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

PROYECTO FINAL: DASHBOARD PARA VISUALIZACIÓN DE DATOS DE ELECCIONES POLÍTICASEN COLOMBIA, SEGMENTACIÓN DE POBLACIONES Y ESTUDIO DE TENDENCIAS SOCIODEMOGRAFICAS RELACIONADAS

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

1. Definición de problemática y entendimiento del negocio

En colaboración con Ingenial Media, una empresa comprometida con la misión de asegurar el éxito electoral de candidatos, partidos y movimientos políticos se está llevando a cabo este proyecto que se enfoca en abordar el desafío de proporcionar información precisa para una gestión efectiva y eficiente de las campañas políticas. Nuestra solución se fundamenta en la creación de un dashboard y la implementación de modelos de aprendizaje automático, lo que permite analizar cómo las variables demográficas influyen en las tendencias políticas de los votantes. Este enfoque posibilita la adaptación de discursos y estrategias de campaña a nivel municipal y departamental al comprender las variables que inciden en la tendencia política de la población, así como anticipar posibles escenarios que puedan afectar los resultados electorales. Con esta herramienta, se podrán gestionar las campañas de manera más eficiente, asignar recursos de forma adecuada y diseñar estrategias personalizadas. En resumen, nuestra solución busca reducir la incertidumbre en las campañas políticas y maximizar las oportunidades de éxito en las elecciones, a través de la generación de Insights valiosos para diversos partidos políticos, basándose en información sociodemográfica. Los objetivos definidos para abordar este desafío son:

- Realizar un análisis exploratorio de los datos proporcionados por Ingenial con el fin de comprender su estructura y contenido, identificando las variables sociodemográficas relevantes para el estudio de las tendencias políticas.
- Aplicar modelos de aprendizaje automático para clasificar y predecir las tendencias políticas basándose en las variables sociodemográficas, lo que permitirá comprender cómo un cambio en estas variables afecta la preferencia de los votantes.
- Desarrollar un Dashboard interactivo que visualice de manera efectiva los resultados obtenidos a través del modelo de machine learning de mayor precisión, facilitando a los usuarios la interpretación y exploración de la información.

Dado el gran número de factores y variables que pueden influir en los resultados electorales, es retador determinar una métrica que permita establecer el éxito del producto, sabiendo que el producto final debe ser la generación de Insights en términos de variables sociodemográficas para las posibles campañas. Teniendo esto en cuenta, se partirá de modelos de clasificación que permitan determinar la posible elección de un partido, y entender cuáles son las variables que más contribuyen a este resultado. Dado esto, el objetivo es lograr un resultado de al menos 70% en métricas como el F1, precisión y cobertura.

2. Ideación

Los usuarios finales de este producto son clientes de Ingenial, cuyo interés radica en obtener observaciones precisas sobre la intención de voto para guiar sus estrategias y acciones. Actualmente, estos clientes acceden a información visual que se enfoca

en tendencias de elecciones pasadas. Esta información resalta los departamentos y municipios más relevantes según la cantidad de votos. Entre las cifras presentadas están el total de votos, la proporción de votos obtenidos por partido, los municipios con más votos, el partido ganador, los votos en consulados y comparativas de resultados en segundas vueltas de elecciones específicas.

Para mejorar la experiencia del usuario, el producto final se centrará en presentar datos de manera dinámica. Esto se logrará al incorporar un mapa interactivo que mostrará los resultados de votaciones pasadas a nivel departamental (Figura 1) y municipal (Figura 2). Este enfoque permitirá a los clientes explorar información relevante sobre elecciones pasadas, así como las variables sociodemográficas pertinentes y su posible aporte a la elección de un cierto partido político. Además, la herramienta identificará cuáles de estas variables son las más relevantes.

En cuanto a los componentes tecnológicos, el producto comprenderá: Una interfaz de usuario en HTML que facilite la interacción y el filtrado de datos en el mapa y un modelo de machine learning (clasificación) integrado que permitirá obtener información crucial acerca del impacto de diferentes variables sociodemográficas a una votación.

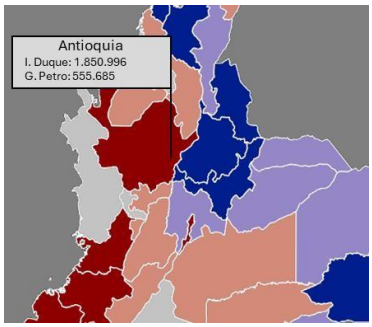

Figura 1: Prototipo de visualización de resultados anivel departamental.

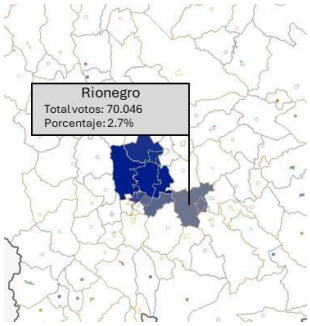

Figura 2: Prototipo de visualización de resultados a nivel municipal

3. Responsable

En la Constitución Política de Colombia establece el voto secreto como un principio fundamental. Por lo tanto, se impone la restricción de no recopilar, asignar o divulgar datos personales que permitan la identificación de un individuo en función de sus votos. En lo que respecta a la información contenida en las bases de datos proporcionadas por Ingenial Media, se debe destacar que esta información no entraña riesgos para la privacidad ni la confidencialidad, ya que se trata de datos públicos que incluyen la identidad de los candidatos, los nombres de los partidos inscritos en cada proceso y los votos obtenidos. Estos datos son accesibles a través de fuentes de la Registraduría Nacional del Estado Civil.

Por otra parte, los datos demográficos provienen de bases de datos de acceso público proporcionadas por el Departamento Administrativo Nacional de Estadística (DANE). Estos datos se limitan a la presentación de estadísticas colectivas relacionadas con la distribución demográfica, el desarrollo socioeconómico, la educación y la salud.

Finalmente, se destaca que la aplicación de modelos de machine learning a estos datos no conlleva ningún riesgo ético, regulatorio o de privacidad, ya que no se

recopilarán ni se utilizarán datos personales para el entrenamiento o la validación del rendimiento del modelo.

4. Enfoque analítico

El enfoque del proyecto será la elección presidencial en segunda vuelta del 2018; presenta una elección binaria y polarizada, y la consistencia con la información sociodemográfica obtenida por el DANE, atribuible al censo electoral de vivienda de 2018. Este proyecto podrá ampliarse a más procesos electorales en diferentes espectros de tiempo, si se tiene la información sociodemográfica respectiva.

Se propone una solución que consta de dos componentes esenciales: visualización de información y modelos de machine learning (clasificación). Estos elementos están diseñados para proporcionar información valiosa que permita orientar y planificar la estrategia de las campañas políticas para los clientes de Ingenial.

En términos de visualización de datos, se presentará una interfaz gráfica que mostrará un mapa de Colombia dividido en departamentos y, a su vez, subdividido en municipios (consulte la Sección 2: Ideación, Figuras 1 y 2). Esta interfaz superpondrá información detallada sobre los votos registrados en cada zona. Además, se exhibirá información sociodemográfica relevante correspondiente a esas áreas geográficas para identificar factores relevantes y patrones.

Los modelos estadísticos propuestos para descubrir información valiosa en el contexto de las campañas políticas se centrarán en la clasificación de tendencias políticas, particularmente entre opciones de derecha e izquierda, dada la polaridad que presenta el contexto colombiano. Se utilizarán tres enfoques analíticos principales: regresión logística, Random Forest y árbol de decisión, basados en información sociodemográfica a nivel de municipio. Estos modelos permitirán analizar hipótesis acerca de cómo las variaciones en diversas condiciones influyen en la probabilidad de que un votante elija una tendencia política específica.

El enfoque de regresión logística proporcionará información valiosa sobre la importancia de las variables sociodemográficas en la decisión de voto entre tendencias políticas. Además, ofrecerá insights clave sobre las interacciones y relaciones entre estas variables, lo que facilitará la comprensión de su influencia en la variable dependiente, es decir, la elección de una tendencia política específica.

Por otro lado, el modelo Random Forest y el árbol de decisión también se aplicarán para evaluar hipótesis y descubrir patrones relevantes en la elección de tendencias políticas, centrándose en la clasificación de derecha e izquierda. Estos enfoques proporcionarán información adicional y perspectivas útiles para el análisis de datos sociodemográficos y su influencia en las decisiones de voto.

Como resultado de estos análisis, se obtendrán medidas importantes, como los Odd Ratios, que indican cómo cambia la probabilidad de un evento en función de las variaciones en las variables independientes. Estas métricas son esenciales para la toma de decisiones estratégicas en el contexto de la dirección de campañas políticas.

5. Recolección de datos

Los datos proporcionados por Alianza CAOBA e Ingenial Media se dividen en 7 bases de datos, de las cuales 3 son consideradas como "templates" y no contienen información relevante. Las tablas utilizadas son las siguientes:

Tabla	**Descripción:**

data_candidatos (267035, 15)	Registro histórico de candidatos que han participado en 18 elecciones. Incluye número de cédula, nombre, apellido, código de partido, entre otros, y representan tanto a organizacionespolíticas 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, fue necesario recolectar datos sociodemográficos de municipios y departamentos del territorio nacional para identificar patrones entre la elección de partidos y variables sociodemográficas. Esto permitirá identificar patrones y oportunidades relacionados con la elección de partidos políticos y variables sociodemográficas como la educación, la composición familiar, la edad, el género, el acceso a servicios públicos o sanitarios, etnia, entre otros. La información sociodemográfica fue obtenida del DANE, basándonos en el censo de 2018,a través de la plataforma “REDATAM”, como se evidencia en el siguiente enlace: http://systema59.dane.gov.co/bincol/RpWebEngine.exe/Portal?BASE=CNPVBASE4V2&lang=esp

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:

Después de realizar un análisis de Pareto en cuanto a la cantidad de mesas de votación, identificamos que los departamentos que representan más del 80% de estas mesas son: Bogotá, Antioquia, Valle, Cundinamarca, Atlántico, Santander, Bolívar, Córdoba, Nariño, Norte de Santander, Tolima, Cauca, Boyacá y Magdalena, como se videncia en la figura 3

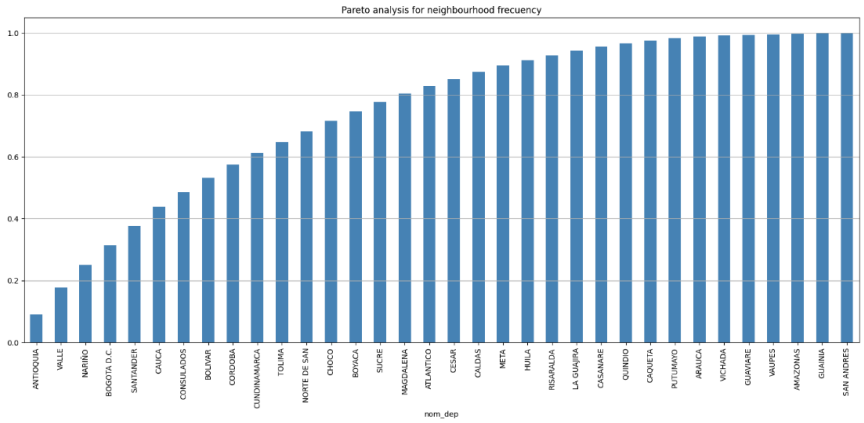

Figura 3: Analisis de Pareto, mesas de votación

Posterior a la revisión de las 4 tablas presentadas por Ingenial, se determina que la base de datos “data_votaciones” presenta la información más importante, de modo que esta la utilizada en la siguiente fase de entrenamiento.

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

Para realizar una evaluación precisa, especialmente considerando las novedades y particularidades presentes, es fundamental analizar detenidamente las tablas proporcionadas. Esto nos permitirá obtener una visión completa, identificar las variables relevantes y asegurar la consistencia de la información. Un ejemplo concreto de este enfoque se refleja en la tabla "data_candidatos", que requiere un análisis conjunto con la tabla "data_votacion" para determinar los registros válidos en las elecciones, es decir, aquellos candidatos elegibles. Durante este análisis, 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. En vista de lo anterior, es necesario segmentar la tabla "data_votacion" debido a su volumen y la complejidad que presenta su análisis con los recursos tecnológicos disponibles. Por lo tanto, se propone segmentar la base de datos "data_votacion" por tipo de elección (por ejemplo, Asamblea – 2015, Alcaldía – 2019, etc.) y por departamento (por ejemplo, Bogotá D.C., Córdoba, Magdalena, etc.).

8. Preparación de datos:

En un primer momento, nos enfocaremos en el proceso electoral de la segunda vuelta presidencial del año 2018 para entrenar nuestros modelos. Este proceso electoral se destaca por ser una decisión de voto binaria, lo que simplifica la asignación de etiquetas y facilita la observación del impacto de las variables sociodemográficas en un conjunto de opciones más reducido. Como en estas elecciones participaron dos candidatos, uno de derecha y otro de izquierda, esta elección es relevante para comenzar, considerando la polarización política actual en Colombia. Además, es importante señalar que la información sociodemográfica recopilada por el DANE corresponde específicamente a este año, lo que garantiza una alta coherencia entre las variables sociodemográficas y la información electoral.

Basándonos en el punto 6, de las cuatro tablas disponibles, la que se considera más relevante para el análisis es "data_votaciones". Esta tabla destaca por su mayor consistencia en la información presentada, que proporciona el número de votos reportados en cada uno de los puestos de votación, abarcando todos los municipios del país y en cada proceso electoral mencionado. Es así como, el producto desarrollado para este proyecto se basa en dos conjuntos de bases de datos: aquellas proporcionadas por Ingenial, que contienen información electoral, y datos sociodemográficos obtenidos del DANE.

Con respecto a Ingenial (información electoral). Inicialmente, incorporamos la información provista por Ingenial sobre elecciones a dos niveles de profundidad: departamental y municipal, generando dos bases de datos. La primera de ellas se diseñó específicamente para satisfacer los requisitos de visualización en el Panel de Control, concentrándose en los datos departamentales. La segunda base de datos resolvió el desafío de disponer de solo 32 departamentos en Colombia para el entrenamiento de nuestros modelos. Al extender nuestro enfoque al nivel municipal, abarcando los 1102 municipios del país, obtuvimos una cantidad considerablemente mayor de registros para nutrir el entrenamiento de los modelos. Este enfoque fortalece la capacidad de los modelos para detectar patrones y tendencias, lo que se traducirá en análisis y predicciones de mayor precisión. Ambas tablas contienen información sobre el total de votos, votos nulos, votos en blanco, votos no válidos, votos válidos y la proporción de cada partido político en el proceso electoral. En la segunda tabla, se asigna una etiqueta "Izquierda" o "Derecha" a cada municipio en función del partido ganador: "Coalición Petro presidente" representado por Gustavo Petro o "Centro Democrático" representado por Iván Duque.

Para el aspecto sociodemográfico, contamos con datos del censo de vivienda del DANE en 2018. Estos datos están distribuidos en cuatro bases de datos que contienen información sobre viviendas, hogares y personas. Hemos extraído datos de interés tanto a nivel departamental para su visualización en el Panel de Control como a nivel municipal para el entrenamiento de nuestros modelos de clasificación. Estos datos se centran en 72 variables de interés que abarcan categorías como Alfabetismo, Edad, Educación, Estado Civil, Estrato, Etnia, Sanitario, Servicios, Sexo y Tipo de Hogar; variables que son de interés en algunos estudios sobre la elección de partidos de derecha o izquierda en Latinoamérica [1]. Es importante destacar que los datos del DANE requerían ciertas transformaciones para adecuarlos a un formato de archivo CSV compatible con Python.

Estas transformaciones incluyeron la eliminación de formatos no legibles, la corrección de celdas combinadas y la normalización de nombres de variables. Dado que trabajaremos con información a nivel municipal, hemos normalizado cada variable en función de la población total de cada municipio, para evitar que el tamaño de la población influya en nuestras mediciones de tendencia.

Finalmente, la información sociodemográfica se incorpora a la tabla de votaciones municipales para entrenar nuestros modelos de clasificación. Antes de iniciar el proceso de entrenamiento y pruebas, reducimos el número de variables de 76 a 52, eliminando aquellas que presentan una alta colinealidad, lo que podría haber llevado a la multicolinealidad en nuestros modelos, afectando su capacidad de predicción, así como aumentar la complejidad de los modelos sin agregar información adicional significativa. Además, aplicamos la estrategia de Kolmogórov-Smirnov, utilizando el percentil del 10% como referencia, lo que equivale a un valor de 0.109. Todas las variables inferiores a este valor se excluyen del conjunto de datos usado para los modelos de clasificación, lo que nos permitió conservar solo las variables que tenían un mayor poder discriminativo en el contexto de las tendencias políticas en Colombia. El conjunto final consta de 46 variables sociodemográficas para los 1102 municipios. El proceso de transformación de datos se puede apreciar en la Figura 4.

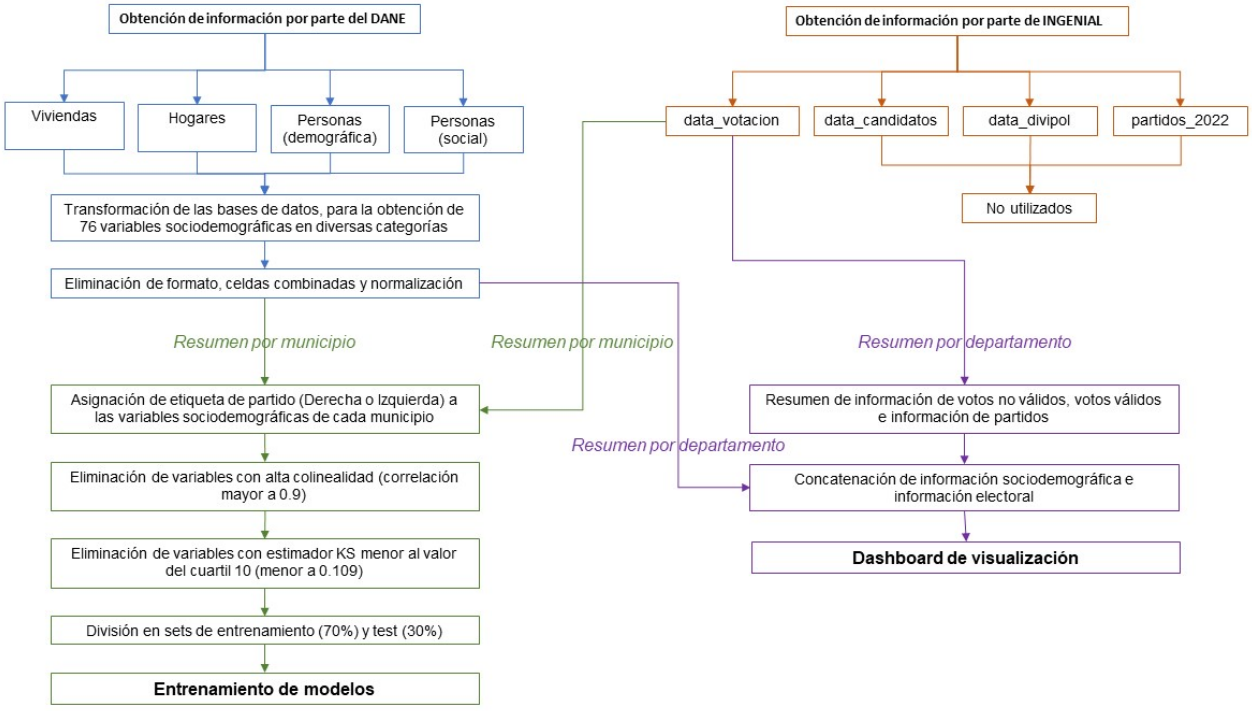

Figura 4. Esquema de transformación de datos

3. Estrategia de validación y selección de modelo:

Considerando los objetivos planteados, hemos determinado una estrategia analítica que involucra la utilización de dos componentes fundamentales propios de modelos de clasificación: los *Odds Ratios* y los *Shap Values*. Los Odds Ratios permiten cuantificar las relaciones entre las variables sociodemográficas y las elecciones, como se detalla en nuestra sección de enfoque analítico. Los Shap Values, por otro lado, ofrecen información esencial sobre la contribución de cada variable a resultados de clasificación específicos. Un aspecto relevante es que los Shap Values pueden aplicarse a los tres modelos de clasificación sin depender de uno en particular, a diferencia de los Odds Ratios. Por lo tanto, procedimos a entrenar y evaluar tres

modelos de clasificación distintos: Random Forest, Regresión Logística (utilizando Odds Ratios) y Árbol de Decisión.

Elegiremos el mejor modelo considerando métricas clave como F1, precisión y cobertura. Optimizamos Random Forest y Regresión Logística mediante una búsqueda exhaustiva de hiperparámetros con GridSearchCV. Para el Árbol de Decisión, el modelo define las variables pertinentes automáticamente.

Para realizar este proceso de entrenamiento, prueba y selección de modelos, dividimos el conjunto de datos construido en dos partes principales: un conjunto de entrenamiento, que comprende el 70% de los datos, y un conjunto de prueba, que representa el 30% restante. Mantuvimos la proporción de etiquetas de interés (Izquierda y Derecha) en ambos conjuntos, asegurando que la distribución se conserve, como se ilustra en la figura 5:

Set de entrenamiento:

	ET_NOETNIA_Norm	EDAD_20-24_Norm	EST_Viudo_Norm	ET_Indigena_Norm	TH_Pareja_mayor_sin_hijos_Norm	EDAD_55-59_Norm	ED_Primaria_Completa_Norm	EDAD_50-54_Norm	EDAD_25-29_Norm	EDAD_60_o_mas
count	785.000000	785.000000	785.000000	785.000000	785.000000	785.000000	785.000000	785.000000	785.000000	785.000000
mean	0.836235	0.080522	0.043405	0.076251	0.141855	0.049782	0.151935	0.055799	0.072092	0.143417
std	0.291581	0.013418	0.014294	0.195997	0.050131	0.011760	0.034725	0.010827	0.011378	0.051184
min	0.002100	0.038900	0.008100	0.000000	0.026000	0.010600	0.072000	0.020700	0.027800	0.026100
25%	0.858500	0.073000	0.033400	0.000300	0.103900	0.041500	0.126200	0.049200	0.064900	0.105600
50%	0.981500	0.081700	0.042000	0.001400	0.132800	0.049600	0.152000	0.056500	0.073500	0.139000
75%	0.991000	0.088600	0.052500	0.028300	0.171400	0.057900	0.175900	0.064200	0.080100	0.174200
max	0.999100	0.250800	0.110200	0.985800	0.332600	0.088700	0.281700	0.082900	0.104100	0.333900

Set de test:

	ET_NOETNIA_Norm	EDAD_20-24_Norm	EST_Viudo_Norm	ET_Indigena_Norm	TH_Pareja_mayor_sin_hijos_Norm	EDAD_55-59_Norm	ED_Primaria_Completa_Norm	EDAD_50-54_Norm	EDAD_25-29_Norm	EDAD_60_o_mas
count	337.000000	337.000000	337.000000	337.000000	337.000000	337.000000	337.000000	337.000000	337.000000	337.000000
mean	0.860950	0.079259	0.044769	0.070499	0.145384	0.051057	0.149891	0.057093	0.071358	0.147964
std	0.267206	0.012245	0.014905	0.186388	0.050793	0.011483	0.032721	0.010999	0.011245	0.051194
min	0.002400	0.038200	0.006000	0.000000	0.036400	0.013500	0.063400	0.018000	0.036900	0.048200
25%	0.918200	0.071800	0.035100	0.000300	0.105700	0.043500	0.127700	0.050700	0.064800	0.112100
50%	0.982800	0.081000	0.043500	0.001400	0.137100	0.050600	0.150000	0.058500	0.072400	0.141000
75%	0.991800	0.087000	0.053700	0.024100	0.177400	0.059700	0.172000	0.065200	0.078300	0.187200
max	0.998600	0.119200	0.091400	0.976500	0.295500	0.077400	0.277300	0.084400	0.101700	0.316000

Set de labels de entrenamiento:

```
Valores unicos en el set de entrenamiento

tendencia
Derecha      0.756688
Izquierda    0.243312
Name: proportion, dtype: float64
```

Set de labels de test:

```
Valores unicos en el set de test

tendencia
Derecha      0.756677
Izquierda    0.243323
Name: proportion, dtype: float64
```

Figura 5: Distribución test y entrenamiento, para sets de variables y de labels

10. Construcción del modelo:

Como se mencionó en el literal anterior, los modelos a utilizar son Random Forest, Regresión Logística y Árbol de decisión. Los hiperparametros para los primeros dos modelos fueron encontrados a través de una búsqueda por GridSearch, utilizando la métrica de F1 como el parámetro de interés. Sin embargo, el Árbol de Decisión es un caso particular, ya que su estructura se basa en su profundidad máxima y el criterio de división. Para encontrar los mejores hiperparámetros para el Árbol de Decisión, implementamos un bucle que evaluó diferentes combinaciones de profundidad máxima y criterio de división. Usamos la validación cruzada para evaluar el rendimiento de cada configuración. El conjunto de hiperparámetros en los cuales se llevó a cabo la búsqueda y los valores encontrados, así como las métricas que respaldan esta decisión, se pueden evidenciar claramente en la tabla 1 (el hiperparámetro óptimo se resalta en **negrita**):

Tabla 1: Valores de rangos hiperparametros por cada modelo de clasificación

Modelo	Hiper Parámetro	Valores de búsqueda	Métricas
Random Forest	Numero de estimadores	5, 10, 50, **100**	F1: **0.96** Precisión: **1.00** Cobertura: **0.92**
	Profundidad máxima	**None**, 10, 20, 30	
	Mínimo número de muestras para *split*	2, 5, **10**	
	Mínimo número de muestras en *leaf*	**1**, 2, 4	
Regresión Logística	Variable de regularización	0.1, 1, 10, **100**	F1: **0.62** Precisión: **0.76** Cobertura: **0.52**
	Tipo de penalidad	**L1**, L2	
Árbol de decisión	Profundidad máxima	None, 3, **5**, 7 y 10	F1: **0.65** Precisión: **0.68** Cobertura: **0.62**
	Criterios	"**Gini**" y "Entropy"	

11. Evaluación del modelo:

A partir de estos resultados, podemos concluir que el modelo de Árbol de Decisión tiene un desempeño inferior en términos de precisión, cobertura y F1 en comparación con los otros dos modelos. Esto lo coloca como la opción menos adecuada para nuestro producto de clasificación.

Por otro lado, los valores de regresión logística presentan mejor desempeño que el valor de árbol de decisión, sin embargo, está muy por debajo de los valores encontrados para Random Forest, que se hallan todos por encima del 90%. Cabe señalar que el modelo de regresión logística es de particular interés, dado que su propio fundamento matemático, permite el cálculo de Odd Ratios, para las variables sociodemográficas.

Se recomienda utilizar el modelo Random Forest debido a su sobresaliente desempeño en términos de precisión, cobertura y F1, con todas las métricas superando el 90%. Además, se sugiere emplear los valores de SHAP del modelo Random Forest para analizar las contribuciones de las variables sociodemográficas en las elecciones. Se recomienda comparar estos SHAP values con los de la Regresión Logística para evaluar posibles diferencias significativas. Esta comparación permitirá determinar si la información generada por el modelo de Regresión Logística, en términos de Odds Ratios, puede mejorar la interpretabilidad del modelo y la comprensión de las variables sociodemográficas influyentes.

12. Conclusiones:

Durante el transcurso de este sprint, hemos logrado importantes avances en la identificación y procesamiento de variables sociodemográficas obtenidas del DANE. Hemos extraído un conjunto inicial de 76 variables que abarcan categorías fundamentales, como Alfabetismo, Edad, Educación, Estado Civil, Estrato, Etnia, Sanitario, Servicios, Sexo y Tipo de Hogar. Estas variables se han resumido tanto a nivel municipal como departamental, para el entrenamiento de los tres modelos de clasificación y el Dashboard de visualización respectivamente.

Uno de los desafíos significativos que enfrentamos durante este sprint fue la

disponibilidad limitada de datos de Ingenial. La plataforma de origen no permitía un acceso continuo, lo que nos llevó a enfocarnos en los datos específicos de la segunda vuelta presidencial de 2018. A pesar de esta limitación, logramos ajustar nuestra estrategia para hacer un uso eficiente de los datos disponibles.

Otro desafío actual es la manera en que los polígonos interactivos se generan en el mapa del país, ya que el programa le asigna un ID arbitrario a cada polígono, agnóstico a los datos con los que está asociado el departamento. Este problema se amplifica cuando tratemos de visualizar municipios, por lo que una posible manera de mitigar este problema sea buscar en qué momento se asignan estos IDs, para poder sobrescribirlos con valores que coincidan con las bases de datos.

Por otro lado, los modelos de clasificación mostraron resultados prometedores. Sin embargo, el enfoque del proyecto es la generación de Insights, más que un estimador de clasificación, de modo que la interpretabilidad de los shap values y Odd Ratios del modelo seleccionado es crucial para este proyecto, por lo que se debe generar una discusión que permita determinar qué modelo puede ser más útil, considerando la interpretabilidad requerida. En lo que respecta a los datos, hemos logrado avances significativos, pero cabe destacar que existe la posibilidad de incorporar información adicional de variables sociodemográficas y evaluar si esto agrega valor a nuestro proyecto. Esta es una oportunidad para futuros trabajos que podría enriquecer aún más nuestro análisis.

Bibliography

[1] M. A. Torrico Terán y D. Solís Delgadillo, «Voto ideológico, ¿por qué los latinoamericanos votan por la izquierda o la derecha?,» *Redalyc,Org,* 2019.