CIENCIA DE DATOS APLICADA – MINE-4101 – 2023-20

PROYECTO FINAL

DASHBOARD PARA VISUALIZACIÓN DE DATOS DE ELECCIONES POLÍTICAS EN COLOMBIA, SEGMENTACIÓN DE POBLACIONES Y ESTUDIO DE TENDENCIAS SOCIODEMOGRAFICAS RELACIONADAS

Bohórquez, Rubén; Parrado, Sebastián; Ramos, Daniel.

1. Definición de problemática y entendimiento del negocio

Este proyecto se desarrollará en conjunto con Ingenial Media, una empresa que crea e implementa líneas de acción orientadas a generar oportunidades de participación electoral para sus clientes, velando por sus garantías, la protección de sus intereses, y la prevención de cualquier alteración en los resultados, y cuya necesidad a resolver es falta de información precisa y confiable para la toma de decisiones en campañas políticas.

Para solucionar esta problemática, se plantearon los siguientes objetivos:

- Realizar una exploración inicial de los datos para entender su estructura y contenido
- Analizar las características sociodemográficas y de participación política en diferentes niveles de organización política: departamentos, y municipios.
- Utilizar modelos estadísticos para encontrar tendencias y patrones de voto de acuerda a condiciones sociodemográficas.
- Desarrollar herramientas gráficas y de visualización para representar los resultados obtenidos.

Una vez se hayan cumplido los objetivos, la efectividad del modelo propuesto podrá ser evaluada aprovechando el gran volumen de la base de datos, por ende, se planea dividir la base de datos en subconjuntos de entrenamiento, validación y prueba.

2. Ideación

Los usuarios finales del producto serían los clientes de Ingenial Media: organizaciones y partidos políticos interesados en observaciones confiables sobre intención de voto, con el fin de generar planes de acción y determinar segmentos poblacionales sobre los cuales orientar sus esfuerzos.

Actualmente, a estos clientes solo se les presenta información visual de tendencias en elecciones anteriores, enfocado inicialmente en puntualizar los departamentos y municipios más relevantes según el número de votos, y luego mostrar segmentos de resultados por departamento, métricas como total de votos, proporción de votos ganados por partido, municipios con más votos, partido ganador, votos en consulados, y comparativos de resultados en segunda vuelta de ciertas elecciones.

Basados en esto, determinamos que el producto final debe presentar información similar de una manera más dinámica: sobreponiendo los resultados en un mapa interactivo que muestre estos resultados a nivel departamental (Figura 1) y municipal (Figura 2), el cliente podrá visualizar zonas de influencia y de tendencia.

Además, se espera incluir una funcionalidad para seleccionar características relevantes de una población, y que el modelo de IA pueda generar determinar la

probabilidad de que un cambio en esa población aumente o reduzca la intención de voto por un candidato o partido específico.

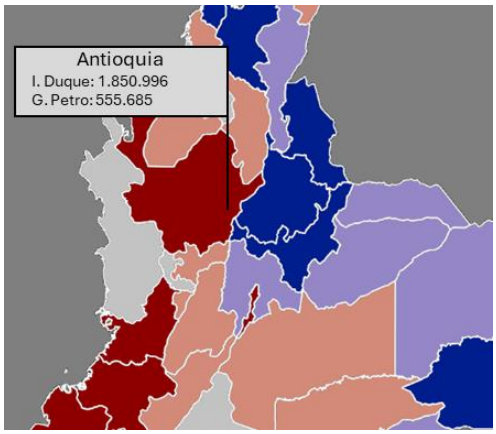


Figura 1: Prototipo de visualización de resultados a nivel departamental.

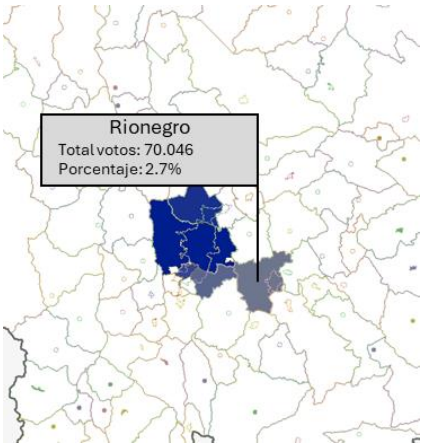


Figura 2: Prototipo de visualización de resultados a nivel municipal.

3. Responsible

Se sienta la base de que la Constitución Política de Colombia establece que el voto se debe ejercer en secreto, por lo que una primera restricción es que no se recoja, asigna o divulga algún dato personal que permita identificar a un individuo según sus votos.

Con respecto a la información disponible en las bases de datos dispuestas por Ingenial Media, ésta no presenta ningún riesgo de privacidad o confidencialidad al ser información respecto a la identidad de los candidatos, nombres de partidos inscritos a cada proceso, y los votos obtenidos por cada uno, que puede ser obtenida desde fuentes de la Registraduría Nacional del Estado Civil.

Los datos demográficos son provenientes de bases de datos de acceso público del Departamento Administrativo Nacional de Estadística (DANE), y se limitan a la presentación de datos colectivos relacionados a distribución demográfica, desarrollo socioeconómico, educación y salud.

Finalmente, la aplicación de modelos de IA sobre estos datos no debe representar ningún riesgo ético, regulatorio, o de privacidad, dado que no se ingresarán ni utilizarán datos personales para entrenar, o validar el desempeño del modelo.

4. Enfoque analítico

Respondiendo a la necesidad de Ingenial Media, la solución propuesta consta de dos partes: visualización de datos, y modelos estadísticos para encontrar *insights* valiosos al momento de organizar las campañas de los partidos políticos

Para la visualización de datos se propone una interfaz gráfica que muestre el mapa de Colombia dividido por departamentos y a su vez estos divididos en municipios (sección **2. Ideación**, Figuras 1 y 2). A éstos se sobrepondrá la información respectiva de los votos para esta zona, de acuerdo con los diversos procesos electorales de los cuales se dispone información (Congreso, Concejo, Alcaldía, Gobernación y Presidenciales en los años 2015, 2018, 2019 y 2022):

Asimismo, se mostrará la información sociodemográfica relevante de dichas zonas, lo cual permitirá al cliente ver la información segmentada que permita tomar decisiones informadas durante las campañas políticas y visualizar tendencias, buscando comprender mejor los comportamientos electorales

El modelo estadístico propuesto para encontrar *insights* valiosos en las campañas políticas es el de regresión logística binaria y multinomial, utilizando información sociodemográfica por departamento y municipio. Esto permitirá generar y comprobar hipótesis acerca de cómo el cambio de diferentes condiciones afecta la probabilidad de votar por cierto partido. En este orden de ideas, el análisis para este estudio deberá hacerse por partido, dividiendo la selección entre votar o no por cada uno, y realizando una regresión logística binomial. Para esto se propone utilizar los 5 partidos más representativos del país, siendo los que cuenten con mayor número de votos.

El cálculo de la regresión logística multinomial permitirá obtener información similar acerca de la importancia de cada una de las variables sociodemográficas sobre la decisión de votos a cada uno de los partidos, no obstante, también nos entrega *insights* clave sobre las relaciones o interacciones que tiene cada una de las variables entre ellas. De igual manera, facilita la obtención de información sobre la significancia de las mismas en el comportamiento de una variable dependiente, así que se descubrirá qué tan significante son estas variables sociodemográficas sobre la votación de algunos partidos.

Como resultados de estas regresiones, se obtendrán los Odd Ratio, una medida que indica cuánto cambia la probabilidad de un evento en relación con un cambio en una de las características independientes, algo particularmente importante a la hora de dirigir una campaña política

5. Recolección de datos

Los datos son provistos por Alianza CAOBA y Ingenial Media. Estos se encuentran divididos en 7 bases de datos, de las cuales 3 corresponde a "*templates*", de modo que no tienen información relevante. Las tablas usadas son: ***data_candidatos, data_divipol, data_votacion*** y ***partidos_2022***

Tabla	**Descripción:**
data_candidatos (267035, 15)	Registro histórico de candidatos que han participado en 18 elecciones. Incluye información como el número de cédula, nombre, apellido, código de partido, entre otros, y representan tanto a organizaciones políticas como a candidatos individuales.
data_divipol (85327, 19)	Puestos de votación habilitados para diversos tipos de elecciones, incluyendo detalles como su ubicación geográfica, el número de mesas disponibles, y votantes registrados en cada una.
data_votacion (45.487.984, 20)	Registro de las votaciones por mesas de todos los departamentos para los procesos electorales objeto de este proyecto. Las votaciones se discriminan por los votos obtenidos por cada uno de los candidatos y su partido en cada mesa, para cada departamento, para cada proceso electoral.
partidos_2022 (341, 3)	Información sobre los partidos políticos activos durante el año 2022. Incluye un indicador asignado a cada partido y una columna adicional llamada "Otro".

Por otro lado, es necesario realizar la recolección de datos sociodemográficos de los diferentes municipios o departamentos del territorio nacional, dado que se requiere de estos para encontrar oportunidades y patrones claros entre la elección de diversos partidos y diferentes condiciones sociodemográficas como pueden ser la educación, el ingreso, la cantidad de miembros familiares, la edad, el sexo, etc. Como se mencionó anteriormente, estos serán encontrados gracias al análisis de variables entre las condiciones (expuestas anteriormente) y la información de votación por sectores.

La información sociodemográfica será obtenida del DANE, acerca del censo de 2018, a través de la plataforma “REDATAM”. Esta permite hacer consultas de diferentes tipos de información del sector colombiano, como puede ser la distribución de edad, de sexo, educación, adultos mayores, tipos de vivienda, etc.; discriminado por diferentes niveles de territorio como departamentos o municipios

6. Entendimiento de los datos

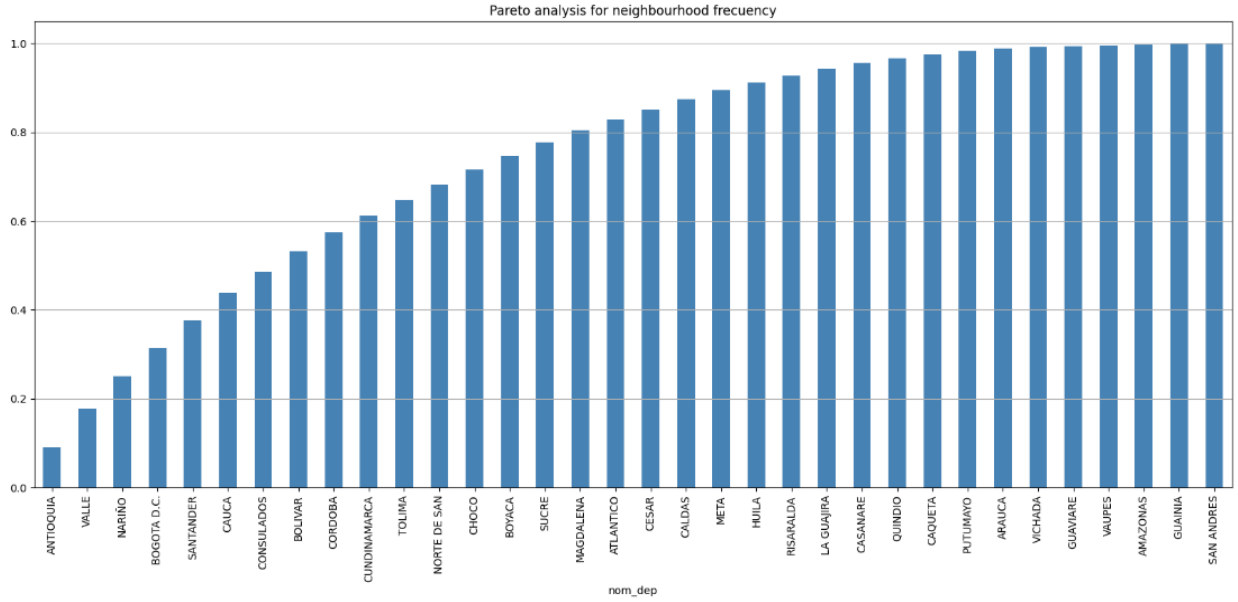
Dentro de las 4 tablas que son objeto de este análisis se encuentra la siguiente información respecto de las dimensiones y descripción:

6.1. Calidad de los datos.

- Duplicidad:
  - Se encontró duplicidad de los datos para la tabla partidos_2022 con 26 duplicados
- Completitud:
  - Los campos "apellido," "cedula," "genero," y "ganador" presentaron una falta de información que supera el 10% en tabla_candidatos
  - De data_divipol se excluyeron varias variables del análisis. Las variables "nom_comuna," "dir_puesto", "zonificado", "codigo_comuna", "latitud" y "longitud" al presentar más del 45% de registros nulos
- Relevancia:
  - En data_candidatos la variable "ganador" tiene la mayoría de los registros nulos (99.97%) y se considera no relevante
  - En data_candidatos la variable "cedula" también tiene una alta proporción de registros nulos, pero es relevante para identificar candidatos.
  - En data_votacion la variable "votos" tiene la información referente a la cantidad de votos registrados a cada candidato en cada mesa de votación.
- Conformidad:
  - En data_candidatos se identificó un problema de conformidad en la variable "nombres" debido a la presencia de estatus como "RETIRADO" o "REVOCADO", lo que llevó a la eliminación de 954 registros con esta problemática.
  - En partidos_2022 y data_votación la relación entre 'nom_partido' y 'cod_par' no es uno a uno. Esto quiere decir que hay diferentes códigos de partido asignados a un mismo partido.
  - Las tablas de Ingenial Media y las extraídas de la DANE identifican los departamentos y municipios con nombres y códigos diferentes. Estos datos deberán ser estandarizados manualmente.

6.2. Información clave:

Posterior a un análisis de Pareto sobre la cantidad de mesas de votación, encontramos que los departamentos que muestran más del 80% de las mismas son Bogotá, Antioquia, Valle, Cundinamarca, Atlántico, Santander, Bolívar, Córdoba, Nariño, Norte de Santander, Tolima, Cauca, Boyacá y Magdalena



7. Primeras conclusiones, *insights* y acciones próximas a ser ejecutadas.

Para obtener una evaluación más precisa, ya que presentan novedades y particularidades, se evaluarán las tablas proporcionadas. Es esencial tener una visión completa para identificar las variables relevantes y garantizar la consistencia de la información. Un ejemplo concreto de esto se observa en la tabla "data_candidatos", que requiere un análisis conjunto con la tabla "data_votacion" para determinar los registros válidos para las elecciones, es decir, aquellos candidatos elegibles. Tras el análisis realizado, se identificaron inconsistencias, como la presencia de candidatos registrados en más de un partido político para la misma elección, así como estados transaccionales como "RETIRADO" o "REVOCADO". Estas observaciones subrayan la importancia de considerar ambas tablas en conjunto para obtener una comprensión precisa de los datos.

Considerando lo anterior, se debe llevar a cabo una segmentación de la tabla data_votación dado que esta presenta la mayor cantidad de información y su análisis se torna complicado con los recursos tecnológicos con los que se cuenta actualmente esto. Dado esto se propone segmentar la base de datos data_votación por tipo de elección (Ej.: Asamblea – 2015, Alcaldía – 2019, etc.) y departamento (Ej.: Bogotá D.C., Córdoba, Magdalena, etc.).

Por otro lado, dentro de las primeras acciones a ejecutar es la obtención de la información sociodemográfica producto de la DANE, la cual puede ser obtenida a través de REDATAM como se mencionó en la sección No 5. De aquí se deberá obtener información de la mayor cantidad de variables sociodemográficas posibles, como sexo, edad, estrato, educación, etc.