UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Trabajo de grado
Enero 2025

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 Dario Pineda Rios.[c]
wilmarpineda@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

[a]Estudiante, Maestría en Estadística Aplicada, Universidad Santo Tomás, Bogotá
[b]Estudiante, Maestría en Estadística Aplicada, Universidad Santo Tomás, Bogotá
[c]Director, Profesor, Departamento de Estadística, Universidad Santo Tomás, Bogotá

compositional. Specifically, it forecasts the vote distribution in each municipality along a unidimensional 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 and 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 and 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 and 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 and 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.

Este documento está estructurado de la siguiente manera: la primera sección corresponde a la introducción, la segunda presenta los referentes teóricos relacionados con el análisis de variables políticas, enfocándose en la predicción de resultados electorales. La tercera sección muestra información sobre el sistema político colombiano, incluyendo las responsabilidades del presidente de la república, los partidos políticos y los espectros ideológicos. La cuarta sección aborda los datos composicionales, mientras que en la quinta sección se detalla la información sobre los cinco modelos de *machine learning* utilizados. La sexta sección presenta el análisis propuesto, abarcando desde la recolección de datos hasta la predicción de los resultados electorales y la última sección expone las conclusiones del estudio.

2. Estadística en las ciencias políticas

La estadística desempeña un papel fundamental en el campo de las ciencias políticas en todo el mundo. A medida que las sociedades evolucionan y los sistemas políticos se vuelven más complejos, la necesidad de recopilar, analizar e interpretar datos se vuelve imprescindible para comprender y explicar los fenómenos políticos.

En la actualidad, las bases de datos son cada vez más grandes y los recursos computacionales sin precedentes, y en ese sentido las técnicas de *machine learning* se están convirtiendo en herramientas comunes para los científicos sociales. Estas herramientas están siendo usadas para una variedad de propósitos, incluyendo la clasificación de textos políticos y materiales audiovisuales, aprendizaje sobre la opinión de las élites y de la población en general con base en la actividad en redes sociales, predicción electoral, detección del fraude electoral, entre otros. A continuación, se relacionan algunas investigaciones desarrolladas en el mundo que relacionan técnicas estadísticas tradicionales o métodos de *machine learning* en las ciencias políticas.

- Durante la campaña electoral previa a las elecciones presidenciales de 2012 en México, Corona and Sánchez (2012) analizaron las opiniones que los usuarios de Twitter publicaban sobre los candidatos cada minuto durante tres meses. Para ello, utilizaron clasificadores bayesianos y de máxima entropía para catalogar las opiniones como positivas o negativas siempre que era posible, elaborando así índices de sentimiento sobre los candidatos que reflejaban con precisión toda la información discutida en las redes sociales. Los resultados mostraron que la promoción de los candidatos presidenciales se organizaba en las redes sociales por sus equipos de campaña y la masa de seguidores que tenían. Además, se estableció un criterio estadístico para identificar a los seguidores reales y a los trolls.
- En 2013, Deltell et al. (2013) llevaron a cabo un estudio en España que utilizó técnicas de análisis de sentimientos para analizar el comportamiento de los partidos políticos andaluces en Twitter. El objetivo fue describir dicho comportamiento y cuantificar el impacto a través de medidas como

seguidores, menciones, retweets y otras métricas. Además, se compararon los resultados y el impacto de los partidos políticos y sus líderes con los resultados de las votaciones. Los resultados mostraron que el modelo propuesto, que consta de dos partes: una cuantitativa basada en el análisis del número de seguidores de los perfiles de los partidos políticos y de las cuentas de los líderes de dichas agrupaciones, y un análisis cualitativo del impacto de los *tweets* y de los líderes de opinión generados a partir de estos *tweets*, fue más preciso que las encuestas tradicionales para predecir la tendencia de voto de los partidos mayoritarios. De hecho, todas las encuestas tradicionales realizadas durante la campaña electoral se equivocaron en sus predicciones, con un margen de error que osciló entre el $8,69\,\%$ al $13,30\,\%$ para los dos principales partidos. En cambio, utilizando el modelo propuesto basado en Twitter, el margen de error se redujo al $3,01\,\%$.

- En el mismo año, Makazhanov and Rafiei (2013) desarrollaron un estudio para demostrar que se puede predecir la preferencia política de los usuarios en Twitter a partir de su interacción con los partidos políticos. Para lograr esto, utilizaron modelos predictivos basados en diferentes características contextuales y de comportamiento. Los modelos fueron entrenados mediante un enfoque de supervisión a distancia, considerando que los candidatos de los partidos tenían una preferencia establecida hacia sus respectivos partidos. Para evaluar el modelo propuesto, se utilizó el contexto de las elecciones generales de la provincia de Alberta en Canadá en 2012. Los resultados mostraron que el modelo de árboles de decisión supera en términos de la prueba F, a los enfoques de clasificación de sentimientos y texto.

- Jesús Aguilar y Marco Antonio Aquino en 2015 usaron la regresión de dirichlet para predecir las elecciones municipales de León de los Aldama en México a partir de los resultados electorales de las últimas seis elecciones. El objetivo del análisis fue ofrecer una herramienta científica, para que dejando de lado cualquier interés particular, se pudiera establecer un escenario probable tomando en cuenta el comportamiento registrado en la estadística electoral oficial. Los resultados muestran que a partir de las estimaciones (valor medio), el partido Acción Nacional obtendría el primer lugar en la elección con el $48.53\,\%$ frente a un $31.89\,\%$ de su principal contendiente, el partido Revolucionario Institucional; una diferencia de 16.55 puntos porcentuales, (Aguilar López and Aquino López, 2015).

- En 2016, el estudio realizado por Borges et al. (2016) se centró en analizar y comparar los algoritmos de K-vecinos, criterios de convergencia y la metodología de clasificación LAMDA para predecir las elecciones del gobernador del estado de Quintana Roo en México. Los investigadores utilizaron los *softwares* Weka y SALSA para llevar a cabo este análisis. Al examinar los datos de manera cualitativa y cuantitativa en conjunto, se llegó a la conclusión de que el algoritmo de LAMDA, que hace uso implícito del *software* SALSA, obtuvo los mejores resultados y se ajustó de manera más precisa a los criterios establecidos.

- Ante las dudas recurrentes sobre la capacidad de predecir los resultados de las elecciones presidenciales a partir de las encuestas de opinión pública, del Tronco Paganelli et al. (2016) propusieron una estimación de los resultados electorales de la segunda vuelta en las elecciones presidenciales de Argentina en 2015. Para llevar a cabo esta estimación, utilizaron la información proporcionada por el Latinobarómetro 2015, una encuesta anual realizada nueve meses antes de las elecciones. A través de un análisis cuantitativo que empleó una regresión logística para estimar los perfiles de los posibles votantes, se demostró que hubiera sido posible predecir los resultados electorales tanto en la primera como en la segunda vuelta, en la cual resultó electo Mauricio Macri.

- En un estudio realizado en Chile por Santander et al. (2017), se utilizó información de la red social Twitter para lograr las predicciones más precisas posibles de los resultados electorales de las primarias. Para ello, se emplearon técnicas de inteligencia computacional y se supervisó la interacción de todos los usuarios chilenos que mencionaron al menos una vez a alguno de los cinco candidatos en competencia. El objetivo fue diseñar un modelo de predicción que tuviera en cuenta el contexto específico de la comunicación política en Twitter. Los resultados obtenidos fueron sorprendentes, ya que las prediciciones presentaron un error absoluto medio (MAE) inferior al $2\,\%$, lo que significa que se logró una mayor precisión en comparación con las encuestas electorales

tradicionales. Este enfoque innovador demuestra el potencial de las redes sociales y la inteligencia computacional para predecir los resultados electorales con mayor precisión.

- En 2017, David Perea a partir de la inestabilidad política que estaba sufriendo en esos meses la democracia española, propuso desarrollar un análisis usando técnicas de *machine learning* no supervisadas y supervisadas junto con técnicas de procesamiento de lenguaje natural. El análisis se dividió en dos partes: el estudio de las elecciones generales del 26 de junio del 2017 para explicar la obtención del poder político y el estudio de *tweets* sobre la independencia de Cataluña para explicar la difusión de la comunicación política en este asunto, según el origen de los usuarios. En el primer caso, obtuvo información relevante para que los partidos políticos establecieran estrategias electorales y en el segundo, esbozó la gravedad del proceso independentista catalán que transmiten los medios de comunicación, en especial los internacionales, (El Khalifi, 2017).

- En los últimos años, se ha despertado un creciente interés en el estudio de las opiniones políticas utilizando datos a gran escala generados en medios sociales. En este contexto, Arcila-Calderón et al. (2017) describieron y evaluaron el uso de la técnica de análisis supervisado de sentimientos en la comunicación política mediante un clasificador en tiempo real de opiniones políticas en *tweets* escritos en español. Utilizando técnicas de *machine learning*, tanto en un ordenador local como en entornos de computación distribuida comercialmente para lidiar con grandes volúmenes de datos, se logró obtener resultados más precisos. El estudio evidenció que el modelado de las opiniones políticas en Twitter a través del análisis supervisado de sentimientos es una metodología complementaria y necesaria para la contrastación y predicción de resultados electorales en español, tanto en España como en países latinoamericanos.

- En Ecuador, la plataforma de redes sociales Twitter se ha convertido en un canal clave para la interacción directa entre las figuras políticas y la población. Conscientes de esta realidad, Gómez-Torres et al. (2018) llevaron a cabo un estudio con el objetivo de analizar los sentimientos expresados en Twitter, considerando los modismos propios de cada región. Esto brindó una valiosa oportunidad para examinar la relación entre el nivel de aceptación de un candidato en Twitter y los resultados electorales. El trabajo se centró en el desarrollo de un análisis de sentimientos (AS) utilizando una herramienta de procesamiento del lenguaje natural (PNL) adaptada a la variación del español. Sin embargo, los resultados obtenidos revelaron que las elecciones de 2017 no coincidieron con los resultados obtenidos en Twitter. En consecuencia, se determinó que la relación entre las tendencias en las redes de microblogging y los resultados electorales implica más variables de las consideradas en dicho estudio.

- En 2019, Berta López llevó a cabo una interesante predicción de la ideología política de los usuarios de Twitter. Su enfoque involucró el desarrollo de una aplicación en Python que recopiló los *tweets* de cuentas oficiales de los cuatro partidos políticos más influyentes en España: partido Popular, partido Socialista Obrero Español, Ciudadanos y Podemos. A través del entrenamiento de varios modelos, pudo determinar cuál de ellos era el más preciso, y concluyó que el modelo *Naive Bayes* con distribución Bernoulli fue el más efectivo. Posteriormente, para hacer uso de este modelo, creó una aplicación web en la que los usuarios podían ingresar el nombre de usuario de Twitter que deseaban analizar y predecir su ideología política. Como resultado de este proceso, la aplicación generaba un gráfico que mostraba la predicción de la ideología política del usuario en cuestión, basándose en los *tweets* analizados por el modelo, (López Medel, 2019).

- En México, se llevó a cabo un estudio que analizó los tweet compartidos en los días previos a las elecciones presidenciales de julio de 2018. Este trabajo empleó diferentes técnicas de clasificación, como *support vector machines*, *k-nearest neighbors*, *decision trees* y *random forest*, con el objetivo de identificar si el contenido de cada tweet estaba relacionado o no con el proceso electoral (clasificación binaria), y en caso afirmativo, a cuál de los candidatos participantes en las elecciones hacía referencia (clasificación multiclase). Tanto en la clasificación binaria como en la clasificación multiclase, se obtuvieron los mejores resultados de precisión con *support vector machines*, alcanzando valores de 0.860 y 0.8289, respectivamente, superando a las demás técnicas de clasificación. Por otro lado, se encontró que *decision trees* mostró los peores resultados, (González Franco et al., 2019).

- En algunos lugares del mundo, se han llevado a cabo numerosas investigaciones que analizan y hacen predicciones sobre los resultados electorales de varios países, utilizando datos proporcionados por plataformas de redes sociales como Facebook y Twitter. En el año 2021, Khan et al. (2021) realizaron un estudio en el que identificaron y categorizaron 787 investigaciones en este campo. Los resultados revelaron que la mayoría de los estudios se enfocaron en el análisis de sentimientos, seguidos por enfoques basados en el volumen de datos y el análisis de redes sociales.

  Asimismo, en el mismo año, Cabrera-Tenecela (2021) presentaron un estudio que proporciona una descripción detallada sobre el origen de la información, el contexto, el nivel de error y la predicción estadística que emplea el *big data* para pronosticar resultados electorales. Esta revisión bibliográfica se llevó a cabo utilizando Google Académico, y encontraron 34 estudios que cumplían con los criterios de selección. De estos, 12 utilizaron métodos computacionales, 18 se centraron en análisis de sentimientos y 4 en análisis de sentimientos supervisados.

- En el año 2021, se llevó a cabo un estudio que realizó el análisis de sentimientos para las elecciones públicas de Ecuador utilizando la red social Twitter. El objetivo de esta investigación fue comprender el sentir de los ciudadanos respecto a los candidatos presidenciales mediante el uso de técnicas de procesamiento de lenguaje natural y un método de predicción de aprendizaje automático durante el proceso electoral. Los resultados obtenidos mostraron el nivel de aceptación de cada candidato presidencial, lo que podría ser útil para revelar las tendencias y decisiones del voto electoral por parte de los ciudadanos ecuatorianos, (Macias et al., 2021).

- En el mismo año, se estudiaron las elecciones presidenciales de noviembre de 2020 en Estados Unidos mediante el análisis de *tweets* obtenidos a través de un hashtag neutral y menciones a los dos principales candidatos. El objetivo del estudio fue detectar perfiles relevantes dentro de esta muestra y realizar un análisis de los mismos para determinar posibles diferencias en el uso de estos mecanismos de participación en Twitter. Los resultados obtenidos confirmaron la hipótesis inicial de investigación, demostrando que existen diferencias significativas en el contenido de las menciones para cada conjunto de datos, especialmente entre los perfiles relevantes que más mencionaron a cada uno de los candidatos, (Sánchez Navarro, 2021).

  En paralelo, Garcia-Moreno et al. (2021) diseñaron un modelo basado en redes neuronales *Long-Short Term Memory* (LSTM) y lo aplicaron en un análisis de sentimientos con el propósito de predecir los resultados electorales de los Estados Unidos del año 2016. Para ello, procesaron un número significativo de *tweets* que se obtuvieron de un dataset público proporcionado por la plataforma Kaggle. Los resultados obtenidos mostraron que las redes neuronales profundas, específicamente las LSTM, demostraron ser una de las técnicas más efectivas en el análisis de sentimientos, superando a otras en términos de precisión y rendimiento.

- En el año 2022, se desarrolló un estudio que utilizó bosques aleatorios de clasificación y regresión para analizar 314 variables potencialmente explicativas del comportamiento político. Estas variables fueron recogidas en la séptima ola de la *World Value Survey*. El objetivo del estudio fue investigar qué teorías y mecanismos son más efectivos para explicar la participación política, tanto en elecciones como en otros ámbitos, en las regiones de América del Norte y América Latina. Los resultados obtenidos mostraron que en la política electoral, los mecanismos explicativos de las diferencias de ocupaciones laborales parecen estar relacionados con la adopción de valores emancipatorios y la importancia que se da a la política. Lo mismo ocurre, pero de manera aún más clara, con el nivel educativo. Se encontró que las personas con una mejor educación tienden a dar más importancia a la política y tienen una mayor probabilidad de adoptar valores postmaterialistas, (Romero García et al., 2022).

- Posteriormente, Mariela Lucina et al. (2023), analizaron las ediciones del periódico peruano "La República" publicadas entre el 5 de mayo y el 6 de junio de 2021, durante las campañas políticas de la segunda vuelta electoral para la presidencia de Perú. El objetivo principal fue descubrir información relevante para comprender el uso de terminologías utilizadas en el lenguaje escrito de ese periódico, mediante técnicas de minería de texto y algoritmos de aprendizaje automático para el

análisis de datos no estructurados. Los primeros resultados del análisis mostraron que las palabras que aparecían con mayor frecuencia y tenían una relación más fuerte eran "pedrocastillo", "candidato" y "keikofujimori". Se concluyó, que a Pedro Castillo se le relacionaban con temas específicos relacionados con la actividad política y social, mientras que a la candidata Keiko Fujimori se le asoció con su oponente y se le otorgó un tratamiento distinto.

Hasta el momento, se han revisado trabajos e investigaciones que analizan los procesos electorales de diversos países, como Ecuador, España, México, entre otros, y la tendencia predominante es el análisis de sentimientos como herramienta común y precisa para predecir las elecciones, principalmente utilizando la red social Twitter. Ahora, se presentan algunos trabajos desarrollados en Colombia que también emplean el análisis de sentimientos, técnicas de *machine learning* y modelos estadísticos tradicionales para analizar el comportamiento electoral.

- En 2016, se implementó un sistema de análisis de sentimientos siguiendo un enfoque de clasificación supervisada que se aplicó en las elecciones presidenciales colombianas de 2014 para investigar el potencial de las redes sociales para inferir la intención de voto. Los resultados del método propuesto fueron bastante buenos en la primera vuelta de las elecciones, con un error absoluto medio (MAE) de cuatro puntos porcentuales (el más bajo) y los candidatos con mayor número de votos clasificados correctamente. Sin embargo, los resultados del método propuesto en la segunda vuelta de las elecciones fueron peores que los de los métodos de referencia, siendo la inferencia basada en encuestas la que clasificó correctamente a todos los candidatos con el MAE más bajo (4.84 %), (Cerón-Guzmán and León-Guzmán, 2016).
- En 2018, Jaramillo Guerra et al. (2018) analizaron el uso de contenidos emocionales en las campañas de los precandidatos de La Gran Consulta por Colombia. Para esto, utilizaron un estudio de caso y la herramienta Python para establecer una polaridad en una escala del 1 al 3, siendo uno negativo y tres positivo, a partir de los *tweets* realizados por los tres precandidatos como unidad de estudio. Los resultados de la investigación mostraron que los tres candidatos estudiados utilizaron sus cuentas de Twitter para emitir mensajes que apelaban a contenidos emocionales con el objetivo de influir en los juicios de los electores. Asimismo, las apelaciones emocionales se enmarcaron sobre todo en emociones de carácter positivo, derivado de la necesidad de los candidatos por mantener una buena relación con sus contrincantes.
- En el año 2019, Baquero and Rosero (2019) propusieron evaluar tres modelos de *machine learning* para pronosticar los resultados de las votaciones a nivel municipal en Colombia utilizando como variables principales el comportamiento histórico de las elecciones para identificar si un municipio tendría una votación que favorecería a un candidato presidencial más cercano a una ideología política de derecha o izquierda. Para lograr el objetivo, emplearon modelos de redes neuronales artificiales, *random forest* y *gradient boosting*. Entre estos modelos, el *gradient boosting* demostró ser el más efectivo, proporcionando los mejores resultados en sus predicciones. De esta manera, su investigación arrojó luz sobre la relación entre el comportamiento histórico de votación y las preferencias políticas a nivel municipal en Colombia.
- Cuervo and Guerrero (2019) presentaron un modelo híbrido para predecir el resultado de la primera vuelta de las elecciones presidenciales de 2018 en Colombia, con el objetivo de minimizar el error absoluto y mejorar la calidad de las predicciones. Para lograr esto, registraron y analizaron las actividades de los usuarios en Twitter y Facebook mediante algoritmos de inteligencia artificial, lo que resultó en una predicción precisa y coherente con la realidad. Los resultados principales destacan que el modelo híbrido propuesto tiene un RMSE aproximado del $2,47\,\%$, superando en promedio al RMSE de las firmas encuestadoras tradicionales más prominentes del país. Además, pronosticaron el valor del abstencionismo electoral con un error diferencial del $1,72\,\%$ respecto al valor real, lo que demuestra la confiabilidad de la metodología propuesta.
- En el mismo año, Castaño-Gómez et al. (2019) desarrollaron un modelo predictivo que permitió inferir quién sería el próximo presidente de Colombia, utilizando un vocabulario ontológico. Además,

crearon una aplicación móvil híbrida que permite a cualquier persona ver en tiempo real las opiniones generadas en Twitter acerca de su candidato favorito. Los resultados del proyecto mostraron que el modelo de árboles de decisión favoreció al candidato electo Iván Duque, lo que sugiere que, a pesar de que han pasado varios meses desde la candidatura presidencial, su imagen sigue siendo positiva.

- Simultáneamente, Romero Moreno et al. (2019) buscaron determinar si el contenido generado por los usuarios en Twitter podía ser considerado como un factor de medición válido para predecir resultados electorales en campañas presidenciales. Para ello, se analizaron *tweets* generados en la campaña electoral para la presidencia de Colombia en 2018. Los *tweets* fueron clasificados automáticamente en términos de polaridad y emoción utilizando los clasificadores *naïve bayes* y *support vector machine*. También, analizaron la cantidad de menciones en Twitter que tuvo cada uno de los candidatos durante el periodo de estudio con un modelo ARIMA. Los resultados mostraron que el promedio de precisión del modelo *support vector machine* utilizado para clasificar el nivel de polaridad de los mensajes generados para cada uno de los candidatos es de $64,04\,\%$. De igual manera, las menciones realizadas a un candidato en esta red social analizadas en un periodo de tiempo a través del modelo ARIMA, permitió predecir el comportamiento futuro que tendrán las menciones que se realicen a un determinado personaje.

- Posteriormente, en 2020 Atencia et al. (2020) diseñaron y prototiparon una aplicación que pudiera identificar la polarización en los *tweets* utilizando herramientas de análisis de texto y la construcción de redes neuronales para determinar los patrones que conforman las prácticas que afectan gravemente la percepción política de Colombia. Los resultados del estudio mostraron que la precisión general del modelo osciló entre el $75\,\%$ y el $80\,\%$ en los casos relevantes, lo que confirmó la eficacia del mismo.

- En 2021, se destaca el trabajo titulado "Métodos Bayesianos para caracterizar el comportamiento legislativo del Senado colombiano en el periodo 2010–2014" (Luque Zabala, 2021), en el cual se implementó el estimador unidimensional de punto ideal bayesiano estándar utilizando algoritmos de cadenas de Markov Monte Carlo para operacionalizar la conducta electoral parlamentaria. Los resultados se destacan en dos áreas principales: la dimensión del espacio político y la identificación de legisladores pivote. Los puntos ideales estimados revelan un rasgo latente no ideológico (oposición-no oposición) subyacente en las votaciones de los senadores.

A continuación, se presentan algunos trabajos relacionados con el análisis de datos composicionales, un enfoque que ha tomado relevancia en múltiples disciplinas. En la literatura internacional, existen diversos estudios que involucran datos composicionales y modelos de *machine learning*. Entre ellos, se destaca el trabajo de las profesoras Zhang and Shi (2019), que sirvió de inspiración para el desarrollo de este estudio. En su investigación, las profesoras compararon cinco modelos de *machine learning* para clasificar e interpolar las partículas del suelo utilizando datos composicionales. Los resultados revelaron que el modelo *support vector machine* obtuvo la mayor precisión global, mientras que el modelo *XGBoost* generó el mayor coeficiente Kappa en cuanto a la clasificación directa de la textura del suelo. Además, en el caso de la interpolación, se observó que los modelos *random forest* superaron a otros en términos de los indicadores RMSE, MAE, R2, AD y STRESS, entre otros resultados obtenidos en el estudio.

Como se mencionó anteriormente, los modelos de *machine learning* están siendo cada vez más populares, integrando datos composicionales. Sin embargo, la mayoría de las contribuciones que utilizan algoritmos de aprendizaje automático en datos composicionales simplemente aplican el método relevante a una versión transformada aditiva, centrada o isométrica de los datos de entrenamiento, sin considerar las propiedades de la composición. Por ejemplo, Tolosana-Delgado et al. (2019) realizaron una breve revisión de la construcción fundamental de estos métodos y examinaron cómo pueden ser ajustados o adaptados para tener en cuenta la escala composicional de los datos. Por otro lado, Rodriguez (2022) presentaron en su trabajo soluciones que abordan dos tipos de problemas: cuando los datos composicionales son la entrada de un modelo y cuando son su salida. Para el primer caso, desarrollaron un algoritmo eficiente

basado en el marco log-ratio, mejorando la velocidad sin sacrificar calidad. Para el segundo, propusieron una nueva familia de distribuciones exponenciales sobre el simplex, con buenas propiedades matemáticas.

En la literatura colombiana, se pueden encontrar varios estudios en ciencias políticas que analizan el comportamiento de los resultados electorales utilizando datos composicionales como fuente de información. Entre ellos, se destaca el trabajo de Liscano Fierro (2017), quien analizó los procesos electorales en Colombia, como el plebiscito por la paz y las elecciones presidenciales de segunda vuelta en 2014, mediante el uso de la regresión dirichlet, un modelo lineal multivariado y un modelo mixto multivariado desde una perspectiva composicional. Los resultados obtenidos al ajustar el modelo mixto se acercaron de manera más precisa a los valores reales del plebiscito por la paz, identificando cómo las variables trabajadas influyeron en el resultado de las votaciones.

De igual forma, en el mismo año, se destaca el trabajo realizado por Plata Rincón (2017), quien también analizó los resultados del plebiscito por la paz en Colombia y los resultados de las elecciones presidenciales en la primera vuelta de 2014 desde una perspectiva composicional. Los resultados del modelo, el análisis de clúster y el análisis discriminante resaltaron las ventajas y capacidades significativas de las técnicas para datos composicionales, ya que las predicciones se acercaron más a la realidad que las realizadas por empresas encuestadoras.

3. Sistema político colombiano

Colombia se configura como un Estado social de derecho, conformado como una república unitaria y descentralizada, con entidades territoriales autónomas. Se fundamenta en principios democráticos, participativos y pluralistas, cimentados en el respeto a la dignidad humana, el trabajo y la solidaridad entre sus ciudadanos, así como en la prevalencia del interés general. Estos valores esenciales, consagrados en el artículo 1 de la Constitución Política, definen a Colombia como una nación democrática. Esta característica democrática ha sido una constante desde su independencia, lo que implica que el poder político emana directamente de los ciudadanos, quienes lo determinan mediante su participación en las elecciones y ejercen su derecho a influir en su configuración, ejercicio y control, (De Colombia et al., 1991).

Uno de los pilares fundamentales de la participación ciudadana es el voto. Mediante este acto y los procesos electorales asociados, los ciudadanos contribuyen activamente a fortalecer y perfeccionar el sistema democrático, facilitando la toma de decisiones que reflejen la voluntad popular. En esencia, el voto representa la herramienta principal para dar forma a decisiones vinculantes que afectan a la sociedad en su conjunto.

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).

3.1. 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. A continuación, se mencionan algunas de estas responsabilidades.

Figura 1: Algunas responsabilidades del Presidente de la República de Colombia. Elaboración propia.

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.

3.2. Partidos políticos

Los partidos políticos desempeñan un rol esencial en el sistema democrático colombiano, siendo entidades de interés público que operan bajo principios e ideales particulares que guían sus normativas, requisitos para su reconocimiento legal, y modalidades específicas de participación en el proceso electoral. Asimismo, definen los derechos, deberes y privilegios que les son otorgados por la ley y actúan como representantes de diversos sectores de la sociedad siendo actores clave en la toma de decisiones.

Históricamente, se han considerado los partidos como organizaciones con cierto grado de estructura, implicando la deliberación entre sus miembros y la adhesión a una ideología con principios y valores definidos. Sin embargo, fue solo en 1848 cuando este concepto comenzó a tomar forma; el 16 de julio de ese año, en el periódico "El Aviso" se publicó un artículo escrito por Ezequiel Rojas titulado "La Razón de mi Voto" que explicaba las razones por las cuales él y sus seguidores planeaban votar por el General José Hilario López en las elecciones presidenciales de 1849. Rojas abogaba por el liberalismo y delineaba una serie de principios que aún tienen relevancia en la actualidad (Isaza, 2009). En el siguiente año, el periódico "La Civilización" del jueves 4 de octubre publicó el "Programa Conservador" redactado por Mariano Ospina Rodríguez y José Eusebio Caro. En este programa, se presentaba un manifiesto que contenía el ideario fundamental 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. También, la Universidad de los Andes en su artículo "Partidos políticos en Colombia: definición, funciones y lista actualizada" (Andes Universidad, 2023) define un partido político como una organización conformada por un grupo de personas que comparten una ideología, visión, política y objetivos comunes. Sus principales funciones son:

1. **Representación ciudadana.** Estas organizaciones actúan como intermediarias entre los ciudadanos y el Estado, sirviendo como vehículos para canalizar las demandas y necesidades de la sociedad hacia las instituciones gubernamentales, y facilitando la participación de la población en el proceso político.
2. **Participación en elecciones.** Los partidos políticos presentan candidatos en las elecciones para ocupar cargos públicos a nivel local, regional y nacional. Estas elecciones permiten a la ciudadanía elegir a sus representantes, quienes ejercerán el poder político y tomarán decisiones sobre el rumbo del país.
3. **Formulación de políticas públicas.** Desempeñan un papel crucial en la formulación y promoción de políticas públicas en áreas como la economía, la educación, la salud, entre otras. A través de su participación en el Congreso y otros espacios de debate político, contribuyen a la construcción del marco normativo y la agenda de gobierno.

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).

La dirección de Vigilancia e Inspección Electoral del Consejo Nacional Electoral certifica que conforme al Registro Único de Partidos, Movimientos Políticos y Agrupaciones Políticas (RUPYM), se encuentran registradas en Colombia treinta y tres (33) organizaciones políticas con personería jurídica vigente (ver Tabla 1).

Tabla 1: Partidos o movimientos políticos según el CNE.

Organización política	**Resolución que otorga personería**
Partido Liberal Colombiano	4 de 1986
Partido Conservador Colombiano	7 de 1986
Partido Cambio Radical	1305 de 1997
Partido Alianza Verde	8 de 1991
Movimiento Autoridades Indígenas de Colombia "AICO"	20 de 1991
Partido Alianza Social Independiente "ASI"	18 de 1992
Partido Político MIRA	0476 de 2000
Partido de la Unión por la Gente - Partido de la U	4423 de 2003
Partido Polo Democrático Alternativo	4426 de 2003
Partido Político Unión Patriótica "UP"	37 de 1986
Partido Centro Democrático	3035 de 2014
Movimiento Alternativo Indígena y Social "MAIS"	1708 de 2015
Partido Político Comunes	2691 de 2017
Partido Colombia Justa Libres	3198 de 2018
Partido Colombia Renaciente	575 de 2019
Movimiento Alianza Democrática Amplia "ADA"	1748 de 2019
Partido Político Dignidad & Compromiso	1291 de 2021
Movimiento Político Colombia Humana	7417 de 2021
Partido Nuevo liberalismo	7822 de 2021
Movimiento de Salvación Nacional	8804 de 2021
Partido Verde Oxígeno	8805 de 2021
Partido Comunista Colombiano	8842 de 2021
Partido Liga Gobernantes Anticorrupción - Liga	3750 de 2022
Partido Demócrata Colombiano	4033 de 2022
Partido Ecologista Colombiano	5107 de 2022
Partido Político la Fuerza de la Paz	5436 de 2022
Partido Político Nueva Fuerza Democrática	1549 de 2023
Partido Político Esperanza Democrática	1544 de 2023
Independientes	1545 de 2023
Partido Político Gente en Movimiento	2938 de 2023
Partido del Trabajo de Colombia	4354 de 2023
Movimiento Político Poder Popular	6886 de 2023
Movimiento Político Soy Porque Somos	16313 de 2023

Nota. Esta tabla muestra los partidos o movimientos políticos con personería jurídica según el Consejo Nacional Electoral de Colombia (CNE) para mayo de 2024. Tabla tomada de `https://www.cne.gov.co/index.php/partidos-movimientos-politicos-y-grupos-significativos/778`[Página Web en línea]

3.3. 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 2), 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).

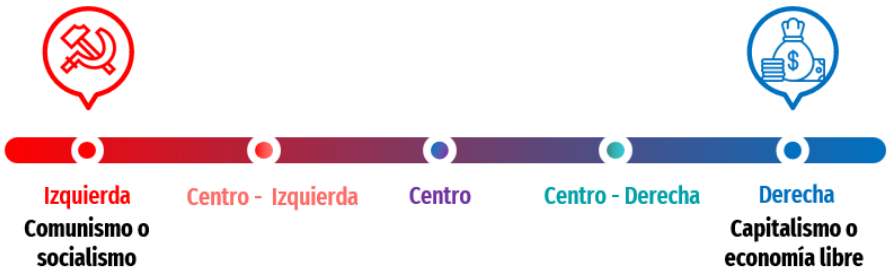


Figura 2: Representación del eje unidimensional de Izquierda - Derecha. Elaboración propia.

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 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 que consideran varios ejes relevantes, como economía, cultura, libertad, autoritarismo, socialismo y capitalismo. El gráfico de Nolan, es una representación visual del espectro político que se utiliza para clasificar las posiciones políticas en dos dimensiones: económica y social. Este fue desarrollado por David Nolan en la década de 1960 y se basa en el concepto de "prueba del mundo real" para evaluar las posiciones políticas, (Roca, 2021).

En el gráfico de Nolan (Ver Figura 3), el eje horizontal representa la dimensión económica, que va desde la izquierda (progresista) hasta la derecha (conservador). El eje vertical representa la dimensión social, que va desde arriba (liberal) hasta abajo (totalitario).

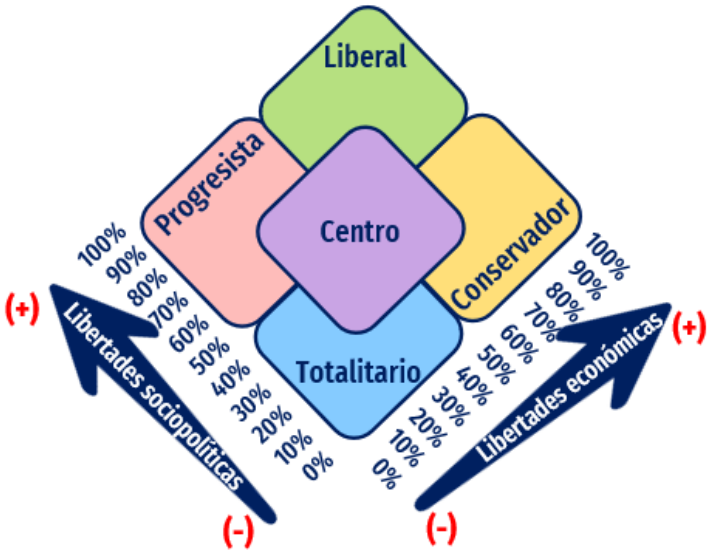


Figura 3: Representación del "Gráfico de Nolan". Elaboración propia.

Por último, se encuentra un enfoque tridimensional que presenta tres niveles distintos, cada uno con diferencias significativas e independientes del modelo unidimensional. Uno de estos enfoques es el "Modelo de Friesian" (Ver Figura 4), propuesto por el filósofo político estadounidense Kelley L. Ross. Este modelo

busca clasificar las ideologías políticas en tres ejes principales: las preferencias políticas, las económicas y las sociales. En el eje político, se encuentran las posturas entre el autoritarismo y el libertarismo; en el económico, entre el intervencionismo y el libre mercado; y en el social, entre el progresismo y el continuismo, (Gentilhombre, 2023).

Esta clasificación tridimensional permite una mayor complejidad y matices en la comprensión de las posturas políticas, en comparación con los enfoques unidimensionales más simples.

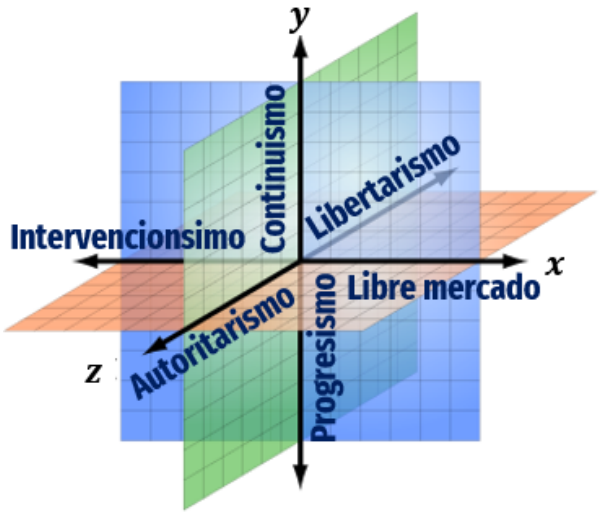


Figura 4: Representación gráfica del "Modelo de Friesian". Elaboración propia.

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).

A pesar de esto, el panorama político en las últimas elecciones presidenciales ha mostrado un sistema polarizado; es decir, una sociedad dividida en porcentajes similares en torno a posiciones opuestas y medidas irreconciliables. Un claro ejemplo es el resultado de la segunda vuelta de las elecciones presidenciales de 2022, donde el candidato ganador obtuvo el 49.77 % de los votos contra el 46.74 %. De manera similar, pero menos marcada, fue la segunda vuelta del 2018, donde el candidato ganador obtuvo el 53.98 % frente al 41.81 %. Otro ejemplo notable de esta división en la sociedad colombiana, aunque no en una elección presidencial, es el plebiscito por la paz, donde el "No" ganó con el 50.23 % de los votos frente al 49.76 %.

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.

4. 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.

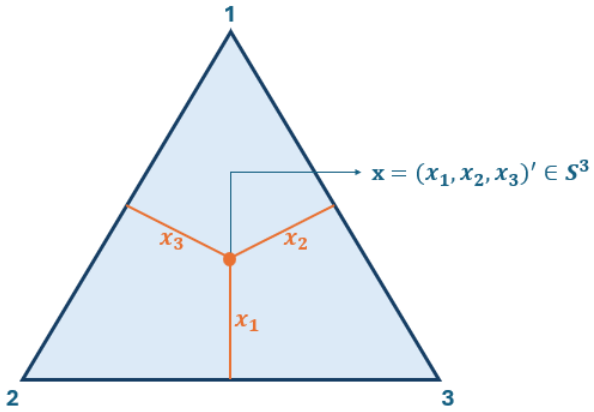


Figura 5: Representación de un dato composicional $(x_1, x_2, x_3)^{'}$ en el símplex S^3. Elaboración propia.

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. Por tal razón, se hace necesario enunciar las siguientes definiciones.

- **Operador clausura C**. El operador clausura $\mathbf{C}$ es una transformación que hace corresponder a cada vector $\mathbf{x} = (x_1, x_2, ..., x_D)^{'}$ de $\mathbb{R}^D$ su dato composicional asociado $\mathbf{C}(\mathbf{x}) = k\mathbf{x}/(x_1+x_2+...+x_D)$ de S^D, con $k > 0$ la constante de clausura. Entonces, dos datos composicionales serán equivalentes si sus cierres para la suma k son iguales:

$$\mathbf{x} \sim \mathbf{y} \Leftrightarrow \mathbf{C}(\mathbf{x}) = \mathbf{C}(\mathbf{y}), \forall k > 0 \tag{2}$$

- **Noción de subcomposición**. Se define como subcomposición $\mathbf{x}_T \in S^T$ de un dato composicional $\mathbf{x} = (x_1, x_2, ..., x_D) \in S^D$, a un subvector formado por T partes seleccionadas de entre las D partes de $\mathbf{x}$, tal que $T \leqslant D$ y $\mathbf{x}_T \subset \mathbf{x}$.

4.1. Principios básicos

A continuación, se presentan tres principios básicos propuestos por Aitchison (1994) de la serie de principios que los datos composicionales deben seguir.

- **Invarianza a la escala**. Los datos de composición sólo aportan información relativa, por lo que cualquier cambio de la escala de los datos originales no tiene ningún efecto. Por ejemplo, $f : S^D \longrightarrow \mathbb{R}^n$ presenta invarianza a la escala si $\forall \lambda \in \mathbb{R}^+, \mathbf{x} \in S^D \Longrightarrow f(\lambda \mathbf{x}) = f(\mathbf{x})$.

- **Invarianza por permutación**. Las conclusiones obtenidas de los análisis composicionales no dependen del orden de las partes. Es decir, si las partes de una composición están ordenadas y/o clasificadas a través de categorías, esta información de ordenación no debe influir en los resultados del análisis composicional.

- **Coherencia subcomposicional**. Los resultados obtenidos para una subcomposición de una composición deben ser consistentes con los resultados de la composición original. Para ello, las subcomposiciones deben comportarse como proyecciones ortogonales en un espacio real, lo que implica que la distancia entre dos datos composicionales debe ser mayor o igual que la distancia entre sus subcomposiciones y debe mantener la invarianza escalar; es decir, las relaciones de proporción entre cualquier par de partes en una subcomposición deben ser iguales a las relaciones correspondientes en el dato composicional original.

4.2. 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.

Las operaciones básicas necesarias para una estructura de espacio vectorial del símplex son las siguientes como lo mencionan Pawlowsky-Glahn et al. (2011) en "Lecture Notes on Compositional Data Analysis".

- **Perturbación.** Es la encargada de medir el cambio composicional en el símplex. Dada dos composiciones de D partes $\mathbf{x}$, $\mathbf{y} \in S^D$ la perturbación se define como:

$$\mathbf{x} \oplus \mathbf{y} = \mathbf{C}(x_1y_1, x_2y_2, ..., x_Dy_D) = \frac{(x_1y_1, x_2y_2, ..., x_Dy_D)}{(x_1y_1 + x_2y_2 + ... + x_Dy_D)} \tag{3}$$

  Esta operación cumple con las propiedades de un grupo abeliano. Primero, satisface la propiedad conmutativa, ya que para cualquier par de composiciones $\mathbf{x}$ y $\mathbf{y}$, se cumple que $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$. También, es asociativa porque para cualquier composición $\mathbf{x}$, $\mathbf{y}$ y $\mathbf{z}$, se cumple $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$. Existe un elemento neutro, que es la composición $\mathbf{e} = (\frac{1}{D}, \frac{1}{D}, ..., \frac{1}{D})$, tal que para cualquier composición $\mathbf{x}$, se cumple $\mathbf{x} \oplus \mathbf{e} = \mathbf{x}$. Finalmente, para cada composición $\mathbf{x}$, existe un elemento simétrico $-\mathbf{x}$ tal que $\mathbf{x} \oplus -\mathbf{x} = \mathbf{e}$.

- **Potenciación.** Permite escalar una composición mediante un exponente, preservando la estructura composicional. Dada la composición $\mathbf{x} \in S^D$ y un escalar $\alpha \in \mathbb{R}$ la potenciación se define como:

$$\alpha \odot \mathbf{x} = \mathbf{C}(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha) = \frac{(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha)}{(x_1^\alpha + x_2^\alpha + ... + x_D^\alpha)} \tag{4}$$

  Esta operación cumple con las propiedades de un producto externo. En primer lugar, satisface la propiedad asociativa porque $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \odot \beta) \odot \mathbf{x}$ para cualquier composición $\mathbf{x}$ y escalares α

y $\beta \in \mathbb{R}$. Además, la potenciación es distributiva respecto de la perturbación, lo que implica que $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$. También, es distributiva respecto de la suma escalar; es decir, $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$. Finalmente, existe un elemento neutro en la potenciación, ya que $1 \odot \mathbf{x} = \mathbf{x}$ para cualquier composición $\mathbf{x}$.

- **Producto interno.** Para dos composiciones $\mathbf{x}$, $\mathbf{y} \in S^D$, el producto interno de Aitchison (a) se define como:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} ln\left(\frac{x_i}{x_j}\right) ln\left(\frac{y_i}{y_j}\right) \tag{5}$$

  Los productos internos actúan como logaritmos de contraste, también denominados "combinaciones lineales" compositivas, esenciales en diversos tipos de análisis de datos composicionales.

- **Norma.** Para una composición $\mathbf{x} \in S^D$, la norma de Aitchison (a) se define como:

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(ln\left(\frac{x_i}{x_j}\right)\right)^2} \tag{6}$$

  La norma de Aitchison es no negativa ya que toma valor cero únicamente cuando el vector es cero. También, la norma del múltiplo escalar de un vector $\alpha\mathbf{x}$ con $\alpha \in \mathbb{R}$ y $\mathbf{x} \in S^D$, es α veces la norma de $\mathbf{x}$. Además, el valor absoluto del producto interno de $\mathbf{x}$ e $\mathbf{y}$ no excede al producto de las normas de $\mathbf{x}$ e $\mathbf{y}$. Finalmente, la norma de la suma de $\mathbf{x}$ e $\mathbf{y}$ no excede la de la suma de sus respectivas normas.

- **Distancia.** Para una composición $\mathbf{x}$ e $\mathbf{y} \in S^D$, se define la distancia de Aitchison entre $\mathbf{x}$ e $\mathbf{y}$ como:

$$\begin{aligned} d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a &= \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(ln\left(\frac{x_i}{x_j}\right) - ln\left(\frac{y_i}{y_j}\right)\right)^2} \\ &= \sqrt{\sum_{i=1}^{D} \left(ln\left(\frac{x_i}{g(\mathbf{x})}\right) - ln\left(\frac{y_i}{g(\mathbf{y})}\right)\right)^2} \end{aligned} \tag{7}$$

  Con $g(\mathbf{x})$ e $g(\mathbf{y})$ las medias geométrica de las D partes de $\mathbf{x}$ e $\mathbf{y}$ respectivamente.

4.3. 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.

4.3.1. Transformación log-cociente aditiva

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

$$\begin{aligned} \mathbf{w} = alr(\mathbf{x}) &= \left(ln\left(\frac{x_1}{x_D}\right), ln\left(\frac{x_2}{x_D}\right), ..., ln\left(\frac{x_{D-1}}{x_D}\right)\right) \\ &= (ln(x_1) - ln(x_D), ln(x_2) - ln(x_D), ..., ln(x_{D-1}) - ln(x_D)) \end{aligned} \tag{8}$$

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

$$\begin{aligned} \mathbf{x} = alr^{-1}(\mathbf{w}) &= \mathbf{C}(exp(w_1), exp(w_2), ..., exp(w_D), 1) \\ &= \mathbf{C}(\mathbf{x}) \end{aligned} \tag{9}$$

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.

4.3.2. Transformación log-cociente centrada

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

$$\begin{aligned} \mathbf{z} = clr(\mathbf{x}) &= \left(ln\left(\frac{x_1}{g(\mathbf{x})}\right), ln\left(\frac{x_2}{g(\mathbf{x})}\right), ..., ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right) \\ &= (ln(x_1) - ln(g(\mathbf{x})), ln(x_2) - ln(g(\mathbf{x})), ..., ln(x_D) - ln(g(\mathbf{x}))) \end{aligned} \tag{10}$$

donde

$$g(\mathbf{x}) = \left[\prod_{j=1}^{D} x_j\right]^{1/D} \tag{11}$$

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:

$$\begin{aligned} \mathbf{x} = clr^{-1}(\mathbf{z}) &= \mathbf{C}(exp(z_1), exp(z_2), ..., exp(z_D)) \\ &= \mathbf{C}(\mathbf{x}) \end{aligned} \tag{12}$$

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.

4.3.3. 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:

$$\mathbf{y} = ilr(\mathbf{x}) = (\langle \mathbf{x}, e_1 \rangle_a, \langle \mathbf{x}, e_2 \rangle_a, ..., \langle \mathbf{x}, e_{D-1} \rangle_a) \tag{13}$$

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:

$$\mathbf{x} = ilr^{-1}(\mathbf{y}) = \bigoplus_{j=1}^{D-1} y \otimes e_j \tag{14}$$

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.

5. *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)).

5.1. Pasos para construir un modelo de *machine learning*

Cualquier tarea de *machine learning* se puede dividir en una serie de tareas más sencillas. Marsland (2011) describen los 5 pasos para aplicar el aprendizaje automático.

- **Recolectar los datos**

  Si los datos están escritos en papel o registrados en archivos de texto y hojas de cálculo, o almacenados en una base de datos SQL, deberán recopilarse en un formato electrónico adecuado para el análisis. Este paso parece obvio, pero es uno de los que más complicaciones trae y más tiempo consume.

- **Explorar y preparar los datos**

  La calidad de cualquier aprendizaje automático se basa en gran medida en la calidad de los datos. Este paso en el proceso de aprendizaje automático tiende a requerir una gran cantidad de intervención humana. Una estadística que se cita con frecuencia sugiere que el 80 % del esfuerzo en el aprendizaje automático se dedica a los datos.

- **Entrenar el modelo**

  Cuando los datos se hayan preparado para el análisis, es probable que se tenga una idea de lo que espera aprender de los datos. La tarea específica de aprendizaje automático informará la selección de un algoritmo apropiado, y el algoritmo representará los datos en forma de modelo.

- **Evaluar el rendimiento del modelo**

  Debido a que cada modelo de aprendizaje automático da como resultado una solución sesgada al problema de aprendizaje, es importante evaluar qué tan bien aprendió el algoritmo de su experiencia.

- **Mejorar el rendimiento del modelo**

  Si se necesita un mejor rendimiento, es necesario utilizar estrategias más avanzadas para aumentar el rendimiento del modelo. A veces, puede ser necesario cambiar por completo el modelo. Es posible que se deba complementar los datos con datos adicionales o realizar un trabajo preparatorio adicional como en el paso dos de este proceso.

5.2. Tipos de *machine learning*

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 la distinción más importante es: ¿el modelo tiene un número fijo de parámetros o el número de parámetros aumenta con la cantidad de datos de entrenamiento? El primero se llama **modelo paramétrico** y el segundo se llama **modelo no paramétrico**. 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 a menudo computacionalmente intratables para grandes conjuntos de datos.

Otro enfoque de clasificación de algoritmos, que a veces se emplea, consiste en dividirlos en tres categorías: supervisados, no supervisados y por refuerzo. Este trabajo se enfoca en los algoritmos supervisados que tienen como objetivo según Murphy (2012) en su libro "Machine Learning A Probabilistic Perspective" aprender a mapear desde las entradas x hasta las salidas y, dado un conjunto de pares de entrada-salida $R = \{(x_i, y_i)\}_{i=1}^{N}$ donde R se denomina **conjunto de entrenamiento** y N es el número de muestras de entrenamiento.

En el caso más simple, cada entrada de entrenamiento x_i es un vector de números R-dimensional, que representa características, atributos o covariables. Sin embargo, en general, x_i podría ser un objeto estructurado complejo, como una imagen, una oración, un mensaje de correo electrónico, una serie temporal, una forma molecular, un gráfico, etc.

De manera similar, la forma de la variable de salida puede ser en principio cualquier cosa, pero la mayoría de los métodos asumen que y_i es una variable categórica o nominal de algún conjunto finito, $y_i \in \{1, ..., C\}$, o que y_i es un escalar de valor real. Cuando y_i es categórica, el problema se conoce como clasificación o reconocimiento de patrones, y cuando y_i tiene un valor real, el problema se conoce como regresión.

5.3. *Support Vector Machines* (SVM)

El algoritmo de *Support Vector Machines* o máquinas de vectores de soporte, fue introducido por Vladimir Vapnik y su equipo en los laboratorios AT&T Bell en la década de los años 90 como una solución a problemas en el campo computacional, (Cortes and Vapnik, 1995). Inicialmente se desarrolló un algoritmo de clasificación binaria; sin embargo, su desarrollo ha sido tal que actualmente se utiliza también para problemas de clasificación multiclase y regresión, (Fernández Casal, 2020). A partir de un conjunto de entrenamiento $S = \{(\mathbf{x_1}, y_1), ..., (\mathbf{x_p}, y_p)\}$, donde $\mathbf{x_i} \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, el objetivo es encontrar un hiperplano óptimo de separación en un espacio $p-$dimensional (p - el número de características) que clasifica claramente los datos. Para una mejor comprensión del modelo, a continuación se presentan algunas definiciones clave y su representación gráfica.

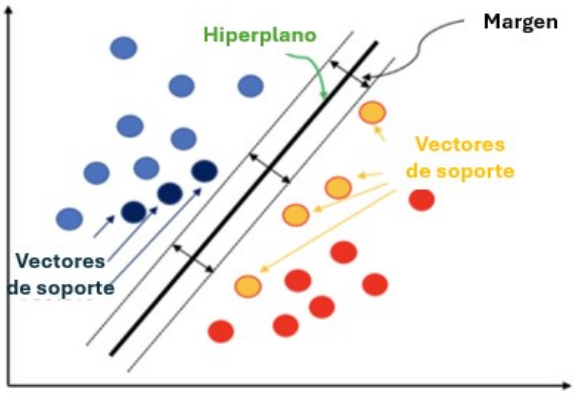


Figura 6: Representación gráfica de un hiperplano entre dos clases, el margen y sus vectores de soporte. Imagen adaptada de `https://datatron.com/what-is-a-support-vector-machine/` [Página Web en línea].

- **Hiperplano.** En un espacio $p-$dimensional, un hiperplano es un subespacio plano y afín de dimensión $p-1$. Es decir, una superficie de $p-1$ dimensiones que divide el espacio $p-$dimensional en dos partes. Por ejemplo, en un espacio de dos dimensiones el hiperplano es una recta y en un espacio tridimensional el hiperplano es un plano convensional. Matemáticamente, se define mediante la siguiente ecuación lineal:

$$w_1x_1 + w_2x_2 + ... + w_px_p + b = 0 \tag{15}$$

  donde $\mathbf{w} = (w_1, w_2, ..., w_p)$ es un vector normal al hiperplano, $\mathbf{x} = (x_1, x_2, ..., x_p)$ es un punto en el espacio $p-$dimensional y b es un término de sesgo que determina la distancia del hiperplano al origen.

- **Margen.** Es la distancia perpendicular desde el hiperplano a las observaciones más cercanas. Es decir, que dado el hiperplano anterior (Ecuación 15), el margen M se define como:

$$M = \frac{1}{\|\mathbf{w}\|} \tag{16}$$

- **Vectores de soporte.** Son las observaciones que se encuentran en el margen y, por lo tanto, son las más cercanas al hiperplano. Es decir, son vectores en el espacio $p-$dimensional que "soportan" el hiperplano óptimo de separación.

5.3.1. *Support Vector Regression* (SVR)

Support Vector Regression es una extensión de los SVM utilizada para tareas de regresión en lugar de clasificación. Mientras que SVM se enfoca en encontrar un hiperplano que separe las clases en un espacio de características, el SVR se enfoca en encontrar un hiperplano de regresión que prediga valores continuos con un margen de tolerancia especificado.

Dado el conjunto de datos $\{(x_1, y_1), ..., (x_i, y_i)\}$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$ y $\Phi : \mathbb{X} \longrightarrow \mathcal{F}$ la función que hace corresponder a cada punto de entrada x un punto en el espacio de características $\mathcal{F}$ (espacio de Hilbert). El objetivo es trasladar el conjunto de entrenamiento a un nuevo espacio de características, para encontrar la función $f(x) = \langle w, \Phi(x) \rangle + b$ que mejor se aproxime a los datos del conjunto.

Se plantea la formulación original del problema y para este caso no depende directamente de las muestras del conjunto, sino de sus imágenes por una cierta función Φ como se observa a continuación:

$$min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{17}$$

$$\text{s.a}\begin{cases} y_i - \langle w, \phi(x_i)\rangle \leqslant \varepsilon + \xi_i; & i = 1, ..., n \\ \langle w, \phi(x_i)\rangle - y_i \leqslant \varepsilon + \xi_i^*; & i = 1, ..., n \\ \xi_i \geq 0, \xi_i^* \geq 0; & i = 1, ..., n \end{cases}$$

La constante $C > 0$ determina el equilibrio entre la regularidad de $f(x)$ y la cantidad hasta la cual se toleran las desviaciones mayores que ε. Se consideran ξ_i y ξ_i^* las variables que controlan el error cometido por la función de regresión al aproximar el i–ésimo ejemplo. Si el valor de C es muy grande se considera que el conjunto está perfectamente representando por el hiperplano predictor ($\xi_i \longrightarrow 0$) y si el valor de C es muy pequeño entonces se estarían admitiendo un número significativo de datos mal representados, (Campo León and Alcalá Nalvaiz, 2017).

Este problema depende de la dimensión en la que se encuentre el estudio y en este caso al transformarse por la función Φ se tiene una dimensión mayor, lo que complica la solución del problema original. Entonces, se plantea el problema dual y para esto se construye la función lagrangiana con la función objetivo y las restricciones correspondientes, mediante la introducción de un conjunto de variables duales (Campo León and Alcalá Nalvaiz, 2017). El objetivo es obtener el problema dual a partir de las condiciones de dualidad; obteniendo el siguiente problema de optimización siendo α_i y α_i^* los multiplicadores de Lagrange:

$$max -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*)(\phi(x_i), \phi(x_j)) - \varepsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{18}$$

$$\text{s.a}\begin{cases} \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

En ocasiones, no es posible ajustar los datos con una función lineal. En tales casos, mediante un *kernel* apropiado, se induce un espacio de Hilbert, también conocido como espacio de características, donde los datos transformados pueden ajustarse con un regresor lineal.

5.3.2. El truco del *kernel*

Un *kernel* (K) es una función que devuelve el resultado del producto escalar entre dos vectores realizado en un nuevo espacio dimensional distinto al espacio original en el que se encuentran los vectores. El truco del *kernel* consiste en transformar los datos a este espacio de mayor dimensión, donde es posible ajustar una función lineal a la relación no lineal presente en los datos originales. En la Figura 7, se observa cómo al añadir una dimensión nueva, se puede separar fácilmente las dos clases con una superficie de decisión.

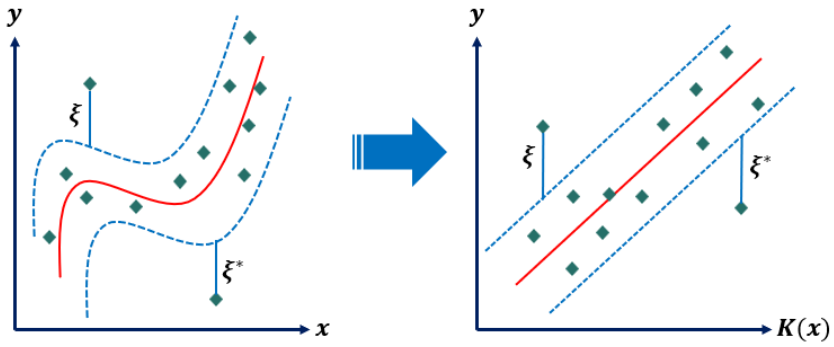


Figura 7: Truco del *kernel*. Elaboración propia.

A continuación, se presentan algunas de las funciones kernel más utilizadas:

- ***kernel* lineal**. Cuantifica la similitud de un par de observaciones usando la correlación de Pearson. Siendo $\mathbf{x}$ un vector, se define el *kernel*lineal como:

$$K(\mathbf{x}, \mathbf{x}^{'}) = \mathbf{x} \cdot \mathbf{x}^{'} \tag{19}$$

- ***kernel* polinómico**. Un *kernel* polinómico de grado d (siendo $d > 1$) permite un límite de decisión mucho más flexible.

$$K(\mathbf{x}, \mathbf{x}^{'}) = (\mathbf{x} \cdot \mathbf{x}^{'} + c)^{d} \tag{20}$$

- ***kernel* Gaussiano**. El valor de γ controla el comportamiento del *kernel*, cuando es muy pequeño, el modelo final es equivalente al obtenido con un *kernel* lineal, a medida que aumenta su valor, también lo hace la flexibilidad del modelo.

$$K(\mathbf{x}, \mathbf{x}^{'}) = exp(-\gamma||\mathbf{x} - \mathbf{x}^{'}||^{2}) \tag{21}$$

5.4. *Random Forest (RF)*

El modelo *Random Forest* fue propuesto por primera vez por Tin Kam Ho en 1995 y posteriormente desarrollado por Leo Breiman y Adele Cutler, (Urbina et al., 2019). Este modelo está compuesto por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída del conjunto de datos de entrenamiento original mediante *bootstrapping*; esto significa que cada árbol se entrena con datos ligeramente distintos. Dentro de cada árbol, las observaciones se dividen en nodos, formando la estructura del árbol hasta llegar a los nodos terminales. La predicción para una nueva observación se obtiene combinando las predicciones de todos los árboles individuales que componen el modelo, (Amat Rodrigo, 2020).

5.4.1. Árboles de regresión

Un árbol es un grafo conexo y acíclico dirigido, compuesto por un conjunto finito de nodos conectados. Su representación gráfica toma la forma que se muestra en la Figura 8, con los nodos dirigidos hacia abajo, (Carbonell, 2020).

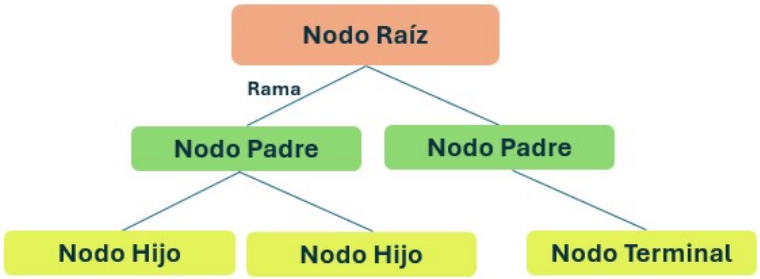


Figura 8: Esquema de un árbol. Elaboración propia.

A continuación, se relacionan las definiciones relacionadas con los árboles.

- **Nodo.** Es la unidad sobre la que se construye un árbol.
- **Nodos hijos.** Es un nodo que desciende directamente de otro nodo, conocido como nodo padre.
- **Nodos padre.** Es un nodo que posee al menos un nodo hijo.
- **Nodo raíz.** Representa el punto de inicio o origen de la jerarquía del árbol.
- **Nodos hermanos.** Son nodos que comparten el mismo nodo padre. Es decir, son nodos que están al mismo nivel de la jerarquía del árbol y descienden directamente del mismo nodo.
- **Nodos terminales u hojas.** Son aquellos que no tienen ramificaciones o nodos hijo.
- **Rama.** Es el camino(unión de arcos simples) desde la raíz a un nodo hoja.
- **Grado de un nodo.** Es el número de descendientes directos que posee. El grado de un árbol es el máximo de los grados de sus nodos.
- **Nivel de un nodo.** Es el número de arcos que han de recorrerse desde la raíz. La altura de un árbol es el máximo de los niveles.

Un árbol de decisión se construye mediante un proceso de particionamiento recursivo de los datos de entrenamiento con el objetivo de crear nodos que maximicen la pureza (en el caso de clasificación) o minimicen la varianza (en el caso de regresión) de las respuestas en cada partición. En el caso de la regresión se requiere realizar una división $R_1, R_2, ..., R_M$, dependientes de los valores de las variables explicativas $x_1, x_2, ..., x_p$ para obtener, (Fernández Villafañez et al., 2022):

$$f(x) = E(Y|X_1 = x_1, X_2 = x_2, ..., X_p = x_p) \tag{22}$$

Mediante funciones de la siguiente forma se quiere expresar que la función regresora tendrá un valor constante para todo x asignado a R_m:

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{C}_m I_{x \in R_m} \tag{23}$$

A través del conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^n$, con $y_i \in R$, la división $R_1, R_2, ..., R_M$ es conocida y se realiza el criterio de mínimos cuadrados $\sum_{i=1}^n (y_i - \hat{f}(x_i))^2$, para apreciar que el mejor valor que puede tomar $\hat{C}_m$ es la media de las respuestas y_i para todas las observaciones que están en R_m, (Fernández Villafañez et al., 2022).

$$\hat{C}_m = \frac{1}{n(m)} \sum_{i \in R_m} y_i \tag{24}$$

En este caso, no se utilizan las mismas ecuaciones de impureza para los árboles de clasificación. Por el contrario, se emplean el error promedio del árbol T en el nodo m:

$$Q_m(T) = \frac{1}{n(m)} \sum_{i \in R_m} (y_i - \hat{C}_m)^2 \tag{25}$$

La impureza total del árbol, se calcula a través de la siguiente fórmula:

$$Q(T) = \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{C}_m)^2 \tag{26}$$

El término de penalización tiene su importancia en la disminución del error total Q sin aumentar de manera drástica el número de hojas del árbol. Su formulación es la siguiente:

$$CP_\alpha(T) = Q(T) + \alpha|T| \tag{27}$$

Donde $|T|$ es el número de nodos terminales tras realizar la poda del árbol T_{max} y α es una parámetro real a escoger.

5.4.2. *Random Forest Regression* (RFR)

Random Forest Regression funciona mediante la creación de múltiples árboles de decisión utilizando *bagging*. En este enfoque, se generan varios subconjuntos del conjunto de datos original mediante muestreo con reemplazo (*bootstrap*); cada uno de estos subconjuntos se utiliza para entrenar un árbol de decisión independiente y en cada nodo de estos árboles, se selecciona aleatoriamente un subconjunto de características para determinar la mejor división, lo que introduce más variabilidad y reduce la correlación entre los árboles. Una vez que todos los árboles han sido entrenados, el modelo predice el valor de una nueva observación promediando las predicciones de todos los árboles en el bosque. Para entender mejor el modelo, a continuación se presentan algunas definiciones clave.

- ***Bootstrap.*** Es un procedimiento propuesto inicialmente por Efron (1992), que se basa en crear un número B de muestras con reemplazo del mismo tamaño que el conjunto original y ajustar un estadístico para cada una de ellas. En este contexto, se ajusta un árbol sin poda para reducir el sesgo al máximo. Aunque la varianza de cada árbol sea alta, al combinarlos entre sí, se logra reducir dicha varianza.

- ***Bagging.*** Proviene del concepto de agregación de *bootstrap*. Es un metaalgoritmo de aprendizaje automático diseñado para mejorar la estabilidad y precisión de algoritmos de *machine learning*. El principal objetivo es la reducción de la varianza de las predicciones promediando las estimaciones de distintos modelos o algoritmos como se muestra en la Figura 9.

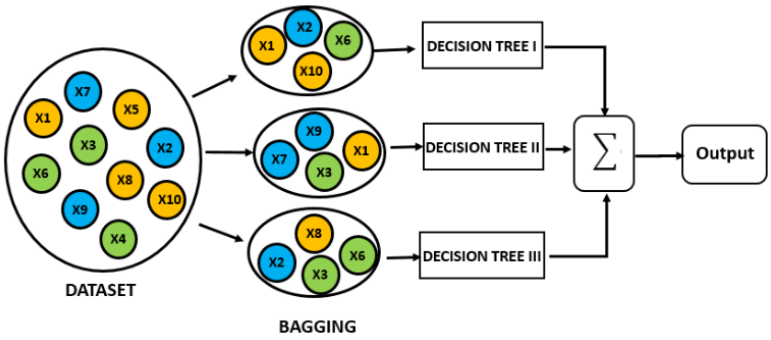


Figura 9: Proceso de *bagging*. Imagen tomada de `https://machinelearningparatodos.com/` [Página Web en línea].

Siendo $\hat{f}_1(x), ..., \hat{f}_B(x)$ los predictores en un problema de regresión con B observaciones independientes, se tiene:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x) \tag{28}$$

Esta predicción es una estimación de la función de regresión $f(x)$ pero con una variabilidad menor. Este predictor no se utiliza de forma práctica porque no es habitual contar con B conjuntos de entrenamiento independientes. Por lo tanto, se emplean B muestras "*bootstrap*" del conjunto de entrenamiento.

A partir de la muestra $Z = \{(x_i, y_i)\}_{i=1}^{n}$ se pueden obtener muestras "*bootstrap*" $Z = \{(x_i^*, y_i^*)\}_{i=1}^{n}$ en las que $x_i^* = x_{i1}$, $y_i^* = y_{i1}, ..., x_n^* = x_{in}$ e $y_n^* = y_{in}$ para $i_1, ..., i_n$. Estas son muestras aleatorias con un reemplazamiento tamaño n para $1, ..., n$. Entonces, el estimador "*bagging*" para un problema de regresión en el que se pueden obtener B muestras de *bootstrap* del conjunto de entrenamiento $\{Z^{*b}\}_{b=1}^{B}$ para ajustar los predictores $\{\hat{f}_{z^{*b}}(x)\}_{b=1}^{B}$, se define (Fernández Villafañez et al., 2022):

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}z^{*b}(x) \tag{29}$$

5.4.3. Algoritmo para *Random Forest*

Se inicia con un conjunto de entrenamiento que tiene la variable de interés $\mathbf{Y}$, las variables predictoras $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_p}$ y el número de repeticiones "*bootstrap*" M que se quieren crear. Luego, se aplican los siguientes pasos, (Carbonell, 2020):

Algorithm 1 *Random Forest*

1: Se construyen M muestras *bootstrap* $B_1^*, B_2^*, B_3^*, ..., B_M^*$.
2: Para cada una de las muestras se construye el árbol $g_1^*, g_2^*, g_3^*, ..., g_M^*$, donde en cada nodo que se construye:
3: Se escogerá entre los $m = p/3$ predictores elegidos de forma aleatoria de entre los p predictores considerados.
4: Se elige el mejor punto de corte para las m variables.
5: Se añaden dos nodos terminales no visitados, si se cumple el criterio de subdivisón de un nodo.
6: Finalmente, el estimador *Random Forest* es $f_{RF}(x) = \frac{\sum_{m=1}^{M} g_m^*(x)}{M}$.

5.5. *Gradient Boosting* (GB)

El *Gradient Boosting* es un método de *machine learning* muy utilizado tanto para tareas de clasificación como de regresión y su desarrollo se atribuye a Jerome H. Friedman, quien desarrollo los primeros algoritmos de aumento de gradiente de regresión explícita, (Friedman, 2001). Este modelo combina el principio de *boosting* con técnicas de optimización de gradientes para producir predicciones de alta precisión mediante la agregación secuencial de modelos débiles, generalmente árboles de decisión. Para comprender mejor el modelo, a continuación se explican algunas definiciones clave.

- ***Boosting.*** Es una técnica que combina múltiples modelos débiles, como árboles de decisión simples, en un modelo final más robusto y preciso a través de un proceso iterativo; cada modelo se entrena para corregir los errores de los anteriores, enfocándose en las observaciones más difíciles de predecir al asignarles un peso mayor en cada paso. Esto permite que los modelos subsiguientes concentren su atención en los errores residuales, mejorando gradualmente las predicciones. Al final, se ponderan los modelos individuales en función de su rendimiento, creando un modelo final capaz de capturar patrones complejos y reducir significativamente el error global.

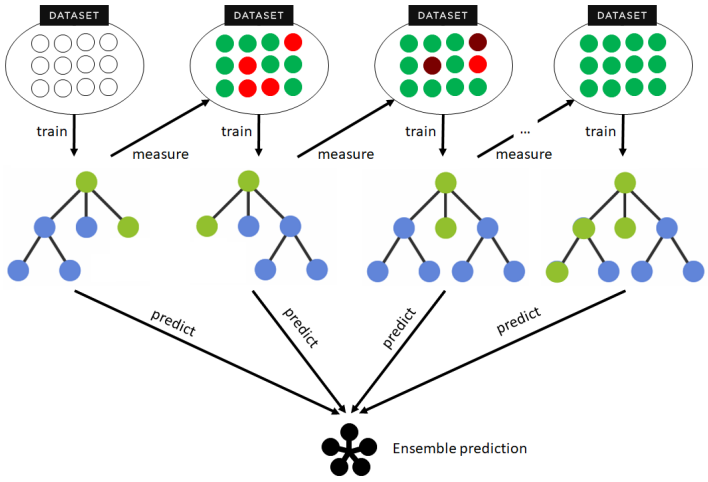


Figura 10: Proceso de *boosting*. Imagen tomada de `https://4geeks.com/es/lesson/boosting-de-algoritmos` [Página Web en línea].

- ***Gradient.*** Es la derivada de la función de pérdida respecto a las predicciones actuales del modelo. Esta derivada indica la dirección en la que deben realizarse ajustes para reducir los errores residuales en cada etapa del modelo. En cada iteración de *gradient boosting*, el modelo nuevo se entrena para aproximar el gradiente de la función de pérdida calculado con las predicciones anteriores. Así, el gradiente sirve como una "guía" para que el siguiente modelo en la secuencia se enfoque en corregir los errores específicos del modelo previo, optimizando progresivamente la precisión del modelo final.

5.5.1. *Gradient Boosting Regression* (GBR)

El objetivo del *gradient boosting* en regresión es construir un modelo que pueda predecir un valor numérico minimizando una función de pérdida, generalmente el error cuadrático medio (MSE) o cualquier otra función de pérdida diferenciable. La función de pérdida general para un conjunto de datos de entrenamiento con n observaciones se define como:

$$L(y, F(\mathbf{x})) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, F(x_i)) \tag{30}$$

Donde $\mathbf{y} = (y_1, y_2, ..., y_n)$ representa los valores de la variable dependiente, $F(\mathbf{x})$ es la predicción del modelo para el conjunto de características $\mathbf{x}$ y $l(y_i, F(x_i))$ es la pérdida de la observación i, que mide la diferencia entre el valor real de y_i y la predicción $F(x_i)$.

El modelo final $F(\mathbf{x})$ se construye agregando secuencialmente modelos débiles, generalmente árboles de decisión, en M iteraciones como se muestra en la siguiente ecuación.

$$F(\mathbf{x}) = F_0(\mathbf{x}) + \sum_{m=1}^{M} \gamma_m h_m(\mathbf{x}) \tag{31}$$

Donde $F_0(\mathbf{x})$ es la predicción inicial, generalmente la media de $\mathbf{y}$, $h_m(\mathbf{y})$ es el modelo débil en la iteración m y γ_m es el parámetro de peso que determina cuánto contribuye el modelo débil $h_m(\mathbf{y})$ en cada iteración.

Para mejorar el modelo en cada paso, *gradient boosting* utiliza el gradiente de la función de pérdida con respecto a las predicciones actuales $F_{m-1}(\mathbf{x})$. Este gradiente indica la dirección de ajuste necesaria para reducir el error de predicción en cada iteración. En la iteración m, el gradiente para cada observación i se define como:

$$g_i^{(m)} = -\frac{\partial l(y_i, F(x_i))}{\partial F(x_i)}\bigg|_{F(\mathbf{x})=F_{m-1}(\mathbf{x})} \tag{32}$$

Este gradiente representa el residuo o error residual de la predicción anterior y se utiliza como objetivo para entrenar el próximo modelo $h_m(\mathbf{x})$, cada modelo débil se entrena para aproximar el gradiente $g_i^{(m)}$. Esto implica ajustar $h_m(\mathbf{x})$ de tal manera que sus predicciones se alineen con los errores residuales de la iteración anterior buscando reducir al máximo esos errores.

Una vez se ha entrenado el modelo $h_m(\mathbf{x})$ en la iteración m, se calcula el valor óptimo del peso γ_m que minimiza la función de pérdida global de la siguiente manera:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{33}$$

Ahora, con el nuevo valor de γ_m, el modelo $F(\mathbf{x})$ se actualiza en cada iteración como:

$$F(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}) \tag{34}$$

Este proceso continua hasta completar las M iteraciones o hasta que se cumpla un criterio de convergencia, como la minimización de la función de pérdida o una mejora insignificante en el error. Finalmente, el modelo para la regresión se expresa como:

$$F(\mathbf{x}) = F_0(\mathbf{x}) + \sum_{m=1}^{M} \gamma_m h_m(\mathbf{x}) \tag{35}$$

Donde cada $h_m(\mathbf{x})$ ha sido entrenado para minimizar los errores residuales de las predicciones anteriores y cada γ_m ha sido optimizado para reducir la función de pérdida global.

5.6. *k-nearest neighbors (k-NN)*

El algoritmo k-NN o k Vecinos Más Cercanos fue desarrollado por Evelyn Fix y Joseph Hodges en 1951 (Fix, 1985) y pertenece al conjunto de técnicas del *machine learning* supervisado. Este modelo trata de buscar los k puntos más cercanos a un punto concreto para poder inferir su valor.

En la clasificación por k-NN, la salida es la pertenencia a una clase. Un objeto es clasificado como perteneciente a una clase si la mayoría de sus k vecinos pertenecen a esa clase: si se quiere clasificar un objeto, con $k = 5$ vecinos, con 3 vecinos de la clase A y 2 de la clase B, por mayoría el objeto se clasifica como clase A. Para evitar empates, se prefiere que el número de vecinos k seleccionado sea un número impar. En la regresión por k-NN, la salida es un valor de una propiedad del objeto; ese valor se puede calcular usando el promedio de los valores de los k vecinos.

5.6.1. *K-Nearest Neighbors Regression*

El algoritmo *K-Nearest Neighbors* (K-NN) para regresión predice el valor de una nueva instancia calculando las distancias a todas las instancias en el conjunto de entrenamiento, seleccionando las k más cercanas y promediando sus valores. La elección de la métrica de distancia para medir la similitud entre los puntos de datos es un aspecto crucial en este algoritmo. Además, elegir las métricas de distancia adecuadas define los límites de decisión, que a su vez crean diferentes regiones de puntos de datos. Existen varios métodos para calcular estas distancias, (IBM, 2024):

- **Distancia euclidiana.** Es la medición de distancia más común, la cual mide una línea recta entre el punto de búsqueda y el otro punto que se está midiendo. Se calcula de la siguiente manera siendo $\mathbf{x}_i$ y $\mathbf{x}_j$ puntos en un espacio de características de d dimensiones:

$$Dist_{euclidiana}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{d}(x_{il} - x_{jl})^2} \tag{36}$$

- **Distancia Manhattan.** Es una métrica de distancia popular que mide el valor absoluto de las diferencias entre dos puntos. Se representa en una cuadrícula y, a menudo, se le conoce como la geometría del taxi. Se calcula de la siguiente manera:

$$Dist_{Manhattan}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{d}|x_{il} - x_{jl}| \tag{37}$$

- **Distancia Minkowski.** Es una generalización de las métricas de distancia euclidiana y Manhattan, que permite la creación de otras métricas de distancia. Se calcula en un espacio de vectores normalizado, donde p es el parámetro que define el tipo de distancia utilizado. Si $p = 1$, se emplea la distancia Manhattan; si $p = 2$, se utiliza la distancia euclidiana. Se calcula de la siguiente manera:

$$Dist_{Minkowski}(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d}|x_{il} - x_{jl}|^p\right)^{1/p} \tag{38}$$

Es común normalizar los datos de cada una de las dimensiones del objeto, para evitar que un cambio en la escala en alguna de las dimensiones, afecte el resultado. Para seleccionar la k que es adecuada para los datos, se ejecuta el algoritmo k-NN varias veces con diferentes valores de k y se elige la k que reduce la cantidad de errores que se encuentran, mientras se mantiene la habilidad del algoritmo para hacer predicciones con precisión cuando se le dan datos que no ha visto antes.

5.6.2. Algoritmo para *k-nearest neighbors*

Se inicia con un conjunto $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$, con $\mathbf{x}_i$ que representa un vector de características en $\mathbb{R}^d$ y y_i la variable de interés y se aplican los siguientes pasos, (García González, 2023):

Algorithm 2 *k-nearest neighbors*

1: Se determina el valor óptimo de k.
2: Se selecciona la métrica de distancia.
3: Para cada ejemplo en los datos: se calcula la distancia entre el ejemplo de consulta y el ejemplo actual a partir de los datos y se agrega la distancia y el índice del ejemplo a una colección ordenada.
4: Se ordenan las distancias e índices desde el más pequeño hasta el más grande (en orden ascendente) por las distancias.
5: Se eligen las primeras k entradas de la colección ordenada.
6: Se obtienen las etiquetas de las entradas k seleccionadas.
7: Finalmente, se obtiene la media de las etiquetas k. Es decir, $\hat{y}_j = \frac{1}{k}\sum_{i \in I_K} y_i$ con I_k el número de vecinos más cercanos.

5.7. Redes neuronales

Las redes neuronales o *neural networks* en inglés es un modelo de aprendizaje automático inspirado en el cerebro humano, que se compone de una serie de neuronas interconectadas organizadas en capas. El primer modelo de red neuronal fue propuesto por McCulloch y Pitts (1943) en términos de un modelo computacional de actividad nerviosa. Este modelo era un modelo binario, donde cada neurona tenía un escalón o umbral prefijado, y sirvió de base para los modelos posteriores, (Barrera, 2012).

Las redes neuronales procesan la información a través de la propagación hacia adelante (*forward pass*) de la entrada a través de las capas ocultas y emiten una salida en la capa de salida. Durante el entrenamiento, los parámetros de la red (como los pesos y los sesgos de las neuronas) se ajustan de forma iterativa para minimizar la función de pérdida y mejorar la precisión del modelo. Para comprender mejor la estructura y funcionamiento de una red neuronal, es fundamental definir qué es una neurona artificial.

5.7.1. Neurona artificial

La neurona artificial es una unidad computacional básica que procesa y transmite información. Cada neurona recibe una serie de entradas, las procesa y a continuación, emite una salida. La salida puede ser enviada a otras neuronas o ser la salida final del modelo.

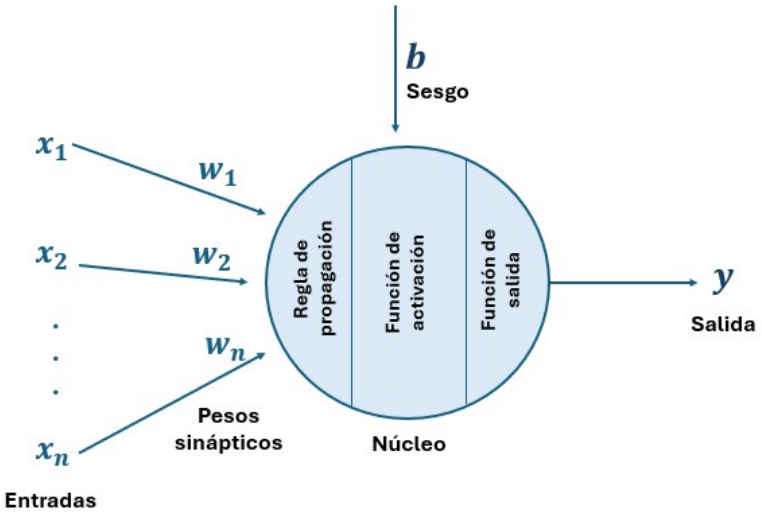


Figura 11: Esquema de una neurona artificial. Imagen adaptada de `https://www.cucei.udg.mx/sites/default/files/pdf/toral_barrera_jamie_areli.pdf` [Página Web en línea].

Una neurona está compuesta por tres partes: entrada, núcleo y salidas como se muestra en la Figura 11. Las entradas $x_1, x_2, ..., x_n$ reciben los datos que le permiten decidir a la neurona si estará activa o no, entre la entrada y el núcleo se tienen los pesos $(w_1, w_2, ..., w_n)$ que representan la memoria. En el núcleo se realizan todas las operaciones necesarias para determinar la salida de la neurona y las salidas devuelven la respuesta de la neurona; es decir, si está activa o no, representadas comúnmente como $y_1, y_2, ..., y_n$.

En el núcleo se llevan a cabo tres tipos de operaciones para determinar la salida de la neurona, las cuales se describen a continuación:

- **Regla de propagación**. La información que recibe la neurona artificial, son variables independientes. Los n-valores de entrada son multiplicados por sus respectivos pesos; es decir, en la sinopsis el vector entrada es multiplicado por el vector peso, dando como resultado una combinación lineal de las entradas y los pesos, adicionando el sesgo b como se muestra en la siguiente ecuación:

$$\mathbf{x} * \mathbf{w} = \sum_{i=1}^{n} x_i w_i + b \tag{39}$$

- **Función de activación**. Es una función matemática que transmite la información generada por la combinación lineal de los pesos y las entradas; es decir, permite la transmisión de información a través de las conexiones de salida. Esta información puede transmitirse sin modificaciones, en cuyo caso se utiliza una función identidad. Dado que se busca que la red sea capaz de resolver problemas cada vez más complejos, las funciones de activación generalmente introducen no linealidad en los modelos. Entre las funciones de activación más conocidas se encuentran, (Ramírez Vicente, 2019):

  - **Función lineal.** Esta función permite que la salida sea idéntica a la entrada. Si la neurona utiliza esta función de activación, su comportamiento será equivalente al de una regresión lineal.
  - **Función sigmoide.** Esta función es interesante porque permite que los valores grandes converjan a 1 y los valores pequeños a -1, lo que facilita la representación de probabilidades. Sin embargo, tiene el inconveniente de que su rango no está centrado en el origen. Este problema puede solucionarse mediante la función tangente hiperbólica, que ofrece un comportamiento muy similar.
  - **Función ReLU (unidad rectificada lineal).** Es una de las funciones de activación más comunes. Actúa como una constante cuando los valores son negativos y como una función lineal cuando los valores son positivos.

- **Función de salida.** La función de salida proporciona el valor de salida de la neurona, en base al estado de activación de la neurona. Es decir:

$$y_i = f(\mathbf{x} * \mathbf{w}) = f\left(\sum_{i=1}^{n} x_i w_i + b\right) \tag{40}$$

Después de comprender qué es una neurona artificial y cómo funciona, se puede entender mejor el funcionamiento de una red neuronal. Esta se puede definir como un grafo, una capa de entradas recibe la señal de entrada y la envía mediante estímulos a la siguiente capa oculta, que se encarga de procesar la información y transmitirla a la siguiente capa. Este proceso continúa hasta llegar a la última capa, la capa de salida, que proporciona la respuesta, (Ramírez Vicente, 2019).

A continuación, se detallan las principales partes de una red neuronal: las capas de entrada, las capas ocultas, las capas de salida y las conexiones entre estas capas. En la Figura 12, se puede observar una representación gráfica de una red neuronal artificial.

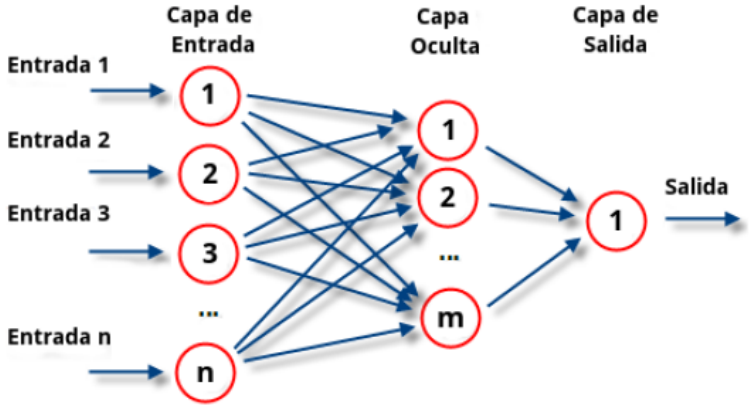


Figura 12: Red neuronal artificial. Imagen tomada de `https://es.wikipedia.org/wiki/PerceptrÃşn_multicapa` [Página Web en línea].

- **Capas de entrada**. Es la capa que recibe directamente la información proveniente de las fuentes externas de la red.
- **Capas ocultas**. Son internas a la red y no tienen contacto directo con el entorno exterior. El número de capas ocultas puede variar desde cero hasta un número elevado. Las neuronas en las capas ocultas pueden estar interconectadas de diversas formas, lo que, junto con su cantidad, determinan las diferentes topologías de redes neuronales.
- **Capas de salidas**. Transfieren información desde la red hacia el exterior.

Entre los resultados más destacados en el campo de las redes neuronales artificiales se encuentra el Teorema de Aproximación Universal. Este proporciona una base teórica sólida para entender la capacidad de las redes neuronales artificiales de aproximar funciones complejas. Este asegura que existe una red neuronal artificial hacia adelante, con sólo una capa oculta y una función de activación no polinomial, que puede aproximar una función continua en un conjunto compacto de $\mathbb{R}^n$. Sin embargo, estos resultados sólo prueban la existencia y no nos proporcionan un método para encontrar dicha red neuronal artificial, (Miranda, 2020). A continuación, se presenta la definición matemática del Teorema de Aproximación Universal.

Teorema de Aproximación Universal. Sea K un subconjunto compacto de $\mathbb{R}^n$, para cualquier $\epsilon > 0$ y cualquier función $g \in C(K)$, siendo $C(K)$ el conjunto de funciones continuas en K, existe una red neuronal hacia delante con una capa oculta $N : \mathbb{R}^n \longrightarrow \mathbb{R}$ definida como $N(x) = f_2(x) \circ f_1(x)$ con $f_1(x) = \phi_{w^{(1)},b_1}(x)$ y $f_2(x) = \phi_{w^{(2)},b_2}(x)$ tal que N es una aproximación de la función g de manera que para todo $x \in K$.

$$|N(x) - g(x)| < \epsilon \tag{41}$$

5.7.2. *Feedforward Neural Networks (FNN)*

Feedforward Neural Networks o redes neuronales hacia adelante o prealimentadas es un tipo de red neuronal en la que las conexiones entre las neuronas no forman ningún ciclo. La información fluye desde la entrada de la red hacia la salida a través de varias capas ocultas. La arquitectura *feedforward* es la más común en las redes neuronales y se utiliza ampliamente en aplicaciones de clasificación, predicción, reconocimiento de patrones y procesamiento de señales (Huet, 2023).

Una red neuronal *feedforward* puede definirse como una función compuesta que mapea un vector de entrada a un vector de salida a través de una serie de capas ocultas. Matemáticamente, se define como

un mapeo $f(\mathbf{x}; \mathbf{w}) = y$, tal que $f : \mathbb{R}^n \longrightarrow \mathbb{R}^l$, mediante el ajuste del valor $\mathbf{w}$ que resulta en la mejor aproximación de f^*, siendo $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ el conjunto de características, $\mathbf{y} = \{y_1, y_2, ..., y_l\}$ el conjunto de etiquetas generados por una función $f^*(x) = y$, y $\mathbf{w} = \{w_1, w_2, ..., w_k\}$ un conjunto de pesos (Valenzula Chaparro, 2020).

La estructura de esta red se suele representar como una cadena de composición de funciones $f_1, f_2, f_3 ... f_n$, de tal manera que se tiene $f(\mathbf{x}) = f_n(...f_3(f_2(f_1(\mathbf{x}))))$, siendo f_1 la capa de entrada, f_2 la segunda capa, f_3 la tercera capa y así sucesivamente hasta la capa de salida f_n. Su representación gráfica se muestra a continuación:

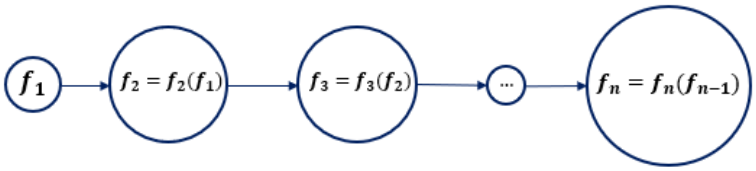


Figura 13: Estructura gráfica de las composiciones de funciones en una red neuronal *feedforward* hasta llegar a la salida f_n. Imagen adaptada de `http://repositorioinstitucional.uson.mx/bitstream/20.500.12984/6994/1/valenzuelachaparrohugodejesusl.pdf`

Los algoritmos de *machine learning* seleccionados y anteriormente expuestos (*random forest*, *gradient boosting*, *support vector machines*, *K-nearest neighbors* y *feedforward neural networks*) fueron seleccionados por sus características específicas y capacidades para resolver el problema propuesto. Los modelos *random forest* y *gradient boosting* fueron elegidos debido a su capacidad para manejar datos con alta variabilidad y su efectividad en la captura de interacciones no lineales entre las proporciones de votos, lo cual es fundamental en un problema composicional. El *support vector machines* y *K-nearest neighbors* se seleccionaron por su capacidad para identificar patrones complejos en los datos, incluso en situaciones con alta dimensionalidad, lo que es típico en datos composicionales y la *feedforward neural networks* se eligió por su habilidad para modelar interacciones complejas, adaptándose bien a los datos composicionales del problema electoral.

6. Predicción de las elecciones presidenciales en Colombia

A continuación, se detallan los pasos propuestos en la Sección 5.1 para implementar modelos de *machine learning*. 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.

6.1. 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.

A continuación, se presenta la clasificación ideológica de los candidatos presidenciales en cada año electoral, basada 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.

- **Elecciones presidenciales 1994**. Estas elecciones fueron particularmente significativas y llenas de controversia. En esta contienda, Ernesto Samper ganó la presidencia, superando a Andrés Pastrana en una reñida segunda vuelta. Este proceso electoral estuvo influenciado por el dinero del narcotráfico en la política ya que se reveló que la campaña de Samper había recibido aportes significativos del Cartel de Cali, lo cual generó una crisis política conocida como el Proceso 8000. Además, estas elecciones marcaron la primera vez que Colombia utilizó el sistema de dos vueltas para elegir presidente, lo que fomentó nuevas alianzas y estrategias políticas, (Comisión de la Verdad, 2023).

Tabla 2: Clasificación ideológica de los candidatos presidenciales de 1994.

Candidato	**Partido o movimiento político**	**Espectro ideológico**
Guillermo Alemán	Movimiento Orientación Ecológica	Centro Izquierda
Luis Rodríguez	Movimiento Nacional Progresista	Centro
José Cortés	Compromiso Cívico Cristiano	Centro Derecha
José Galat	Frente Moral	Centro Derecha
Andrés Pastrana	Partido Conservador Colombiano	Centro Derecha
Enrique Parejo	Movimiento Alternativa Democrática Nal.	Centro Izquierda
Gloria Gaitán	JEGA	Izquierda
Doris de Castro	Movimiento Cristiano Independiente	Centro Derecha
Efrain Torres	Crea- No a la Guerra	Centro
Ernesto Samper	Partido Liberal Colombiano	Centro Izquierda
Jorge Barbosa	Organización para la Paz Nal.	Centro Izquierda
Alberto Mendoza	Convergencia Nacional	Centro Derecha
Óscar Rojas	Somos Libres	Centro
Mario Diazgranados	C.G.T Cristiana	Centro
Miguel Maza	Movimiento Concentración Cívica Nacional	Derecha
Miguel Zamora	Protestemos	Centro Derecha
Antonio Navarro	Compromiso Colombia	Izquierda
Regina Betancourt	Movimiento Unitario Metapolítico	Centro

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://pdba.georgetown.edu/Elecdata/Col/pres94_1.html`), el espectro ideológico e hipervinculado la fuente de información utilizada para atribuir dicho espectro ideológico (en azul están identificados los espectros ideológicos que tienen asignación neutral).

- **Elecciones presidenciales 1998**. En estas elecciones, Andrés Pastrana del Partido Conservador y Horacio Serpa del Partido Liberal fueron los principales candidatos. Pastrana ganó la presidencia en una segunda vuelta muy reñida. Esta elección se destacó por la polarización política y las expectativas de cambio en un país afectado por la violencia y la corrupción. Además, un aspecto notable fue la alianza de Noemí Sanín y Antanas Mockus, quienes representaron una opción independiente a los partidos tradicionales, aunque no lograron llegar a la segunda vuelta, (El Tiempo, 1998).

Tabla 3: Clasificación ideológica de los candidatos presidenciales de 1998.

Candidato	**Partido o movimiento político**	**Espectro ideológico**
Beatriz Cuellar	Movimiento Unión Cristiana	Centro Derecha
Guillermo Alemán	Movimiento Orientación Ecológica	Centro Izquierda
Germán Rojas	Movimiento 19 de Abril	Izquierda
Jorge Betancur	Movimiento Unitario	Centro
Jesús Lozano	Movimiento por las Comunidades Negras	Centro Izquierda
Jorge Pulecio	Movimiento Participación Popular	Izquierda
Harold Bedoya	Movimiento Fuerza Colombia	Centro Derecha
Noemí Sanín	Movimiento Sí Colombia	Centro
Guillermo Nannetti	Movimiento Nacional Progresista	Centro
Efrain Díaz	Movimiento Político Ciudadanos en Formación	Centro
Francisco Córdoba	Movimiento Séptima Papeleta	Centro Izquierda
Horacio Serpa	Partido Liberal Colombiano	Centro Izquierda
Andrés Pastrana	Partido Conservador Colombiano	Centro Derecha

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://www.eltiempo.com/archivo/documento/MAM-758608`), el espectro ideológico e hipervinculado la fuente de información utilizada para atribuir dicho espectro ideológico (en azul están identificados los espectros ideológicos que tienen asignación neutral).

- **Elecciones presidenciales 2002**. El candidato Álvaro Uribe ganó con una rotunda mayoría, obteniendo más del 50 % de los votos en la primera vuelta. Uribe, quien se postuló bajo el movimiento Primero Colombia, centró su campaña en promesas de mano dura contra las guerrillas y una política de seguridad democrática. Su victoria reflejó el deseo de los colombianos por un cambio ante la violencia y la inseguridad que azotaban al país, (El Tiempo, 2002).

Tabla 4: Clasificación ideológica de los candidatos presidenciales de 2002.

Candidato	**Partido o movimiento político**	**Espectro ideológico**
Luis Garzón	Frente Soc. y Pol-Via Alterna-UD-Anapo-PSD-ASI-PSOC	Izquierda
Álvaro Uribe	Primero Colombia	Derecha
Noemí Sanín	Movimiento Si Colombia	Centro
Harold Bedoya	Movimiento Fuerza Colombia	Centro Derecha
Francisco Tovar	Movimiento Defensa Ciudadana	Centro Derecha
Guillermo Cardona	Movimiento Político Comunal y Comunitario Colombia	Centro Izquierda
Horacio Serpa	Partido Liberal Colombiano	Centro Izquierda
Augusto Lora	Movimiento 19 de Abril	Izquierda
Álvaro Cristancho	Movimiento Participación Común	Centro
Ingrid Betancourt	Partido Verde Oxigeno	Centro Izquierda
Rodolfo Rincón	Movimiento Participación Común	Centro

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://pdba.georgetown.edu/Elecdata/Col/pres02.html`), el espectro ideológico y la fuente de información utilizada para atribuir dicho espectro ideológico.

- **Elecciones presidenciales 2006**. En estas elecciones presidenciales, el candidato Álvaro Uribe fue reelegido con más del 60 % de los votos, logrando una contundente victoria en la primera vuelta. Esta reelección fue posible gracias a una reforma constitucional que permitió la reelección inmediata. La campaña de Uribe se centró en continuar su política de seguridad democrática y fortalecer la economía. Su principal oponente, el candidato Carlos Gaviria del Polo Democrático Alternativo, obtuvo alrededor del 22 % de los votos que fue insuficiente para generar una segunda vuelta, (El Mundo, 2006).

Tabla 5: Clasificación ideológica de los candidatos presidenciales de 2006.

Candidato	Partido o movimiento político	Espectro ideológico
Álvaro Uribe	Primero Colombia	Derecha
Carlos Gaviría	Polo Democrático Alternativo	Izquierda
Horacio Serpa	Partido Liberal Colombiano	Centro Izquierda
Antanas Mockus	Movimiento Alianza Social Indígena	Centro
Enrique Parejo	Movimiento Reconstrucción Democrática Nacional	Centro Izquierda
Álvaro Leiva	Movimiento Nacional de Reconciliación	Centro Derecha
Carlos Rincón	Movimiento Comunal y Comunitario de Colombia	Centro

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://pdba.georgetown.edu/Elecdata/Col/pres06.html`), el espectro ideológico y la fuente de información utilizada para atribuir dicho espectro ideológico.

- **Elecciones presidenciales 2010**. Juan Manuel Santos del Partido de la U ganó la presidencia en una segunda vuelta electoral después de que en la primera vuelta no se lograra una mayoría absoluta. Santos, quien había ocupado el cargo de Ministro de Defensa durante el mandato de Álvaro Uribe, derrotó a Antanas Mockus del Partido Verde, quien lideraba en la primera ronda. Estas elecciones se caracterizaron por su enfoque en temas de seguridad y relaciones internacionales, así como por mantener las políticas de seguridad democrática instauradas por Uribe, (Romero and Montaño, 2010).

Tabla 6: Clasificación ideológica de los candidatos presidenciales de 2010.

Candidato	Partido o movimiento político	Espectro ideológico
Juan Manuel Santos	Partido Social de Unidad Nacional	Centro Derecha
Antanas Mockus	Partido Verde	Centro
Germán Vargas	Partido Cambio Radical	Derecha
Gustavo Petro	Polo Democrático Alternativo	Izquierdo
Noemí Sanín	Partido Conservador Colombiano	Centro Derecha
Rafael Pardo	Partido Liberal Colombiano	Centro
Jairo Calderón	Movimiento Apertural Liberal	Derecha
Robinson Devia	Movimiento La Voz de la Consciencia	Centro
Jaime Araujo	Alianza Social Afrocolombiana	Centro Derecha

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://pdba.georgetown.edu/Elecdata/Col/pres10_1.html`), el espectro ideológico y la fuente de información utilizada para atribuir dicho espectro ideológico.

- **Elecciones presidenciales 2014**. En estas elecciones, Juan Manuel Santos, candidato del Partido de la U, ganó la reelección en segunda vuelta con el 50.95 % de los votos, venciendo a Óscar Iván Zuluaga del Centro Democrático. La campaña estuvo marcada por la polarización en torno al proceso de paz con las FARC y temas de seguridad y economía. La victoria de Santos permitió la continuación de las negociaciones de paz, (Molano, 2014).

Tabla 7: Clasificación ideológica de los candidatos presidenciales de 2014.

Candidato	Partido o movimiento político	Espectro ideológico
Óscar Iván Zuluaga	Centro Democrático	Derecha
Juan Manuel Santos	Partido Social de la Unidad Nacional	Centro Derecha
Marta Ramírez	Partido Conservador Colombiano	Centro Derecha

Candidato	Partido o movimiento político	Espectro ideológico
Clara López	Polo Democrático Alternativo	Izquierda
Enrique Peñalosa	Partido Alianza Verde	Centro

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://urosario.edu.co/observatorio-legislativo/elecciones-presidenciales-2014`), el espectro ideológico y la fuente de información utilizada para atribuir dicho espectro ideológico.

- **Elecciones presidenciales 2018**. Las elecciones presidenciales de Colombia en 2018 estuvieron marcadas por una notable diversidad de candidatos, incluyendo figuras independientes y seudocandidatos que jugaron un papel crucial en la contienda. Iván Duque, del partido Centro Democrático, resultó ganador en una segunda vuelta contra Gustavo Petro, líder de la Colombia Humana. La elección reflejó una polarización significativa en el país, con debates centrados en temas como el acuerdo de paz con las FARC y la economía, (Daza, 2020).

Tabla 8: Clasificación ideológica de los candidatos presidenciales de 2018.

Candidato	Partido o movimiento político	Espectro ideológico
Iván Duque	Partido Centro Democrático	Derecha
Gustavo Petro	Coalición Petro Presidente	Izquierda
Sergio Fajardo	Coalición Colombia	Centro
Germán Vargas	Coalición #MejorVargasLleras Ante Todo Colombia	Centro Derecha
Humberto de la Calle	Coalición Partido Liberal & Alianza Social Independiente	Centro Izquierda
Jorge Antonio Trujillo	Movimiento Político Todos somos Colombia	Centro Derecha
Viviane Morales	Partido Somos	Centro Derecha

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://pdba.georgetown.edu/Elecdata/Col/pres10_1.htmlhttps://moe.org.co/wp-content/uploads/2018/11/Resultados-Electorales-Elecciones-Presidenciales-2018_Digital.pdf`), el espectro ideológico y la fuente de información utilizada para atribuir dicho espectro ideológico.

- **Elecciones presidenciales 2022**. En las elecciones presidenciales de Colombia de 2022, Gustavo Petro del Pacto Histórico, obtuvo una histórica victoria, convirtiéndose en el primer presidente de izquierda en la historia del país. Petro, exalcalde de Bogotá y exguerrillero, venció en segunda vuelta a Rodolfo Hernández, un empresario populista, en un contexto marcado por el descontento social y las demandas de cambio estructural. Las elecciones reflejaron un país polarizado, con debates centrados en la desigualdad, la corrupción y la paz, (The Carter Center, 2022).

Tabla 9: Clasificación ideológica de los candidatos presidenciales de 2022.

Candidato	Partido o movimiento político	Espectro ideológico
Enrique Gómez	Partido Movimiento de Salvación Nacional	Centro Derecha
Federico Gutiérrez	Coalición Equipo por Colombia	Derecha
Gustavo Petro	Coalición Pacto Histórico	Izquierda
Ingrid Betancourt	Partido Verde Oxígeno	Centro
Jhon Milton Rodríguez	Colombia Justa Libres	Centro Derecha
Luis Pérez	Colombia Piensa en Grande	Centro Derecha
Rodolfo Hernández	Liga de Gobernantes Anticorrupción	Centro
Sergio Fajardo	Coalición Centro Esperanza	Centro

Nota. Esta tabla presenta el candidato presidencial, el partido o movimiento político (Tomado de `https://www.moe.org.co/wp-content/uploads/2022/11/DIGITAL-Resultados-Presidenciales-2022.pdf`), el espectro ideológico y la fuente de información utilizada para atribuir dicho espectro ideológico.

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

Inicialmente, se analizó la distribución de observaciones (municipios) con registro de votos por año electoral, como se muestra en la Figura 14; en esta 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.

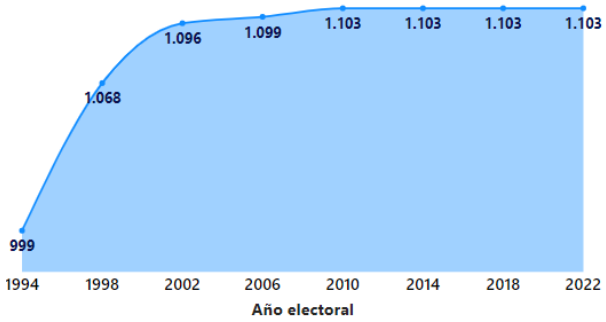


Figura 14: Distribución de municipios con registro de votos por año electoral. Elaboración propia.

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.

En la Figura 15, se observa que en 2002 y 2010 los votos por la Izquierda presentan una distribución estrecha y concentrada hacia valores bajos, lo que indica un apoyo relativamente limitado en esos años; un fenómeno comprensible en el contexto del conflicto armado y la política de Seguridad Democrática implementada por el presidente Álvaro Uribe, que consolidó una tendencia de apoyo a propuestas de Derecha y Centro-Derecha. Sin embargo, en 2006 se aprecia una mayor dispersión en la distribución,

reflejando que, aunque la Izquierda no obtuvo un porcentaje significativo en la mayoría de los municipios, en algunos alcanzó valores altos. Este comportamiento puede estar relacionado con el crecimiento de sectores que cuestionaban el impacto de la estrategia militar en la población civil y el surgimiento de líderes de izquierda como Carlos Gaviria, candidato del Polo Democrático Alternativo, quien alcanzó un apoyo notable en ese año.

A partir de 2014, la distribución comienza a ampliarse, un cambio que podría estar vinculado al inicio de las negociaciones de paz con las FARC bajo el gobierno de Juan Manuel Santos, lo cual generó un clima más propicio para que las ideas de izquierda ganaran terreno en la opinión pública. En 2018 y 2022, se observa un crecimiento considerable en los valores medianos y una dispersión más amplia, lo que indica un aumento significativo del apoyo a la Izquierda en varios municipios. En 2022, se registra la mayor mediana y la mayor dispersión, reflejando un cambio trascendental en el panorama político colombiano. Este cambio puede atribuirse al liderazgo de Gustavo Petro, quien representó una alternativa de izquierda consolidada y logró articular temas sociales como la reducción de la desigualdad, la justicia social y la implementación de políticas progresistas en favor de sectores históricamente marginados.

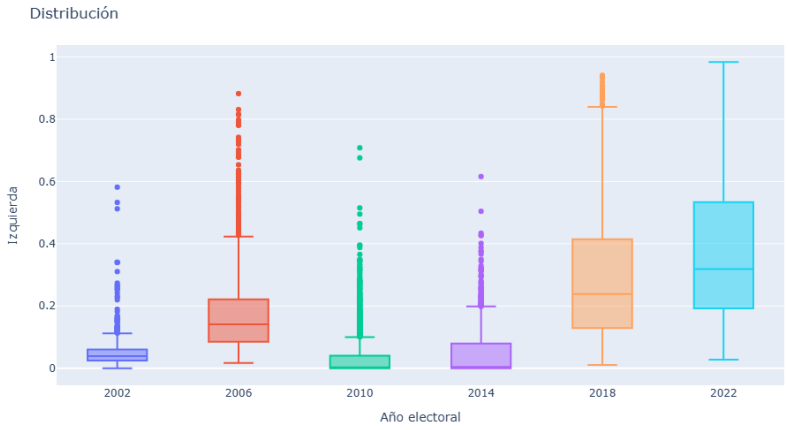


Figura 15: Distribución de votos del espectro ideológico de Izquierda en elecciones presidenciales (2002-2022). Elaboración propia.

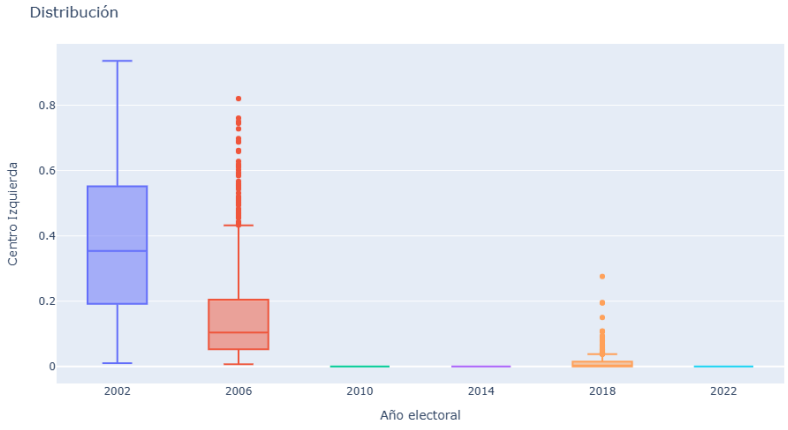


Figura 16: Distribución de votos del espectro ideológico de Centro-Izquierda en elecciones presidenciales (2002-2022). Elaboración propia.

La distribución de votos del espectro ideológico de Centro-Izquierda se muestra en la Figura 16. En

2002, los votos presentan una mayor dispersión y una mediana relativamente alta, lo que indica un apoyo importante en ese año, influenciado por figuras como Ingrid Betancourt, quien como candidato de una coalición alternativa, captó la atención de votantes interesados en propuestas de cambio sin una inclinación radical. Sin embargo, en 2006, aunque la dispersión se mantiene amplia, la mediana disminuye significativamente lo que sugiere una caída en el apoyo, posiblemente debido a la consolidación del Polo Democrático Alternativo como la principal fuerza de izquierda, lo cual atrajo a muchos votantes de Centro-Izquierda hacia esta opción.

También, se puede observar que a partir de 2010, los votos de Centro-Izquierda prácticamente desaparecen, con valores muy bajos y concentrados cerca de cero, que puede interpretarse como el fortalecimiento de propuestas de Derecha y Centro-Derecha lideradas por el expresidente Uribe y su sucesor Juan Manuel Santos en su primer periodo. Además, la polarización generada por las negociaciones de paz y el crecimiento de la Izquierda en los últimos años absorbieron el potencial electorado de Centro-Izquierda, contribuyendo a su debilitamiento y prácticamente a su desaparición en las elecciones más recientes.

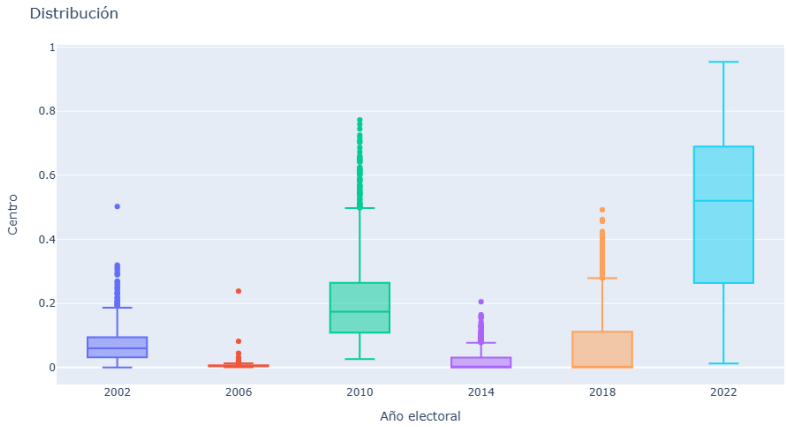


Figura 17: Distribución de votos del espectro ideológico de Centro en elecciones presidenciales (2002-2022). Elaboración propia.

La Figura 17, muestra que en 2002 y 2006 los votos para el espectro ideológico de Centro son bajos y presentan poca dispersión, lo que indica un apoyo limitado y poco significativo en un contexto donde las posiciones políticas estaban polarizadas y la preferencia se inclinaba hacia candidatos de Derecha y en menor medida de Izquierda. Sin embargo, en 2010 se observa un aumento notable tanto en la mediana como en la dispersión, lo que refleja un crecimiento considerable del apoyo al Centro en varios municipios que puede estar relacionado con la candidatura de Antanas Mockus, quien con su enfoque alternativo y su propuesta de lucha contra la corrupción, logró captar la atención de votantes que buscaban una opción distinta a los extremos políticos.

A partir de 2014, el apoyo disminuye nuevamente con una distribución más concentrada y valores bajos en la mediana; este descenso coincide con la reelección de Juan Manuel Santos quien aunque gobernó en el espectro de Centro-Derecha, orientó su segunda administración hacia un enfoque de paz reduciendo la visibilidad de otras propuestas de Centro. En 2018, la tendencia se mantiene con un leve aumento en el apoyo, probablemente impulsado por la candidatura de Sergio Fajardo, quien representó a un sector moderado sin lograr un apoyo masivo. Finalmente, en 2022 el apoyo al Centro se dispara, mostrando una amplia dispersión y una mediana mucho más alta, que se explica con el creciente descontento con la polarización política y el deseo de una opción que representara una visión equilibrada.

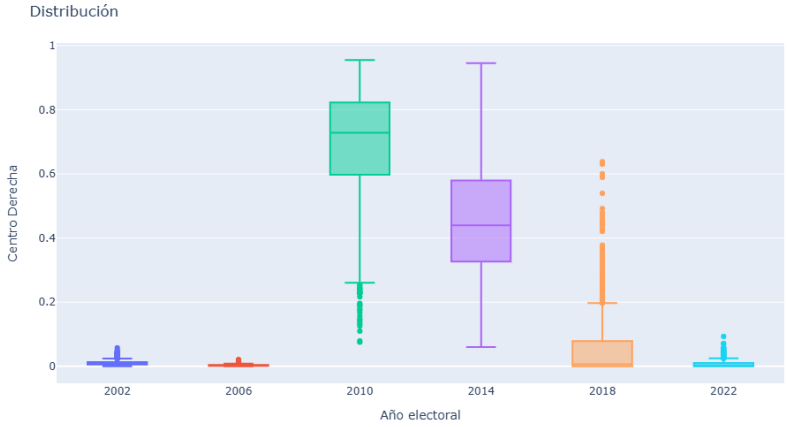


Figura 18: Distribución de votos del espectro ideológico de Centro-Derecha en elecciones presidenciales (2002-2022). Elaboración propia.

El comportamiento de votos del espectro ideológico de Centro-Derecha se observa en la Figura 18. En 2002 y 2006, la proporción de votos es baja y con poca dispersión, lo que indica un apoyo muy limitado en esos años, posiblemente debido a la fuerte polarización hacia opciones de Derecha más contundentes y la falta de un referente claro en el Centro-Derecha. Sin embargo, en 2010 se observa un gran aumento y una amplia dispersión, lo cual refleja un fuerte respaldo en varios municipios. Este crecimiento puede estar relacionado con la candidatura de Juan Manuel Santos considerado de Centro-Derecha logrando atraer a porción grande de votantes con una postura de continuidad de las políticas de Álvaro Uribe y una retórica moderada que le permitió conectar con sectores diversos.

En 2014, aunque la dispersión sigue siendo amplia la mediana disminuye levemente, posiblemente como resultado de la orientación hacia el proceso de paz en su segundo mandato, lo que generó tensiones con los sectores de Derecha y Centro-Derecha que apoyaban una posición más estricta respecto al conflicto. Para 2018 y 2022, el apoyo cae considerablemente y se puede explicar como el auge de candidatos de Derecha e Izquierda más radicales dejando a la Centro-Derecha en una posición de menor relevancia frente a una polarización intensa en el panorama político colombiano.

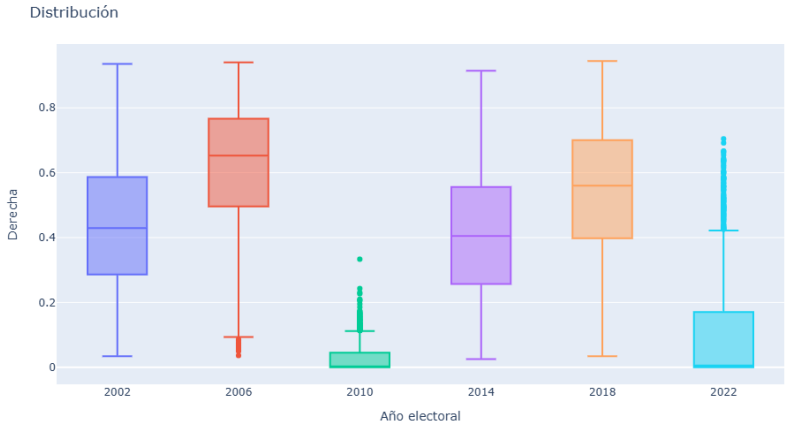


Figura 19: Distribución de votos del espectro ideológico de Derecha en elecciones presidenciales (2002-2022). Elaboración propia.

La Figura 19, presenta la evolución del apoyo al espectro ideológico de Derecha a lo largo de las elec-

ciones. En 2002 y 2006, el respaldo a la Derecha es sólido, con una distribución amplia y una mediana relativamente alta que evidencian un fuerte apoyo en la mayoría de los municipios. Pero en 2010, la situación cambia drásticamente: la concentración de votos disminuye notablemente y la mediana baja considerablemente reflejando una caída importante en el respaldo a candidatos de Derecha. Este cambio puede estar relacionado con el relevo en el liderazgo de Álvaro Uribe y el surgimiento de nuevas posturas políticas que reconfiguraron el panorama.

Desde 2014, el apoyo vuelve a crecer aunque sin llegar a los niveles previos de 2006, posiblemente influido por el contexto de las negociaciones de paz y las divisiones dentro del electorado de Derecha. En 2018, se observa otro repunte en la mediana, indicando una recuperación del respaldo a la Derecha, impulsada en gran parte por el regreso de discursos de seguridad y orden. Sin embargo, en 2022 la tendencia vuelve a caer y la dispersión de los votos es amplia, lo que refleja no solo una reducción en el apoyo general, sino también una mayor variabilidad en los resultados entre municipios, sugiriendo que el electorado de Derecha se fragmentó frente a una oferta electoral más diversa y polarizada.

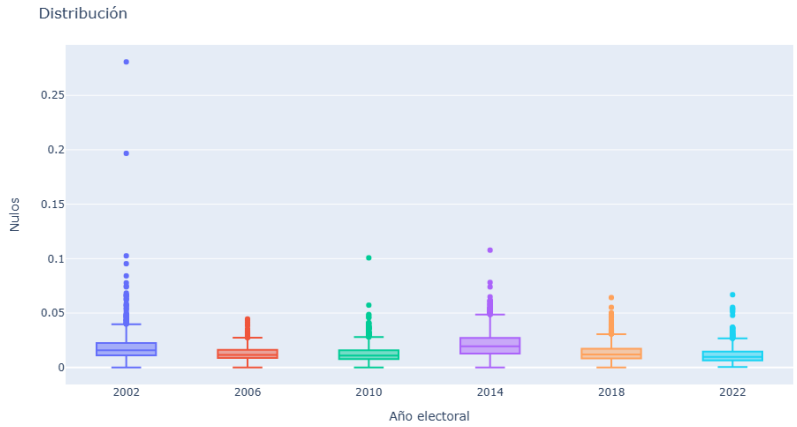


Figura 20: Distribución de votos nulos en elecciones presidenciales (2002-2022). Elaboración propia.

La Figura 20, presenta la distribución de votos nulos. En 2002, se observa una alta dispersión con algunos municipios registrando porcentajes de votos nulos elevados, lo cual podría estar relacionado con la falta de claridad en las opciones electorales y el contexto de polarización durante el gobierno de Álvaro Uribe, que pudo generar descontento o confusión entre los votantes. A partir de 2006, la tendencia se estabiliza con porcentajes bajos y consistentes en la mayoría de los municipios, posiblemente debido a mejoras en la educación electoral y campañas de sensibilización.

La distribución de votos en blanco se presenta en la Figura 21. En 2002, el comportamiento de los votos muestran una notable dispersión y altos porcentajes en algunos municipios, reflejando el descontento y la incertidumbre generados por el conflicto armado y la falta de opciones representativas, en un contexto donde la llegada de Álvaro Uribe y su política de Seguridad Democrática polarizaban el electorado. Desde 2006, la distribución se vuelve más concentrada con porcentajes bajos lo cual sugiere que la consolidación de fuerzas políticas más definidas redujo el uso del voto en blanco al ofrecer opciones claras. En 2014, aunque hay un leve aumento en la dispersión, probablemente relacionado con las reacciones al proceso de paz de Juan Manuel Santos, los votos en blanco permanecen en niveles bajos. Esta tendencia continúa en 2018 y 2022, con una concentración de votos en blanco y poca variabilidad, posiblemente debido a la creciente polarización que movilizó al electorado hacia opciones específicas, disminuyendo el atractivo del voto en blanco como alternativa significativa.

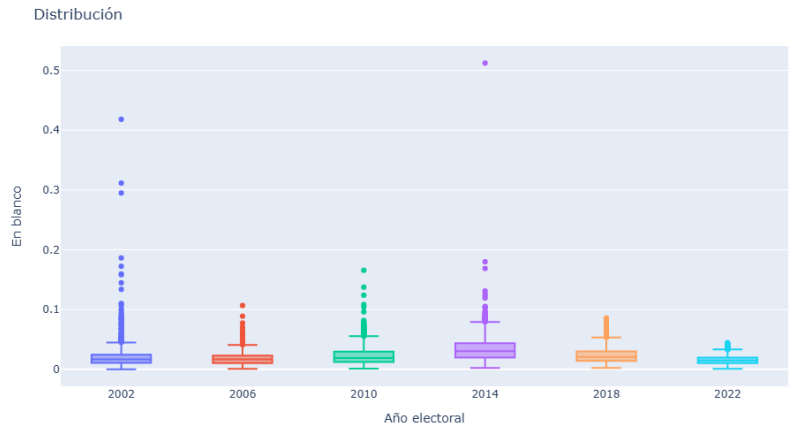


Figura 21: Distribución de votos en blanco en elecciones presidenciales (2002-2022). Elaboración propia.

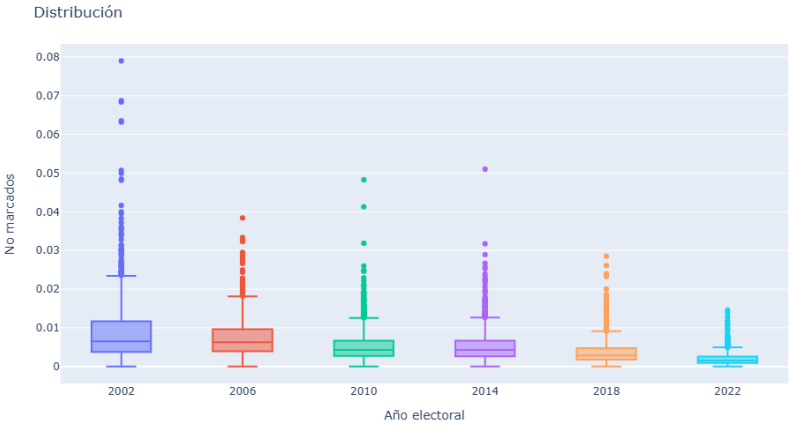


Figura 22: Distribución de votos no marcados en elecciones presidenciales (2002-2022). Elaboración propia.

La Figura 22, muestra la distribución de los votos no marcados en el que se destaca un patrón similar en los comportamientos a lo largo de los años. En 2002, se observa una mayor dispersión y algunos municipios con porcentajes relativamente altos de votos no marcados, posiblemente por descontento en el electorado en un periodo de gran polarización política. A partir de 2006, la distribución se concentra más con una disminución en los porcentajes de votos no marcados en la mayoría de los municipios, mostrando un comportamiento similar en los años posteriores. Aunque en elecciones como 2014 y 2018 aparecen valores atípicos, el comportamiento general de los votos no marcados se mantiene estable y en niveles bajos, lo cual sugiere una consistencia en la comprensión del proceso electoral y en la decisión del electorado de marcar sus votos.

A partir de los análisis expuestos anteriormente, se observa que la proporción de votos en blanco, nulos y no marcados es consistentemente baja en las diferentes elecciones analizadas, por lo que hemos decidido excluir estos tipos de votos de la composición de los datos. En lugar de ello, nos enfocaremos exclusivamente en las componentes del espectro ideológico unidimensional de Izquierda-Derecha. Esta decisión nos permite centrar el análisis en los patrones de apoyo ideológico, asegurando que las variaciones en la distribución de votos reflejen únicamente las preferencias ideológicas de los votantes en cada elección.

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 (Ver Anexo 1). 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. Para estimar los valores faltantes se utilizó 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 10: 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 indicadores varían en función de las posiciones ideológicas, permitiendo identificar algunas tendencias.

La Figura 23, presenta la relación entre los deciles de la tasa de mortalidad por cada 1,000 habitantes y el promedio de votos en los cinco espectros ideológicos. Este análisis revela patrones distintos en cada espectro, influenciados por el contexto socioeconómico y las prioridades de los votantes en áreas con diferentes niveles de mortalidad. El espectro de Izquierda muestra una caída inicial en los primeros deciles, posiblemente porque en zonas con menor mortalidad los votantes priorizan políticas económicas antes que cambios sociales profundos. Sin embargo, hacia los deciles más altos, la Izquierda gana apoyo, lo cual puede reflejar la atracción de votantes en áreas de mayor vulnerabilidad que buscan reformas sociales significativas. El Centro Izquierda mantiene una tendencia descendente uniforme, pero presenta una leve recuperación en los últimos deciles quizás captando a quienes buscan un cambio moderado en áreas más afectadas. El Centro, por otro lado, muestra un aumento constante en el apoyo, especialmente en los deciles altos, lo que podría estar relacionado con su enfoque pragmático que resuena en contextos de alta mortalidad donde los votantes prefieren propuestas equilibradas y no polarizadas. La Centro Derecha exhibe un pico temprano alrededor del decil 3, seguido de una caída lo que sugiere un apoyo

inicial en zonas de mortalidad intermedia, donde prevalece la preocupación por el orden y la estabilidad antes de dar paso a opciones más a la Derecha o Izquierda. Finalmente, el espectro de Derecha muestra un crecimiento inicial, alcanzando su punto máximo cerca del decil 7, para luego descender, reflejando que en áreas con mortalidad creciente el apoyo a políticas de Derecha disminuye conforme se hacen más visibles las necesidades de cambio estructural.

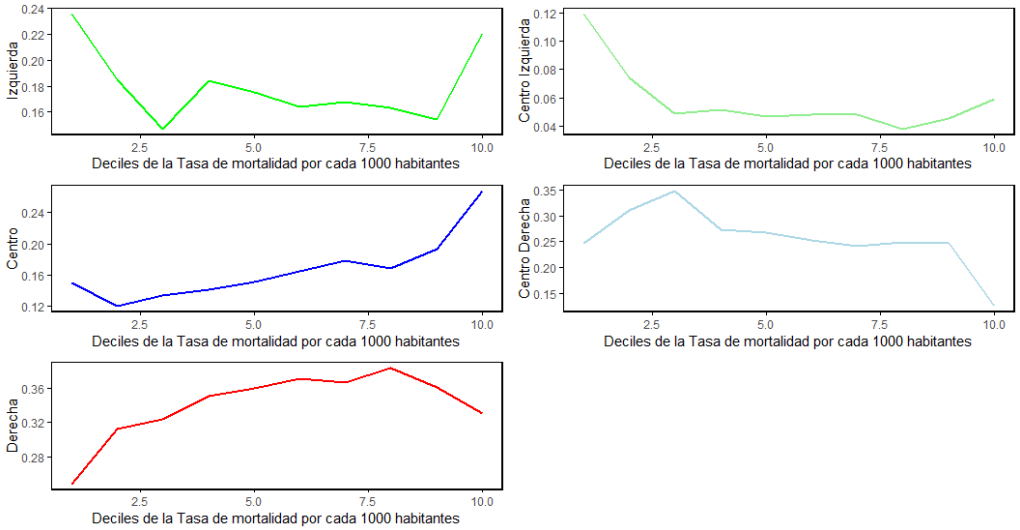


Figura 23: Relación entre la tasa de mortalidad y el promedio de votos de cada espectro ideológico. Elaboración propia.

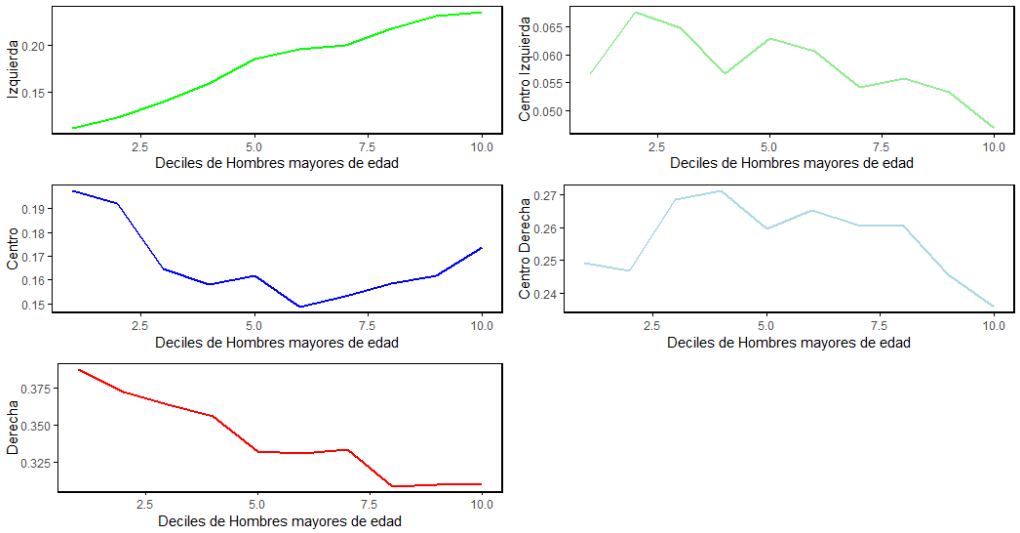


Figura 24: Relación entre el número de hombres mayores de edad y el promedio de votos de cada espectro ideológico. Elaboración propia.

En la Figura 24, se observa la relación entre el número de hombres mayores de edad y el promedio de votos de cada espectro ideológico. La Izquierda experimenta un aumento constante en el apoyo conforme se avanza a los deciles más altos, lo que sugiere que los hombres mayores de edad tienden a identificarse cada vez más con propuestas de cambio social y redistribución económica. En contraste, el Centro Izquierda presenta un patrón más irregular, con una tendencia descendente en los deciles superiores, lo cual podría indicar una menor conexión de esta ideología con las prioridades de los hombres de mayor edad en

comparación con otras opciones. El Centro, muestra una recuperación gradual tras una disminución inicial, sugiriendo que las posturas equilibradas ganan atractivo en los deciles más altos, posiblemente por su enfoque pragmático. Por su parte, el espectro de Centro Derecha exhibe un pico en los primeros deciles, seguido de una caída sostenida lo que sugiere que su enfoque no es tan atractivo para hombres de edad. Finalmente, la Derecha muestra una disminución constante en los deciles superiores, lo cual puede reflejar un distanciamiento progresivo de los hombres mayores de edad hacia las políticas de esta ideología, prefiriendo enfoques que atiendan mejor las necesidades de estabilidad y apoyo social en esta etapa de la vida.

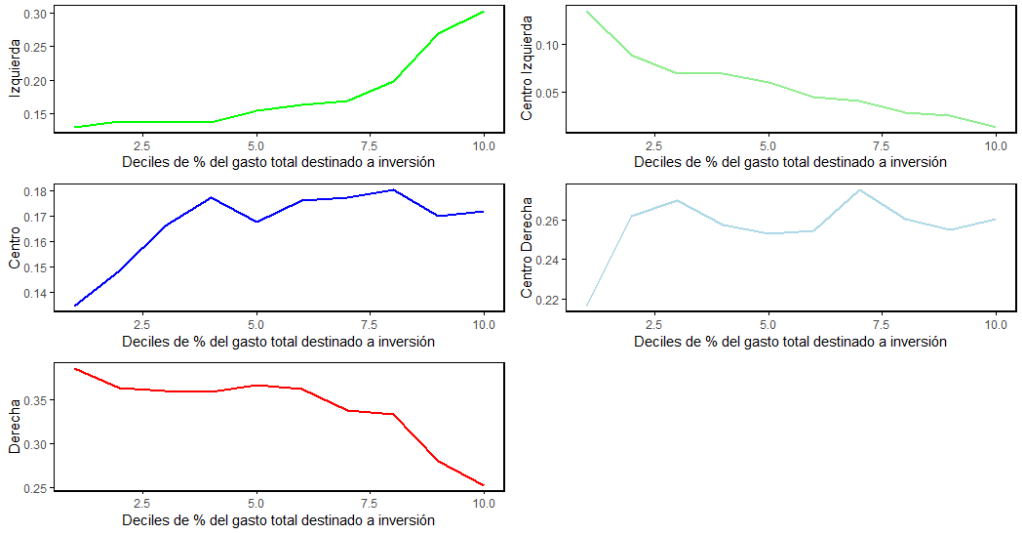


Figura 25: Relación entre el porcentaje del gasto total destinado a inversión y el promedio de votos de cada espectro ideológico. Elaboración propia.

La Figura 25, muestra cómo varía el apoyo a cada espectro ideológico en función de los deciles del porcentaje de gasto destinado a inversión. El espectro de Izquierda destaca con un aumento significativo, duplicando su apoyo al pasar de un 0.15 en los deciles iniciales a aproximadamente 0.30 en el decil más alto lo cual sugiere que mayores inversiones públicas están asociadas a un crecimiento en el respaldo a políticas de corte izquierdista. Por otro lado, el Centro Izquierda experimenta una caída constante, pasando de 0.11 a prácticamente 0, lo que indica una pérdida casi total de apoyo conforme aumenta el gasto en inversión, posiblemente reflejando una percepción de insuficiencia en sus propuestas ante políticas de inversión más robustas. El Centro muestra un incremento moderado desde 0.14 hasta 0.18 en los primeros deciles, para luego estabilizarse y caer ligeramente, lo cual sugiere que su enfoque es atractivo en contextos de inversión media. Centro Derecha alcanza su punto máximo alrededor del decil 3 con un valor de 0.27, pero luego tiende a estabilizarse y a descender levemente lo que puede indicar una pérdida de atractivo en escenarios de inversión elevada. Finalmente, el espectro de Derecha muestra una tendencia de caída continua, bajando de 0.36 a 0.25, posiblemente porque la ideología derechista pierde atractivo en situaciones de mayor inversión pública.

La relación entre los ingresos tributarios per cápita y los cinco espectros ideológicos se presentan en la Figura 26. La Izquierda muestra un incremento sostenido, pasando de aproximadamente 0.15 a 0.25 en los deciles superiores, indicando que en contextos de mayor capacidad recaudatoria las propuestas de Izquierda encuentran mayor afinidad. En contraste, el Centro Izquierda experimenta una caída continua desde alrededor de 0.12 hasta casi desaparecer en los deciles altos, reflejando una pérdida total de apoyo y sugiriendo una desconexión con regiones de altos ingresos tributarios. El Centro, por su parte, muestra un incremento importante, triplicando su apoyo de 0.1 a 0.30 lo cual sugiere que posturas pragmáticas y moderadas son cada vez más atractivas en contextos de mayores recursos fiscales. La Centro Derecha alcanza su máximo apoyo en el decil 3 rondando los 0.43, pero su respaldo disminuye hacia los deciles

superiores lo que indica una preferencia inicial que se desvanece en entornos de alto ingreso tributario. Finalmente, la Derecha exhibe un patrón fluctuante, comenzando cerca de 0.39 y oscilando con subidas y bajadas hasta caer abruptamente a 0.24 en los últimos deciles. En general, Izquierda y Centro emergen como los espectros ideológicos ganadores en escenarios de mayores ingresos tributarios, mientras que Centro Izquierda, Centro Derecha y Derecha experimentan pérdidas significativas en estos contextos.

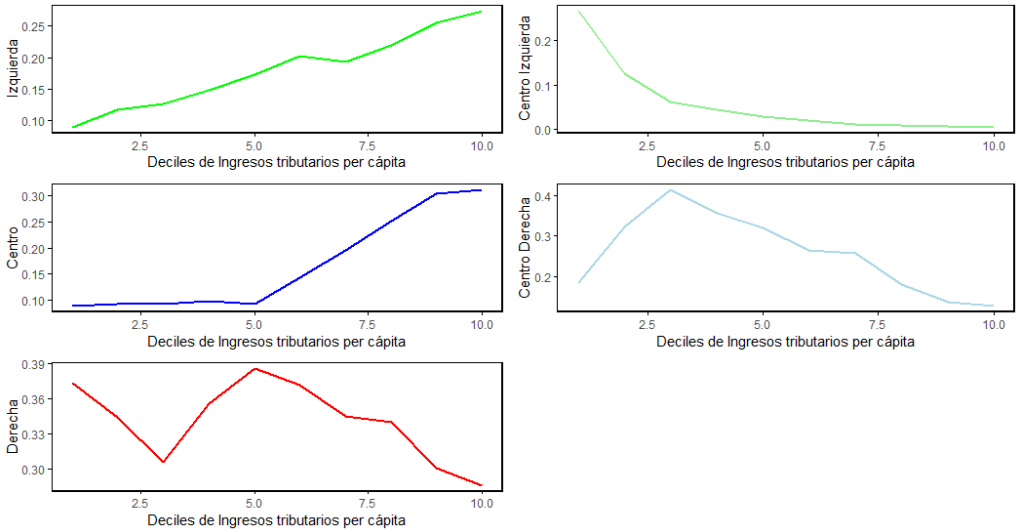


Figura 26: Relación entre los ingresos tributarios per cápita y el promedio de votos de cada espectro ideológico. Elaboración propia.

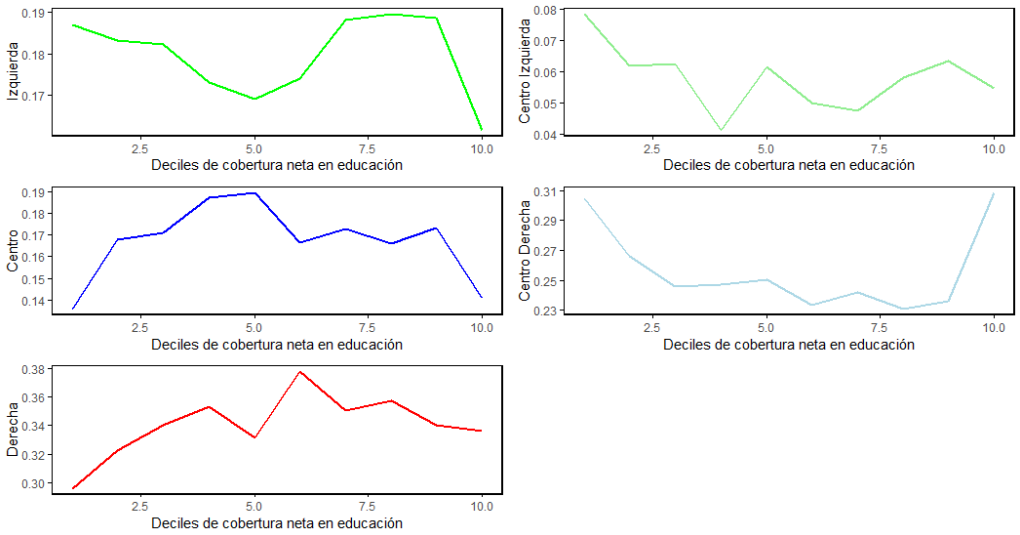
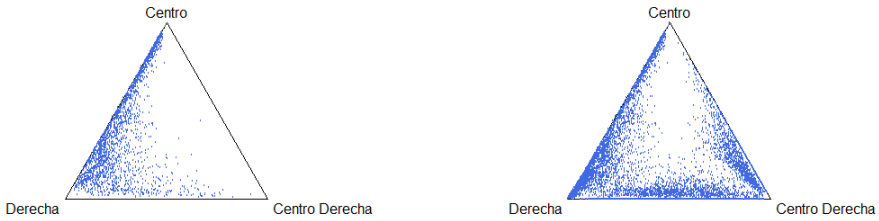


Figura 27: Relación entre la cobertura neta en educación y el promedio de votos de cada espectro ideológico. Elaboración propia.

En la Figura 27, se observa la relación entre los cinco espectros ideológicos y la cobertura neta en educación. La Izquierda muestra estabilidad hasta el decil 7, pero experimenta una caída notable en los deciles más altos, sugiriendo que en contextos de mayor cobertura educativa, su atractivo disminuye. El Centro Izquierda enfrenta una caída temprana y no logra recuperarse, lo que indica una pérdida de apoyo en áreas con una educación más accesible. En cambio, el Centro experimenta un crecimiento moderado y constante lo que sugiere que las posturas centristas resuenan mejor en regiones con amplia cobertura

educativa. Tanto el Centro Derecha como la Derecha presentan un patrón mixto, con picos tempranos y caídas leves hacia el final, lo cual sugiere que su apoyo es más variable en función de la cobertura educativa. En general, el Centro parece ser el espectro mejor posicionado en áreas con mayor acceso a la educación, mientras que los extremos ideológicos, tanto de Izquierda como de Derecha, muestran una mayor variabilidad en su respaldo en estos contextos.

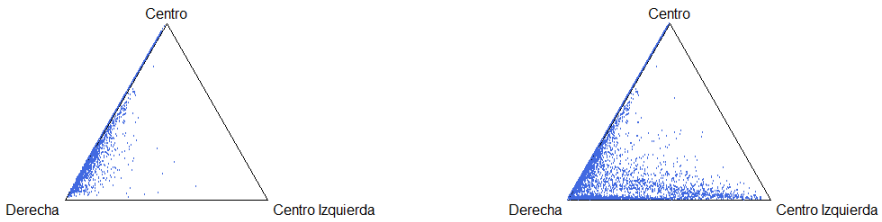
Ahora, se presenta el análisis de las elecciones presidenciales desde una perspectiva composicional permitiendo explorar cómo se distribuyen las preferencias electorales entre los distintos espectros ideológicos.



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

Figura 28: 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.

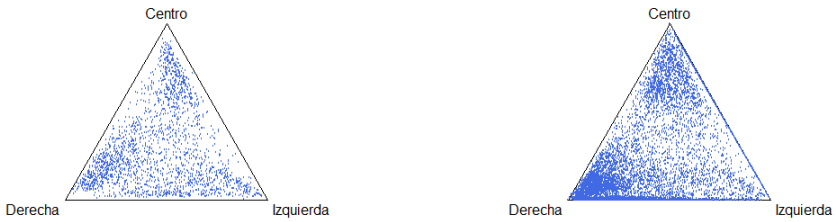
La Figura 28, presenta los diagramas ternarios que muestran las dinámicas ideológicas de Centro, Centro Derecha y Derecha en las elecciones de 2018 y 2022 (a) frente al total histórico de elecciones (b). En el gráfico de 2018 y 2022, se observa una dispersión significativa de los votos hacia las posiciones 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 de 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 29: Diagramas ternarios de distribución ideológica (Centro, Derecha, Centro Izquierda) para las elecciones 2018-2022 y totales, tomando el espectro de Centro como el punto fijo. Elaboración propia.

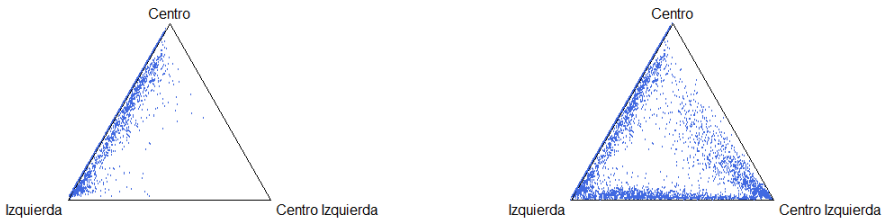
En el primer diagrama ternario (a) que se presenta en la Figura 29, se observa una concentración marcada hacia la Derecha, lo que indica un fuerte apoyo hacia esta posición en ese periodo específico, con menor dispersión hacia el Centro Izquierda. En cambio, el segundo diagrama (b), que representa las elecciones históricas, muestra una distribución más amplia, con los votos repartidos no solo hacia la Derecha sino también hacia el Centro y, en menor medida, el Centro Izquierda. Esto sugiere que en elecciones recientes, los votantes han preferido opciones de Derecha, mientras que históricamente ha habido una mayor diversidad en las preferencias ideológicas, incluyendo un apoyo significativo al Centro y algo de respaldo al Centro Izquierda.



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

Figura 30: 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 30, 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.

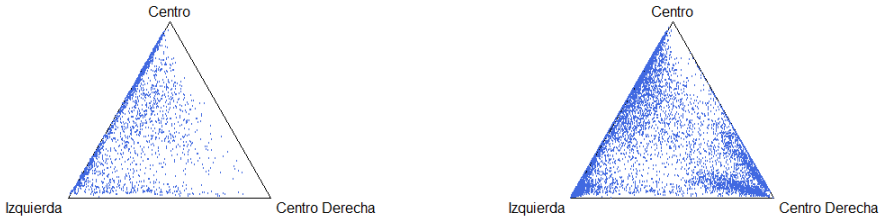


(a) Elecciones 2018-2022. (b) Elecciones totales.

Figura 31: 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 31, muestra que los votos están concentrados hacia el vértice del Centro e Izquierda, indicando una inclinación mayoritaria del electorado hacia

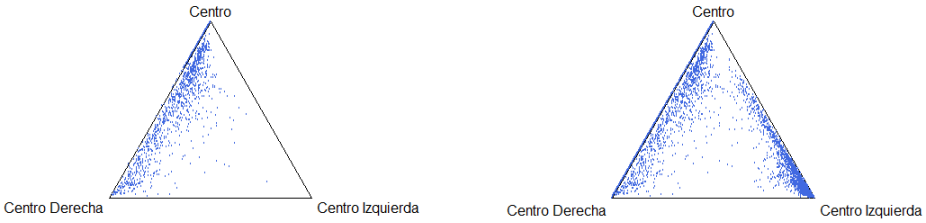
posiciones de Izquierda en este periodo reciente. Por otro lado, el segundo diagrama (b) que representa las elecciones históricas, la dispersión es más amplia extendiéndose de manera más uniforme entre los tres espectros ideológicos. Esto refleja que a lo largo del tiempo, los votantes han mostrado una 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.



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

Figura 32: Diagramas ternarios de distribución ideológica (Centro, Izquierda, Centro Derecha) para las elecciones 2018-2022 y totales, tomando el espectro de Centro como el punto fijo. Elaboración propia.

La Figura 32, presenta en el primer diagrama ternario (a) una fuerte concentración de votos hacia el Centro y Centro Derecha, lo cual indica una preferencia marcada por posiciones moderadas en este periodo reciente, con menor dispersión hacia la Izquierda. Por otro lado, el segundo diagrama (b) muestra una distribución más homogénea, extendiéndose de forma más equilibrada entre los tres espectros ideológicos. Esto muestra que a lo largo del tiempo los votantes han mostrado una mayor diversidad en sus preferencias aunque en menor valor para la Izquierda. En términos generales, los diagramas muestran que en las elecciones recientes, los votantes han optado predominantemente por el Centro e Izquierda, mientras que en las elecciones históricas han exhibido una pluralidad ideológica más amplia entre estos tres espectros.



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

Figura 33: Diagramas ternarios (Centro, Centro Derecha, Centro Izquierda) para las elecciones 2018-2022 y totales, tomando el espectro de Centro como el punto fijo. Elaboración propia.

Los diagramas ternarios en la Figura 33, muestran que históricamente el electorado ha mostrado una

diversidad de preferencias ideológicas más equilibrada entre estos tres espectros, mientras que en el periodo 2018-2022 los votantes se concentraron más en el Centro. En particular, el primer diagrama (a) refleja una fuerte concentración de puntos hacia el vértice del Centro indicando una preferencia destacada por posturas moderadas en este periodo reciente, con menos dispersión hacia el Centro Derecha y Centro Izquierda. En cambio, en el segundo diagrama (b) la dispersión es más uniforme con una mayor distribución de votos hacia las posiciones de Centro y Centro Izquierda además del Centro.

6.3. Entrenamiento y evaluación: estimando la predicción electoral

En esta sección, 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 el 70 % de los datos de las elecciones presidenciales entre 2002 y 2022, y se evaluó su rendimiento en el 30 % restante. Es importante mencionar que para las transformaciones *alr* e *ilr*, se optó por tomar el espectro de Centro como componente de referencia y base ortonormal, respectivamente, debido a su posición intermedia en el espectro ideológico unidimensional facilitando la interpretación de las transformaciones log-cocientes de las demás categorías ideológicas en función de su proximidad o divergencia respecto al Centro.

6.3.1. *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 11: 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.

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

En la Figura 34, 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.

6.3.2. *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 12: 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 12, 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 13.

Tabla 13: 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.

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 35: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. 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 35. 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.

6.3.3. *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 14: Configuraciones óptimas para cada espectro ideológico en cada transformación.

Modelo	Espectro ideológico	Transformación	*n trees*	*interaction. depth*	*n. minobsinnode*	*shrinkage*	*bag. fraction*
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. Elaboración propia.

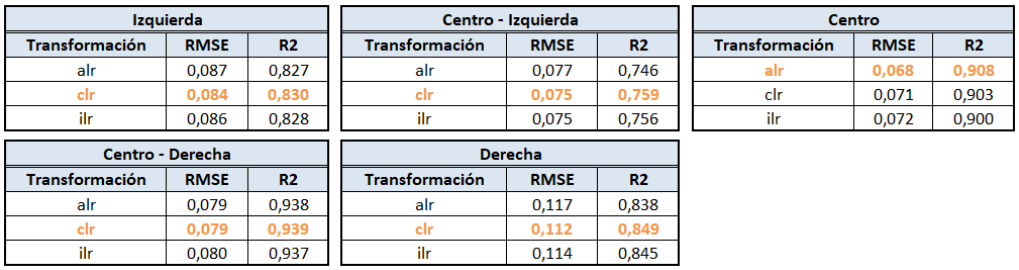


Figura 36: Métricas R^2 y $RMSE$ para cada transformación en cada espectro ideológico. Elaboración propia.

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 36. 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.

6.3.4. *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 15: 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.

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

La Figura 37, 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.

6.3.5. *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 dropout*. A continuación, se presentan las configuraciones óptimas seleccionadas para cada espectro ideológico con cada una de las transformaciones.

Tabla 16: 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.

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

Las tablas en la Figura 38, 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.

Modelo	Izquierda			Centro - Izquierda			Centro			Centro - Derecha			Derecha		
	Transformación	RMSE	R2	Transformación	RMSE	R2	Transformación	RMSE	R2	Transformación	RMSE	R2	Transformación	RMSE	R2
Support	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
Vector	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
Regression	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	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
Forest	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
Regression	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	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
Boosting	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
Regression	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	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
Neighbors	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
Regression	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	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
Neural	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
Network	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

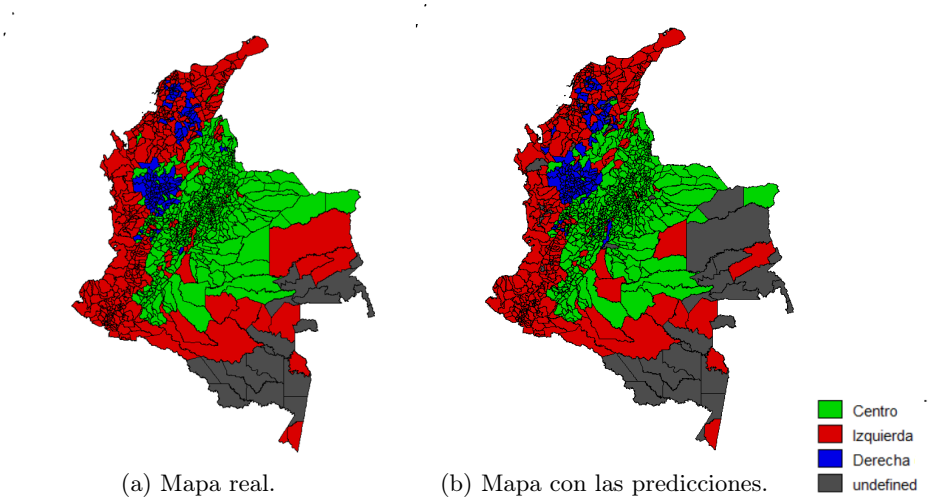
Figura 39: Consolidado de las métricas R^2 y $RMSE$ de los modelos para cada transformación en cada espectro ideológico. Elaboración propia.

La Figura 39, 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*.

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 40, 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 donde diferentes sectores de la población tienen posiciones ideológicas muy marcadas generando patrones de voto que varían significativamente entre localidades. Este comportamiento se puede observar al

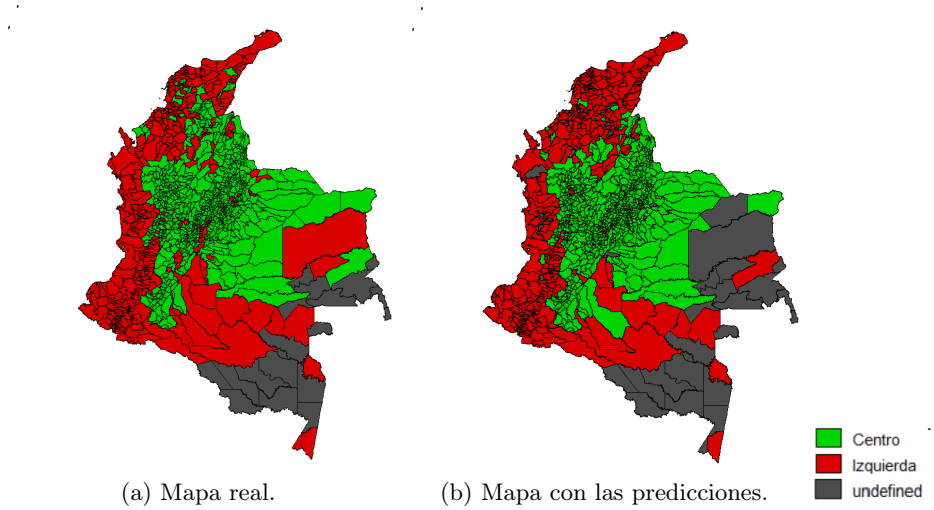
analizar las estimaciones; la estimación de votos para el espectro de Izquierda de 40.8 % y para Derecha de 45.5 %. De igual forma, en Tunja se observa polarización con una estimación de votos del 39 % para el espectro de Izquierda y un 45.7 % para el espectro de Centro. En cambio, para Armenia se observa 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 40: 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.



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

Figura 41: Mapas del espectro ideológico ganador en cada municipio de Colombia para la segunda vuelta de las elecciones presidenciales 2022. Elaboración propia.

La Figura 41, 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.

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.

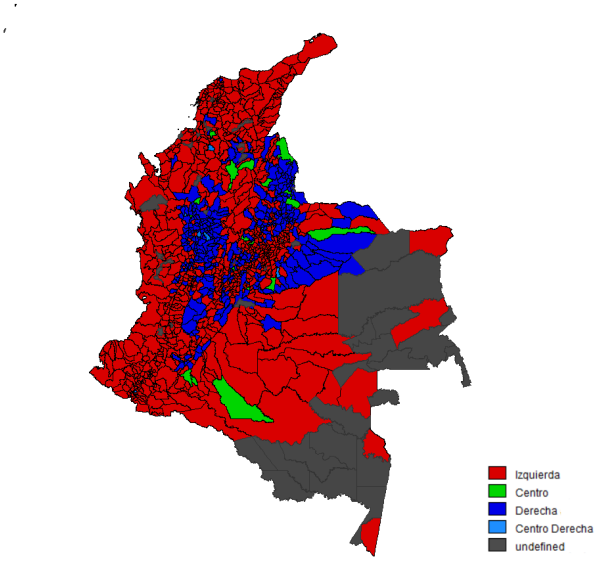


Figura 42: Mapas del espectro ideológico ganador en cada municipio de Colombia para la primera vuelta de las elecciones presidenciales 2026. Elaboración propia.

La Figura 42, 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.

7. 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 and 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.

8. Limitaciones

Una de las principales limitaciones de este estudio fue la dificultad para acceder a una mayor cantidad de variables auxiliares a nivel municipal, las cuales podrían haber aportado información relevante y enriquecido los modelos predictivos. Además, nos enfrentamos a la incompletitud de algunas variables auxiliares, lo que afectó la precisión de los análisis realizados. En cuanto a la representación ideológica, se utilizó un espectro unidimensional para clasificar a los municipios, lo que pudo haber sesgado los resultados, ya que no se capturó la complejidad y diversidad de las posiciones políticas que podrían existir dentro de cada espectro. Otra limitación significativa fue la asignación del espectro ideológico a los presidentes, la cual estuvo influenciada por su variabilidad a lo largo del tiempo, lo que dificultó una categorización precisa y consistente.

Asimismo, debido a la naturaleza composicional de los datos, nos enfrentamos a la necesidad de completar los municipios donde no se presentaba votación para algún espectro ideológico. Esto fue necesario, ya que, conforme a las propiedades de los datos composicionales, ninguna componente puede ser igual a cero, lo que obligó a realizar ajustes en los datos. Además, las transformaciones log-coientes aplicadas en los análisis, que son fundamentales para trabajar con datos composicionales, enfrentan una limitación inherente: los logaritmos no están definidos para cero, lo que introduce dificultades adicionales en la interpretación y manipulación de los datos. Las restricciones y limitaciones asociadas a estas transformaciones también pudieron haber afectado la representatividad de los resultados obtenidos. Finalmente, la falta de completitud de los datos electorales para periodos más antiguos limitó nuestra capacidad para realizar un análisis temporal más amplio y detallado, restringiendo la cobertura histórica del estudio.

9. Trabajos futuros

En futuros trabajos, se podría explorar el uso de técnicas de *machine learning* más complejas y avanzadas que permitan mejorar la capacidad predictiva del modelo, como redes neuronales profundas, que integren múltiples algoritmos para obtener un rendimiento más robusto y preciso. Además, sería importante considerar enfoques alternativos para la representación del espectro ideológico de los candidatos presidenciales, como modelos bidimensionales o tridimensionales, que puedan capturar de manera más precisa la complejidad y variabilidad de las posiciones políticas, en lugar de emplear un enfoque unidimensional Izquierda-Derecha que podría simplificar en exceso las dinámicas ideológicas.

En cuanto al tratamiento de los datos composicionales, sería conveniente explorar el uso de modelos especializados en este tipo de datos, como aquellos basados en la teoría de composiciones, que permiten modelar las relaciones proporcionales entre las categorías ideológicas de manera más eficiente. También, el uso de técnicas de partición más avanzadas, como la validación cruzada estratificada, sería una opción prometedora para evaluar de forma más robusta el modelo, asegurando que cada conjunto de datos mantenga la distribución adecuada de las proporciones ideológicas.

Finalmente, es importante tener en cuenta el posible desbalanceo en las clases al aplicar estos enfoques, ya que algunas categorías ideológicas podrían estar sobrerrepresentadas o subrepresentadas en ciertos municipios. El uso de técnicas como el ajuste por ponderación de clases o el muestreo estratificado podría ayudar a mitigar los efectos de este desbalanceo, mejorando la generalización del modelo y asegurando una representación más justa de todos los espectros ideológicos en el análisis.

Referencias

Aguilar López, J. and 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.

Aitchison, J. (1994). Principles of compositional data analysis. *Lecture Notes-Monograph Series*, pages 73–81.

Amat Rodrigo, J. (2020). Árboles de decisión, random forest, gradient boosting y c5.0. `https://cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#Random_Forest`.

Andes Universidad, U. (2023). Partidos políticos en colombia: definición, funciones y lista actualizada. `https://programas.uniandes.edu.co/blog/partidos-politicos-de-colombia`.

Arcila-Calderón, C., Ortega-Mohedano, F., Jiménez-Amores, J., and Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *Profesional de la Información*, 26(5):973–982.

Atencia, W., Rambal, J., Bustillo, J., et al. (2020). Analizador de tweets asociados a la política y polarización colombiana. .

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

Barrera, J. A.-T. (2012). Redes neuronales. *Universidad de Guadalajara Disponible en: http://www. cucei. udg. mx/sites/default/files/pdf/toral_ barrera_ jamie_ areli. pdf [Visitada en octubre de 2016]*.

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

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

Cabrera-Tenecela, P. (2021). Revisión bibliográfica del pronóstico electoral a través del big data. *South American Research Journal*, 1(2):27–35.

Campo León, E. and Alcalá Nalvaiz, J. T. (2017). Introducción a las máquinas de vector soporte (svm) en aprendizaje supervisado. *Trabajo de Fin de Grado en Matemáticas, Universidad de Zaragoza. Obtenido de https://zaguan. unizar. es/record/59156/files/TAZ-TFG-2016-2057. pdf.*

Carbonell, D. G. E. (2020). Ärboles de regresión. algunos algoritmos y extensiones a métodos de consenso. `https://cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#Random_Forest`.

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. and 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)*, volume 2016, pages 250–257. IEEE.

Comisión de la Verdad (2023). Las elecciones presidenciales de 1994. https://www.comisiondelaverdad.co/las-elecciones-presidenciales-de-1994.

Corona, R. M. and Sánchez, R. M. (2012). Las elecciones presidenciales de 2012 vistas desde twitter. *Virtualis*, 3(6):30–41.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Cuervo, M. C. and 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.

Daza, J. D. (2020). Las elecciones presidenciales de colombia en 2018: Candidatos, autocandidatos y seudocandidatos. *Revista Colombiana de Ciencias Sociales*, 11(1):234–266.

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

del Tronco Paganelli, J., Flores Ivich, G., and Madrigal Ramírez, A. (2016). La utilidad de las encuestas en la predicción del voto. la segunda vuelta de argentina 2015. *Revista mexicana de opinión pública*, .(21):73–92.

Deltell, L., Claes, F., and Osteso, J. M. (2013). Predicción de tendencia política por twitter: Elecciones andaluzas 2012. *Ámbitos. Revista internacional de comunicación*, .(22).

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

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

El Khalifi, D. P. (2017). Aplicación del aprendizaje automático en dos casos de política española: Elecciones 26j e independencia de cataluña. *University of Huelva & International University of Andalusia*, .(.):.

El Mundo (2006). Álvaro uribe, reelegido presidente de colombia con más del 60 % de los votos. https://www.elmundo.es/elmundo/2006/05/27/internacional/1148762529.html.

El Tiempo (1998). Las elecciones presidenciales de 1994. https://www.eltiempo.com/archivo/documento/MAM-811673.

El Tiempo (2002). Arrollador triunfo de uribe. https://www.eltiempo.com/archivo/documento/MAM-1315988.

Fernández Casal, R. (2020). Svm - máquinas de vectores de soporte. `https://rubenfcasal.github.io/aprendizaje_estadistico/svm.html`.

Fernández Villafañez, S. et al. (2022). Métodos de regresión y clasificación basados en árboles. .

Fix, E. (1985). *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Garcia-Moreno, V. E., Alvarez-Caicedo, C. R., and Vásquez-Saenz, N. G. (2021). Análisis de sentimientos en la predicción de resultados de elecciones presidenciales. *Revista de investigación de sistemas e informática*, 14(1):69–81.

García González, A. (2023). Modelado matemático del algoritmo knn (k-nearest neighbors). https://panamahitek.com/modelado-matematico-del-algoritmo-knn-k-nearest-neighbors/.

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

Gentilhombre, E. (2023). Espectro político. `https://www.elgentilhombre.com/espectro-politico/`.

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

González Franco, N., Tecnológico, D., González Serna, J. G., Astiazarán Yépiz, G. J., and Castro Sánchez, N. A. (2019). Las benditas redes sociales: Twitter y las elecciones presidenciales méxico 2018. *Congreso Estudiantil de Inteligencia Artificial Aplicada a la Ingeniería y Tecnología, UNAM, FESC*.

Gómez-Torres, E., Jaimes, R., Hidalgo, O., and Luján-Mora, S. (2018). Influencia de redes sociales en el análisis de sentimiento aplicado a la situación política en ecuador (influence of social networks on the analysis of sentiment applied to the political situation in ecuador). .

Huet, P. (2023). Qué son las redes neuronales y sus aplicaciones. https://openwebinars.net/blog/que-son-las-redes-neuronales-y-sus-aplicaciones/.

IBM (2024). ¿qué es el algoritmo de k vecinos más cercanos? `https://www.ibm.com/es-es/topics/knn`.

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

Jaramillo Guerra, M. C. et al. (2018). Las emociones en la política: el caso de las campañas de los precandidatos de la gran consulta por colombia. B.S. thesis, Universidad de La Sabana.

Khan, A., Zhang, H., Boudjellal, N., Ahmad, A., Shang, J., Dai, L., and 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. .

Luque Zabala, C. M. (2021). *Métodos Bayesianos para caracterizar el comportamiento legislativo del Senado colombiano en el periodo 2010-2014*. PhD thesis, Universidad Santo Tomás.

López Medel, B. (2019). Estudio de ideología política en redes sociales a través de machine learning. .

Macias, A. C. R., Cobeña, L. E. S., and Valle, J. E. P. (2021). Análisis de sentimientos de las elecciones públicas del ecuador basado en la red social twitter. *Informática y Sistemas: Revista de Tecnologías de la Informática y las Comunicaciones*, 5(1):7–16.

Makazhanov, A. and Rafiei, D. (2013). Predicting political preference of twitter users. In ., pages 298–305.

Mariela Lucina, C. C., David Alejandro, C. S., and Rubén, U.-A. (2023). Elecciones presidenciales en el perú: minería de textos de los editoriales del diario la república. *Revista de Comunicación*, 22(1):71–87.

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

Miranda, M. V. (2020). Teorema de aproximación universal: prueba constructiva basada en conjuntos semisimpliciales. .

Molano, J. O. S. (2014). Elecciones presidenciales en colombia: 2014-2018. *Revista de la Facultad de Derecho y Ciencias Políticas*, 44(120):11–15.

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`.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana Delgado, R. (2011). Lecture notes on compositional data analysis. .

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

Ramírez Vicente, F. (2019). Las matemáticas del machine learning: Redes neuronales (parte i). https://telefonicatech.com/blog/las-matematicas-del-machine-learning-redes-neuronales-parte-i.

Roca, P. E.-S. (2021). Sobre las ideologías. `https://www.nodulo.org/ec/2021/n194p14.htm`.

Rodriguez, E. G. (2022). *Advances in Machine Learning for Compositional Data*. Columbia University.

Romero, B. and Montaño, L. (2010). Elecciones presidenciales en colombia 2010. .

Romero García, M. et al. (2022). La decisión de participar: testando las teorías del comportamiento político en américa mediante técnicas de machine learning. .

Romero Moreno, F. Y. et al. (2019). Las redes sociales como factor de predicción de resultados electorales en campañas presidenciales. .

Santander, P., Elórtegui, C., González, C., Allende-Cid, H., and 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.

Sánchez Navarro, A. (2021). Uso de twitter durante los últimos días de una campaña electoral. elecciones presidenciales en eeuu 2020. .

The Carter Center (2022). Informe de la misión de expertos: Elecciones presidenciales de colombia 2022. `https://www.cartercenter.org/resources/pdfs/news/peace_publications/election_reports/colombia-expert-mission-report-2022-spanish.pdf`.

Tolosana-Delgado, R., Talebi, H., Khodadadzadeh, M., and Van den Boogaart, K. (2019). On machine learning algorithms and compositional data. In *Proceedings of the 8th International Workshop on Compositional Data Analysis, Terrassa, Spain*, pages 3–8.

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

Urbina, S. L., Zayas, H. A. V., and López, O. G. T. (2019). Algoritmo random forest para la detección de fallos en redes de computadoras. *Serie Científica de la Universidad de las Ciencias Informáticas*, 12(8):27–41.

Valenzula Chaparro, H. d. J. (2020). Exploración de métodos en aprendizaje automatizado y su uso en física de altas energías. .

Zhang, M. and Shi, W. (2019). Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data. *Hydrology and Earth System Sciences Discussions*, pages 1–39.

10. Anexos

10.1. Anexo 1.

Se presenta la matriz de correlaciones de los 38 indicadores clasificados en finanzas públicas, educación, conflicto armado y seguridad, salud, demografía y población tomados de la *TerriData*.

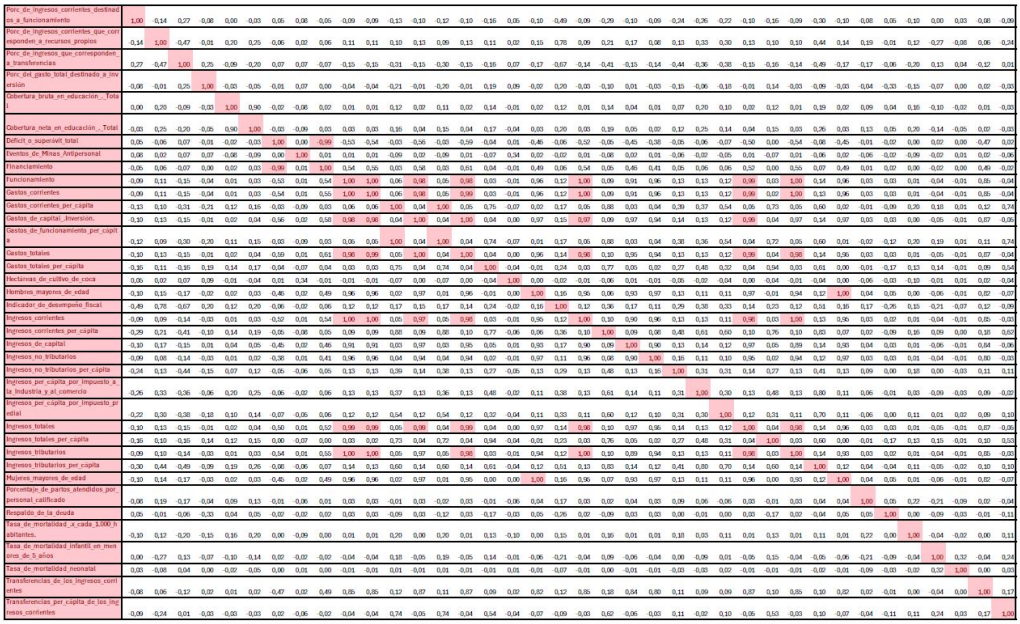


Figura 43: Matriz de correlaciones de las variables auxiliares. Elaboración propia.