Trabajo Final del Curso

Sistemas de Información y Bases de Datos

Maestría en Generación y Análisis de Información Estadística

UNTREF

Estudiante: Manuel Miller

Título del Trabajo Final: Clasificar y agrupar: Machine Learning aplicado a un caso de encuesta de opinión política en Argentina.

Marzo 2025

1. Objetivos

La autopercepción es un concepto ampliamente estudiado en la psicología y la sociología. En este trabajo me propongo realizar sociología política basada en evidencia en búsqueda de ampliar nuestro conocimiento sobre la forma en la que las identidades políticas se aglutinan en la argentina actual. guía este trabajo la presunción de un alto nivel de solapamiento en estas "fronteras identitarias" que tantas veces son el génesis de encarnecidas discusiones en la opinión publica.

El **objetivo general** consiste en evaluar la capacidad de distintos modelos predictivos de Machine Learning para predecir la identificación política de los encuestados, en base a preguntas diseñadas específicamente para actuar como clivajes identitarios.

Esto trae aparejados los siguientes objetivos específicos:

- Realizar un análisis exploratorio de los resultados obtenidos para identificar patrones entre variables clave.
- Aplicar un modelo de regresión logística para evaluar su eficacia en la predicción de la identificación política del encuestado
- Entrenar el algoritmo XGBoost con el mismo objetivo para comparar su performance frente a la regresión logística
- Obtener Clusters "Naturales" en base al patrón de respuestas de los encuestados utilizando el algoritmo K-means para contrastar estos resultados con sus autopercepciones.

2. Metodología desarrollada

2.1. Herramientas utilizadas

Para el desarrollo del presente trabajo, se emplearon diversas herramientas digitales. La recolección de datos se realizó mediante Google Forms [1]. Para la limpieza de los datos crudos se utilizó R [2], y RStudio [3] sirvió como entorno de desarrollo para su implementación. El procesamiento y análisis de datos fueron llevados a cabo en Python [4] en el entorno de ejecución Google Colaboratory [5].

2.2. Origen y estructura de los datos utilizados

Se realizó una encuesta auto administrada por medio de Google Forms, la cual obtuvo 134 respuestas. La misma relevó desde ciertos hábitos cotidianos, consumos y características socio-demográficas, hasta el sentimiento frente a determinados países o posicionamientos político-culturales acera de ciertos "dilemas.º debates que candidatos a dividir el campo de posicionamiento ideológico argentino. Se limpiaron los datos en R generando 3 revisiones de dataframes distintas para los mismos datos:

Cuadro 1: Archivos Generados

Nombre	**Descripción**	**Finalidad**
134NODUMMY.xlsx	Las variables se presentan en su nivel de medición original	Análisis descriptivo
D134.xlsx	Ciertas variables categóricas se presentan como DUMMY eliminando una de sus modalidades para evitar colinealidad	Modelos de clasificación
kmeans134.xlsx	Igual a D134.xlsx pero sin borrar modalidades	Modelos de clusterización

A continuación, se presenta la estructura del dataframe D134

Cuadro 2: Tabla de Atributos de la Encuesta

ATRIBUTO	TIPO DATO	RANGO	DESCRIPCIÓN
HIJOS	bool	0-1	Tiene hijos o no
N_HIJOS_HOY	int	0-3	Numero de hijos que tiene actualmente
BUSCA_HIJOS	bool	0-1	Piensa tener (más) hijos o no
N_BUSCA_HIJOS	bool	0-1	Cuántos hijos (más) piensa tener
TH_EDAD	int	19-37	A quienes tuvieron hijos ¿A qué edad tuvo el primero?
UP	bool	0-1	¿La universidad pública debe ser gratuita?
MASCOTHIJOS	bool	0-1	¿Una mascota es un hijo?
GORRA	bool	0-1	Acuerdo con la frase "Muerte a la gorra"
EF	bool	0-1	Acuerdo con la frase "Necesitamos más educación financiera"
EJERCITO	bool	0-1	Acuerdo con la frase " El ejército es necesario"
TARIFAS	bool	0-1	Acuerdo con la frase " Está bien que aumenten los servicios y el transporte. Pagábamos muy poco."
EMPRESARIOS	bool	0-1	Acuerdo con la frase "Los empresarios son el problema"
MALVINAS	bool	0-1	Los ex combatientes son (1) Héroes (0) Victimas
EF_QUEES	bool	0-1	La educación financiera tiene más que ver con (1) Saber comprar y vender acciones, criptomonedas o bonos (0) Saber ahorrar y no endeudarse
EL_PROBLEMA	bool	0-1	El problema de Argentina es que: (1) Hay pocos que tienen mucho dinero(0) Hay mucha gente que no piensa

Continúa en la siguiente página

ATRIBUTO	TIPO DATO	RANGO	DESCRIPCIÓN
DESIGUALDAD	bool	0-1	esos pocos con mucho dinero son sobretodo: (0)Gente que evade impuestos y contrata en negro (1) Políticos que se la robaron
NO_PIENSAN	bool	0-1	Esa gente que no piensa, por lo general: (0) Tiene menos educación o dinero que yo (1) Tiene mas educación o dinero que yo
EEUU	int	-2,2	¿Qué opinión te generan los siguientes países? Rechazo - Negativa - Indiferencia/Desconocimiento - Positiva - Cariño
PALESTINA	int	-2,2	Mismas preguntas y categorías que EEUU
ISRAEL	int	-2,2	Mismas preguntas y categorías que EEUU
UCRANIA	int	-2,2	Mismas preguntas y categorías que EEUU
RUSIA	int	-2,2	Mismas preguntas y categorías que EEUU
BOLIVIA	int	-2,2	Mismas preguntas y categorías que EEUU
CHINA	int	-2,2	Mismas preguntas y categorías que EEUU
INGLATERRA	int	-2,2	Mismas preguntas y categorías que EEUU
GENERO	bool	0-1	Género del encuestado (bool porque nadie marcó “otros”)
EDAD	int	10-78	Edad del encuestado
ESTUDIO	int	1-4	Máximo nivel de estudio alcanzado
SOCIECON	int	0-2	clase autopercibida
TRABAJA	bool	0-1	Trabaja o no actualmente
MILEI	int	0-4	Nivel de satisfacción con el actual gobierno
ETIQUETA	string		Peronista - Radical - De derecha - De izquierda - Liberal - Apolítico - No sabe/No contesta
PROLE	DUMMY	0-1	Se considera clase trabajadora o no
SS	DUMMY	0-1	Situación sentimental

Continúa en la siguiente página

ATRIBUTO	TIPO DATO	RANGO	DESCRIPCIÓN
PROGRAMA	DUMMY	0-1	Que programa preferiría ver dentro de un set definido de opciones
NOTICIAS	DUMMY	0-1	Que noticiero preferiría ver dentro de un set definido de opciones
GRUPO	DUMMY	0-1	Se considera empresario, Emprendedor, Empleado, profesional u otro
NO_HIJOS_PQ	DUMMY	0-1	Razón que mas lo detiene de tener hijos

En el desarrollo del código puede encontrarse el análisis descriptivo en el cual se estudió:

- La relación entre el consumo de ciertos medios de información y la opinión de los encuestados frente a ciertos países.
- La convivencia entre etiquetas políticas y noticieros consumidos.
- Las diferencias significativas halladas en la opinión sobre ciertos países y la etiqueta política del encuestado.
- La relación observada entre el deseo de no tener hijos y la etiqueta política.

2.3. Descripción de Metodología utilizada

Luego de realizar One Hot Encoding en R y codificar las variables, los 3 archivos son procesados en Python donde se filtran las respuestas que representan ruido en los datos y se realiza un análisis exploratorio. Comencé probando el algoritmo XGBoost con todas las variables presentes y comencé a eliminar variables de menor ganancia de información. Luego eliminé variables por su baja relación teórica con el problema. La eficacia en la predicción de clases no estaba siendo la esperada.

En ese momento logré verificar que la sobre-representación de peronistas sesgaba el modelo, volviendo a esta clase la más fácil de predecir, mientras había muy pocos casos para predecir otras. En ese momento, elegí probar un algoritmo de regresión logística, que se verificó como el

método más eficaz, logrando una precisión del 53 % entrenando al modelo con el 60 % de los datos y un 60 % de precisión al entrenarlo con el 70 % de los datos. Se utilizó Recursive Feature Elimination y K-folds para obtener el mejor de los modelos alcanzados el cual logra predecir la identificacíon política del encuestado con tan solo 19 variables.

3. Resultados encontrados

3.1. Modelo de regresión logistica

Se presenta a continuación las 19 mejores variables encontradas por el modelo para predecir la identificación política ordenas por sus coeficientes de regresión logística, las cuales corresponden a tan solo 14 preguntas, siendo algunas variables dummy una modalidad especifica de una variable categórica original.

Cuadro 3: Importancia promedio de las variables seleccionadas (Ranking 1)

Variable	**Importancia**
TARIFAS	0.633160
GORRA	0.519790
MASCOTHIJO	0.504816
CHINA	0.485430
EJERCITO	0.446338
EMPRESARIOS	0.415394
PROGRAMA_Un programa o video de ciencia o historia	0.407701
GENERO	0.405529
EF_QUEES	0.403795
PALESTINA	0.400569
EL_PROBLEMA	0.381331
BOLIVIA	0.335578
Continúa en la siguiente página	

Variable	**Importancia**
NOTICIAS_Prefiero no ver nada	0.332024
TRABAJA	0.319705
MALVINAS	0.310580
NOTICIAS_Cronica	0.307757
EF	0.306920
EEUU	0.291156
NOTICIAS_TN	0.278714

A partir de dichas variables, se obtiene esta matriz de confusión al entrenar al modelo con el 70 % de los datos:

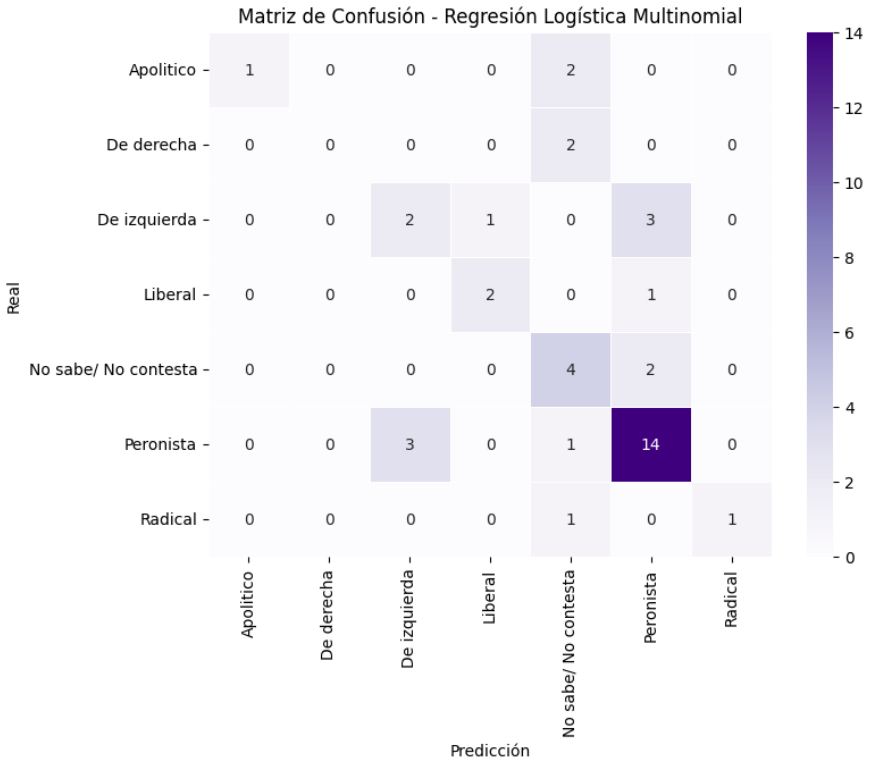

Figura 1: Matriz de confusión del modelo L3

A continuación se muestran las métricas del modelo:

Cuadro 4: Métricas del Modelo de Clasificación

Clase	**Precisión**	**Recall**	**F1-Score**
Apolítico	1.00	0.33	0.50
De Derecha	0.00	0.00	0.00
De Izquierda	0.40	0.33	0.36
Liberal	0.67	0.67	0.67
No sabe/No contesta	0.40	0.67	0.50
Peronista	0.70	0.78	0.74
Radical	1.00	0.50	0.67
Exactitud (Accuracy)		**60.00 %**	
Macro Promedio	0.60	0.47	0.49
Promedio Ponderado	0.61	0.60	0.58

Los *odds ratios* permiten interpretar cómo la presencia de una variable influye en la probabilidad de pertenecer a una categoría política determinada en comparación con una categoría de referencia.

3.1.1. Variables con mayor impacto en cada grupo político (>1.5)

Clase "De Izquierda"

- **GORRA (3.50):** Estar de acuerdo con la frase "Muerte a la gorra" aumenta significativamente la probabilidad de identificarse con la izquierda.
- **PALESTINA (2.70):** Tener una opinión positiva sobre Palestina está fuertemente relacionada con la izquierda.
- **EL_PROBLEMA (2.50):** Creer que el problema de Argentina es la concentración de riqueza (y no la falta de pensamiento) es un fuerte predictor de identificación con la izquierda.

Clase "De Derecha"

- **MALVINAS (1.77):** Considerar a los excombatientes como héroes se asocia con la identidad de derecha.
- **EJERCITO (1.62):** Apoyar la frase "El ejército es necesario" está positivamente relacionado con la identidad de derecha.
- **EMPRESARIOS (1.61):** Creer que "Los empresarios no son el problema" aumenta la probabilidad de identificarse con la derecha.
- **NOTICIAS_Prefiero no ver nada (1.45):** No consumir noticias con frecuencia está relacionado con la identidad de derecha.

Clase "Liberal"

- **TARIFAS (3.20):** Apoyar aumentos en servicios públicos y transporte es un factor clave de identificación liberal.
- **NOTICIAS_Cronica (2.40):** Preferir consumir *Crónica* como fuente de noticias está altamente asociado con el liberalismo.

- **GENERO (2.10):** Ser hombre aumenta la probabilidad de identificarse como liberal.

Clase "Peronista"

- **BOLIVIA (2.30):** Tener una visión positiva sobre Bolivia se correlaciona con una mayor probabilidad de identificarse como peronista.
- **CHINA (1.90):** Tener una opinión positiva sobre China es un fuerte predictor de identificación peronista.
- **EF_QUEES (1.79):** Creer que la educación financiera está más ligada al ahorro y la administración del dinero que a inversiones y criptomonedas es un indicador fuerte del peronismo.

Clase Radical"

- **MASCOTHIJO (2.28):** Considerar que una mascota es equivalente a un hijo está fuertemente asociado con la identidad radical.

3.1.2. Variables con menor impacto en ciertas clases (<0.5)

Clase "De Izquierda"

- **EJERCITO (0.48):** Apoyar la necesidad del ejército se asocia con una menor probabilidad de ser de izquierda.
- **EEUU (0.53):** Tener una visión positiva sobre Estados Unidos disminuye la probabilidad de identificarse como de izquierda.

Clase "De Derecha"

- **GORRA (0.65):** Tener una visión negativa sobre la frase "Muerte a la gorra" disminuye la probabilidad de identificarse con la derecha.

Clase "Peronista"

- **TARIFAS (0.23):** No estar de acuerdo con aumentos en servicios públicos reduce considerablemente la probabilidad de ser peronista.
- **NOTICIAS_Prefiero no ver nada (0.32):** No consumir noticias disminuye la probabilidad de identificarse con el peronismo.

Clase Radical"

- **EMPRESARIOS (0.50):** Creer que "Los empresarios son el problema" reduce la probabilidad de identificarse como radical.

3.2. Modelo XGBoost

El uso de **XGBoost** permitió analizar diversas variables en términos de **peso e impacto en la reducción del error**, identificando aquellas más relevantes para la clasificación política. Sin embargo, los resultados mostraron que las **mismas variables clave utilizadas en la regresión logística** fueron las más efectivas, sin que XGBoost aportara mejoras significativas en precisión, alcanzando un **42.5 % de precisión como**

máximo. Dado esto, **resulta más conveniente y eficiente proceder con la regresión logística** en este contexto.

Se excluyó la variable **MILEI**, ya que, si bien demostró ser un fuerte predictor (incrementando la precisión máxima hasta **45 %**), es demaciado autoexplicativa . Finalmente, las **coincidencias observadas entre la izquierda y el peronismo en el análisis descriptivo** fueron corroboradas por XGBoost, dado que los **mayores desafíos de clasificación** se concentraron en la diferenciación entre estos dos grupos, confirmando la similitud en sus patrones de respuesta.

Para evaluar más a fondo esta dificultad en la clasificación, se realizó un ejercicio en el que **se agruparon las etiquetas de Peronistas e Izquierda en una misma clase**. Al realizar esta modificación, el modelo mejoró su desempeño:

- Precisión con división 30/70: **52.5 %**
- Precisión con división 40/60: **55 %** (con 500 rondas de boosting)

Este incremento en la precisión puede relacionarse al hecho de que ahora Peronistas e Izquierda conformaron una sola clase, lo que implica que esta categoría aglutina aproximadamente el **40 % de los casos**.

4. Análisis de Clustering con K-Means

Para realizar la segmentación de los encuestados mediante la técnica de K-Means, se eliminaron las variables relacionadas con atributos demográficos como edad, género e hijos, con el fin de evitar que estos factores exógenos sesgaran la agrupación en torno a ideas y posturas. Asimismo, se excluyó la etiqueta política de los encuestados para no condicionar los clusters con la variable que posteriormente se contrastaría con los resultados de la segmentación.

El número óptimo de clusters due buscado mediante el método de elbow y el coeficiente de silueta, estableciendo el valor más adecuado en torno a cuatro grupos. No obstante, se optó por una clusterización dicotómica el objetivo de analizar cómo se polarizaban los encuestados y qué patrones emergían de la división binaria.

4.1. Resultados de la Clusterización General

Se observaron dos grandes grupos en la segmentación:

- El **Cluster 0** agrupa a la gran mayoría de los encuestados identificados como peronistas y de izquierda. Sus características reflejan una tendencia más progresista en diversas posturas económicas, sociales e internacionales.
- El **Cluster 1** es más heterogéneo y está compuesto por una combinación de encuestados apolíticos, de derecha, liberales y radicales. Este grupo muestra una mayor diversidad ideológica y no se asocia tan claramente a una única orientación política.

Se realizó un join de los clusters con la etiqueta política y género. Se confirmó que el Cluster 0 tiene una alineación más clara con posturas progresistas, mientras que el Cluster 1 contiene una variedad de tendencias políticas más amplia.

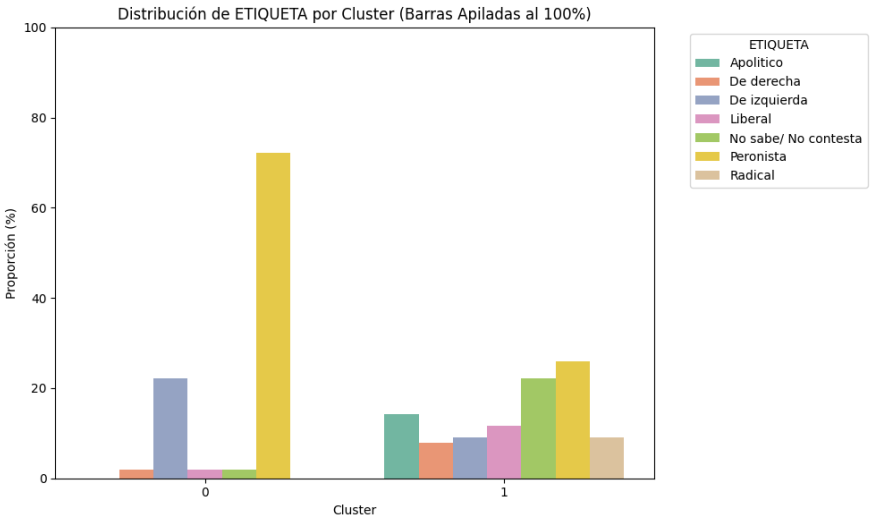

Figura 2: Distribución de etiquetas políticas por Cluster General

Cluster 0: Se encuentra **dominando el peronismo y la izquierda**, con una baja representación de otras etiquetas políticas. Esta agrupación parece indicar una **concepción compartida del mundo y de la política**, separada del resto de los encuestados.

Cluster 1: Presenta una composición **heterogénea**, donde conviven **apolíticos, liberales, radicales, de derecha y tambien peronista**, mostrando una cohesíon ideologica con diversidad de etiquetas.

Estos resultados sugieren que la izquierda y la mayor parte del peronismo parecen estar agrupados en un **bloque ideológico diferenciado**, con una lógica que se distancia del resto de la sociedad, como si existiera una **brecha cultural e interpretativa** entre ambos grupos.

4.2. Clusterización Interna Peronistas

Dado que la mayoría de los peronistas se agrupaban en el Cluster 0, se realizó una segunda segmentación interna dentro de este grupo. Se identificaron dos subclusters:

- **Peronistas del Grupo 0:** Más alineados con una visión económica más estatalista, con posiciones más críticas hacia sectores privados y mayor afinidad con posturas progresistas en temas internacionales.
- **Peronistas del Grupo 1:** Con una visión más moderada en lo económico y geopolítico, manteniendo elementos tradicionales del peronismo sin una inclinación tan marcada hacia la izquierda.

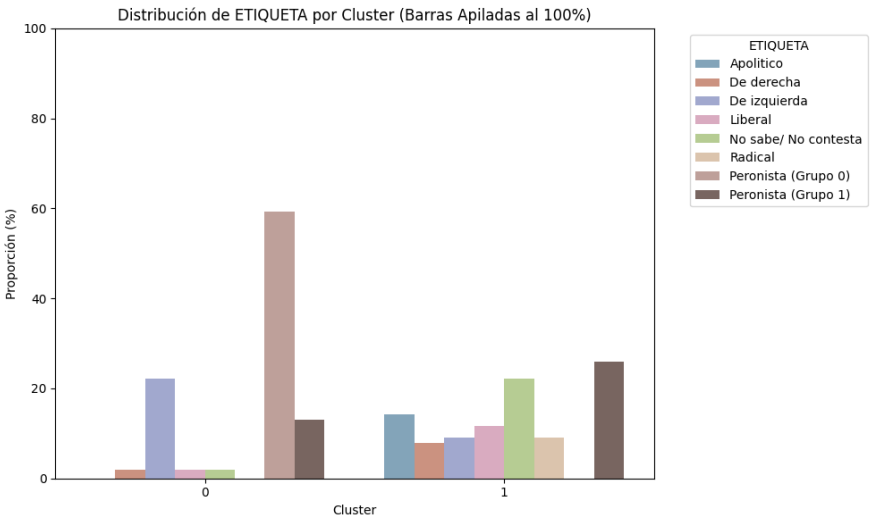

Figura 3: Distribución de etiquetas políticas considerando la clusterización interna de peronistas

Cuando se reintegraron estos subclusters en la segmentación general, se verificó que los peronistas del Grupo 0 se integraban completamente con el Cluster 0 original, mientras que los peronistas del Grupo 1 se distribuían en ambos clusters, aunque mayormente en el Cluster 1, mostrando una mayor diversidad interna dentro del espacio peronista.

Los resultados sugieren que al forzar una polarización al interior del Peronismo, la misma tiene un considerable nivel de coincidencia con la imagen que tenemos al polarizar el conjunto del arco político. Habiendo un peronismo que dialoga en los terminos politico- ideologicos y culturales de la izquierda mientras mientras hay otro sector que está mas cercano al resto de la sociedad en termnos ideologicos, de forma mas congruente con la idea de Peronismo transversal y de "Movimiento nacional".

4.3. Autocríticas al Estudio

Para mejorar este ejercicio en el futuro, se identifican dos aspectos clave a mejorar:

- **Balancear mejor las clases:** En futuras iteraciones, sería ideal incrementar la cantidad de respuestas en las clases menos representadas. Un mejor equilibrio en la distribución de las etiquetas mejorará la capacidad del modelo para realizar predicciones más precisas y evitará sesgos en la clasificación.
- **Mejorar la categorización de la variable objetivo:** Con el objetivo de abarcar el amplio espectro político, se definieron múltiples etiquetas que, en retrospectiva, podrían presentar cierto solapamiento. Una mejor alternativa habría sido estructurar la variable de la siguiente manera:
    - Utilizar una **escala continua** del 1 al 10 para medir el posicionamiento político en un eje de izquierda a derecha.
    - Incorporar una pregunta separada sobre **afinidad con polos ideológicos**, como nacionalismo vs. globalismo o liberalismo vs. socialdemocracia.

- Incluir una pregunta específica sobre **identificación con movimientos políticos concretos** (Peronismo, Radicalismo, PRO, partidos de izquierda, entre otros).

Esta reformulación habría permitido mayor flexibilidad en el análisis, evitando superposiciones entre categorías y habilitando una segmentación más precisa para futuras investigaciones.

5. Código desarrollado

Para acceder al repositorio del proyecto, visita: **Repositorio en GitHub**

Referencias

[1] Google LLC. *Google Forms*. Accedido: 23 de febrero de 2025. 2025. URL: `https://docs.google.com/forms/`.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. Accedido: 23 de febrero de 2025. R Foundation for Statistical Computing. Vienna, Austria, 2025. URL: `https://www.R-project.org/`.

[3] RStudio Team. *RStudio: Integrated Development Environment for R*. Accedido: 23 de febrero de 2025. RStudio, PBC. Boston, MA, 2025. URL: `https://www.rstudio.com/`.

[4] Python Software Foundation. *Python: A Dynamic, Open Source Programming Language*. Accedido: 23 de febrero de 2025. 2025. URL: `https://www.python.org/`.

[5] Google Research. *Google Colaboratory*. Accedido: 23 de febrero de 2025. 2025. URL: `https://colab.research.google.com/`.