al trabajar con datos composicionales, lo que sugiere que este tipo de redes neuronales están ganando relevancia en este contexto. A pesar de ello, el *gradient boosting* se posicionó como el segundo modelo con mejores resultados en nuestra investigación.

El modelo propuesto *feedforward neural network* con la transformación *clr* ofreció en general predicciones precisas para los resultados electorales en Colombia debido a que este modelo de *machine learning* es capaz de capturar relaciones no lineales complejas entre las variables, lo que le otorga una gran capacidad

predictiva. Al aplicar la transformación *clr*, que elimina la colinealidad y proporciona una representación equilibrada y simétrica de los datos composicionales, se mejoró la estabilidad y precisión del modelo. Además, la *clr* facilitó la comparación directa de las proporciones de votos en el espectro ideológico evitando sesgos que podrían haber surgido con otras transformaciones como la *alr* o la *ilr*, resultando en predicciones más robustas y coherentes para los distintos espectros ideológicos. Sin embargo, enfrenta desafíos relevantes en la estimación de resultados específicos para Bogotá. Esta dificultad puede atribuirse a las condiciones únicas y complejas de la capital, que incluyen una diversidad socioeconómica, polarización política y factores locales únicos que afectan los patrones de voto. Las variables generales utilizadas para describir los municipios pueden no captar adecuadamente la heterogeneidad de Bogotá, dado que pueden ocultar variable importante de las diferentes localidades de la ciudad. Por lo tanto, se recomienda llevar a cabo análisis adicionales segmentados por UPZ o Localidad para Bogotá, permitiendo capturar esta variabilidad y mejorar la precisión del modelo.

Finalmente, este trabajo representa un aporte significativo al conocimiento existente en predicción electoral al proponer un enfoque innovador que integra técnicas *machine learning* y datos composicionales, lo cual permitió modelar y entender de manera más precisa las dinámicas electorales en un contexto complejo como el colombiano. Este estudio enfatiza la importancia de abordar la predicción de elecciones presidenciales desde perspectivas nuevas y multidimensionales, incorporando factores contextuales como la polarización regional y el impacto de dinámicas locales en el comportamiento de los votantes. Este trabajo no solo amplía las herramientas disponibles en el campo de las ciencias políticas, sino que también ofrece una base metodológica sólida para investigaciones futuras; se sugiere considerar la posibilidad de reducir los espectros ideológicos a una clasificación simplificada de Izquierda, Centro y Derecha lo cual podría mejorar la interpretación y reducir la complejidad del análisis sin sacrificar precisión.

Recibido: Noviembre de 2024
Aceptado: Febrero de 2025

Referencias

Aguilar López, J. & Aquino López, M. A. (2015), 'Modelo de predicción electoral: el caso de la elección municipal 2015 de león de los aldama, guanajuato', *Estudios políticos (México)* **2015**(35), 87–101.

Aitchison, J. (1982), 'The statistical analysis of compositional data', *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(2), 139–160.

Baquero, K. S. M.-P. X. & Rosero, A. B. (2019), 'Aprendizaje de máquinas para la predicción de elecciones presidenciales en colombia', . .

Barrios, A., Montoya, N. & Mancera, C. (2018), *Sistema electoral - elecciones generales*, Misión de Observación Electoral.

Borges, J. A. L., Balam, R. I. N., Gómez, L. R. & Strand, M. P. (2016), 'The machine learning in the prediction of elections', *ReCIBE* **4**(2).

Castaño-Gómez, I. M. et al. (2019), 'Modelo predictivo para inferir en el próximo presidente de estado a través de un vocabulario ontológico en twitter', . .

Cerón-Guzmán, J. A. & León-Guzmán, E. (2016), A sentiment analysis system of spanish tweets and its application in colombia 2014 presidential election, *in* '2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (socialcom), sustainable computing and communications (sustaincom)(BDCloud-socialcom-sustaincom)', Vol. 2016, IEEE, pp. 250–257.

Cuervo, M. C. & Guerrero, M. A. V. (2019), 'Predicción electoral usando un modelo híbrido basado en análisis sentimental y seguimiento a encuestas: elecciones presidenciales de colombia', *Revista Politécnica* **15**(30), 94–104.

De Colombia, A. C. et al. (1991), *Constitución política de Colombia*, leyfacil. com. ar.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. (2003), 'Isometric logratio transformations for compositional data analysis', *Mathematical geology* **35**(3), 279–300.

Gechem Sarmiento, C. E. (2009), 'Los partidos políticos en colombia: entre la realidad y la ficción', . .

González, V. (2019), 'Una breve historia del machine learning', `https://empresas.blogthinkbig.com/una-breve-historia-del-machine-learning/`.

Isaza, R. L. (2009), 'Historia resumida del partido liberal colombiano', *Bogotá, Colombia: Partido Liberal Colombiano* .

Khan, A., Zhang, H., Boudjellal, N., Ahmad, A., Shang, J., Dai, L. & Hayat, B. (2021), 'Election prediction on twitter: A systematic mapping study', *Complexity* **2021**.

Liscano Fierro, J. M. (2017), 'Modelos mixtos para datos composicionales: Una aplicación con resultados electorales en colombia', . .

Marsland, S. (2011), *Machine learning: an algorithmic perspective*, Chapman and Hall/CRC.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT press.

Orjuela Escobar, L. J. (2022), 'Quién es quién en el espectro político colombiano', `https://cerosetenta.uniandes.edu.co/quien-es-quien-en-el-espectro-politico-colombiano/`.

Partido Conservador, C. (2021), 'Manual del conservador', `https://www.partidoconservador.com/wp-content/uploads/2021/04/Manual-del-Conservador-1.pdf`.

Plata Rincón, C. (2017), 'Plebiscito por la paz en colombia: análisis estadístico a partir de datos composicionales', . .

Santander, P., Elórtegui, C., González, C., Allende-Cid, H. & Palma, W. (2017), 'Redes sociales, inteligencia computacional y predicción electoral: el caso de las primarias presidenciales de chile 2017', *Cuadernos. info* **2017**(41), 41–56.

Triglia, A. (2015), 'Los ejes políticos (izquierda y derecha). portal psicología y mente', `https://psicologiaymente.com/social/ejes-politicos-izquierda-derecha`.