Escuela de Negocios
Tipo de documento: Tesis de maestría

Master in Management + Analytics

Modelado Multinivel y de Machine Learning aplicado al análisis de la participación electoral en la provincia de Buenos Aires

Autoría: Di Biase, Bruno
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

MODELADO MULTINIVEL Y DE MACHINE LEARNING APLICADO AL ANÁLISIS DE LA PARTICIPACIÓN ELECTORAL EN LA PROVINCIA DE BUENOS AIRES

TESIS

Bruno Di Biase

Mayo de 2025

Tutores: Magdalena Cornejo – Javier Marenco

Resumen

Esta tesis examina el incremento sistemático en la participación electoral entre las Primarias Abiertas Simultáneas y Obligatorias (PASO) y las elecciones generales en la Provincia de Buenos Aires entre 2011 y 2023. Se emplea un enfoque metodológico mixto que combina modelos estadísticos explicativos de regresión multinivel con algoritmos de aprendizaje automático (Elastic Net, Random Forest, XGBoost) aplicados a un extenso conjunto de datos de más de 240.000 mesas de votación. Los resultados señalan que la participación en las PASO es el principal predictor de la participación en las elecciones generales, junto con variables demográficas (como la edad poblacional), el nivel socioeconómico y características del contexto político local. Asimismo, se desarrollan modelos predictivos orientados a optimizar estrategias de campaña electoral o de intervenciones de políticas públicas basados en la identificación de estos factores clave.

Abstract

This thesis examines the systematic increase in voter turnout between the open, simultaneous, and compulsory primaries (PASO) and the general elections in the Buenos Aires Province from 2011 to 2023. It employs a mixed-method approach combining explanatory multilevel regression models with machine learning algorithms (Elastic Net, Random Forest, XGBoost) applied to a large dataset of over 240.000 polling stations. The results indicate that PASO turnout is the main predictor of the turnout during the general elections, along with demographic variables (such as age distribution), socioeconomic level, and local political context factors. Additionally, predictive models are developed to optimize campaign strategies or public policies based on identifying these key factors.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

1.1. Contextualización de la participación electoral y el sistema PASO argentino

La participación electoral es ampliamente reconocida como un pilar fundamental de un sistema democrático robusto y representativo. En particular, el acto de votar empodera a los ciudadanos para moldear la dirección de su gobierno y asegura que diversas voces sean escuchadas en el proceso político. En este sentido, un alto nivel de participación no solo otorga legitimidad a los funcionarios electos, sino que también contribuye a la estabilidad y consolidación general de las instituciones democráticas, funciona como la base de las estructuras democráticas, y afirman el principio de autogobierno a través de la elección colectiva de representantes (Douglas, 2013).

Argentina presenta un caso de estudio único en la dinámica electoral debido a su implementación del sistema de Primarias Abiertas, Obligatorias y Simultáneas (PASO), establecido por la Ley 26.571 en 2009. Este sistema exige que todos los votantes habilitados participen en las elecciones primarias que se celebran simultáneamente en toda la nación, independientemente de su afiliación partidaria, para nominar candidatos para las elecciones generales. En consecuencia, los ciudadanos argentinos están obligados a votar al menos dos veces en cada ciclo electoral: una vez en las PASO, que generalmente se celebran en agosto, y nuevamente en las elecciones generales en octubre.

La naturaleza obligatoria de las PASO[1], incluso para los partidos políticos que presentan una sola lista de candidatos, sugiere que estas primarias cumplen un doble propósito. Más allá de la selección de candidatos, funcionan como una encuesta preelectoral integral a nivel nacional, que ofrece una indicación confiable de las preferencias de los votantes y la posible distribución de votos en las próximas elecciones generales. Esto se debe a que todos los partidos, independientemente de la competencia interna, deben participar, generando datos extensos sobre el apoyo de los votantes a diversas fuerzas políticas meses antes de las elecciones generales reales (Gallo, 2018). Esta información disponible públicamente puede influir significativamente en el comportamiento de los votantes y en la planificación estratégica de las campañas en las elecciones generales posteriores.

Además, el sistema PASO incorpora un umbral del 1,5% de los votos que los partidos políticos deben superar para calificar para la boleta de las elecciones generales. Este requisito actúa como un filtro sustancial, que potencialmente conduce a una consolidación del panorama político a medida que los partidos minoritarios que no

[1] En Argentina el voto es obligatorio para ciudadanos entre 18 y 70 años, y optativo para aquellos entre 16 y 17 años, y para mayores de 70 años de edad (Ley 26.774 y Ley 19.945)

cumplen con este umbral quedan excluidos de las elecciones generales. Esto puede incentivar el voto estratégico durante las PASO y resultar en un campo de candidatos más simplificado y competitivo en las elecciones generales.

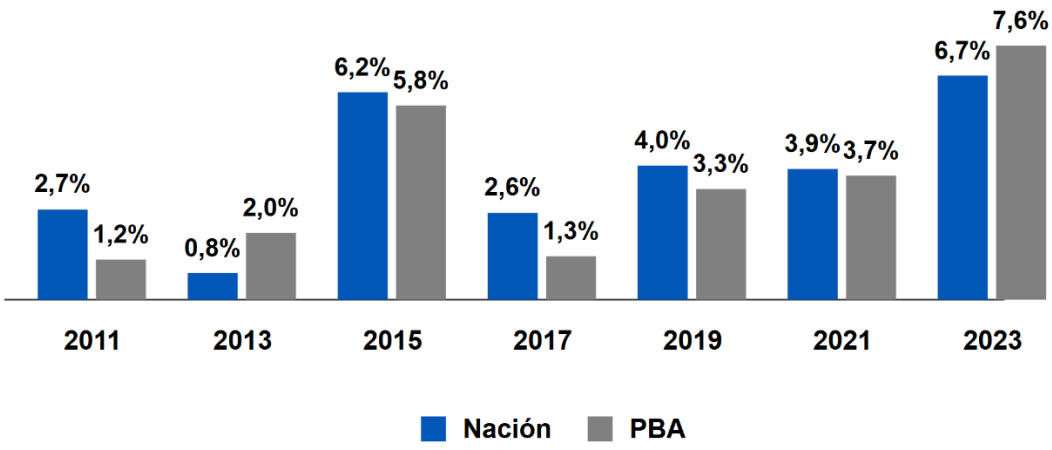
El Patrón Distintivo del Aumento de la Participación Electoral en Argentina

Una tendencia notable y consistente observada en Argentina desde la implementación del sistema PASO es un aumento en la participación total de votantes en las elecciones generales en comparación con las PASO precedentes dentro del mismo ciclo electoral. Este fenómeno ha sido evidente en los últimos siete ciclos electorales, que abarcan tanto las elecciones legislativas como presidenciales realizadas bajo el marco de las PASO desde su inicio en 2009.

Esta investigación se centrará específicamente en la Provincia de Buenos Aires, que tiene un peso electoral significativo como el distrito más grande de Argentina, representando aproximadamente el 37% del electorado nacional. La gran concentración de votantes en esta provincia la convierte en un área crítica para comprender las tendencias y patrones electorales nacionales.

Los datos sobre la participación electoral en las elecciones presidenciales y legislativas argentinas entre 2011 y 2023 ilustran claramente este patrón de aumento de la participación desde las PASO hasta las elecciones generales en cada ciclo.

Figura 1: Diferencia en Participación Electoral en las Elecciones Presidenciales y de Diputados (2011-2023): Elecciones Generales vs PASO – Nación y PBA (en %)



Fuente: elaboración propia en base a datos de la Comisión Nacional Electoral (CNE)

El aumento constante en la participación electoral entre las PASO y las elecciones generales a lo largo de múltiples ciclos sugiere que existen factores sistemáticos que influyen en este patrón en lugar de una mera variación aleatoria.

Antecedentes y marco conceptual

Aunque la literatura existente no aborda directamente el aumento de participación PASO-General en PBA, diversos estudios sobre comportamiento electoral y referidos al contexto político argentino proporcionan un amplio marco de referencia:

Factores Determinantes de la Participación Electoral

La literatura internacional sobre comportamiento electoral ha debatido extensamente los factores que impulsan la participación. En particular, la investigación sobre el comportamiento electoral en Estados Unidos ha demostrado consistentemente que los factores demográficos, aunque importantes, son insuficientes por sí solos para predecir las decisiones de los votantes. Ya desde trabajos fundacionales como *The American Voter* (Campbell et al., 1960) y reforzado por investigaciones posteriores que subrayan la identificación partidista como una identidad social clave (p. ej., Green et al., 2002), se advierte sobre la necesidad de considerar elementos más profundos y factores coyunturales para explicar el voto.

Por otro lado, enfoques como la regresión multinivel con postestratificación (MRP), utilizada por Ghitza y Gelman (2013), demuestran la importancia de analizar la interacción compleja entre múltiples características demográficas y geográficas para comprender patrones de votación a nivel subnacional. Un aspecto adicional es la importancia de los recursos individuales (como el tiempo, el dinero y las habilidades cívicas) y las redes de reclutamiento que facilitan el involucramiento cívico (Verba et al., 1995). Otros trabajos enfatizan el papel crucial de la movilización por parte de partidos, campañas y organizaciones para superar los costos individuales de la participación e incentivar el voto (Rosenstone & Hansen, 1993).

En el contexto regional, análisis comparados buscan explicar las variaciones en la participación electoral en América Latina considerando factores institucionales (como la obligatoriedad del voto o el tipo de sistema electoral) y socioeconómicos específicos de los países (Fornos et al., 2004). En particular, en este último caso, los autores destacan que el caso latinoamericano difiere de otros países de occidente en el hecho de que el diseño institucional y las variables de índole política juegan un rol mucho más relevante que los factores socioeconómicos.

Investigaciones en diversos contextos han identificado variables claves asociadas a la participación. Estudios comparados en América Latina (Strnad, 2022) y análisis en otras regiones como Corea del Sur (Yoon, 2018) consistentemente señalan la influencia de factores institucionales, el nivel educativo y variables socioeconómicas sobre la

decisión de votar, si bien la magnitud y dirección de estas relaciones pueden variar significativamente entre países.

El Barómetro de las Américas *(Latin American Public Opinion Project* [LAPOP], 2022) confirma para América Latina la tendencia general de que la edad, la riqueza y la educación se correlacionan positivamente con la participación, aunque enfatiza las importantes variaciones contextuales dentro de la región. Estos hallazgos sugieren líneas de análisis relevantes para explorar si tendencias similares se observan en relación con el aumento de participación en la Provincia de Buenos Aires.

Contexto Argentino y Sistema PASO

Buquet y Gallo (2022) efectúan un análisis comparativo entre el sistema de primarias de Argentina y Uruguay. Argumentan que características distintivas del diseño institucional argentino –como el voto ciudadano obligatorio (a diferencia de Uruguay), la presentación de fórmulas presidenciales cerradas (presidente y vicepresidente definidos de antemano) y la estricta cláusula que impide a los perdedores de la interna competir por otros cargos en la elección general generan una lógica de "suma cero". Esta configuración, sostienen los autores, desincentiva fuertemente la competencia real dentro de los partidos o coaliciones en la instancia de las PASO. En lugar de funcionar como un mecanismo para dirimir liderazgos internos mediante el voto popular, las PASO argentinas tienden a promover la coordinación temprana de las élites, la negociación anticipada de lugares y la formación de grandes frentes electorales que presentan candidaturas únicas o con competencia simbólica.

Un aspecto interesante que Buquet y Gallo (2022) resaltan es la consecuencia paradójica de este fenómeno: aunque la competencia interna es mínima o inexistente, los resultados de las PASO en Argentina adquieren una notable capacidad predictiva para la elección general, operando en la práctica más como una encuesta nacional a gran escala o una primera vuelta anticipada que como una genuina instancia de selección interna de candidatos (Buquet & Gallo, 2022).

Desde una perspectiva institucional, las particularidades del sistema de primarias argentino –su carácter obligatorio, simultáneo y abierto– ha sido relevada en diversos documentos del Proyecto ACE (ACE Project, 2022), enfocándose, entre otros aspectos, en el efecto sobre la participación electoral respecto a otros países de Latinoamérica.

Otro aspecto relevante dentro de la literatura específica sobre Argentina es el trabajo de Brusco et al. (2004), quienes analizan el fenómeno de la compra de votos en las estrategias de movilización partidaria, especialmente en contextos con presencia de clientelismo. Encuentran que, en Argentina, este efecto es significativo en la capacidad

de influenciar el apoyo electoral, especialmente entre segmentos sociales de bajos ingresos.

Una forma útil de interpretar la menor concurrencia en las PASO y su posterior repunte en octubre es enmarcar las primarias como “elecciones de segundo orden”. Reif y Schmitt (1980) analizan este caso en el contexto de las elecciones al parlamento europeo. Encuentran que cuando los votantes perciben que una instancia no define directamente quién gobernará, tienden a participar menos, aun bajo voto obligatorio. Bajo esta óptica, las elecciones generales argentinas serían las verdaderas “elecciones de primer orden” y, por tanto, concentran la expectativa de decisión final, la cobertura mediática y el esfuerzo de movilización partidaria. Este lente comparado permite alinear el caso argentino con hallazgos europeos y latinoamericanos sobre los diferenciales de turnout entre comicios de distinta jerarquía.

Por último, Lucardi et al. (2024) investigan cómo la información pública generada por las PASO sobre la viabilidad de los candidatos influye en la coordinación electoral en las elecciones generales. Su trabajo demuestra que las PASO actúan como una gran encuesta nacional que altera el panorama informativo y estratégico para los votantes. Argumentan que los electores cuyos candidatos preferidos muestran un bajo desempeño en las PASO son más propensos a reevaluar su decisión y votar estratégicamente por candidatos más viables en la elección general, buscando evitar el "voto desperdiciado" (Lucardi et al., 2024). Si bien su enfoque principal es la coordinación del voto (cambio de preferencia), identifican un mecanismo relacionado con la reconfiguración del escenario estratégico post-PASO basada en la información sobre viabilidad. Este cambio podría influir no solo en a quién se vota, sino también en la motivación para participar en la elección general, por ejemplo, al clarificarse las opciones competitivas o aumentar la percepción sobre lo que está en juego, factores que podrían incidir en el aumento de la participación. Las características institucionales específicas de las PASO (como la obligatoriedad, el umbral de proscripción, y su rol informativo) son, por lo tanto, elementos centrales que podrían estar explicando, al menos parcialmente, los patrones diferenciales de participación observados entre las dos instancias electorales.

El conjunto de variables y dinámicas relevantes analizado —desde factores sociodemográficos hasta las particularidades institucionales del sistema PASO y sus efectos—son pertinentes para comprender el fenómeno del aumento de la participación en la Provincia de Buenos Aires. El desarrollo subsiguiente procederá teniendo presente este diverso marco contextual provisto por los estudios existentes.

1.2. Descripción del Problema

El Fenómeno a Investigar: El Aumento de Participación entre PASO y Generales

El problema fundamental es la falta de una comprensión detallada de los factores que explican el aumento constante de la participación electoral en la Provincia de Buenos Aires entre las PASO y las elecciones generales posteriores. Si bien se han ofrecido explicaciones anecdóticas, como la mayor importancia percibida de las elecciones generales o la intensificación de las actividades de campaña, estas no se han examinado ni cuantificado rigurosamente dentro de la literatura académica existente.

Marco Analítico Propuesto: Integración de Dimensiones Sociodemográficas, Económicas y Políticas

Para abordar esta cuestión, se propone un marco analítico que modela la dinámica del aumento de la participación electoral en la Provincia de Buenos Aires mediante la integración de una variedad de variables sociodemográficas, económicas y políticas. El objetivo es identificar las posibles interacciones entre estos factores y su influencia en los cambios observados en la participación electoral.

Adoptar un enfoque multidimensional que considere los factores sociodemográficos, económicos y políticos contribuirá a lograr una comprensión integral del aumento de la participación electoral. Es probable que estos diferentes tipos de factores estén interconectados y puedan influir en el comportamiento de los votantes de maneras complejas. Por ejemplo, el nivel socioeconómico de un individuo podría afectar su nivel de compromiso político y su percepción de la importancia del voto.

Alcance Temporal: Examen de Siete Ciclos Electorales (2011-2023)

El análisis en esta investigación se basará en los siete ciclos electorales que ocurrieron entre 2011 y 2023. Se tomarán los datos de la categoría Presidente y Vice para las elecciones de 2011, 2015, 2019 y 2023, y los de Candidatos a Diputados Nacionales para las elecciones legislativas de 2013, 2017 y 2021. Este período representa toda la duración durante la cual ha estado vigente el sistema de PASO actual, y permitirá analizar la identificación de tendencias y patrones consistentes en el comportamiento de los votantes.

1.3. Objetivo

a) Pregunta Principal de Investigación

Se buscará responder dos interrogantes complementarios: ¿Cuáles son los determinantes sociodemográficos, económicos y políticos clave de la participación electoral en las elecciones generales en la Provincia de Buenos Aires? Y, adicionalmente: ¿Cómo puede desarrollarse un modelo predictivo, que utilice los datos disponibles tras la celebración de las PASO, para estimar con precisión la participación en las elecciones

generales y, de este modo, optimizar la asignación de recursos para intervenciones de políticas públicas o estrategias de campaña orientadas a incrementar la participación?

b) Objetivos Específicos:

Identificar Determinantes Clave de la Participación Electoral:

El primer objetivo específico de esta investigación es analizar cómo diversas variables sociodemográficas, económicas y políticas influyen en la participación electoral en las elecciones generales, centrándose particularmente en los resultados de las PASO. Esto implicará la realización de análisis estadísticos de los datos electorales de la Provincia de Buenos Aires entre 2011 y 2023, en conjunto con datos socioeconómicos disponibles públicamente. Para guiar este análisis, se parte de un conjunto de hipótesis:

Hipótesis 1 (Participación en las PASO): En la lógica de las elecciones de segundo orden (Reif y Schmitt, 1980) las PASO convocan sobre todo al electorado más politizado y con menores costos de información. Una mesa con alta participación en las primarias refleja la presencia de redes partidarias activas y de votantes habituados al proceso, configurando un piso organizativo que facilita la movilización posterior.

Por ello, se espera que haya una relación positiva con la participación en las PASO al explicar la participación en la elección general: cada punto adicional registrado en las PASO anticipa un aumento —aunque con rendimientos marginales decrecientes— en la asistencia a la contienda de mayor jerarquía.

Hipótesis 2 (Nivel Socioeconómico): De acuerdo con la literatura sobre comportamiento electoral que asocia un mayor estatus socioeconómico con mayores niveles de participación cívica (Strnad, 2022), se hipotetiza una relación positiva entre el nivel socioeconómico municipal y la participación en la elección general. Se espera que los municipios con mayores recursos exhiban, en promedio, una mayor afluencia a las urnas.

Hipótesis 3 (Contexto Político Local): Se postula que los municipios gobernados por el partido oficialista a nivel provincial disponen de mayores recursos institucionales y redes para la movilización de votantes. Por consiguiente, se espera que, *ceteris paribus*, la participación en las elecciones generales sea menor en los distritos gobernados por la oposición en comparación con los oficialistas.

Desarrollar un Modelo Predictivo para la Participación Futura:

El segundo objetivo es utilizar técnicas estadísticas apropiadas para construir un modelo predictivo capaz de estimar la participación electoral en futuras elecciones generales basándose en los determinantes identificados y los datos disponibles después de las PASO.

Formular Recomendaciones Basadas en Evidencia:

El objetivo final es generar recomendaciones específicas y prácticas para partidos políticos, autoridades electorales y organizaciones de la sociedad civil que operan en la Provincia de Buenos Aires en base a los resultados previos, tomando como métrica la participación electoral en las elecciones generales.

2. Datos

2.1 Integración y Preparación Inicial de los Datos

El proceso metodológico comenzó con la obtención e integración de los registros oficiales de votación proporcionados por la Cámara Nacional Electoral correspondientes a las elecciones PASO y generales desde 2011 hasta 2023, para la provincia de Buenos Aires, con aproximadamente 38.000 mesas por elección. Además, se incorporó un conjunto de estadísticas sociodemográficas de los 135 municipios bonaerenses, incluyendo información económica y demográfica, alcanzando alrededor de 30 variables exógenas.

Inicialmente se cargaron las bases de datos individuales por elección y año. Posteriormente, para homogeneizar los conjuntos de datos, se añadieron columnas que identificaban claramente el año electoral y el tipo de elección (PASO o General). Luego, se identificaron columnas comunes a todos los datos electorales y se ajustaron para mantener sólo estas columnas, consolidándolos en una única base denominada `dataset_completo`.

Cabe destacar que la información relevada de la Comisión Nacional Electoral está compuesta de bases de datos electorales provisorios y, por tanto, pasibles de contener errores de carga, e información inconsistente e incompleta. La cobertura de los datos provisorios, adicionalmente, no llega a ser del 100% de los votos registrados (que se contabilizan en el escrutinio definitivo), ascendiendo sin embargo a más del 90% en todos los casos.

2.2 Normalización y Corrección de Datos

Con la base consolidada, se ejecutó una revisión exhaustiva para garantizar la consistencia de variables categóricas como tipo de elección y nombres de municipios y secciones electorales. Se corrigieron inconsistencias tales como la variabilidad en la escritura de "GENERAL" (originalmente como "GENERALES") y otros nombres con errores de escritura o diferencias tipográficas mediante funciones condicionales y de normalización.

Se realizó también la conversión explícita del identificador de agrupaciones (`agrupacion_id`) a formato texto para asegurar uniformidad, y se estandarizaron las identificaciones numéricas (circuitos electorales), eliminando ceros iniciales que generaban inconsistencias en la comparación posterior de mesas electorales.

2.3 Detección y Tratamiento de Mesas con datos aberrantes o erróneos.

Para asegurar la validez del análisis comparativo entre elecciones PASO y generales, se aplicó un procedimiento de detección y eliminación de mesas con datos aberrantes o inconsistentes. Se generaron dos bases de datos por separado para cada año electoral, uno correspondiente a PASO y otro a Generales. La comparación exacta entre ambas instancias electorales se realizó generando claves únicas combinando `seccion_id`, `circuito_id y mesa_id`.

Las mesas presentes únicamente en una de las instancias electorales fueron consideradas aberrantes y potencialmente erróneas, y fueron identificadas mediante la función `anti_join()`. Estas mesas se eliminaron del análisis para mantener únicamente registros comparables. En total, se eliminaron menos del 7,1% de las observaciones para una determinada instancia electoral, como se detalla en la figura 2. Los causales de supresión, por su parte, se observan en la figura 3.

2.4 Manejo de Valores Cero y Votos No Positivos

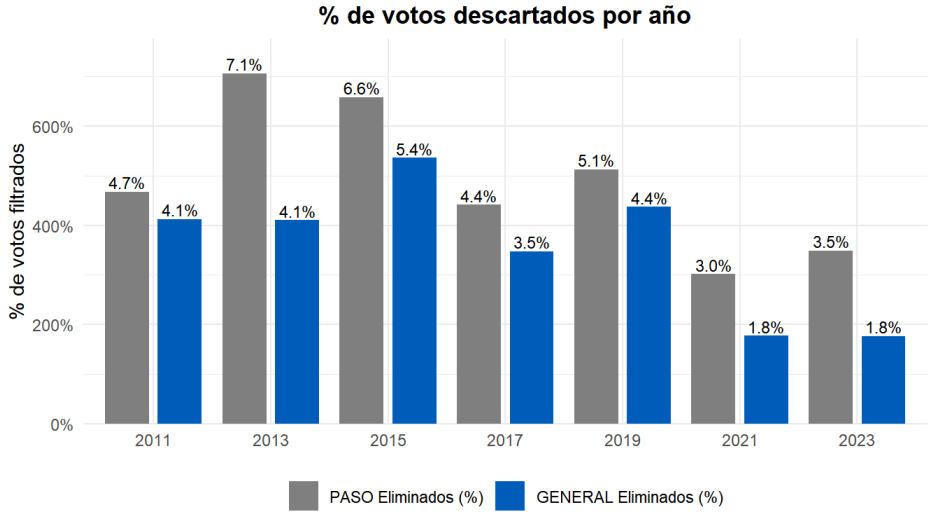
El tratamiento de valores cero y votos no positivos se efectuó mediante revisiones detalladas y sumas agregadas por año y tipo de elección. Se analizaron específicamente columnas asociadas a votos en blanco, impugnados, nulos y recurridos. Estos valores fueron revisados para garantizar que no existiesen errores de carga o inconsistencias internas.

Los criterios para el manejo de valores cero fueron los siguientes: se eliminaron o corrigieron aquellos registros que presentaban inconsistencias evidentes, como mesas con cero votos totales o mesas cuyo total de votos positivos y no positivos no coincidía con el número registrado de electores por mesa.

2.5 Construcción de la base de datos definitiva

Como consecuencia del pre procesamiento previo se eliminaron las siguientes observaciones:

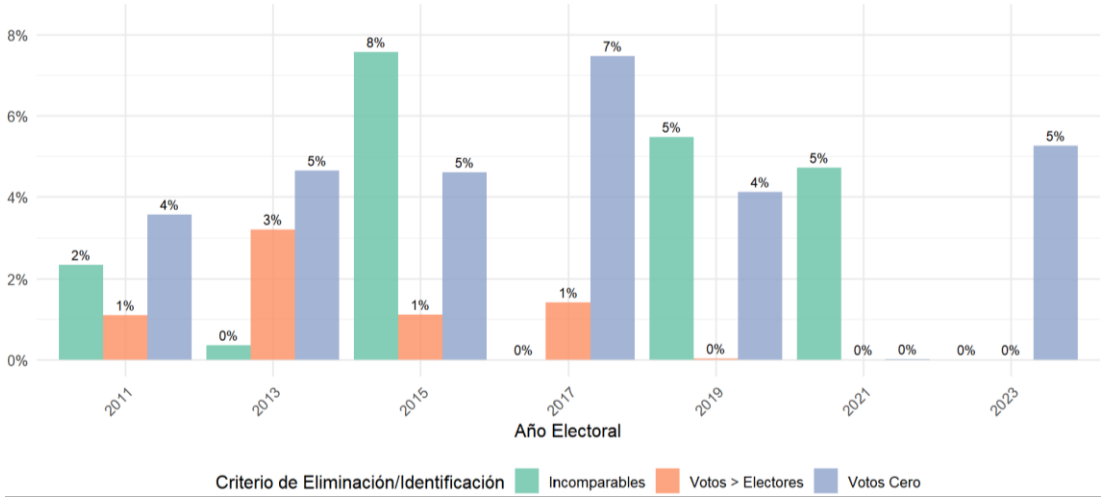
Figura 2: Votos filtrados de la base de datos final (% por año)



Fuente: elaboración propia

Figura 3:Porcentaje de mesas eliminadas por criterio y año

Respecto al total de mesas únicas originales por año



Fuente: elaboración propia

La base de datos final se construyó sobre la base consolidada y depurada `votos_wide`, que ya contenía todos los controles mencionados anteriormente. Se procedió a generar columnas específicas para cada agrupación política y tipo de voto, mediante un formato ancho para facilitar el análisis comparativo directo.

El proceso incluyó un emparejamiento de las mesas electorales presentes en elecciones PASO y Generales, para garantizar la existencia de datos válidos para cada

una de las mesas. Este cálculo se efectuó para cada mesa individualmente, preservando así la precisión espacial y temporal del análisis.

2.6 Caracterización de los datos y análisis descriptivo

La base de datos principal para el análisis se construyó adicionando variables exógenas de carácter sociodemográfico y económico a nivel municipal, obtenidas de fuentes como el INDEC y la Dirección Provincial de Estadística de la Provincia de Buenos Aires. Estas variables exógenas fueron vinculadas a cada mesa según su municipio y año correspondiente.

La base final contiene 229.996 observaciones, donde cada observación representa una mesa electoral única en un año electoral determinado. El Cuadro 1 provee un resumen detallado de las principales variables incluidas en este *dataset* final, especificando su descripción, tipo y nivel de agregación. Se incluyen identificadores (Año, Municipio), la variable dependiente (`Participacion_generales`), variables exógenas municipales (originales y transformadas), datos electorales base (electores y votos por tipo de elección a nivel mesa), y variables calculadas (participación, cuantiles, contexto político).

Tabla 1: Descripción de variables en la base de datos

Nombre Variable	Descripción	Tipo	Nivel
Participacion_General_pct	Tasa de participación en la mesa en la ELECCIÓN GENERAL (puntos %).	Numérico	Mesa-Año
Participacion_PASO_pct	Tasa de participación en la mesa en la PASO (puntos %).	Numérico	Mesa-Año
Año	Año del ciclo electoral (categoría base = 2011).	Factor	Mesa-Año
Municipio	Mesa agrupada por municipio (efecto aleatorio).	Factor	Municipal
Contexto_Politico	1 = oficialismo prov./nac.; 0 = oposición. Categoría omitida = **0 (oposición)**.	Factor	Municipal
log_PBG_pc	Logaritmo del PBG per cápita municipal.	Numérico	Municipal
log_Poblacion	Logaritmo de la población total municipal.	Numérico	Municipal
Edad_0_14_pct	Porcentaje de población de 0-14 años.	Numérico	Municipal
Edad_65_mas_pct	Porcentaje de población de 65+ años.	Numérico	Municipal
Pisos_revest	Porcentaje de hogares con pisos con revestimiento.	Numérico	Municipal
Agua_caneria	Porcentaje de hogares con agua por cañería dentro de la vivienda.	Numérico	Municipal

Tabla 2: Métricas de distribución de las variables continuas

	n	mean	sd	median	min	max	skew	kurtosis
Participacion_General_pct	229996	79.88	5.64	80.75	0.22	100.00	-2.48	22.85
Participacion_PASO_pct	229996	75.43	7.66	76.64	0.23	100.00	-1.91	11.28
log_PBG_pc	229996	9.43	0.48	9.47	8.13	11.29	0.00	0.02
log_Poblacion	229996	12.51	1.23	12.76	7.49	14.70	-0.74	0.32
Edad_0_14_pct	229996	24.11	3.27	24.00	16.00	32.86	0.08	-0.62
Edad_65_mas_pct	229996	11.86	3.35	12.00	5.39	23.00	0.02	-1.05
Pisos_revest	229996	83.27	9.27	84.89	49.61	97.80	-0.70	-0.13
Agua_caneria	229996	91.85	5.26	93.38	73.09	98.98	-0.89	0.00

Tabla 3: Distribución de variables categóricas

Variable	Categoria	Frecuencia	Prop_Pct
Año	2011	29339	12.8
Año	2013	31423	13.7
Año	2015	31138	13.5
Año	2017	33909	14.7
Año	2019	32982	14.3
Año	2021	35136	15.3
Año	2023	36069	15.7
Contexto_Politico	Oficialismo	146305	63.6
Contexto_Politico	Opositor	83691	36.4

Las tablas 2 y 3 resumen las distribuciones de las variables estudiadas sobre 229.996 mesas. Se destaca la mayor variabilidad de la participación en las PASO, con una desviación estándar de 7,66 puntos por mesa (vs 5,64 puntos en el caso de las elecciones generales). En cuanto a los períodos electorales considerados, los casos se distribuyen desde 2011 (12,8 %) hasta 2023 (15,7 %) con un aumento gradual (reflejando en parte el crecimiento del padrón electoral en cada elección). Por su parte, el contexto político se compone de 63,6 % mesas bajo municipios oficialistas frente a 36,4 % opositores.

Variable Dependiente: Tasa de Participación en la Elección General

La variable dependiente central de esta tesis es la *Participacion_General_pct*, definida como el porcentaje de electores empadronados que efectivamente votaron en la elección general en cada mesa, respecto a la cantidad total de electores. La elección de esta variable responde a un diseño metodológico que busca explicar y predecir el nivel de participación en la elección de primer orden, utilizando la participación en las primarias (PASO) como un predictor clave.

Desde una perspectiva causal, la participación en las PASO es un evento temporalmente antecedente que provee información y establece un contexto que influye en el comportamiento posterior del electorado en la elección general. Modelar la participación electoral en los comicios generales permite testear hipótesis sobre los mecanismos del comportamiento electoral en sistemas de dos vueltas.

Este modelo se enmarca en teorías estudiadas de la ciencia política:

Teoría de las Elecciones de Segundo Orden: Se argumenta que las PASO operan como una "elección nacional de segundo orden". Estos comicios, al no definir directamente los cargos electivos, se caracterizan por una menor participación. Sin embargo, sus resultados son esenciales porque proveen información que reconfigura el escenario para la "elección de primer orden" (la elección general). De esta forma se enmarca el uso de *participacion_paso* como un predictor de *Participacion_General_pct.*

Información y Comportamiento Estratégico: Las PASO funcionan como una encuesta pública a gran escala que revela la viabilidad de los candidatos y la competitividad de la elección. Esta nueva información puede alterar el cálculo de los votantes, influyendo no solo en *por quién votar*, sino también en *la decisión de votar*. Un resultado muy reñido en las PASO, por ejemplo, puede aumentar la percepción de que el propio voto es decisivo, incentivando la participación en la elección general.

Modelos de Predicción de Participación: El enfoque se alinea con los modelos estándar de predicción de la participación, que consistentemente identifican la participación en elecciones previas como el predictor más potente. En el sistema electoral argentino, la PASO es el evento electoral previo más relevante y próximo en el tiempo a las elecciones generales.

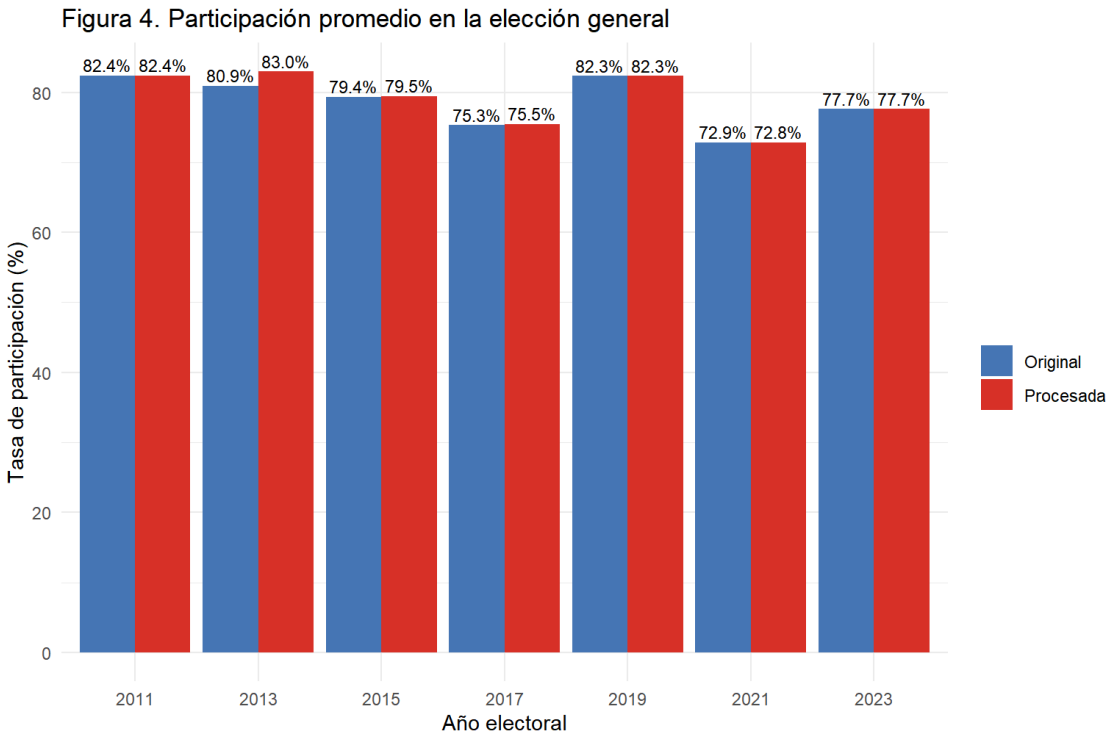
Análisis Descriptivo de la participación en elecciones generales

Representatividad de la base de datos

Antes de analizar la distribución de la variable dependiente, cabe evaluar si el proceso de limpieza y filtrado de mesas afectó la representatividad del fenómeno

estudiado. La Figura 4 compara la tasa de participación promedio en las elecciones generales, calculada con la base de datos original y con la base de datos procesada para el análisis. Se observa una notable similitud entre ambas para todos los años, lo que indica que la eliminación de mesas con datos inconsistentes no alteró sustancialmente la magnitud del nivel de participación promedio.

Figura 4: Tasa de participación promedio en elecciones generales (%)



Fuente: Elaboración propia en base a datos de la CNE.

Distribución de la participación en elecciones generales

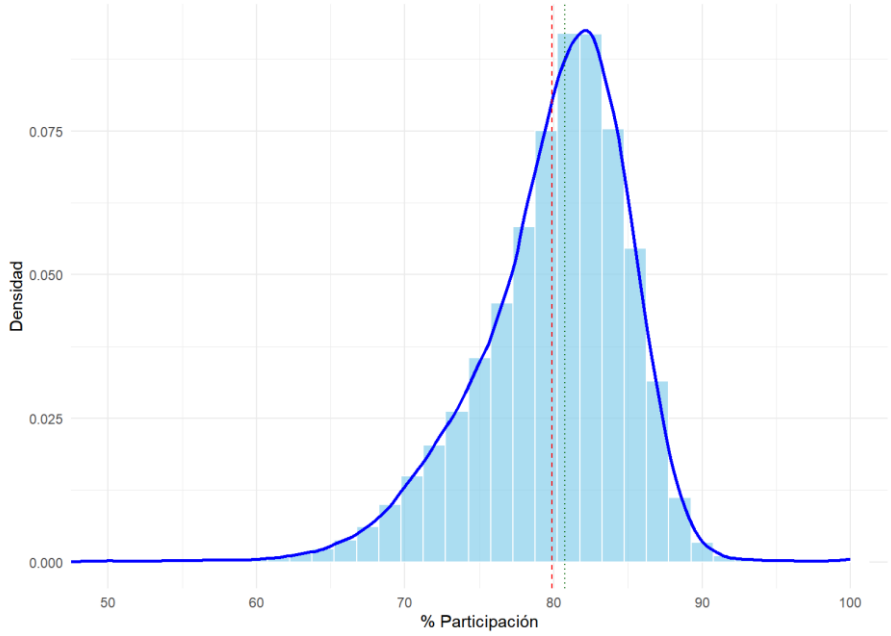
La variable *Participacion_General_pct* presenta una media de 79,85% y una desviación estándar de 11,23%:

Tabla 4: Métricas de distribución de la variable dependiente

Estadístico	Valor
Media	79.9
Desvío estándar	5.6
Mínimo	0.2
Q1	77.1
Mediana	80.7
Q3	83.5
Máximo	100.0

La Figura 5 muestra la distribución de esta variable para el conjunto de todas las mesas y años, con una gran concentración alrededor de la mediana (80,7%).

Figura 5: Distribución de la tasa de participación en elecciones generales por mesa

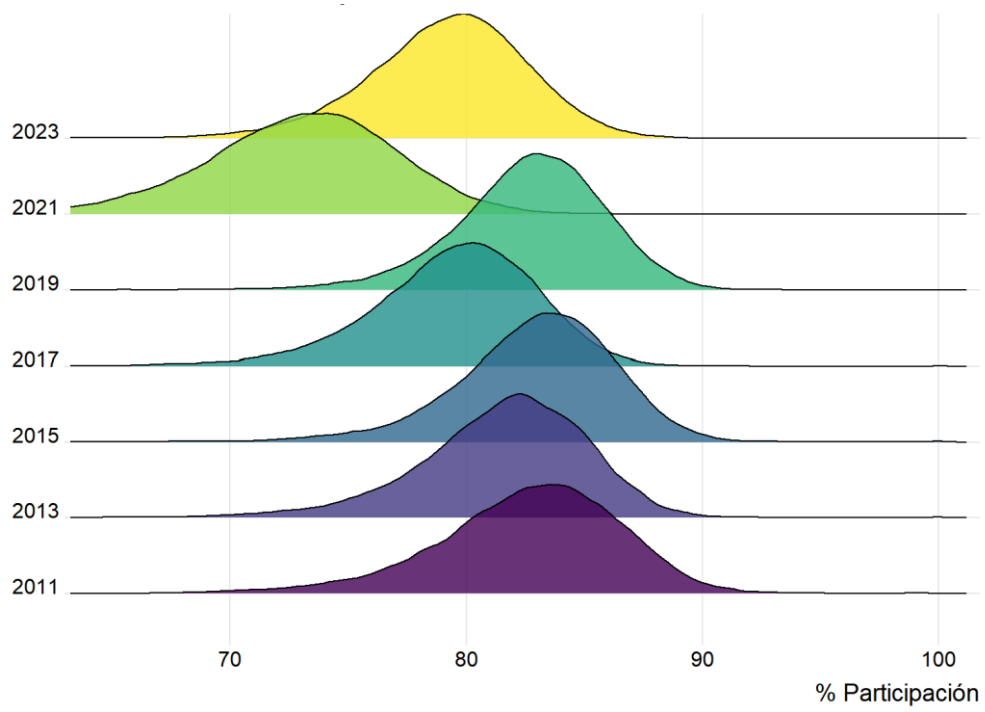


Fuente: Elaboración propia.

La distribución es asimétrica a la izquierda, con una alta densidad de mesas en torno a los valores más elevados de participación, lo cual es consistente con la naturaleza obligatoria del voto en Argentina.

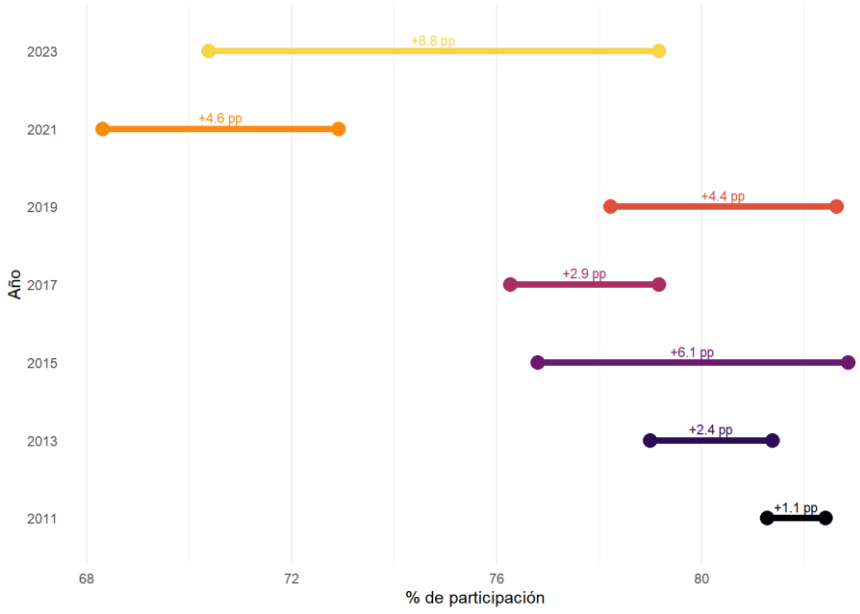
A su vez, las distribuciones por año (Figura 6) muestran distintas configuraciones, reflejando las particularidades de cada ciclo electoral. Por ejemplo, se pueden observar desplazamientos en la media y cambios en la dispersión que coinciden con si la elección era presidencial o legislativa, siendo estas últimas las que tienden a presentar una participación general más baja.

Figura 6*: *Distribución de densidad de participación electoral en elecciones generales por año



Fuente: Elaboración propia.

Figura 7*: *Diferencia entre participación electoral en las PASO y en las elecciones Generales, por año (en puntos porcentuales)



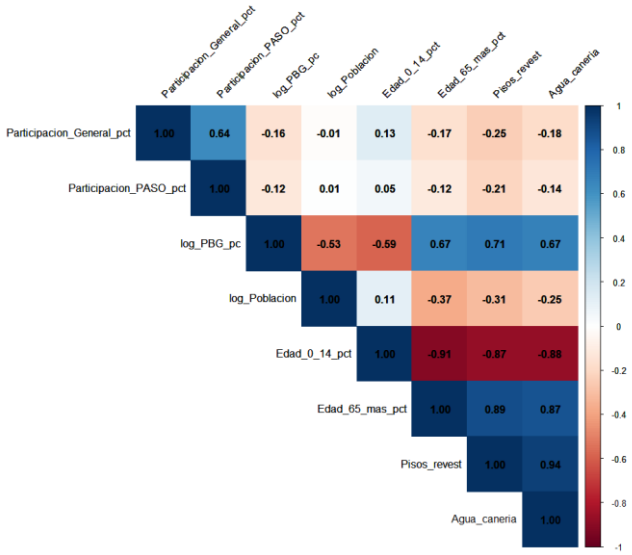
Fuente: Elaboración propia.

Análisis de variables exógenas y su relación con el diferencial de votos

La base de datos incluye variables exógenas a nivel municipal para explorar factores contextuales: PBG per cápita (`PBG_pc`), Población total (población), porcentajes de población joven (`Edad_0_14_pct`) y mayor (`Edad_65_mas_pct`), y *proxies* de condiciones de vida (porcentaje de población con pisos con revestimiento y con acceso a agua corriente). Las variables `PBG_pc` y población mostraron alta asimetría, por lo que se utilizaron sus transformaciones logarítmicas (`log_PBG_pc`, `log_Poblacion`) en los análisis de correlación y regresión.

A continuación, se presenta la matriz de correlación lineal para las variables mencionadas:

Figura 8: Matriz de Correlación



Fuente: Elaboración propia.

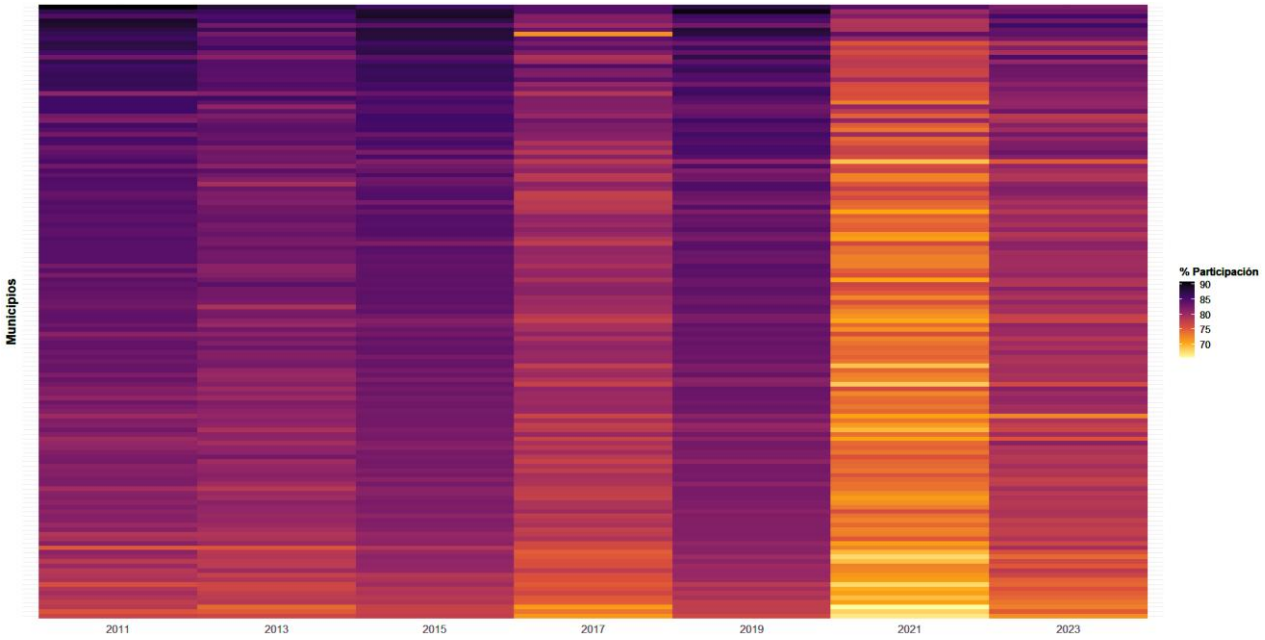
Estos resultados sugieren que las características estructurales socioeconómicas de los municipios, al menos las medidas aquí tienen un poder explicativo lineal limitado sobre la variación del diferencial de votos mesa a mesa. Cabe notar, sin embargo, la existencia de correlaciones significativas entre las variables exógenas, aspecto relevante a considerar por posible multicolinealidad en la etapa de modelado.

Exploración de otras dinámicas asociadas al incremento en participación

Dinámica Temporal: la participación varía significativamente entre años, con picos en 2015, 2019 y 2023, lo que evidencia la importancia de las elecciones presidenciales y la mayor movilización electoral respecto a los ciclos de elecciones legislativas (2013, 2017 y 2021). Se destaca el año 2021 como el de menor participación

de todo el ciclo de elecciones analizado, reflejando probablemente el impacto de la pandemia mundial por Covid-19.

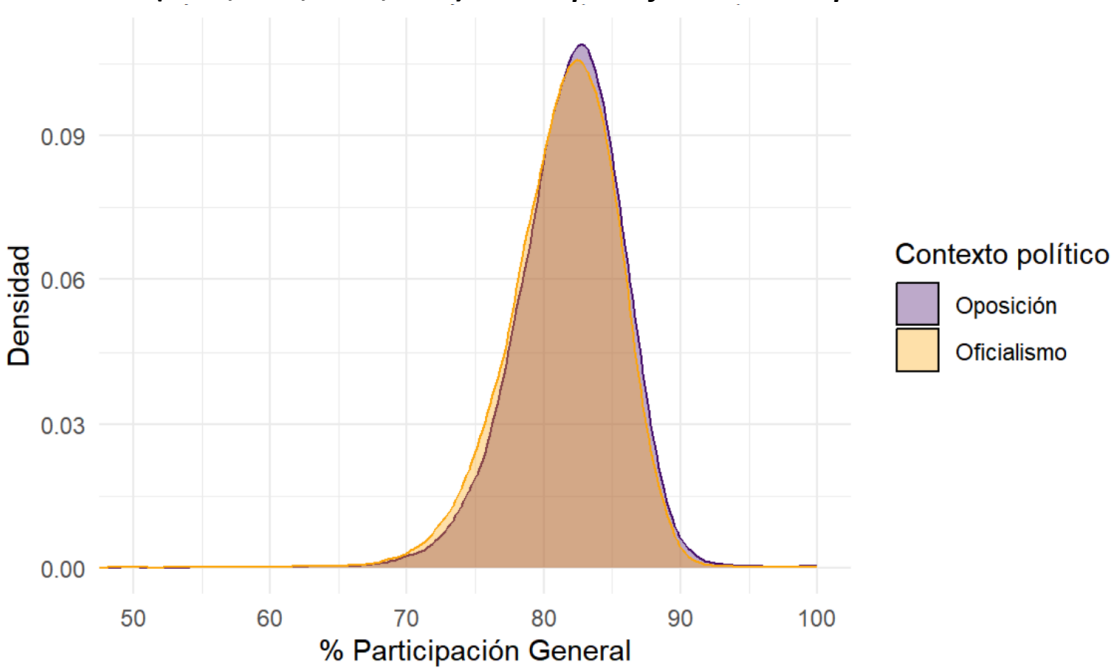
Figura 9: Mapa de Calor - Participación electoral en elecciones generales por municipio y año



Fuente: Elaboración propia en base a datos de la CNE.

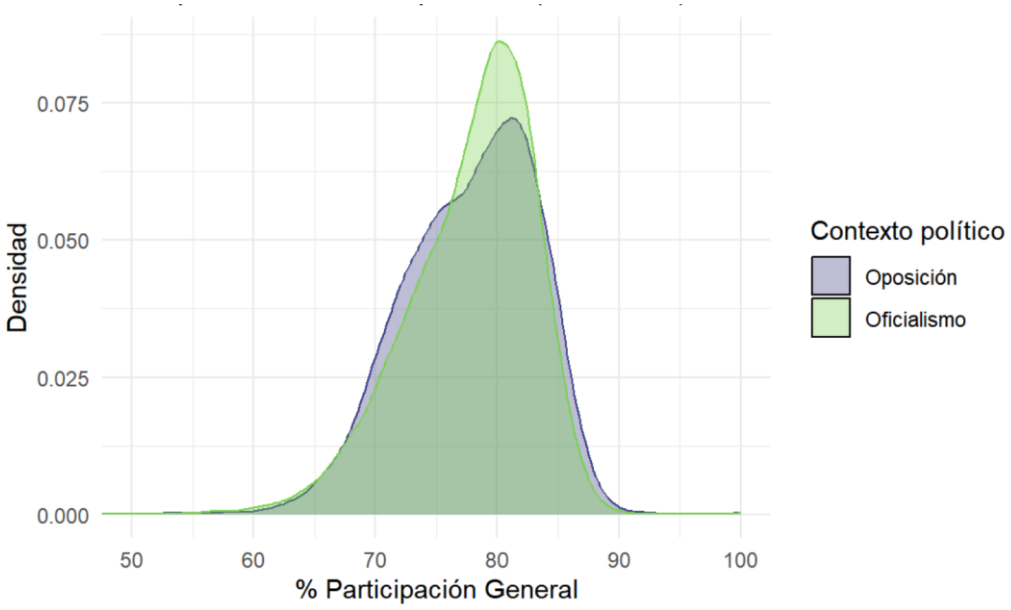
Contexto Político Municipal: El análisis por contexto político sugiere que entre municipios cuyo signo político está alineado con el oficialismo provincial y nacional, la participación en las elecciones generales es más alta en los años de elecciones legislativas.

Figura 10: Distribución de participación electoral en elecciones presidenciales (2011,2015, 2019,2023)- municipios oficialistas vs opositores



Fuente: Elaboración propia en base a datos de la CNE.

Figura 11: Distribución de participación electoral en elecciones legislativas (2013, 2017, 2021)- municipios oficialistas vs opositores



Fuente: Elaboración propia en base a datos de la CNE.

Análisis de *Clustering* a nivel circuito electoral

A efectos de capturar la heterogeneidad territorial relevante para el comportamiento electoral, se optó por emplear el circuito electoral por sobre el municipio como unidad de agrupamiento por tres razones fundamentales. Primero, el circuito constituye la escala operativa de la logística comicial: concentra un conjunto relativamente homogéneo de mesas que comparten accesibilidad física y redes de movilización locales, mientras que el municipio agrega zonas urbanas y rurales con perfiles sociodemográficos muy dispares. Segundo, el número de circuitos (836) provee suficiente variabilidad estadística para estimar efectos aleatorios y, simultáneamente, mantiene la potencia analítica; por contraste, los 135 municipios hubieran reducido drásticamente los grados de libertad del modelo multinivel, dificultando la identificación de patrones intra-municipales. Finalmente, el circuito es la instancia administrativa sobre la cual se reportan los resultados provisionales y se articulan las estrategias de campaña, de modo que analizar este nivel brinda implicancias operativas directas para la intervención territorial.

Sobre esa base, se implementó un *k-means* con k = 3, valor escogido porque maximizó la silueta promedio y correspondió al primer "codo" en la curva WSS, indicando el mejor equilibrio entre cohesión interna y separación entre grupos. Las variables —previamente estandarizadas— incluyeron participación promedio (general y PASO), logaritmos de PBG per cápita y población, composición etaria (0-14 y ≥ 65 años), indicadores de vivienda (pisos revestidos y agua por cañería) y una *dummy* de signo político (oficialismo/oposición).

Los resultados revelan tres tipologías espaciales, detalladas en la tabla siguiente:

Tabla 5: Caracterización estadística de clusters por circuitos electorales

Resumen por Cluster de Circuitos

Promedios de participación y variables contextuales

	Circuitos	% General	% PASO promedio	log PBG per cápita	log Población	% 0-14	% ≥65	% pisos rev.	% agua cañería	% Ofic.	% Opos.
1	124.0	80.0	75.2	9.3	12.2	26.7	9.6	77.6	88.7	91.1	8.9
2	123.0	64.6	56.4	9.9	10.5	23.3	14.7	89.0	94.6	52.8	47.2
3	589.0	79.5	74.7	9.9	11.0	22.5	14.7	89.2	95.1	62.6	37.4

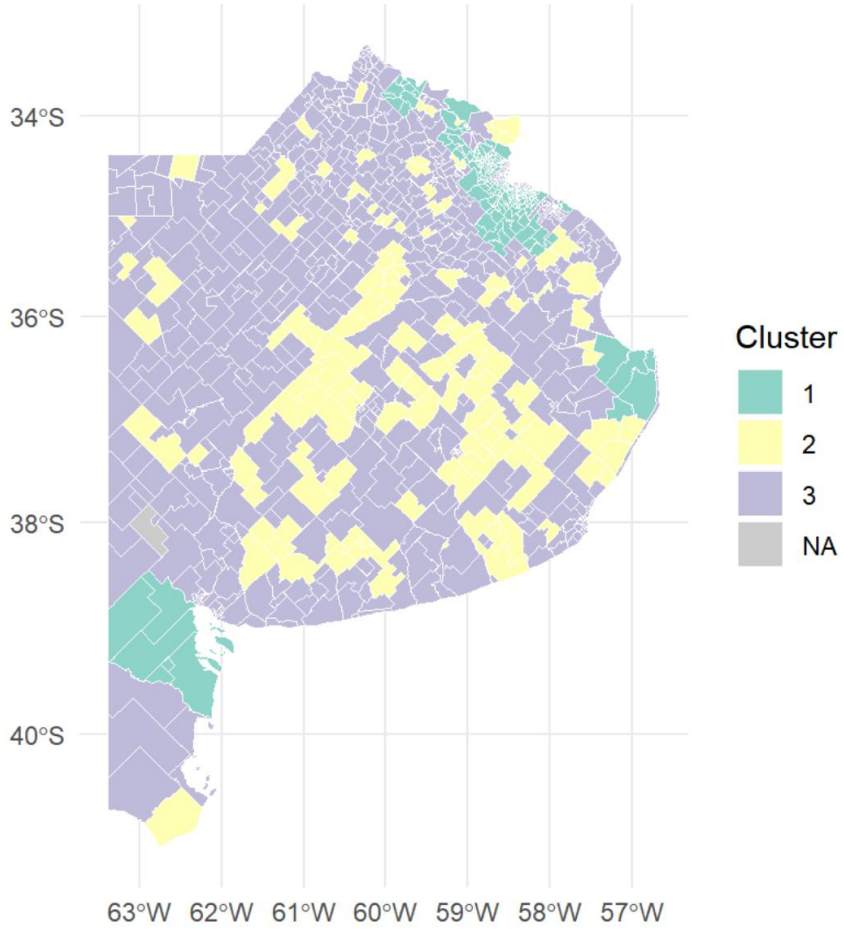
Clúster 1. Reúne 124 circuitos con la mayor participación (80 % en generales, 75 % en PASO). Presenta el menor PBG per cápita y la infraestructura habitacional más rezagada (77,6 % de pisos revestidos, 88,7 % de agua por cañería), junto con la población

más joven (26,7 % de niños) y la menor proporción de adultos mayores. Se destaca la altísima prevalencia oficialista (91 %).

Clúster 2. Con 123 circuitos y la participación más baja (64,6 %), constituye una zona intermedia: combina el PBG per cápita más alto, buena dotación de servicios (94,6 % de agua, 89 % de pisos) y la población menos numerosa. La estructura etaria es más envejecida (14,7 % ≥65) y la distribución política está casi equilibrada entre oficialismo y oposición.

Clúster 3. Abarca la mayor parte del universo (589 circuitos) y muestra niveles de participación similares al Clúster 1 (79,5 % en generales) pero con mejor infraestructura residencial y alto acceso al agua (95,1 %). Mantiene valores medios en PBG y población, una composición etaria intermedia y una ligera ventaja oficialista (62,6%).

Figura 12: Clusterización de circuitos electorales



Síntesis del Análisis Exploratorio

El análisis exploratorio de los datos revela que el aumento de participación entre PASO y Generales en PBA es un hecho consistente pero altamente heterogéneo. El

porcentaje de participación en elecciones generales muestra una gran dispersión a nivel mesa, y su magnitud promedio varía significativamente entre años y municipios.

Las dinámicas temporales (coyuntura de cada elección), geográficas (factores locales no medidos por las variables exógenas estándar), posiblemente el contexto político municipal, las estrategias y performance diferencial de las fuerzas políticas, y quizás características intrínsecas de las mesas o la relación con el nivel de participación inicial, emergen como dimensiones relevantes para incluir en la modelización econométrica. La sección siguiente se dedicará al desarrollo y estimación de modelos que incorporen estas variables para explicar la varianza observada en la participación en las elecciones generales.

3. Metodología

Siguiendo la distinción fundamental en el modelado estadístico entre explicar un fenómeno y predecir su ocurrencia (Shmueli, 2010; Breiman, 2001), se adoptará un enfoque mixto. En este contexto, la literatura enfatiza que estos dos objetivos requieren enfoques distintos: Breiman (2001) describió dos "culturas", una centrada en ajustar modelos paramétricos para entender *el por qué* (modelado de datos) y otra enfocada en el *qué*, maximizando la precisión predictiva con algoritmos flexibles, aunque menos interpretables (modelado algorítmico). Shmueli (2010) advierte que confundir estos propósitos puede llevar a errores, ya que un modelo con alto poder explicativo no necesariamente predice bien, y viceversa.

Considerando lo anterior, se emplearán ambos enfoques de manera complementaria:

1. **Modelo Explicativo:** Se utilizará un modelo estadístico interpretable (Regresión Multinivel Jerárquica) para estimar el efecto de diversos factores sociodemográficos, económicos y políticos sobre la participación electoral, buscando responder al *porqué*.

2. **Modelos Predictivos:** Se aplicarán modelos de aprendizaje automático (*Elastic Net, Random Forest, XGBoost*) para evaluar la capacidad predictiva de dichos factores y del modelo explicativo en su conjunto, abordando el *qué* y validando la robustez de los hallazgos.

Este esquema permite contrastar la comprensión detallada del fenómeno con la eficacia de predicción *out-of-sample*, proporcionando un análisis más completo y riguroso (Shmueli, 2010).

3.1 Modelo Explicativo Principal: Regresión Jerárquica Multinivel (MLM)

Justificación: La estructura de los datos –mesas electorales (i) anidadas dentro de circuitos electorales (j), observadas a lo largo de múltiples años electorales (t)– viola el supuesto de independencia requerido por modelos como mínimos cuadrados ordinarios (MCO). Las observaciones de la misma mesa en distintos años, o de distintas mesas dentro del mismo circuito, tienden a estar correlacionadas. Ignorar esta estructura jerárquica llevaría a errores estándar incorrectos e inferencias inválidas (Gelman & Hill, 2007; Raudenbush & Bryk, 2002). Los MLM están diseñados para manejar explícitamente esta anidación, modelando la varianza en diferentes niveles y permitiendo la inclusión de predictores medidos tanto a nivel de mesa-año como a nivel de circuito electoral.

Se destacan dos principales *features* de este tipo de modelo: en primer lugar, permite capturar la variación entre municipios de forma explícita: algunos circuitos electorales pueden tener consistentemente un mayor aumento de participación entre elecciones (por prácticas locales de movilización, idiosincrasia política, etc.), mientras que otros presenten brechas menores. Estimar interceptos aleatorios para cada circuito facilita cuantificar dichas diferencias residuales una vez controlados los predictores observables

En segundo lugar, el enfoque multinivel realiza un "*pooling* parcial" de la información entre unidades, lo que produce estimaciones más estables, especialmente para municipios con pocas mesas o comportamiento atípico. Este fenómeno, conocido como *shrinkage* o contracción hacia la media, evita estimaciones extremas poco fiables: esencialmente, el modelo "desconfía" de desviaciones muy grandes en municipios con escasa evidencia, ajustándolas hacia el promedio provincial. Adicionalmente, el "*pooling* parcial" inherente a los MLM produce estimaciones más estables y eficientes, especialmente para unidades con pocas observaciones (Gelman & Hill, 2007).

3.2 Modelos Predictivos

Sumado al análisis explicativo, se implementarán tres métodos de aprendizaje automático supervisado para evaluar la capacidad predictiva de las variables consideradas y del modelo en su conjunto: (1) una regresión regularizada *Elastic Net* (2) un modelo de bosque aleatorio (*Random Forest*) y (3) la aplicación del algoritmo *XGBoost*. El propósito es probar en qué medida es posible predecir la participación en una mesa dada, a partir de sus características conocidas antes de la elección general (perfil sociodemográfico de su municipio, participación en PASO, contexto político, etc.).

Estos algoritmos suelen lograr alta precisión predictiva al capturar relaciones complejas y potencialmente no lineales en los datos, a costa de menor interpretabilidad. Su uso aquí responde al doble objetivo de validar la robustez del modelo explicativo (por ejemplo, comprobando si las variables identificadas como significativas también son las

más útiles para la predicción) y de cuantificar el máximo poder predictivo alcanzable con la información disponible. A continuación, se describe la aplicación de cada uno de estos métodos en este contexto:

Regresión Regularizada *Elastic Net*: Método que combina penalizaciones L1 y L2, útil para manejar multicolinealidad entre predictores y realizar selección automática de variables. Se ajustarán los hiperparámetros α y λ mediante validación cruzada. Se analizará el patrón de coeficientes seleccionados como indicador de importancia predictiva (Hastie, Tibshirani, & Friedman, 2009). Esta técnica ha mostrado ser especialmente útil en escenarios con muchas variables explicativas colineales, ya que realiza una selección de atributos mitigando la inestabilidad que provoca la colinealidad. En el contexto del presente trabajo, las variables sociodemográficas municipales presentan correlaciones entre sí (por ejemplo, porcentaje de población joven vs. adulta mayor, o indicadores socioeconómicos relacionados), lo que puede dificultar la estimación fiable de un modelo de Mínimos Cuadrados Ordinarios (MCO) tradicional. *Elastic Net* tiende a mantener en el modelo grupos de predictores correlacionados, asignándoles coeficientes reducidos pero distintos de cero.

Bosque Aleatorio (Random Forest): Un *Random Forest* construye un gran número de árboles de regresión independientes, entrenados sobre diferentes muestras *bootstrap* del conjunto de datos y con subconjuntos aleatorios de variables en cada división de nodo (técnica de *bagging* y *feature subsampling)*. Al promediar las predicciones de muchos árboles, el bosque resultante logra una capacidad predictiva robusta y generalmente superior a la de cualquier árbol individual, mitigando problemas de sobreajuste propios de los árboles únicos (Hastie, Tibshirani, & Friedman, 2009).

En el marco de esta investigación, un árbol de decisión intentaría segmentar iterativamente las mesas en grupos cada vez más homogéneos en cuanto al nivel de participación, eligiendo en cada paso la variable y punto de corte que mejor reduce la incertidumbre (minimizando la varianza intragrupo).

El bosque aleatorio, al promediar cientos de árboles entrenados ligeramente diferente, promedia los errores y captura únicamente los patrones consistentes, mejorando la generalización. Breiman (2001) aduce que los *Random Forests* suelen ser de los algoritmos de mayor precisión predictiva disponible, funcionando bien automáticamente incluso sin mucho ajuste de parámetros. Además, presentan varias ventajas prácticas relevantes para este estudio:

1) **pueden manejar eficientemente grandes bases de dato**s (tanto en número de observaciones –aquí ~240 mil mesas– como de variables) mediante computación paralela de árboles.

2) **no requieren depurar el conjunto de predictores eliminando variables menos significativas**, ya que realizan una forma de selección interna –los predictores menos informativos simplemente rara vez serán elegidos en los nodos de decisión; y

3) **son resistentes a la multicolinealidad y a relaciones no lineales**, dado que los árboles pueden capturar interacciones complejas y efectos de umbral entre variables sin necesidad de especificarlos previamente. Por ejemplo, el modelo podría descubrir automáticamente que el efecto de la participación en PASO sobre la participación en las elecciones generales no es lineal, sino que exhibe un punto de saturación, o que cierta combinación de características municipales conjuntamente produce altos aumentos de participación, algo difícil de captar con una regresión lineal clásica.

La configuración del *Random Forest* para este proyecto se realizará usando todas las variables disponibles como potenciales predictores (incluyendo variables a nivel municipal y a nivel de mesa). Se tratará como un problema de regresión continua, donde la predicción es el nivel de participación (en puntos porcentuales) en las elecciones generales en cada mesa. Antes del entrenamiento, se dividirán los datos en conjuntos de entrenamiento y prueba para validar el desempeño del modelo. Se entrenará el bosque con un número suficientemente grande de árboles (500 árboles) para garantizar la convergencia de las métricas de *error out-of-bag* –los errores de predicción en las muestras no usadas por cada árbol, que sirven como estimación interna del error sin necesidad de un conjunto de validación aparte.

Extreme Gradient Boosting (XGBoost)

XGBoost (*eXtreme Gradient Boosting*) es una implementación altamente optimizada de *Gradient Boosted Decision Trees* (GBDT) desarrollada por Chen y Guestrin (2016) que constituye un ensamblado de árboles de decisión de manera secuencial, corrigiendo los errores del modelo previo mediante el uso de gradientes (primeras derivadas) y Hessianos (segundas derivadas) de la función de pérdida para ajustar cada nuevo árbol. A diferencia de *Random Forest*, que promedia árboles independientes, *XGBoost* añade en cada iteración un árbol que minimiza una función objetivo diferenciable. Sus hiperparámetros clave incluyen la tasa de aprendizaje (`eta`), el número de árboles (`nrounds`), la profundidad máxima de los árboles (`max_depth`), el parámetro `gamma` (para control de ganancia mínima en un Split) y el `subsample` de filas y columnas, permitiendo explorar un compromiso entre sesgo y varianza. El entrenamiento se realiza mediante *boosting* basado en gradiente: se construye un modelo inicial, se evalúan los residuos y se ajusta un nuevo árbol para predecir esos residuos, repitiendo el proceso hasta lograr la convergencia de la función objetivo.

4. Resultados

4.1 Modelo multinivel explicativo

Para analizar los determinantes del fenómeno, se especificó un modelo que aprovecha la estructura anidada de los datos, con observaciones a nivel de mesa electoral (nivel 1) agrupadas dentro de circuitos electorales (nivel 2). La elección de un modelo de regresión multinivel (MLM) se justifica porque las observaciones dentro de un mismo circuito tienden a estar más correlacionadas entre sí, violando el supuesto de independencia de los modelos de regresión tradicionales. Ignorar esta estructura jerárquica podría llevar a estimaciones de error estándar incorrectas e inferencias inválidas.

Se probaron múltiples especificaciones para encontrar el modelo con mejor ajuste y poder explicativo. Se comparó un modelo base solo con efectos aleatorios para circuito y año (m0), un modelo con una estructura anidada de circuito dentro de clúster (m1), y un modelo con clúster como efecto fijo (m2). Como se observa en la Tabla 5, el modelo con efectos anidados (m1) demostró un desempeño superior en términos de AIC, BIC y R^2 condicional.

Tabla 6:Comparación de desempeño entre modelos multinivel

Name	Model	R2_conditional	R2_marginal	ICC	RMSE	Sigma	AIC_wt	AICc_wt	BIC_wt	Performance_Score
m1	lmerMod	0.75	0.17	0.70	3.62	3.62	1	1	1	0.88
m2	lmerMod	0.55	0.32	0.33	3.62	3.62	0	0	0	0.37
m0	lmerMod	0.53	0.31	0.32	3.67	3.67	0	0	0	0.12

La estructura del modelo incluye **efectos fijos** para estimar la relación promedio entre los predictores y la variable de respuesta, y un **efecto aleatorio** para el intercepto a nivel de circuito electoral. Esto último permite que el nivel de participación basal varíe entre los diferentes circuitos, capturando así la heterogeneidad geográfica no explicada por las covariables incluidas.

A continuación, se presenta la formalización matemática del modelo. Para cada mesa i en el circuito j:

$$(1) \quad Y_{ij} = \beta_{0j} + \beta_1 \cdot PASO_{ij} + \beta_2 \cdot ln(PBG\_pc_j) + \beta_3 \cdot ln(Pob_j) + \beta_4 \cdot Edad_{0-14,j} + \beta_5 \cdot Edad_{65^+,j} + \beta_6 \cdot Pisos_j + \beta_7 \cdot Agua_j + \beta_8 \cdot Cluster_{2,j} + \beta_9 \cdot Cluster_{3,j} + \beta_{10} \cdot Opositor_j + \varepsilon_{ij}$$

Donde:

- Y_{ij} es la **tasa de participación en la elección general** (en puntos porcentuales) de la mesa i en el circuito j.

- β_0 es el **intercepto** para el circuito j, que representa la participación general predicha para ese circuito cuando todos los demás predictores son cero. Este intercepto varía aleatoriamente entre circuitos.

- El coeficiente fijo β_1 está asociado a la **tasa de participación en la PASO** correspondiente a la mesa i, y mide el cambio promedio en la participación general por cada punto de aumento en la participación primaria.

- Los términos desde β2 hasta β10 representan los **efectos fijos** de las covariables medidas a nivel de circuito o municipio: los coeficientes β_2 a β_{10} recogen los efectos fijos de las covariables de contexto: el **logaritmo del PBG per cápita** (β_2) y el **logaritmo de la población total** (β_3) ; los porcentajes de población en los rangos **0-14 años** y **65 años o más** $(\beta_4 \, y \, \beta_5$)dos indicadores de condiciones habitacionales —**pisos revestidos** y **agua por cañería dentro del hogar** $(\beta_6 \, y \, \beta_7)$; las variables *dummy* **clúster 2** y **clúster 3** $(\beta_8 \, y \, \beta_9)$, que se interpretan en referencia al clúster 1; y, por último, la *dummy* **ofic_opos** (β_{10}) igual a 1 cuando el municipio es gobernado por la oposición.

- ε_{ij} es el **residuo o error aleatorio a nivel de mesa**, que captura la variabilidad no explicada. Se asume que sigue una distribución normal e independiente: ϵij $\sim \mathcal{N}(0, \sigma^2)$

Ecuación 2: Modelización a Nivel de Circuito (Nivel 2)

El componente que varía entre los circuitos/municipios es el intercepto. Este se modela como la suma de un intercepto promedio global y una desviación aleatoria específica para cada circuito:

(2) $$\beta_{0j} = \gamma_{00} + u_{0j} \, ; \; u_{0j} \sim \mathcal{N}(0, {\tau_{00}}^2) \, ; \; \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Donde:

- $\boldsymbol{\gamma_{\{00\}}}$ es el intercepto fijo global, que representa el promedio de la participación general en todos los circuitos cuando los predictores están en su valor de referencia.

- $\boldsymbol{u_{\{0j\}}}$ es el efecto aleatorio para el circuito j. Este término captura la desviación del intercepto de un circuito específico respecto del promedio global. Refleja la heterogeneidad no observada entre circuitos (factores culturales, geográficos o

políticos locales no medidos por las covariables). Se asume que estos efectos aleatorios siguen una distribución normal con media cero y una varianza constante: $\boldsymbol{u}_{\{0j\}} \sim N(0, \tau_{00}^2)$.

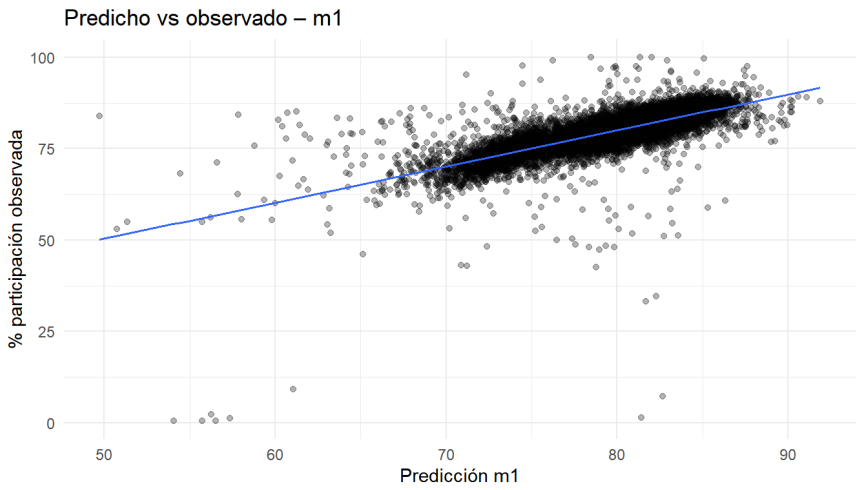
Interpretación de la Especificación del Modelo

Al combinar ambas ecuaciones, el modelo final estima una única relación (pendiente) para cada predictor en todos los circuitos (efectos fijos), pero permite que el punto de partida (intercepto) de esa relación sea diferente para cada circuito. A diferencia de un modelo con pendientes aleatorias, esta especificación asume que el efecto de un aumento en la participación en las PASO, por ejemplo, es constante en todo el territorio. Esta decisión se basa en la comparación de modelos, donde una especificación más parsimoniosa como esta demostró un mejor ajuste y poder predictivo.

En síntesis, el modelo permite responder preguntas como: ¿cuál es el efecto *promedio* de la participación en las PASO sobre la participación general?, mientras simultáneamente cuantifica *cuánta* de la variabilidad total en la participación se debe a diferencias estructurales inherentes a cada circuito electoral.

La Figura 13 contrasta los valores predichos por el modelo anidado (m1) con las tasas de participación observadas a nivel de mesa. Cada punto gris representa una mesa electoral, mientras que la línea azul resume la relación lineal entre ambas magnitudes. La concentración diagonal de los puntos —especialmente entre 70 % y 85 %— indica que el modelo reproduce con fidelidad la mayor parte del rango empírico, aunque la dispersión se amplía en los extremos inferiores y superiores, señalando mayor incertidumbre en mesas atípicas con muy baja o muy alta participación.

Figura 13: Ajuste predictivo: nube de puntos



Fuente: elaboración propia

Efectos fijos: estimaciones e interpretación

A continuación, se presentan los estimadores de los efectos fijos del modelo:

Tabla 7:Efectos fijos- coeficientes, error estándar e intervalos de confianza al 95%

effect	term	estimate	std.error	statistic	conf.low	conf.high
fixed	(Intercept)	93.554	2.507	37.321	88.641	98.467
fixed	participacion_paso_pct	0.375	0.001	267.495	0.373	0.378
fixed	log_pbg_pc	-3.056	0.063	-48.393	-3.179	-2.932
fixed	log_poblacion	-0.936	0.080	-11.691	-1.093	-0.779
fixed	edad_0_a_14_percent	-0.049	0.052	-0.941	-0.151	0.053
fixed	edad_65_o_mas_percent	-0.099	0.032	-3.108	-0.162	-0.037
fixed	pisos_con_revestimiento	-0.147	0.008	-17.468	-0.164	-0.131
fixed	agua_por_caneria_dentro_de_la_vivienda	0.135	0.016	8.445	0.104	0.166
fixed	cluster2	-7.027	0.402	-17.472	-7.815	-6.239
fixed	cluster3	1.459	0.318	4.593	0.837	2.082
fixed	ofic_oposOpositor	-0.610	0.021	-28.625	-0.651	-0.568

El análisis de los coeficientes revela varias dinámicas clave:

- **Participación en PASO**: Esta variable presenta el coeficiente positivo y más robusto del modelo (**0,375**). Esto indica que, controlando por los demás factores, cada punto porcentual de participación en las primarias se asocia con un aumento de 0,375 puntos en la participación de la elección general. Este hallazgo es central y sugiere que la primaria actúa como un evento movilizador que predice una mayor concurrencia posterior.

- **Nivel socioeconómico (PBG per cápita)**: El coeficiente de **-3,056** para el logaritmo del PBG per cápita es negativo y significativo. Esto sugiere que los municipios con mayor nivel de riqueza tienden a tener una participación general comparativamente menor, una vez que se controla por la participación en las PASO y otros factores. Esto contradice los resultados esperados en la literatura, que halla relaciones positivas entre nivel socioeconómico y participación electoral. Esto podría deberse a factores no lineales que no se capturaron correctamente en la especificación del MLM, o bien al hecho de que existen otras variables asociadas al nivel socioeconómico que capturan parte de este efecto sobre la participación (por ejemplo, acceso a agua por cañería, que presenta un coeficiente positivo). Podría deberse también al hecho de que parte de la heterogeneidad socioeconómica se captura vía los efectos fijos a nivel circuito electoral.

- **Variables demográficas**: El logaritmo de la población (**-0,936**) y el porcentaje de población mayor a 65 años (**-0,099**) muestran coeficientes negativos. Esto podría indicar que en municipios más poblados o con una estructura demográfica más

envejecida, el incremento de la participación entre elecciones es menor. La proporción de población joven (0-14 años), en cambio, no resultó estadísticamente significativa.

- **Condiciones de vida y contexto político**: Un mayor porcentaje de hogares con pisos con revestimiento (**-0,147**) se asocia negativamente con la participación, mientras que el acceso a agua por cañería (**0,135**) lo hace positivamente. Por su parte, el hecho de que un municipio sea gobernado por un partido opositor al gobierno provincial (**-0,610**) se asocia con una participación general menor en comparación con los municipios oficialistas.

- **Efecto Clúster**: La pertenencia a los clústeres territoriales identificados en el análisis exploratorio es un predictor fuerte. Tomando el Clúster 1 como base, pertenecer al Clúster 2 se asocia con una disminución de **7,03** puntos en la participación, mientras que pertenecer al Clúster 3 se asocia con un aumento de **1,46** puntos.

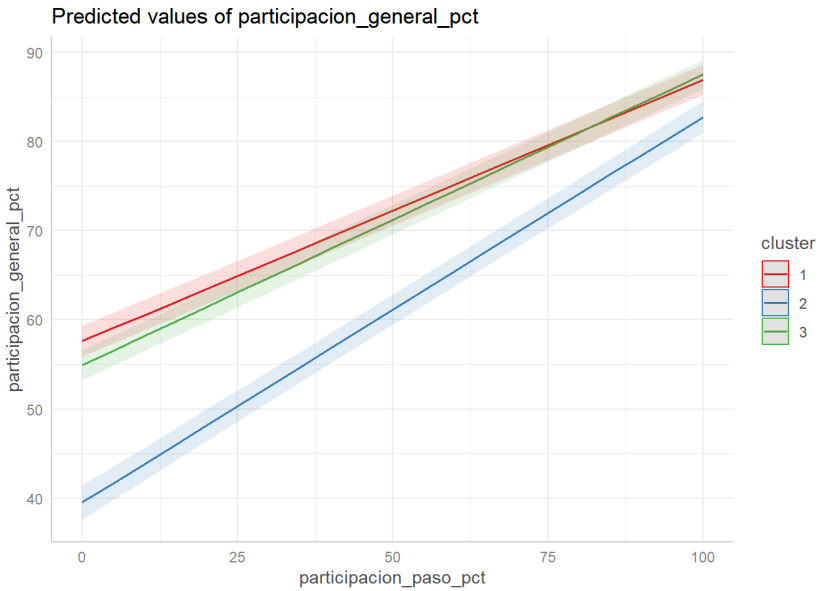
Efectos aleatorios: varianzas y correlaciones

El modelo estima la varianza asociada al intercepto aleatorio por circuito electoral y la varianza residual a nivel de mesa.

***Tabla 8**:Efectos aleatorios, estimación de coeficientes*

effect	group	term	estimate
ran_pars	circuito_num	sd__(Intercept)	2.440
ran_pars	Residual	sd__Observation	4.088

Figura 14: Ajuste predictivo por clúster, efectos fijos y marginales



Fuente: elaboración propia

La Figura 8 muestra los valores predichos de la participación general en función de la participación en las PASO, desagregados por los tres clústeres tipificados en el modelo. Las rectas ascendentes confirman el efecto positivo estimado: a mayor participación primaria, mayor asistencia prevista en la elección de primer orden. La pendiente prácticamente idéntica entre clústeres respalda la especificación sin interacción—el incremento marginal asociado a cada punto adicional en las PASO es homogéneo en todo el territorio.

Las diferencias entre bandas coloreadas obedecen, por tanto, a interceptos distintos: el clúster 1 (rojo) parte de un piso de participación más alto, seguido del clúster 3 (verde), mientras que el clúster 2 (azul) registra los niveles más bajos en todo el rango de la variable independiente (las franjas sombreadas representan intervalos de confianza al 95 %).

La varianza del intercepto a nivel de circuito (2,44) indica que existen diferencias sistemáticas en el nivel de participación entre los distintos circuitos electorales, incluso después de controlar por los efectos fijos. La varianza residual (4,088) representa la variabilidad no explicada a nivel de mesa.

Tabla 9: Efectos aleatorios, ICC, y R2 marginal y condicional

Random Effects	
σ^2	16.71
$\tau_{00\ circuito\_num}$	5.95
ICC	0.26
$N_{\ circuito\_num}$	836
Observations	229971
Marginal R^2 / Conditional R^2	0.370 / 0.535

A partir de estos componentes, se calculó el Coeficiente de Correlación Intraclase (ICC), que resultó ser de 0,26. En cuanto a la bondad de ajuste, el R^2 marginal asciende a 0,37 y representa la proporción de varianza explicada únicamente por los efectos fijos —participación en las PASO, PBG per cápita, composición etaria y demás covariables—. Cuando se incorporan los interceptos aleatorios, el R^2 condicional se eleva a 0,535. El incremento de 0,165 puntos revela que la heterogeneidad territorial, capturada mediante el componente aleatorio a nivel de circuito, aporta cerca de un 17 % adicional de poder explicativo. Esta diferencia subraya la pertinencia de la especificación multinivel: no considerar la estructura geográfica habría dejado sin modelar una fracción sustancial de la varianza asociada a factores culturales, institucionales o logísticos propios de cada circuito electoral.

Métricas de performance y validación del modelo

El desempeño del modelo se contrastó en datos no vistos mediante una validación cruzada estratificada de cinco pliegues. El conjunto de mesas se dividió aleatoriamente en cinco subconjuntos de tamaño similar, manteniendo la distribución de los clústeres territoriales en cada pliegue para preservar la heterogeneidad espacial. En cada iteración, se ajustó el modelo con cuatro pliegues (80 % de los casos) y se evaluó la capacidad predictiva sobre el pliegue restante, calculando el error cuadrático medio (RMSE) y el error absoluto medio (MAE). Al repetir el proceso cinco veces —rotando sucesivamente el pliegue de prueba— se obtuvo una estimación robusta del rendimiento fuera de muestra junto con su variabilidad; la media y la desviación estándar de ambas métricas se reportan en la Tabla a continuación:

Tabla 10: Métricas de error de pronóstico del modelo final (k-fold CV)

RMSE_mean	RMSE_sd	MAE_mean	MAE_sd
3.62	0.06	2.22	0.01

El error cuadrático medio (RMSE) promedio fue de 3.62 y el error absoluto medio (MAE) fue de 2.22, indicando que, en promedio, las predicciones del modelo se desvían entre 2,2 y 3,6 puntos porcentuales del valor real de participación.

Análisis de Multicolinealidad (VIF)

Se realizó un diagnóstico para detectar posible multicolinealidad entre las variables predictoras, la cual puede incrementar los errores estándar y hacer que las estimaciones de los coeficientes sean inestables. Para ello, se calculó el Factor de Inflación de la Varianza (VIF) para cada predictor. Un VIF superior a 5 o 10 suele ser considerado problemático.

Tabla 11: Análisis de multicolinealidad (factor VIF)

Variable (Predictor)	VIF	IC 95% (VIF)	VIF Ajustado
participacion_paso_pct	1,02	[1.01, 1.02]	1,01
log_pbg_pc	1,21	[1.20, 1.21]	1,1
log_poblacion	1,25	[1.25, 1.26]	1,12
edad_0_a_14_percent	1,17	[1.16, 1.17]	1,08
edad_65_o_mas_percent	1,45	[1.45, 1.46]	1,21
pisos_con_revestimiento	4,85	[4.82, 4.89]	2,2
agua_por_caneria_dentro_de_la_vivienda	4,54	[4.50, 4.57]	2,13

Los resultados muestran que la mayoría de las variables tienen valores de VIF muy cercanos a 1, indicando una correlación muy baja con los otros predictores. Las variables *pisos_con_revestimiento* y *agua_por_caneria* presentan los valores más altos (4.85 y 4.54, respectivamente), lo cual es esperable ya que ambas son indicadores de desarrollo socioeconómico municipal. Sin embargo, al estar por debajo del umbral crítico de 5, se concluye que no existe un problema de multicolinealidad severo que comprometa la validez de las estimaciones del modelo.

4.2 Modelos predictivos de *machine learning*

Además del MLM explicativo, se entrenaron modelos predictivos para estimar el mismo objetivo (el porcentaje de participación en las elecciones generales) a nivel de mesa. Estos modelos utilizan validación temporal tipo *Leave-One-Election-Out* (LOEO): se deja fuera cada elección general para validación, entrenando con los años anteriores, de manera de tomar en cuenta el componente temporal de las observaciones. Las

covariables empleadas son las mismas variables socioeconómicas y la participación en PASO, con *dummy* de partido oficialismo-opositor y de clústeres.

En cada iteración de LOEO, se reserva una elección completa como conjunto de validación y se entrena el modelo únicamente con las elecciones anteriores, garantizando que la información futura no "filtre" hacia el ajuste. Se construyeron seis *folds* (particiones) sucesivos: para predecir 2013 se entrena con 2011; para predecir 2015 se entrena con 2011 y 2013; y así sucesivamente:

Se contrastaron tres familias de modelos supervisados:

1) **Elastic Net**: regresión lineal con penalizaciones mixtas L1 y L2; se optimizó la proporción α y el parámetro de regularización λ mediante búsqueda en cuadrícula dentro de cada fold LOEO.

2) **Random Forest**: 500 árboles, criterio de impureza MSE y tamaño de muestreo mtry=√p.

3) **XGBoost**: 150 rondas de *boosting*, tasa de aprendizaje 0,10 y profundidad máxima de los árboles igual a 5.

Cada algoritmo recibió idéntico set de predictores (preprocesados mediante z-score) y contó con un proceso de imputación de valores faltantes por la mediana.

4.2 Desempeño predictivo comparado e importancia de variables

La Tabla 11 detalla, para la mejor configuración de hiperparámetros de cada familia de modelos, el error cuadrático medio (RMSE) promedio entre los siete pliegues y su correspondiente error estándar:

Tabla 12: RMSE comparado por método

Modelo	RMSE medio (p.p.)	Desvío estándar
Elastic Net	4,58	0,43
XGBoost	4,96	0,54
Random Forest	5,01	0,65

La brecha de 0,4 puntos porcentuales en el RMSE entre *Elastic Net* y *XGBoost*—y de casi 0,5 frente a *Random Forest*—se reproduce de manera consistente a lo largo de los siete pliegues LOEO. Esta estabilidad indica que el compromiso entre sesgo y varianza se inclina en favor de la regresión penalizada cuando la relación señal-ruido exhibe un comportamiento predominantemente lineal. En efecto, la penalización mixta L1-L2 de *Elastic Net* actúa de forma adaptativa: atenúa o elimina predictores colineales

(frecuentes en los indicadores socioeconómicos) sin sacrificar variables débiles pero sistemáticas, limitando la varianza sin aumentar de forma apreciable el sesgo.

Esta estrategia de regularización redunda en un modelo parsimonioso—menos de una decena de coeficientes sin penalizar—que explica más del 75 % de la varianza observada, facilitando además la interpretación y la transferencia de los hallazgos al plano operativo. La homogeneidad del error entre elecciones confirma la robustez temporal de los coeficientes estimados: los distritos con baja participación inicial, elevada proporción de votantes jóvenes y contiendas competitivas emergen de manera reiterada como los escenarios con mayor potencial de movilización de cara a la elección definitiva.

Tabla 13: Importancia de variables por método

Importancias de variables: Elastic Net (|coef|), Random Forest (Impurity), XGBoost (Gain, Cover, Frequency)

Variable	ElasticNet	RandomForest	Gain	Cover	Frequency
participacion_paso_pct	3.3013	3205748.22	0.9524	0.6731	0.6667
log_pbg_pc	0.0866	354654.74	0.0000	0.0000	0.0000
log_poblacion	0.2624	303248.51	0.0000	0.0000	0.0000
edad_0_a_14_percent	0.0000	266418.26	0.0206	0.2863	0.1667
edad_65_o_mas_percent	0.0000	288797.38	0.0000	0.0000	0.0000
pisos_con_revestimiento	0.5843	423719.76	0.0000	0.0000	0.0000
agua_por_caneria_dentro_de_la_vivienda	0.0000	284495.86	0.0000	0.0000	0.0000
cluster_X1	0.0011	10646.65	0.0000	0.0000	0.0000
cluster_X3	0.0000	26696.09	0.0000	0.0000	0.0000
cluster_other	0.3657	117944.54	0.0270	0.0406	0.1667
ofic_opos_Oficialismo	0.0577	32466.87	0.0000	0.0000	0.0000
ofic_opos_Opositor	0.0010	32961.59	0.0000	0.0000	0.0000

La tabla precedente sintetiza la relevancia de cada predictor a partir de tres estrategias de modelización con distintas métricas. En el caso de *Elastic Net* se reporta el valor absoluto del coeficiente estandarizado, indicador directamente interpretable como variación porcentual de la variable dependiente ante un cambio de una desviación típica en el predictor. Para *Random Forest* la importancia se resume mediante la *Mean Decrease in Impurity* (MDI), que refleja la contribución acumulada de cada variable a la reducción del error de los árboles; en XGBoost se exponen de forma conjunta el *gain* (aporte promedio a la función objetivo), el *cover* (proporción de observaciones a las que la variable sirve de nodo de partición) y la *frequency* (frecuencia con que aparece en los árboles). Si bien las escalas numéricas no son comparables de forma absoluta, su lectura relativa permite identificar patrones consistentes y divergencias metodológicas.

En los tres algoritmos sobresale la participación en la PASO como predictor, corroborando la hipótesis de que el caudal inicial de votantes es el principal determinante de la concurrencia en la elección general. Los indicadores de estructura

económica —*log_pbg_pc* y *log_poblacion*— ocupan posiciones altas en los modelos basados en árboles, lo que sugiere la presencia de umbrales no lineales vinculados al desarrollo municipal y a la densidad demográfica; su peso es menor, aunque todavía apreciable, en la regresión penalizada al capturar una relación más próxima a la linealidad. Por el contrario, variables demográficas específicas como los porcentajes de niños (*edad_0_a_14_percent*) y de adultos mayores (*edad_65_o_mas_percent*) son eliminadas por *Elastic Net* debido a la penalización L1, mientras que mantienen una influencia moderada en *Random Forest* debido probablemente a la interacción con otros predictores. La condición política del municipio *(ofic_opos_Oficialismo*) muestra una importancia intermedia y estable, indicando que la alineación partidaria incide en la movilización, pero en menor medida que los determinantes estructurales.

Finalmente, los clústeres derivados del análisis no supervisado reciben escaso peso en la especificación lineal, pero alcanzan valores apreciables de *cover* en *XGBoost,* lo que evidencia su utilidad para caracterizar segmentos reducidos pero informativos dentro del espacio de decisión de los árboles. Finalmente, la comparación confirma la importancia de los predictores centrales —participación previa, nivel de desarrollo y densidad— y pone de relieve cómo cada técnica modula la contribución de variables periféricas según su propio régimen de regularización o de partición, aportando así una visión complementaria de los mecanismos subyacentes al comportamiento electoral**.**

4.3 Discusión comparativa

Al comparar los enfoques explicativos (MLM) y predictivos (ML), se observa una complementariedad informativa. El MLM ofrece interpretabilidad: cuantifica cómo cada factor socioeconómico afecta el diferencial de participación y revela heterogeneidad municipal. Por ejemplo, confirma que una mayor participación en PASO tiende a predecir la participación en la elección general (coeficiente fijo positivo) y destaca la importancia de variables demográficas e institucionales (edad, ingresos, características de la vivienda, etc.). Sin embargo, su poder predictivo individual es moderado (varianza residual alta), lo que motivó la prueba de modelos de ML. Los modelos ML, aunque menos transparentes, capturan mejor los patrones complejos: logran predecir con menor error (especialmente *Elastic Net*). No obstante, las señales clave son consistentes con el MLM: en todos ellos, la participación en PASO emerge como el predictor más influyente del diferencial por mesa. En ML se mide su contribución absoluta; en MLM se ve que alta participación en PASO implica una mayor participación electoral en las elecciones generales.

Esta consistencia sugiere que los factores significativos en el análisis explicativo son justamente los que dominan la precisión predictiva. Sin embargo, también hay discrepancias: *Elastic Net* otorga algo más de importancia a las condiciones habitacionales, lo que podría reflejar no linealidades que el MLM no capta. En términos

prácticos, combinar ambos enfoques permite recomendaciones más sólidas. El MLM señala por qué ciertos municipios presentan alta magnitud de participación altos (p.ej. municipios con baja concurrencia en PASO y de orientación oficialista); mientras que los modelos ML pueden utilizar esas mismas variables para estimar cuánto podría aumentar la votación en cada mesa si se interviene. Por ejemplo, si se prioriza campañas en mesas con baja participación en PASO (según señal MLM) y, simultáneamente, se confirma que esa variable es predictiva (según ML), se puede planificar recursos de campaña o intervenciones de políticas públicas con mayor eficacia.

De esta forma, el modelo multinivel aporta explicación y valida la dirección de los efectos estructurales, mientras que los modelos de ML aportan capacidad predictiva en la asignación de esfuerzos. La señal uniforme de las variables clave (participación PASO, orientación política del municipio, edad, etc.) aumenta la confianza en las estrategias derivadas: se recomienda enfocarse en municipios de menor participación en PASO y con perfiles demográficos específicos, pues son aquellos donde un esfuerzo adicional que incentive la participación electoral en las elecciones generales tiene mayor impacto esperado, según ambos tipos de modelos.

4.4 Optimización dinámica de la asignación de recursos

Para traducir los pronósticos de la participación electoral en decisiones concretas de asignación de esfuerzos de campañas de incentivo de movilización ciudadana, se puede formular un problema de optimización que maximice el impacto electoral sujeto a un límite presupuestario. En el modelo actual —un problema de programación entera mixta (MIP) de tipo "*knapsack*" binario—, se podría plantear lo siguiente:

1. **Variables de decisión**

(3) $$x_i \in \{0, 1\}, \quad i = 1, \ldots, n$$

indica si se destinan recursos (por ejemplo, visitas de terreno, spots locales, envío de boletas informativas) a la mesa o circuito i.

2. **Función objetivo**

(4) $$\max \sum_{i=1}^{n} \Delta_i x_i$$

donde Δi es la ganancia esperada, medida en términos de la diferencia entre la la participación en las PASO y la participación predicha en la mesa i.

3. **Restricción presupuestaria**

(5) $$\sum_{i=1}^{n} c_i x_i \leq B$$

con c_i el costo (directo o proxy) de atender la mesa i y B el presupuesto total disponible.

Este esquema binario corresponde a la versión más elemental del problema de la mochila (0–1 *knapsack*), cuya solución exacta se obtiene eficientemente para varios miles de variables con *solvers* como GLPK o CPLEX.

Supuestos y limitaciones

1. **Aditividad y linealidad**. Se supone que las ganancias por incrementar la participación de diferentes mesas se suman sin interacción. Esto descuida posibles efectos de saturación (p. ej. que después de cierto nivel de esfuerzo adicional en un mismo circuito el beneficio marginal disminuya) y externalidades (movilización de votantes vecinales).

2. **Tratamiento uniforme de costos**. Asumir ci=1 es razonable cuando no hay datos fidedignos de costos diferenciales, pero pierde realismo en contextos donde algunas mesas requieren viajes más largos o donde el "costo" de convencer a cada votante varía según perfil sociodemográfico. La estimación de este parámetro excede los límites de este trabajo, pero podrían incluirse variables como costo de movilizar recursos, diferencias en el público objetivo que exijan estrategias diferenciales para poder hacer efectiva la comunicación de campaña, etc.

3. **Un único período estático**. La formulación ignora la dimensión temporal: no contempla que los efectos de una campaña hoy puedan influir en escenarios futuros ni permite repartir el presupuesto en distintos momentos del ciclo electoral.

Simulación

La formulación previa describe un modelo de optimización estático basado en programación entera mixta (MIP) para asignar recursos de manera eficiente, maximizando el incremento predicho en la participación electoral (*turnout*) sujeto a un presupuesto. Para ilustrar la aplicabilidad de este enfoque y explorar sus implicaciones prácticas, se realiza una simulación utilizando las predicciones generadas por el mejor modelo de *Machine Learning* identificado en la sección anterior *(Elastic-Net)*

Metodología de Simulación

La simulación se enfoca en traducir las predicciones de incremento potencial de participación electoral (Δi) para cada mesa electoral (i) en una selección óptima de mesas a intervenir, dadas diferentes restricciones presupuestarias (B).

1. Datos de Entrada: Se utiliza el valor predicho del diferencial de votos (*pred)* obtenido con el modelo *Elastic-Net* final, ajustado sobre todo el conjunto de datos históricos. Para la simulación, se seleccionan las predicciones de nivel de participación en las elecciones generales correspondientes al año 2023, ya que representan la estimación más actualizada del potencial de incremento. Se asume que Δi en el modelo de optimización corresponde directamente al valor predicho para la mesa i en el año seleccionado.

2. Definición de Costos (ci):. Se explora la sensibilidad de la solución del modelo de optimización ante una especificación de costos que depende del nivel de participación en las PASO de cada mesa. En lugar de asignar un costo fijo estándar, se define:

(6) $$c_i = 1 + \alpha \frac{p_i - \bar{p}}{\bar{p}}$$

donde p_i es la participacion observada en las PASO en la mesa i, $\bar{p}$ es la participación promedio en el año 2023 sobre todas las mesas, y α es un parametro de sensibilidad que linealiza la penalizacion o incentivo en funcion de la desviacion relativa. Se mantuvo un presupuesto total de $B = 1000$ unidades de costo para garantizar la comparabilidad entre escenarios, y se exploran dos valores de α :1 y 2. Entre una multiplicidad de factores, podría especularse que un nivel de participación más alto en las PASO deja menos margen de crecimiento para las generales, lo cual implicaría un mayor esfuerzo para captar votantes que, en su mayoría ya votaron en las PASO.

En ambos escenarios, se resolvió el problema de maximización de la ganancia total de incremento de participación basada en las predicciones del modelo, sujeta a la restricción $\sum_i c_i x_i \leq B$. Los resultados se resumen en la Tabla 13 que muestra, para cada α el número total de mesas seleccionadas, la ganancia agregada de participación electoral (medida simplemente como el incremento en participación entre las PASO y las elecciones generales de las mesas tratadas) y el costo promedio por mesa seleccionada.

Tabla 14 ***:Sensibilidad del costo según participación en PASO (Presupuesto = 1000)***

α	Mesas seleccionadas	Incremento en participación	Costo Promedio
1	995	8.214	1,01
2	1.049	8.253	0,95

Con α = 1 el algoritmo elige 995 mesas, lo que proyecta un incremento agregado de 8.214 puntos de participación (es decir, incremento de 8,26 puntos promedio por mesa) y supone un costo promedio de 1,01 por mesa. Al duplicar la penalización (α = 2), la optimización se inclina todavía más hacia mesas con participación PASO relativamente

baja: amplía el conjunto a 1.049 mesas, eleva levemente la ganancia esperada a 8.253 puntos (7,87 puntos de incremento promedio) y, al mismo tiempo, reduce el costo unitario a 0,95.

La comparación confirma que intensificar la penalización sobre las mesas "caras" —aquellas donde la participación en las primarias ya es alta— fomenta la inclusión de un mayor número de mesas de bajo costo relativo, logrando así una ligera mejora en la ganancia total y, sobre todo, un descenso del costo promedio del programa.

Implicaciones prácticas para la estrategia de campaña o políticas públicas orientadas a mejorar la participación electoral:

- **Priorización basada en datos:** El marco de optimización proporciona un método sistemático y basado en evidencia para clasificar y seleccionar mesas o circuitos electorales para esfuerzos de campaña dirigidos (como visitas puerta a puerta, publicidad local, llamadas telefónicas, envío de material informativo). Supera la dependencia exclusiva de la intuición o reglas heurísticas simples, al integrar las predicciones del modelo de *Machine Learning* (ML).

- **Soporte a la presupuestación:** Los resultados ayudan a informar las decisiones sobre la asignación de presupuesto. Permiten estimar el impacto incremental esperado al aumentar o disminuir los recursos dedicados a la movilización, facilitando discusiones más informadas sobre el nivel óptimo de inversión.

- **Operacionalización de la sinergia MLM-ML:** Este ejercicio de optimización culmina el proceso analítico que combina enfoques explicativos (MLM) y predictivos (ML). El MLM ayudó a identificar factores estructurales relevantes y a comprender la heterogeneidad (ej., por qué ciertas características municipales se asocian con mayor grado de participación electoral). El ML cuantificó el potencial de cambio (Δi) a nivel de mesa, capturando interacciones complejas. La optimización utiliza estas predicciones cuantitativas para generar un plan de acción concreto, seleccionando las mesas específicas donde la intervención promete ser más fructífera, de acuerdo con los patrones identificados por los modelos previos.

5 Conclusiones y extensiones futuras

Esta tesis ofrece evidencia empírica sobre los determinantes del incremento en la participación electoral entre las elecciones primarias (PASO) y las generales en la Provincia de Buenos Aires, empleando una combinación de métodos explicativos y predictivos. En primer lugar, se confirma la relevancia central del nivel de participación inicial en las PASO como el principal predictor de la participación observada en los comicios generales posteriores. Las mesas y municipios con baja concurrencia en las

PASO tienden sistemáticamente a exhibir mayores niveles de participación en la elección general, validando empíricamente la hipótesis de que una primaria con escasa afluencia deja un margen amplio para la movilización electoral subsiguiente.

En segundo lugar, la aplicación de un modelo de regresión multinivel jerárquico junto con algoritmos de *machine learning* (*Elastic Net, Random Forest y XGBoost*) permitió capturar la influencia de diversas variables sociodemográficas y políticas en la participación electoral, atendiendo a la complejidad e interacciones no lineales del fenómeno. Variables estructurales como el nivel socioeconómico (por ejemplo, ingresos per cápita), la estructura etaria de la población (particularmente una mayor proporción de adultos mayores) y el contexto político local (como la condición de municipio oficialista u opositor) demostraron desempeñar un rol significativo en el nivel de participación entre elecciones. El enfoque multinivel ofreció interpretabilidad a nivel circuito, revelando que, por ejemplo, en municipios menos poblados, de carácter oficialista o de menor desarrollo socioeconómico se observa un mayor nivel de participación, a la vez que detectó heterogeneidad entre municipios.

Paralelamente, los modelos predictivos de *machine learning* capturaron patrones complejos en los datos a nivel de mesa y lograron una mayor capacidad de predicción, coincidiendo en identificar a la participación en PASO como el factor más influyente. No obstante, las variables que estos algoritmos destacaron como más influyentes no coincidieron íntegramente con las identificadas por el modelo multinivel. Esta falta de solapamiento se explica por la naturaleza metodológica distinta de ambos enfoques: mientras el modelo multinivel busca estimar efectos promedio coherentes con la teoría electoral y con la estructura jerárquica de los datos, los algoritmos de aprendizaje automático priorizan maximizar la exactitud predictiva, incluso si ello implica seleccionar combinaciones de predictores con alto poder discriminatorio pero menor interpretabilidad sustantiva.

Finalmente, la integración de enfoques explicativos y predictivos evidenció una complementariedad que extiende el análisis descriptivo para ofrecer recomendaciones prácticas. El modelo multinivel jerárquico aportó comprensión causal y validó la dirección de los efectos, mientras que los modelos de *machine learnin*g aprovecharon esas mismas variables para estimar con mayor precisión dónde ocurrirían el mayor nivel de participación electoral post-PASO.

Cabe señalar, no obstante, algunas limitaciones del estudio, como el uso de datos de escrutinio provisorio (lo que podría introducir imprecisiones) y la ausencia de información detallada sobre la movilización partidaria o factores contingentes a nivel de mesa, así como la naturaleza estática de los modelos empleados, incapaces de capturar adaptaciones estratégicas en tiempo real durante la campaña.

Asimismo, para las elecciones legislativas de Argentina de 2025 el Congreso aprobó la suspensión de las PASO mediante la Ley de Reforma para el Fortalecimiento Electoral; aunque el Poder Ejecutivo promueve su eliminación definitiva, la normativa que las creó (Ley 26.571) sigue vigente mientras no sea derogada expresamente por el Parlamento. Resta analizarse en el futuro el potencial impacto de su anulación y las nueva dinámicas que podrían darse en el sistema electoral del país (por ejemplo, el efecto del desdoblamiento de las elecciones provinciales y nacionales).

En cuanto a extensiones futuras, se propone incorporar datos definitivos y encuestas post-electorales para validar y refinar las predicciones, así como explorar modelos espaciales que tengan en cuenta la interdependencia entre municipios. Asimismo, sería valioso desarrollar experimentos de campo que evalúen la efectividad de diferentes tácticas de movilización.

Referencias

1. ACE Project. (s.f.). *Primary elections in Latin America*. ACE Electoral Knowledge Network. https://aceproject.org/ace-en/topics/pc/annex/pcy/primary-elections-in-latin-america/
2. Breiman, L. (2001). *Statistical Modeling: The Two Cultures*. Statistical Science, 16(3), 199–231
3. Breiman, L. (1996). *Bagging predictors*. *Machine Learning, 24*(2), 123-140. https://doi.org/10.1007/BF00058655
4. Brusco, V., Nazareno, M., & Stokes, S. C. (2004). *Vote buying in Argentina*. Latin American Research Review, 39(2), 66-88. https://doi.org/10.1353/lar.2004.0022
5. Buquet, D., & Gallo, A. (2022). *Elección presidencial a tres vueltas: efectos de las primarias abiertas, simultáneas y obligatorias en Argentina y Uruguay*. Opinião Pública, 28(2), 292-321. https://doi.org/10.1590/1807-01912022282292
6. Cámara Nacional Electoral de la República Argentina (CNE), *Resultados electorales oficiales: Elecciones nacionales* [Base de datos]. https://www.electoral.gob.ar/.
7. Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American Vote*r. Wiley.
8. Chen, T., & Guestrin, C. (2016). Xgboost: *A scalable tree boosting system*. Proceedings of the 22nd international conference on knowledge discovery and data mining (pp. 785-794).
9. Douglas J. A., (2013). The Foundational Importance of Participation: A Response to Professor Flanders, Oklahoma Law Review 66.
10. Fornos, C. A., Power, T. J., & Garand, J. C. (2004). Explaining Voter Turnout in Latin America, 1980–2000. Comparative Political Studies, 37(8), 909–940.
11. Gallo, Ariadna (2018). Primarias Abiertas Simultáneas y Obligatorias en Argentina. Resultados electorales y coordinación de actores. E-l@tina - Revista electrónica de estudios latinoamericanos, vol. 16, núm. 63, 2018
12. Ghitza, Y., & Gelman, A (2013). Deep Interactions with MRP: Election Turnout and Voting Patterns. American Journal of Political Science, 57(3), 762-776. https://doi.org/10.1111/ajps.12004
13. Gelman A, Hill J. (2007). Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.
14. Green, D. P., Palmquist, B., & Schickler, E. (2002). Partisan Hearts and Minds: Political Parties and the Social Identities of Voters. Yale University Press
15. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Stanford, CA: Stanford University

16. Latin American Public Opinion Project (LAPOP) (2022). *Barómetro de las Américas - Vanderbilt University*. Recuperado de https://www.vanderbilt.edu/lapop
17. Lucardi, A., Vallejo, J., & Feierherd, G. (2024). *Three is a crowd: Information and electoral coordination in Argentina*. https://adrianlucardi.com/wp-content/uploads/2024/07/LucardiVallejoFeierherd-Three-Is-a-Crowd-02.pdf
18. Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models. Applications and Data Analysis Methods (2nd ed.). Thousand Oaks, CA Sage Publications.
19. Reif, K., & Schmitt, H. (1980). Nine second-order national elections – A conceptual framework for the analysis of European election results. European Journal of Political Research, 8(1), 3-44.
20. Strnad, M. (2022). Determinantes de participación electoral en referendos en América Latina. Latin American Politics and Society, 64(1), 118-143.
21. Rosenstone, S. J., & Hansen, J. M. (1993). Mobilization, participation, and democracy in America. Macmillan.
22. Shmueli, G. (2010) To Explain or to Predict? Statistical Science, 25, 289-310.
23. Verba, S., Schlozman, K. L., & Brady, H. E. (1995). Voice and equality: Civic voluntarism in American politics. Harvard University Press.
24. Yoon, J. (2018). Socioeconomic Factors and Voter Turnout in South Korea. Asian Journal of Political Science, 26(2), 123-142.