



UNIVERSITAT OBERTA DE CATALUNYA (UOC)

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (*Data Science*)

TRABAJO FINAL DE MÁSTER

ÁREA: CIENCIA DE DATOS EN SISTEMAS COMPLEJOS, SOSTENIBILIDAD Y
ECOLOGÍA

Desarrollo de un Marco de Planificación Territorial para la Educación Superior

**Uso de Clustering No Supervisado y Análisis Geoespacial para la Expansión de
la Educación Superior en Línea en la Universidad Politécnica Salesiana**

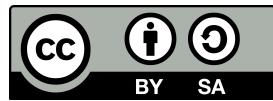
Autor: Christian Rolando Oyola Flores

Tutor: Raúl Parada Medina

Profesor: Susana Acedo Nadal

31 de diciembre de 2025

Créditos/Copyright



Esta obra está sujeta a una licencia de Reconocimiento - CompartirIgual3.0

3.0 España de CreativeCommons.

FICHA DEL TRABAJO FINAL

Título del trabajo:	Desarrollo de un Marco de Planificación Territorial para la Educación Superior: Uso de Clustering No Supervisado y Análisis Geoespacial para la Expansión de la Educación Superior en Línea en la Universidad Politécnica Salesiana
Nombre del autor:	Christian Rolando Oyola Flores
Nombre del Tutor/a de TF:	Raúl Parada Medina
Nombre del/de la PRA:	Susana Acedo Nadal
Fecha de entrega:	12/2025
Titulación o programa:	Máster Universitario en Ciencia de Datos (Data Science)
Área del Trabajo Final:	Ciencia de Datos en Sistemas Complejos, Sostenibilidad y Ecología
Idioma del trabajo:	Castellano
Palabras clave	Planificación educativa geoespacial, Sistemas de apoyo a la decisión en educación, Minería de datos educativos

Abstract

In the context of online higher education in Ecuador, significant challenges persist related to limited access in rural and marginal urban areas, as well as student dropout rates resulting from geographical, technological, and socioeconomic barriers. This study presents a territorial planning framework for the expansion of online higher education at the Salesian Polytechnic University (UPS), based on the CRISP-DM model 1.0 model adapted to the education sector. The study seeks to optimize the implementation of a network of hybrid support centers by analyzing demographic data (INEC 2022) and student academic profiles, using advanced techniques such as unsupervised clustering and geospatial analysis. The objective is to identify strategic areas in the 221 local administrative units LAU2 (in Ecuador, equivalent to cantons) that will reduce access gaps and improve student retention. The proposed methodology not only facilitates evidence-based decision-making, but also contributes to the Sustainable Development Goals by promoting more inclusive and equitable education.

Keywords: Geospatial Educational Planning, Decision Support Systems in Education, Educational Data Mining

Resumen

En el contexto de la educación superior en línea en Ecuador, persisten desafíos significativos relacionados con el acceso limitado en zonas rurales y urbanas marginales, así como tasas de deserción estudiantil derivadas de barreras geográficas, tecnológicas y socioeconómicas. Este estudio presenta un marco de planificación territorial para la expansión de la educación superior en línea en la Universidad Politécnica Salesiana (UPS), fundamentado en el modelo CRISP-DM 1.0 adaptado al sector educativo. El trabajo busca optimizar la implementación de una red de centros de apoyo híbridos mediante el análisis de datos demográficos (INEC 2022) y perfiles académicos de estudiantes, empleando técnicas avanzadas como clustering no supervisado y análisis geoespacial. El objetivo es identificar zonas estratégicas en los 221 unidades administrativas locales LAU2 (en Ecuador, equivalentes a los cantones) que permitan reducir brechas de acceso y mejorar la retención estudiantil. La metodología propuesta no solo facilita la toma de decisiones basada en evidencia, sino que también contribuye a los Objetivos de Desarrollo Sostenible promoviendo una educación más inclusiva y equitativa.

Palabras clave: Planificación educativa geoespacial, sistemas de apoyo a la toma de decisiones en educación, minería de datos educativos.

Índice general

Abstract	V
Resumen del trabajo	VII
Índice	IX
Lista de Figuras	XI
Lista de Tablas	1
1. Introducción	3
1.1. Contexto y justificación del trabajo	3
1.2. Objetivos del trabajo	7
1.3. Impacto en sostenibilidad, ético-social y de diversidad	8
1.4. Enfoque y método seguido	9
1.5. Planificación del trabajo	14
1.6. Sumario de productos obtenidos	17
1.7. Breve resumen de capítulos	17
2. Estado del Arte	18
2.1. Políticas de acceso, QA y servicios integrales de apoyo	18
2.2. Planificación basada en evidencia: analítica, GIS y optimización de cobertura	20
3. Materiales y métodos	23
3.1. Base contextual analítica	25
3.2. Preprocesamiento y Selección de Variables	30
3.3. Distancia de Gower ponderada para variables mixtas	36
3.4. Agrupamiento jerárquico aglomerativo y criterios de enlace	41
3.5. Refinamiento interno mediante PAM	47
3.6. Limpieza de casos fronterizos	49
3.7. Consolidación de clústers	51
4. Experimentos y resultados	53

4.1.	Patrones identificados	53
4.2.	Aprendizaje supervisado: predicción de perfiles a partir de etiquetas de cluster (LightGBM)	57
4.3.	Indicador territorial por cantón	62
5.	Conclusiones y Trabajos Futuros	66
Bibliografía		67
6.	Anexos	75

Índice de figuras

1.	Metodología CRISP-DM 1.0 empleada para el desarrollo de la propuesta. Fuente: Metodologías y estándares, Jordi Gironés Roig, UOC	12
2.	Arquitectura técnica general de la propuesta planteada constituida por 6 etapas	24
3.	Matriz de asociación categórica (V de Cramer) entre las variables preliminares.	33
4.	Matriz reducida de asociación categórica (V de Cramer) para las variables seleccionadas	34
5.	Matriz de coeficiente de Spearman para variables ordinales y continuas	35
42figure.0.6		
7.	Índice de Silhouette global en función del número de clusters k	43
8.	t-SNE sobre matriz de distancias de Gower coloreado por clusters HCA	45
9.	Gráfico de Silhouette para la solución jerárquica de $k = 5$ clusters. La línea vertical indica el Silhouette medio global.	46
10.	Gráfico de Silhouette con refinamiento PAM para la solución jerárquica de $k = 5$. .	49
11.	Gráfico de Silhouette con refinamiento de casos fronterizos para la solución jerárquica de $k = 5$	50
12.	Gráfico de Silhouette con refinamiento de clusters para la solución jerárquica de $k = 5$	51
13.	Proyección t-SNE de la segmentación final	52
14.	Tamaño de los clusters-porcentaje de individuos	53
15.	Matriz de confusión del resultado de clasificación de LightGBM	60
16.	Importancia de variables por ganancia en el modelo LightGBM	61
17.	Flujo ETL en Spoon para homologación, recodificación y control de calidad previo a la consolidación del dataset INEC.	61
18.	Identificación de cantones prioritarios mediante mapa coroplético	64
19.	Isócronas de acceso en tiempo 10, 30 y 60 minutos sobre mapa de densidad por perfil	65

Índice de cuadros

1.	Planificación de actividades del TFM	15
2.	Categorización temática de las variables	31
3.	Diccionario de las variables del dataset final	36
4.	Pesos w_j asignados por variable en la distancia de Gower	38
5.	Índices de Davies–Bouldin (DBI) y Calinski–Harabasz (CH) para distintos valores de k	44
6.	Resumen del coeficiente de Silhouette por clúster ($k = 5$).	47
7.	Matriz de contingencia entre la partición HCA original y la partición refinada con PAM ($k = 5$)	48
8.	Resumen del coeficiente de Silhouette por clúster ($k = 5$). Silhouette global: 0.2609.	49
9.	Resumen del coeficiente de Silhouette por clúster ($k = 5$). Silhouette global: 0.2743.	50
10.	Resumen del coeficiente de Silhouette por clúster ($k = 5$). Silhouette global: 0.2989.	51
11.	Modalidad dominante (top_cat) y proporción (top_pct) por variable y cluster.	54
12.	Diccionario de variables y esquema de codificación	54
13.	Estadísticos descriptivos por cluster para variables numéricas/codificadas.	55
14.	Top 10 cantones con mayor densidad de perfil C3	64
15.	Composición de las Bases de Datos del INEC (Vivienda, Hogar, Mortalidad, Emigración y Población)	75

1. Introducción

1.1. Contexto y justificación del trabajo

La educación superior en América Latina ha experimentado una expansión significativa en cobertura durante las últimas dos décadas; sin embargo, esta expansión cuantitativa no ha traído consigo una distribución equitativa de oportunidades. Según [1], si bien la cobertura de la educación primaria y secundaria en América Latina supera el 90% en la mayoría de los países, la transición hacia la educación superior continúa siendo el principal cuello de botella para la equidad educativa regional.

En el caso de Ecuador, los datos derivados de la *Encuesta Nacional de Empleo, Desempleo y Subempleo* (ENEMDU, 2019) y del *Instituto de Estadística de la UNESCO* ubican la tasa de asistencia a la educación superior en torno al 37% del total de jóvenes en edad universitaria (de 18 a 23 años), lo que sugiere que el reto no reside en el acceso inicial al sistema educativo, sino en la permanencia y transición efectiva hacia el nivel superior. A pesar de la expansión cuantitativa de la matrícula en educación superior en América Latina, múltiples estudios evidencian que el acceso rural sigue siendo un factor determinante que limita una transición efectiva al nivel terciario. En Chile, [2] muestran que los estudiantes rurales tienen menor probabilidad de inscribirse a los exámenes de admisión universitarios y enfrentan barreras adicionales respecto al rendimiento, lo que reduce sus posibilidades de acceder a universidades en comparación con sus pares urbanos. Asimismo, en Colombia según [3], las políticas orientadas a la educación superior rural han mitigado parcialmente el problema, pero aún persisten obstáculos esenciales como la distancia geográfica, la falta de infraestructura y la ausencia de apoyo institucional.

Otro factor crítico es la desigualdad en la conectividad digital entre zonas urbanas y rurales. Estudios asociados muestran que el acceso a internet de calidad es mucho menor en áreas rurales, limitando la posibilidad de aprovechar modalidades en línea o híbridas de educación superior con soporte local. Este desfase según [4] amplifica la desventaja de los estudiantes de los quintiles inferiores, quienes suelen residir en zonas menos urbanizadas.

El desfase tecnológico no solo agrava la desigualdad inicial de acceso, sino que incide en que muchos jóvenes rurales no puedan aprovechar modelos virtuales o híbridos de educación superior; es en este contexto que se identifica una desigualdad de acceso acompañada de tasas más altas de abandono temprano y de menor eficiencia educativa para los estudiantes con menos recursos socioeconómicos. Estas diferencias según [5], no solo se presentan al momento de no acceder, sino también durante el trayecto formativo, lo que exige intervenciones en soporte institucional y pedagógico para mejorar la permanencia y culminación de los estudios.

La complejidad de factores que determinan el éxito en modalidades en línea trasciende dimensiones puramente tecnológicas o pedagógicas, involucrando de forma relevante aspectos de apoyo institucional integral. [6] advirtieron en sus primeros estudios mediante encuestas a profesorado de

educación en línea que muchos docentes atribuyeron pérdida de estudiantes no a deficiencias de diseño curricular o interacción pedagógica, sino a la inestabilidad de la infraestructura tecnológica institucional y a la incapacidad del personal de apoyo técnico del campus para resolver problemas operacionales que frecuentemente frustraban a estudiantes en línea y les impedían tener una experiencia satisfactoria. Este hallazgo subraya que, si bien el personal de tecnologías de información no ejerce funciones docentes directas y puede no tener contacto cotidiano con estudiantes, en entornos de enseñanza en línea sus acciones o inacciones pueden afectar significativamente el éxito o fracaso de programas académicos completos mediante efectos indirectos sobre la satisfacción y persistencia estudiantil.

A estas limitaciones institucionales se suman brechas estructurales de acceso territorial que profundizan las desigualdades en la educación superior latinoamericana. En numerosos contextos rurales o periféricos, la falta de redes locales de apoyo académico y administrativo limitan la participación sostenida de estudiantes en modalidades en línea, incluso cuando logran ingresar al sistema educativo. Esta combinación añadida de aislamiento territorial y soporte institucional insuficiente configura un entorno de vulnerabilidad educativa donde la permanencia depende más de la capacidad individual de superar obstáculos externos que del acompañamiento institucional. La magnitud y persistencia de estas barreras revelan la necesidad de reformular los modelos universitarios tradicionales, dando paso a enfoques organizativos y pedagógicos más flexibles y equitativos que respondan a las condiciones reales de los estudiantes y a las asimetrías territoriales del sistema.

Frente a este contexto multidimensional de inequidad estructural en el acceso y la permanencia en educación superior latinoamericana, han emergido modelos institucionales divergentes que buscan explícitamente democratizar el acceso mediante innovaciones organizativas, pedagógicas y tecnológicas, buscando mayor integración académica, coherencia institucional y, fundamentalmente, dar respuesta a las necesidades de poblaciones estudiantiles no tradicionales que incluyen trabajadores adultos, estudiantes de primera generación universitaria sin capital cultural académico heredado familiarmente, población rural y grupos étnicos históricamente marginados del sistema universitario convencional estructurado según modelos importados que presuponen estudiantes urbanos de tiempo completo con soporte económico familiar. Según [1], estos modelos presentan además bajo nivel de interculturalización, sin responder con pluralidad cultural a las sociedades de las que forman parte. Es en este último enfoque, se busca propiciar un espacio de encuentro fuera de modelos “convencionales”, en donde los idiomas, historias, conocimientos, modos de vida, visiones de mundo, y modos de aprendizaje y de producción de conocimientos de los pueblos originarios, sean contemplados dentro de la planificación y enfoque académico, donde la universidad no se convierta solo en un agente de construcción de representaciones homogeneizadas de las poblaciones, ni en la transformación de los pueblos indígenas en objetos puramente de estudio, incluso en contra de su voluntad, fomentando enfoques etnocéntricos de investigación que producen representaciones descalificadoras de sus “razas”,

formas de vida, visiones del mundo, conocimientos, y proyectos de futuro.

Integrar la expansión de la educación superior en línea con las prácticas tradicionales de aprendizaje de los pueblos originarios convierte a esta modalidad en un canal que no solo transporta contenidos académicos, sino que también puede encaminar modos de transmisión de saberes ancestrales, repensando incluso la arquitectura misma del aula digital.

Este modelo de universidad divergente se fundamenta en tres pilares institucionales interconectados:

Flexibilidad curricular y pedagógica. Reconoce la heterogeneidad de trayectorias estudiantiles, permitiendo combinaciones de modalidades presenciales, semipresenciales y virtuales adaptadas a restricciones laborales y familiares de estudiantes no tradicionales, en contraste con la rigidez de programas presenciales a tiempo completo que presuponen disponibilidad exclusiva para actividades académicas, característica de estudiantes tradicionales jóvenes sin responsabilidades laborales o familiares significativas [7].

Descentralización territorial. Mediante el establecimiento de centros de apoyo, sedes regionales o nodos universitarios en ciudades intermedias y zonas rurales que reducen barreras geográficas de acceso, reconociendo que los costos de oportunidad asociados al desplazamiento hacia centros urbanos metropolitanos constituyen filtros de selección socioeconómica que se reflejan antes de cualquier proceso formal de admisión [8].

Democratización tecnológica. Mediante la inversión institucional en infraestructura digital que garantice accesibilidad de plataformas virtuales independientemente de limitaciones de conectividad o dispositivos de los estudiantes, internalizando en la institución la responsabilidad de proveer condiciones tecnológicas para el aprendizaje en lugar de externalizar esta responsabilidad hacia estudiantes y sus familias [9].

En este marco, la *Universidad Politécnica Salesiana* (UPS) se encuentra en un momento estratégico para materializar los principios de un modelo divergente mediante el establecimiento de una primera red de centros de apoyo para la modalidad de estudios en línea. Según [10] a diferencia de su infraestructura multisede consolidada en las tres principales ciudades del Ecuador (Cuenca, Quito y Guayaquil), que atiende programas presenciales, la modalidad en línea gestionada por la *Unidad de Educación a Distancia y Virtual* (UNADEDVI) opera actualmente sin componente de infraestructura física descentralizada, constituyendo una modalidad completamente remota sin puntos de contacto presencial fuera de las tres sedes principales. Esta ausencia de centros de apoyo específicamente diseñados para estudiantes en línea constituye una oportunidad estratégica única para implementar desde cero una red de soporte territorial fundamentada rigurosamente en evidencia empírica, en lugar de decisiones reactivas.

El análisis de coyuntura institucional revela cuatro factores convergentes y contextuales que crean una ventana de oportunidad para su desarrollo e implementación:

Primero, la UPS ha consolidado durante tres décadas de trayectoria institucional (1994–2024) legitimidad académica y social, incluyendo experiencia pedagógica salesiana distintiva en atención a poblaciones no tradicionales y capacidad administrativa desarrollada para la gestión de complejidad organizativa multisede, lo que garantiza que nuevos centros de apoyo se construyan sobre bases institucionales sólidas en lugar de experimentación con riesgo significativo.

Segundo, la modalidad en línea de la UPS gestionada por la UNADEDVI ha alcanzado madurez operacional con plataformas tecnológicas estables, contenidos curriculares desarrollados y procesos administrativos estandarizados que eliminan incertidumbres de la fase exploratoria inicial y permiten un enfoque estratégico en optimización de alcance territorial y soporte estudiantil.

Tercero, y desde una perspectiva de planificación basada en evidencia, la disponibilidad de datos demográficos granulares provenientes del *Censo de Población y Vivienda 2022* del *Instituto Nacional de Estadística y Censos* (INEC) genera una ventana temporal única para fundamentar decisiones estratégicas de posicionamiento institucional en datos a nivel nacional.

Cuarto, en relación con el crecimiento de capacidades analíticas en el campo del *Educational Data Mining* y la *Ciencia de Datos* aplicada a la educación superior, la disponibilidad de herramientas de código abierto accesibles (*Python*, *R*, *QGIS*) y servicios de análisis [Análisis Geoespacial](#) mediante APIs públicas ([GraphHopper API](#), [OpenStreetMap](#), [Openrouteservice](#), entre otros) democratiza las capacidades de análisis sofisticadas en estos ámbitos.

Considerando esta confluencia de factores estructurales favorables se configura un escenario óptimo para la formulación de una estrategia universitaria de expansión territorial inteligente y basada en datos, que, más allá de responder a una necesidad institucional de fortalecimiento operativo, esta investigación se inserta en el debate regional sobre equidad educativa, aportando una metodología replicable que puede ser adoptada por otras instituciones de educación superior enfrentadas al desafío de democratizar el acceso a la educación en contextos marcados por la desigualdad estructural. Respondiendo además al reglamento expedido en el año 2015 por el Consejo de Educación Superior del Ecuador (CES), para carreras y programas académicos en modalidades en línea, a distancia y semipresencial o de convergencia de medios, cuyo artículo 13 establece:

[...] que las Instituciones de Educación Superior deberán sustentar en ejercicio de su autonomía académica un modelo pedagógico y curricular con pertinencia que promueva el aprendizaje, bajo entornos potencializados por el uso de las tecnologías de la información y la comunicación, propendiendo a una educación personalizada. [11]

En este contexto, el presente estudio propone el diseño de una herramienta analítica de apoyo a la toma de decisiones que permita identificar, justificar y optimizar la implementación de centros de apoyo territorial en zonas con potencial educativo no atendido, considerando como eje central el perfil socio-demográfico, académico y tecnológico de la población académica de tercer nivel de la modalidad en línea de la UPS.

1.2. Objetivos del trabajo

1.2.1. Objetivo General

Desarrollar un marco de análisis impulsado por datos, basado en técnicas de aprendizaje no supervisado y análisis geoespacial, para identificar perfiles de estudiantes en modalidad en línea de la Universidad Politécnica Salesiana y correlacionarlos con datos demográficos del Instituto Nacional de Estadística y Censos (INEC), con el fin de proponer el posicionamiento estratégico de centros de apoyo híbridos en unidades administrativas locales LAU2 o [Cantones](#) prioritarios de Ecuador con alto potencial de similitud demográfica que expandan la matrícula en línea en un rango estadísticamente significativo (superior al crecimiento natural observado en los últimos 10 ciclos académicos), al mismo tiempo, estrechar brechas de acceso reforzando la identidad salesiana que sitúa a los jóvenes de sectores populares en el centro de su misión.

1.2.2. Objetivos Secundarios

OE1 Identificar perfiles de estudiantes en línea de la Universidad Politécnica Salesiana mediante técnicas de aprendizaje no supervisado, utilizando variables demográficas y de acceso a tecnología (ejemplo.: edad, nivel educativo, uso de internet, ocupación, entre otras), con el fin de responder a la pregunta: "¿Cuáles son los patrones conductuales, socioeconómicos y de conectividad de los estudiantes objetivo de la UPS?"

OE2 Correlacionar los perfiles de estudiantes identificados con datos cantonales del INEC, con el fin de identificar áreas con características demográficas y socioeconómicas similares a los perfiles de la UPS permitiendo identificar y analizar una métrica que permita priorizar unidades administrativas locales LAU2 (cantones) potenciales para posicionamiento estratégico.

OE3 Visualizar la distribución geoespacial de los perfiles de estudiantes mediante técnicas de [Choropleth](#), [Isocrona](#) y análisis de accesibilidad, utilizando el clasificador geográfico Estadístico - DPA del INEC ([Shapefiles DPA](#) oficiales de la División Política Administrativa del Ecuador) y evaluar los gaps o carencias en la cobertura actual.

El alcance de este trabajo de fin de máster incluye el desarrollo del marco de análisis y la identificación de unidades administrativas locales LAU2 (en Ecuador, equivalentes a los cantones) prioritarias, el análisis de la correlación entre perfiles de estudiantes y datos demográficos, así como la propuesta de ubicación de centros de apoyo híbridos basada en evidencia cuantitativa. Asimismo, se excluye la implementación física, el análisis económico detallado, el rendimiento de matrícula y la evaluación de impacto post-implementación, considerado dentro de una investigación futura.

1.3. Impacto en sostenibilidad, ético-social y de diversidad

Este trabajo, apegado a la definición del CCEG: “*Actuar de manera honesta, ética, sostenible, socialmente responsable y respetuosa con los derechos humanos y la diversidad, tanto en la práctica académica como en la profesional, y diseñar soluciones para mejorar estas prácticas.*”, y alineado con las dimensiones: I. Sostenibilidad, II. Comportamiento ético y responsabilidad social (RS), III. Diversidad (género, entre otros) y derechos humanos, busca acercar la educación superior y mejorar la retención estudiantil, promoviendo espacios de encuentro y respetando la pluralidad cultural de los pueblos originarios, contribuyendo principalmente a los Objetivos de Desarrollo Sostenible (ODS), 4 (Educación de calidad) y 10 (Reducción de desigualdades), pero además, contribuye a los ODS 5 (Igualdad de género), 11 (Ciudades y comunidades sostenibles) y 16 (Paz, justicia e instituciones sólidas). En este enfoque se detalla la reflexión sobre los impactos en sostenibilidad, ética y diversidad, alineados con la Competencia de compromiso ético y Global (CCEG) en las fases de diseño:

a) Dimensión Sostenibilidad (ODS 11)

El diseño del TFM contempla la *sostenibilidad social* al plantear el posicionamiento estratégico basado en datos de centros de apoyo híbridos que fortalecen comunidades locales, con enfoque en unidades administrativas locales LAU2 (cantones) rurales y marginados, al facilitar el acceso a la educación superior en línea, minimizando desplazamientos estudiantiles. Esto reduce la presión sobre infraestructuras urbanas, fomentando el desarrollo de comunidades sostenibles, alineándose con el *ODS 11*.

Este marco, resalta la vinculación de recursos digitales y presenciales, lo que minimiza el impacto en los recursos naturales al no requerir una infraestructura física integral. La motivación en sostenibilidad radica en *empoderar comunidades locales mediante la educación en línea con centros de apoyo*, promoviendo su desarrollo integral sin comprometer sus entornos.

b) Dimensión Comportamiento Ético y Responsabilidad Social (ODS 4, 16)

Desde el planteamiento o diseño del TFM se consideran principios éticos fundamentales para abordar dos decisiones críticas en el desarrollo de este marco analítico: la definición de las tareas de machine learning y la selección de los conjuntos de datos (registros de estudiantes UPS y Censo INEC 2022). Dado que los datos se consideran según la disponibilidad institucional y determinados a partir de la aplicación de esta propuesta, se reconoce un posible riesgo de sesgos debido a tamaños de muestra desiguales entre grupos demográficos. Para mitigar esto, se plantea el uso de bibliotecas para auditar sesgos algorítmicos en el clustering y la correlación, buscando asegurar que los hallazgos no generen desigualdades sociales. Además, se garantiza la anonimización de los datos, en cumplimiento con la Ley Orgánica de Protección de Datos Personales

del Ecuador y la Política de Tratamiento, Protección y Uso de Datos Personales y Clasificación de la Información de la UPS. Este proceso respeta la autonomía de los individuos representados en los datos, asegurando que no se utilicen en contra de su voluntad. Así mismo, esta propuesta busca fortalecer la infraestructura educativa de la UPS al proponer ubicaciones estratégicas para centros de apoyo, promoviendo el acceso equitativo a la educación (*ODS 4*). La motivación ética se centra en *democratizar la educación superior*, beneficiando a comunidades marginadas y apoyando el *ODS 16*.

c) Dimensión Diversidad, Género y Derechos Humanos (ODS 5, 10)

Este diseño del marco analítico considera la *diversidad de los estudiantes* (género, etnia, condición conómica, acceso TIC) al segmentar perfiles mediante *clustering*, asegurando que las propuestas de centros sean accesibles para colectivos vulnerables, como mujeres rurales, comunidades indígenas y personas con necesidades de aprendizaje y con discapacidad.

Se busca propiciar un espacio de encuentro que contemple los idiomas, historias, conocimientos, modos de vida y visiones de mundo de los pueblos originarios, evitando representaciones homogeneizadas o enfoques etnocéntricos que descalifiquen sus formas de vida o proyectos de futuro.

En el contexto de la implementación, se busca emplear un *lenguaje visual inclusivo y representativo*. La motivación incluye garantizar que el marco beneficie a todos los sectores, promoviendo la *igualdad de género (ODS 5)* y la *reducción de desigualdades (ODS 10)*.

1.4. Enfoque y método seguido

1.4.1. Estrategias de investigación

Durante la elaboración de este Trabajo Final de Máster se analizaron cuatro estrategias operativas para apoyar la localización de centros de apoyo en modalidad en línea.

- a. La primera estrategia consiste en diseñar y ejecutar encuestas de intención de matrícula en los 221 unidades administrativas locales LAU2 (cantones) del país. No obstante, esta alternativa requeriría al menos doce meses para su desarrollo, desde el diseño hasta el análisis, costos logísticos elevados y riesgo de sesgo por deseabilidad social en las respuestas.
- b. En segundo lugar, se evaluó un mapeo comparativo de la oferta de centros de apoyo de universidades equivalentes, con el objetivo de ubicar los puntos de la Universidad Politécnica Salesiana (UPS) en los mismos cantones donde ya operan otras instituciones. Esta alternativa, inicialmente se considera de naturaleza reactiva, no se contemplan los perfiles sociodemográficos

específicos de los estudiantes UPS y podría caer en asignar recursos a nichos ya saturados, incluso limita la diferenciación institucional.

- c. En tercer lugar, se plantea la contratación de una consultoría especializada en planificación territorial, que considera estudios de mercado, análisis de viabilidad y grupos focales. Si bien esta opción proporciona conocimiento experto externo, implica un costo considerable y no incluye mecanismos formales para la transferencia o replicación del conocimiento.
- d. Finalmente, se seleccionó una estrategia fundamentada en *Educational Data Mining*, que integra técnicas de agrupamiento, similitud vectorial y análisis geoespacial aplicadas a datos censales abiertos del Instituto Nacional de Estadística y Censos (INEC). Este enfoque permite identificar unidades administrativas locales LAU2 (cantones) con características demográficas similares a las de los estudiantes actuales de la UPS y con cobertura institucional insuficiente. Así mismo, se reducen costos mediante el uso de software libre y datos públicos, y aporta escalabilidad y replicabilidad para otros contextos educativos como por ejemplo posgrados, proporcionando evidencia objetiva y apegada a perfiles institucionales no genéricos para la toma de decisiones a la alta gerencia.

1.4.2. Justificación de estrategia de investigación

Se seleccionó la estrategia del modelado analítico basado en datos como aproximación metodológica óptima por tres razones fundamentales:

- a. La integración de los datos académicos institucionales durante los procesos de admisión de la Universidad Politécnica Salesiana (UPS) en la modalidad en línea, que abarcan 18 programas académicos y aproximadamente ≈ 380 estudiantes desde el periodo 2025-2026 hasta 2026-2026 [12], con microdatos del Censo del Instituto Nacional de Estadística y Censos (INEC) 2022, que contiene 16.9 millones de registros [13], evita la necesidad de generar datos primarios (o basados en encuestas) mediante mecanismos costosos y complicados a nivel nacional. Esta estrategia permite completar el análisis en un plazo que tiene concordancia con el plan del TFM (Sección 1.5).
- b. En segundo lugar, la reproducibilidad y transparencia metodológica al emplear datos exclusivamente públicos del INEC, combinado con herramientas opensource (Python, GeoPandas, Scikit-learn) y APIs gratuitas (GraphHopper, OpenStreetMap, Openrouteservice, entre otras), lo que permiten desarrollar los procesos analíticos en su totalidad, facilitando sobre todo la trazabilidad, transparencia y control durante las etapas de desarrollo, con la libertad de diseñar y construir una arquitectura modular sin dependencia de software propietario o conocimiento tácito de consultores externos.

- c. Por último, se aprovecha la disponibilidad de recursos institucionales de la UPS dispuestos para análisis de datos, y la posibilidad de que esta solución pueda integrarse con entornos de data warehouse y visualización existentes en la universidad, con lo que se busca agilizar y dinamizar la aplicabilidad de la propuesta en entornos accesibles sin costos adicionales, maximizando incluso el valor de los recursos institucionales y garantizando que el estudio se alinea con las capacidades técnicas y operativas de la UPS.

1.4.3. Descripción general del proceso de trabajo

Este trabajo propone desarrollar un marco de análisis fundamentado en técnicas de ciencia de datos que permita a la Universidad Politécnica Salesiana tomar decisiones estratégicas basadas en evidencia para el establecimiento de su primera red de centros de apoyo híbridos para modalidad de estudios en línea. Este marco se apega a una estrategia metodológica complementaria que, en conjunto, permite identificar, priorizar y validar ubicaciones óptimas que maximicen objetivos institucionales de crecimiento sostenible, retención de matrícula y objetivos sociales de equidad territorial alineados con la misión salesiana.

La metodología propuesta se basa en el framework CRISP-DM 1.0 (*Cross-Industry Standard Process for Data Mining*) (Ver Figura. 1) [14], adaptado al contexto educativo [15], integrando las etapas de comprensión del dominio, identificando las problemáticas del negocio a ser resueltas en el ámbito de la extensión de los centros de apoyo. La comprensión y preparación de datos considerando fuentes académicas y del INEC y plataformas de e-learning, con el objetivo de identificar su procedencia, condiciones, estructura, propiedades, inconvenientes y cómo mitigarlos. La modelización permitirá disponer de un modelo que ayude a alcanzar los objetivos de data mining, con algoritmos de aprendizaje automático para predecir rendimiento, deserción o comportamientos. La evaluación permitirá determinar el grado de acercamiento a los objetivos de negocio mediante la implementación de métricas de precisión y *recall*. Finalmente, la implementación o despliegue puede traducirse en dashboards o sistemas de apoyo a decisiones educativas.

En la Figura 1 podemos evidenciar las fases y su interacción de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), versión 1.0, a ser aplicada. Este diagrama muestra un ciclo de vida compuesto por las seis fases principales, organizadas de manera circular con su naturaleza iterativa.

1.4.4. Descripción componentes metodológicos

- a. El primer componente metodológico consiste en perfilado demográfico, socioeconómico y tecnológico de estudiantes en modalidad en línea de la UPS mediante aprendizaje no supervisado (clustering). Durante esta etapa se consumirá información institucional de la UPS que permita

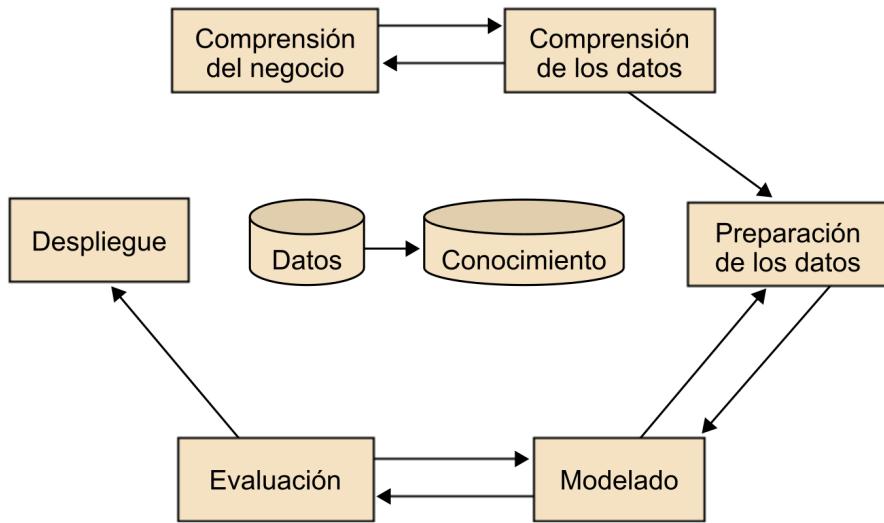


Figura 1: Metodología CRISP-DM 1.0 empleada para el desarrollo de la propuesta.
Fuente: Metodologías y estándares, Jordi Gironés Roig, UOC.

consolidar las dimensiones de información de los estudiantes de grado en linea.

Esta información es registrada por los estudiantes durante sus procesos de admisión mediante mecanismos apropiados y reposa en registros institucionales a los cuales se dispone el acceso mediante un acuerdo de confidencialidad apegado con lineamientos contractuales firmado entre las partes de manera previa. Esta fase se complementa con la aplicación de un agrupamiento jerárquico aglomerativo sobre una matriz de distancias de Gower, seguido de un refinamiento local mediante PAM (Partitioning Around Medoids) [16], para identificar los perfiles caracterizados por combinaciones específicas de variables demográficas (edad, sexo, etnia, máximo nivel académico completado), socioeconómicas (situación laboral, nivel socioeconómico, entre otros) y tecnológicas (acceso a Internet, dispositivos utilizados, entre otros).

Este clustering excluirá la variable de procedencia geográfica con el objetivo de garantizar que los perfiles identificados sean generalizables territorialmente, permitiendo posteriormente buscar unidades administrativas locales LAU2 (cantones) con composiciones demográficas similares independientemente de si actualmente son estudiantes de la institución.

Este primer acercamiento de clustering no supervisado, validado mediante métricas, busca y permite capturar heterogeneidad o variabilidad de la población estudiantil.

- b. El segundo componente metodológico consiste en *matching* demográfico entre perfiles institucionales y poblaciones cantonales mediante cálculo y análisis de similitud. Para cada uno de los doscientos veintiún (221) cantones de Ecuador, se calcularán vectores de características demográfico-tecnológicas comparables con variables utilizadas en el clustering de estudiantes, utilizando microdatos del Censo INEC 2022. Los denominados vectores cantonales contendrán

proporciones poblacionales por grupo de edad, distribución de nivel educativo de población con secundaria completa o superior, indicadores de acceso a TIC (porcentaje de hogares con Internet, distribución de tipo de conectividad, promedio de dispositivos digitales, entre otros), nivel socioeconómico, y variables de equidad (proporción de población indígena, afrodescendiente, con discapacidad, entre otros). Para cada cantón, se calcularán métricas de similitud con cada uno de los clusters de estudiantes o perfiles, identificando el perfil con máxima similitud y asignando este valor como *score* potencial de referencia base.

- c. El tercer componente metodológico consiste en un análisis geoespacial de accesibilidad mediante isocronas para apoyar la identificación de *gaps* (oportunidades de posicionamiento no saturados) territoriales. Los *scores* de potenciales unidades administrativas locales LAU2 (cantones) se integrarán espacialmente con *shapefiles* oficiales de División Político Administrativa del INEC utilizando sistema de coordenadas EPSG:32717 para Ecuador, generando visualizaciones cartográficas mediante *choropleths* temáticos buscando reducir la varianza dentro de las categorías y maximizar la varianza entre clases.

Considerando que actualmente la UPS no cuenta con centros de apoyo para modalidad en línea, el análisis geoespacial no se centrará en evaluar isocronas desde centros existentes sino que se busca identificar directamente unidades administrativas locales LAU2 (cantones) con combinación óptima de tres criterios:

- (1) *Score* de potencial demográfico que presente alta similitud con los perfiles de los estudiantes UPS,
 - (2) Tamaño de población objetivo que justifique el posicionamiento del centro en el punto resultante, y
 - (3) Distancia temporal ampliada respecto a las sedes principales de la UPS en Cuenca, Quito y Guayaquil, para evitar que los centros de apoyo compitan por la misma población que ya tiene acceso a infraestructura presencial, sino que expandan cobertura hacia territorios potencialmente desatendidos.
- d. Como aporte complementario, se propone la incorporación de un análisis de accesibilidad territorial “real” mediante la generación de isocronas de tiempo de viaje (10 min, 30 min y 60 min), en línea con la metodología empleada por [17] para la evaluación de cobertura educativa. El objetivo de este aporte complementario permite identificar y visualizar la accesibilidad real hacia los centros de apoyo, considerando la red vial existente en el país y las condiciones topográficas del territorio donde se podría posicionar.

1.5. Planificación del trabajo

A continuación, se presenta el Cuadro 1 con la temporalización aproximada del proyecto, indicando la fecha de inicio y final de cada actividad, organizadas por categorías para mejorar la interpretación. Los tiempos de revisión para cada actividad se considerarán dentro de los plazos definidos, así como, la comunicación con el Tutor y ajustes requeridos.

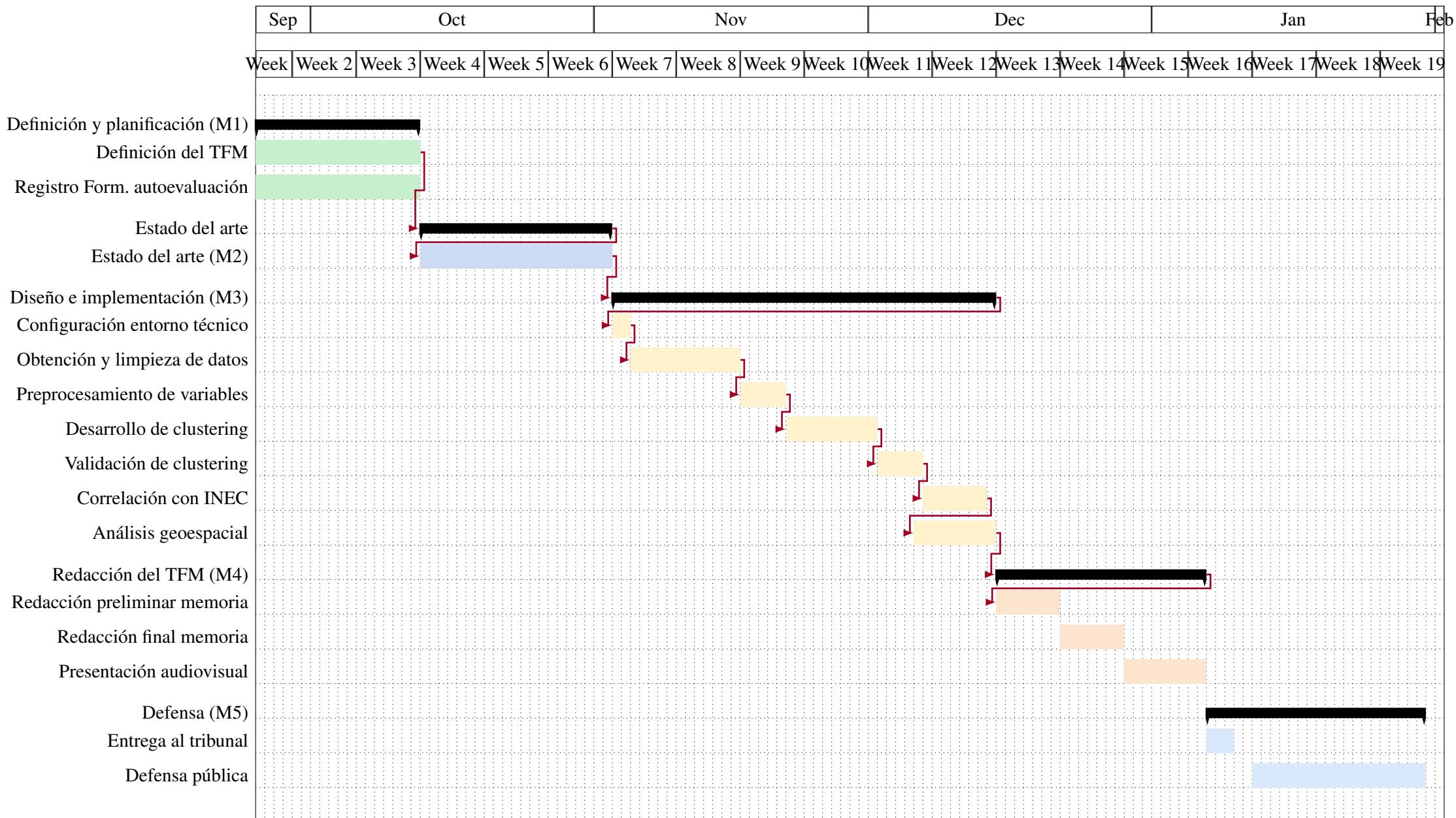
Categoría	#	Actividad	Descripción	Días	Inicio	Fin
Definición y planificación del trabajo final	1	Definición del TFM (M1)	Redacción y registro de los requisitos de presentación del TFM.	17	25/09/2025	12/10/2025
Definición y planificación del trabajo final	2	Registro de formulario de autoevaluación y autorresponsabilidad	Preparación de documentación para el comité de ética.	17	25/09/2025	12/10/2025
Estado del arte o análisis de mercado del proyecto	3	Estado del arte (M2)	Revisión bibliográfica en BD académicas sobre posicionamiento de centros de apoyo, educación en línea y Educational Data Mining.	21	13/10/2025	02/11/2025
Diseño e implementación del trabajo	4	Configuración del entorno técnico (M3)	Instalación y configuración de Python (Scikit-learn, Pandas, GeoPandas, Folium), GraphHopper API (free tier) y QGIS en entornos cloud o locales. Analizar compatibilidades.	2	03/11/2025	04/11/2025
Diseño e implementación del trabajo	5	Obtención y limpieza de datos (M3)	Consolidar datasets UPS (4K registros) e INEC (16.9M registros). Aplicar EDA, normalización y merge con shapefiles DPA.	12	05/11/2025	16/11/2025
Diseño e implementación del trabajo	6	Preprocesamiento de variables (M3)	Selección y transformación de variables para clustering y correlación, considerando codificación y escalado.	5	17/11/2025	21/11/2025
Diseño e implementación del trabajo	7	Desarrollo de clustering (M3)	Implementación de clustering para segmentar perfiles de estudiantes UPS por variables sociodemográficas y tecnológicas.	10	22/11/2025	01/12/2025

Continúa en la siguiente página

Categoría	#	Actividad	Descripción	Días	Inicio	Fin
Diseño e implementación del trabajo	8	Validación de clustering (M3)	Aplicación y análisis de pruebas y ajuste de parámetros si es necesario.	5	02/12/2025	06/12/2025
Diseño e implementación del trabajo	9	Correlación con INEC (M3)	Cálculo y análisis de similitud entre perfiles de la UPS y datos INEC. Análisis de intervalos de confianza para priorizar unidades administrativas locales LAU2 (cantones) que se ajusten al perfil de la UPS.	7	07/12/2025	13/12/2025
Diseño e implementación del trabajo	10	Análisis geoespacial (M3)	Creación de choropleths e isocronas (con GraphHopper API) para identificar unidades administrativas locales LAU2 (cantones) prioritarios considerando accesibilidad 30, 60, 90 min.	9	06/12/2025	14/12/2025
Redacción de la documentación del TFM	11	Redacción preliminar de la memoria (M4)	Borrador inicial del TFM. Planteamiento de resultados preliminares y visualizaciones (mapas).	7	15/12/2025	21/12/2025
Redacción de la documentación del TFM	12	Redacción final de la memoria (M4)	Finalización del informe con resultados, mapas y recomendaciones para centros de apoyo híbridos en la UPS.	7	22/12/2025	28/12/2025
Redacción de la documentación del TFM	13	Presentación audiovisual (M4)	Creación de la presentación resumen del TFM.	9	29/12/2025	06/01/2026
Defensa del proyecto	14	Entrega al tribunal (M5)	Envío de la entrega final y documentación completa al tribunal evaluador para revisión.	3	07/01/2026	09/01/2026
Defensa del proyecto	15	Defensa pública (M5)	Previsión y Presentación oral del TFM ante el tribunal.	19	12/01/2026	30/01/2026

Cuadro 1: Planificación de actividades del TFM

Cronograma de Actividades



1.6. Sumario de productos obtenidos

El estudio generó una cadena completa de productos analíticos y operativos. Primero, se consolidó un dataset final UPS con variables mixtas (nominales, ordinales y numéricas) estandarizadas y documentadas mediante un diccionario de codificación, asegurando consistencia para el análisis. Segundo, se construyó una matriz de distancias de Gower ponderada, donde los pesos se definieron con respaldo estadístico (NMI) y ajuste intencional según el objetivo del estudio, evitando ponderaciones arbitrarias. Tercero, se obtuvo una segmentación no supervisada mediante clustering jerárquico (enlace promedio) y se aplicó un refinamiento local con Partitioning Around Medoids (PAM) para mejorar la coherencia interna, complementado con validación y depuración por Silhouette para controlar casos frontera. Cuarto, se consolidó una tipología final de perfiles representativos ($k=3$) y se construyó su perfilado descriptivo con estadísticos por variable (modalidad dominante y proporción en categóricas; tendencia central y dispersión numéricas), lo que habilitó una lectura ejecutiva interpretable. Quinto, se entrenó un modelo supervisado LightGBM para replicar etiquetas a gran escala, incluyendo probabilidades por clase y un indicador de confianza para control de calidad. Finalmente, se diseñó e implementó una arquitectura de integración INEC 2022 (CSV → Esquema Dimensional → Parquet) y se ejecutó la clasificación masiva sobre millones de registros, derivando productos territoriales como densidad por cantón (LAU2), rankings por perfil, mapas coropléticos y análisis de accesibilidad mediante isócronas para nodos priorizados.

1.7. Breve resumen de capítulos

El documento se organiza en cinco(5) capítulos. La Introducción 1 presenta el contexto y la justificación del trabajo, define los objetivos, incorpora consideraciones de sostenibilidad, ética-social y diversidad, y describe el enfoque metodológico junto con la planificación. El Estado del Arte 2 revisa antecedentes sobre políticas de acceso, aseguramiento de calidad y servicios de apoyo, así como enfoques de planificación basada en evidencia que integran analítica, GIS y optimización de cobertura, fundamentando la pertinencia del estudio. En Materiales y Métodos 3 se detalla la base analítica, el preprocesamiento y la selección de variables, la construcción de la distancia de Gower ponderada, el agrupamiento jerárquico y sus criterios de enlace, el refinamiento interno con (Partitioning Around Medoids) PAM, la limpieza de casos fronterizos y la consolidación de clusters. El capítulo de Experimentos y Resultados 4 reporta los patrones identificados y el perfilado descriptivo, desarrolla el aprendizaje supervisado con LightGBM para predecir perfiles a partir de etiquetas de cluster, y construye el indicador territorial a escala de unidades administrativas locales LAU2 (cantones), incluyendo sus productos cartográficos. Finalmente, Conclusiones y Trabajos Futuros 5 sintetiza los aportes y la escalabilidad alcanzada, y se plantea mejoras vinculadas con la ampliación de la muestra, la actualización del INEC, la gobernanza del esquema de variables, la arquitectura de datos y el ajuste de hiperparámetros.

2. Estado del Arte

La educación superior en línea atraviesa una transformación impulsada por la convergencia de cambios sociales y avances tecnológicos. En la última década, su expansión en América Latina ha reabierto debates sobre acceso, permanencia y equidad territorial, mientras que el desarrollo de la ciencia de datos y tecnologías geoespaciales ofrece nuevas capacidades para analizar y tomar decisiones frente al comportamiento de los sistemas educativos mediante la integración de fuentes diversas de información. Esta intersección configura un escenario complejo que demanda desarrollar marcos analíticos capaces de articular determinantes socioeducativos de la inclusión educativa con herramientas tecnológicas para una gestión basada en evidencia. En este sentido, el marco referencial cumple una doble función: (i) situar la investigación dentro del marco de las políticas y debates contemporáneos sobre el acceso equitativo y permanencia de la educación superior en entornos digitales; y (ii) identificar los desarrollos técnicos que permiten traducir esos principios en estrategias concretas de planificación institucional.

2.1. Políticas de acceso, QA y servicios integrales de apoyo

En este escenario de transformaciones sociales, económicas y territoriales, donde la diversificación de perfiles estudiantiles tensiona los modos de acceso y permanencia, los instrumentos como el financiamiento adquieren un rol central para sostener trayectorias efectivas. Entre ellos, la gratuidad opera como palanca de redistribución, pero su impacto es condicionado por la presencia de apoyos académicos y psicológicos. En Chile, la evidencia reciente muestra que los esquemas de gratuidad han contribuido a elevar la persistencia de estudiantes de menores ingresos, especialmente en el primer tramo del itinerario formativo [18, 19, 20]. No obstante, su efecto no es automático; cuando la expansión de cobertura no se acompaña de apoyos de permanencia (tutorías, nivelación, bienestar) y de una orientación vocacional informada, emergen desplazamientos en las preferencias de postulación hacia ofertas de menor selectividad y desajustes entre trayectorias esperadas y reales. Por su parte, para Colombia, se reporta que la expansión reciente de la gratuidad ensancha el acceso en Instituciones de Educación Superior (IES) públicas y territorios rezagados; sin embargo, su efectividad exige seguimiento de cohortes para focalizar apoyos de conectividad, manutención y tecnología, para traducir el acceso en permanencia sostenida [21].

De manera congruente, surge otro mecanismo de atención basado en admisión contextual con acción afirmativa por subcuotas (origen escolar, ingreso, raza/etnia) la cual, ataca la barrera de la competencia en terreno desigual, diversificando la composición estudiantil pero sin deteriorar el rendimiento promedio; el reto se desplaza a sostener el ingreso y la permanencia, especialmente en contextos de educación a distancia o modelos híbridos heterogéneos; donde resulta crítico combinar cuotas con diseño instruccional exigente, acompañamiento académico intensivo y apoyos económicos focalizados hasta la titulación [22, 23, 24].

De forma complementaria a las palancas de acceso relacionadas con la gratuidad y admisión contextual, la evidencia sugiere que factores no académicos como los costos de manutención, transporte conectividad/dispositivos, responsabilidades de cuidado y necesidades de bienestar psicosocial, condicionan la traducción del ingreso en permanencia sostenida, especialmente cuando la trayectoria ocurre en modalidades en línea o híbridas. Lo que se traduce en focalizar la atención por parte de las (IES) a fortalecer los servicios de apoyo, mejorar la infraestructura física y digital, y garantizar el acceso equitativo a los recursos académicos y extracurriculares, por otro lado, reforzar los programas

de becas, apoyo psicológico y académico y flexibilidad curricular, con especial atención a los estudiantes en situación de vulnerabilidad y garantizar que el prestigio institucional refleje objetivamente la calidad de la oferta académica [25]

Tras financiamiento y admisión contextual, el Aseguramiento de la Calidad (QA) específico para modalidad en línea es la tercera palanca que convierte el acceso en permanencia. En este contexto un QA con lente de equidad se vuelve condición de efectividad. En Chile, la calidad se redefine como transformación con apoyo e inclusión, desplazando la selectividad como métrica principal. Ello introduce matices de equidad en la definición de calidad online y desplaza el énfasis desde indicadores tradicionales hacia una interacción significativa, retroalimentación oportuna, accesibilidad y resultados desagregados por subgrupos [26]. En Ecuador, la evidencia muestra que, sin un diseño instruccional activo y formación docente, la modalidad en línea se percibe como transmisiva [27] y que la percepción de calidad depende de acceso y servicios de apoyo tanto como de lo académico [26]. Así, el QA con enfoque de equidad opera como la bisagra entre políticas de acceso (gratuidad y subcuotas) y trayectorias efectivas, buscando estandarizar interacciones, feedback y accesibilidad.

A esta mirada, se suma la analítica de aprendizaje como componente operativo del QA, [28] destaca que a partir de trazas del LMS (Learning Management System) durante las primeras semanas, los sistemas de alertar temprana (EWS) permiten anticipar el riesgo (inactividad, no-entradas, caídas de conexión) y activar interacciones proactivas (tutorías, consejerías, apoyos económicos/tecnológicos), lo que es coherente con la evidencia institucional sobre los determinantes académicos-organizativos del abandono y con experiencias internacionales de EWS.

Por otro lado, desde una síntesis operativa y perspectiva de diseño institucional, las asimetrías de conectividad, equipamiento y acompañamiento local son factores agregados que condicionan el acceso y la permanencia en la educación superior. No obstante, cuando la oferta en línea se estructura con diseño pedagógico robusto, y soportes organizativos adecuados, sus resultados pueden ser equiparables con la modalidad presencial [29]. Este escenario, propone un debate sobre la necesidad estructurar una infraestructura de apoyo territorial, y de estrategias híbridas que articulen lo mejor de ambos mundos, apuntando a una flexibilidad digital y anclaje comunitario.

Ese enfoque se territorializa en redes de polo o centros locales, dispuestos por laboratorios, bibliotecas, conectividad y servicios que integran espacios de acceso y acompañamiento, acercando recursos y tutores a contextos con restricciones económicas y geográficas desde una perspectiva multicultural y de justicia social. Este andamiaje territorial se concibe normativamente como estructuras de apoyo pedagógico, tecnológico y administrativo que dispuestos por equipamiento, acompañamiento humano y gestión administrativa in situ, buscan mitigar el posible aislamiento de estudiantes en contextos periféricos. Esta lógica territorial no sustituye el QA, en virtud de que lo opera donde la brecha es mayor, y permite enlazar la inclusión digital (micro-becas de datos, préstamo de equipos, acuerdos de zero-rating del LMS) con apoyo personalizado y trámites universitarios asistidos, reduciendo fricciones de ingreso y permanencia, en especial durante el primer año [30].

En conjunto, estos hallazgos delimitan una arquitectura de equidad con tres pilares: (1) inclusión digital como condición material (datos, equipos, conectividad segura) para un acceso efectivo; (2) aseguramiento pedagógico de la modalidad en línea (interacción, evaluación auténtica, formación docente) respaldado por gobernanza de datos y analítica; y (3) modelos híbridos territorializados mediante hubs/centros de apoyo que acerquen servicios académicos y psicosociales a los estudiantes, con prioridad en primer año y en asignaturas 'troncales'. Este encuadre no solo explica por qué la expansión digital reproduce brechas cuando carece de soportes, sino también cómo se pueden mitigar ya sea, territorializando la calidad y la inclusión a través de hubs/centros de apoyo que conecten

infraestructura, servicios y datos para sostener trayectorias académicas hasta la titulación en estas modalidades.

2.2. Planificación basada en evidencia: analítica, GIS y optimización de cobertura

Paralelamente, la consolidación de entornos educativos mediados por tecnología ha generado un volumen creciente de soluciones técnicas basadas en datos sobre los comportamientos, trayectorias y contextos de los estudiantes, abriendo nuevas posibilidades para comprender fenómenos complejos mediante enfoques cuantitativos y analíticos. La disponibilidad de información institucional, sumada a fuentes externas, ha impulsado el surgimiento de metodologías capaces de integrar múltiples dimensiones del aprendizaje y la participación educativa [31, 32]. Este movimiento traduce las aspiraciones de equidad educativa en capacidades técnicas concretas de diagnóstico y gestión, fortaleciendo el principio de planificación informada por evidencia.

Este enfoque integrado facilita comprender cómo las transformaciones sociales y las innovaciones tecnológicas convergen en un terreno común, dando lugar a un campo de investigación emergente que busca transformar los datos en conocimiento accionable para la planificación, la mejora de la calidad y la innovación educativa. Así, la literatura reciente refleja una transición desde modelos descriptivos hacia enfoques analíticos, que no solo buscan explicar las desigualdades, sino también anticiparlas y mitigarlas mediante intervenciones contextualizadas [33, 34]

En continuidad con el tránsito desde modelos descriptivos hacia analíticas operativas, la evidencia reciente muestra cómo las IES pueden articular políticas (gratuidad, admisión contextual, QA con lente de equidad) con técnicas de ciencia de datos e Inteligencia Artificial que impactan tres frentes: (i) acceso/ingreso a programas en línea (atracción, pronóstico de demanda, admisión justa), (ii) permanencia (detección e intervención temprana), y (iii) equidad de acceso (gobernanza, inclusión digital y territorialización).

En este primer marco, la analítica causal, de la mano con tecnologías de aprendizaje profundo y aprendizaje automático configuran un andamiaje técnico-estratégico para hacer frente a la incertidumbre de ingreso durante los primeros años académicos. El planteamiento de este tipo de estrategias aporta significativamente a la planificación académica enfocada por ejemplo en la habilitación de grupos de estudiantes por asignatura, infraestructura y personal académico. Estudios sugeridos por [35] plantean la implementación de árboles de decisión, bosques aleatorios y redes neuronales de retropropagación sobre datos institucionales y contextuales, para pronosticar matrícula en primero años; en este tipo de escenarios complejos se reportan exactitudes superiores al 60% y $F1 \approx 0,70$, óptimos para orientar dimensionamientos de secciones, planificación docente/infraestructura y la asignación temprana de apoyos tanto económicos y tecnológicos.

Esta inteligencia de acceso se articula con la permanencia cuando los mismos paneles alimentan EWS. Es en este contexto, estudios como el presentado por [36], emplean modelos multinivel (regresiones lineales y logísticas), con lo cual, se lleva a cabo análisis objetivo de disparidades en el desempeño académico de estudiantes en un programa de aprendizaje en línea. Evidenciando hallazgos cualitativos significativos en virtud de visibilizar de que aquellos estudiantes con trastornos emocionales, discapacidades específicas de aprendizaje y otras discapacidades de salud, así como aquellos de bajos ingresos procedentes de ciudades medianas o zonas rurales periféricas, muestran un desempeño significativamente inferior en cursos virtuales, incluso cuando el acceso tecnológico está garantizado.

Estos resultados guardan relación con el análisis previamente realizado, evidenciando que las brechas educativas no se limitan al acceso tecnológico, sino que responden a factores estructurales de acompañamiento y entorno. Es aquí donde los modelos analíticos de carácter territorial y social abordados más adelante juegan un rol importante para identificar desigualdades latentes y orientar estrategias focalizadas.

Articulado al marco técnico, estrategias como las propuestas por [37] plantean un mecanismo computacional eficiente que aborda el problema de la deserción estudiantil bajo dos enfoques concretos, (i) como una tarea de clasificación para predecir quién abandonará los estudios y (ii) como un análisis de supervivencia que busca determinar cuándo ocurrirá la deserción; lo que permite orientar recursos focalizados como conectividad, dispositivos, acompañamiento intensivo en los primeros niveles, buscando reducir brechas ante la presencia de una formación académica deficiente relacionada con la educación secundaria. Este tipo de soluciones, de la mano con los sistemas de alerta temprana (Early Warning System) que combinan trazas de los Learning Management System y mensajería personalizada que incluso, son integradas en sistema de apoyo al aprendizaje (LIS) [28], permiten anticipar riesgos de abandono y plantear mecanismo de intervención temprana para la deserción académica (tutorías, consejería, ayudas tecnológicas). En esta misma línea, [38] contemplan que no basta con solo diagnosticar y prevenir el problema de deserción académica, sino es imprescindible proporcionar posibles justificaciones para el abandono. Mediante la utilización de clasificadores con scores de precisión mayores al 95 % y tasas de Kolmogorov-Smirnov (KS) de hasta el 97 %, combinado con técnicas de explicación de modelos, proponen un enfoque para identificar las posibles razones del abandono académicos en etapas preescolares y secundaria para la toma de acciones preventivas.

Por otro lado, una respuesta efectiva demanda un anclaje territorial con enfoque de equidad. Los avances en ciencia de datos y análisis geoespacial ofrecen herramientas especialmente valiosas para la educación superior, al permitir visualizar patrones de concentración y dispersión de la matrícula, identificar perfiles estudiantiles con base en características compartidas y proyectar escenarios de crecimiento bajo criterios de equidad y eficiencia territorial. El uso de algoritmos de agrupamiento, técnicas de minería de datos y sistemas de información geográfica (GIS) no solo fortalece la capacidad diagnóstica de las instituciones, sino que facilita la toma de decisiones estratégicas sustentadas en evidencia empírica y reproducible [39, 40, 11]. Este acoplamiento metodológico habilita decisiones operativas consistentes con el dimensionamiento de oferta online, posicionamiento y acercamiento estratégico donde la accesibilidad es reducida, articulado con estándares QA y sistemas de alerta temprana.

Si el objetivo es territorializar la equidad, la pregunta clave es: ¿Por qué centrarse en datos geoespaciales? Porque una planificación educativa eficaz debe considerar las especificidades locales a nivel comunitario. El cruce de registros del sistema educativo con información georreferencial posibilita políticas con un carácter contextualizado importante, que alinean la oferta con necesidades reales y bajo enfoque culturales; a mayor equidad en la distribución de oportunidades, mejor adecuación de los servicios y uso más eficiente de los recursos disponibles.

Sobre esta base, el diagnóstico regional propuesto por [41] revela que factores como la capacitación y los recursos limitados, así como el acceso a internet y a la infraestructura, son cuellos de botella para la adopción de tecnologías digitales para la educación en las instituciones de educación superior de la región de América Latina. Esto sumado a una dependencia de ecosistemas de soporte como tutorías, hubs de conectividad, asesoría administrativa y bienestar, refuerzan la necesidad de enfoques que integren el análisis de datos con planificación territorial. Desde esta perspectiva, la aplicación de Sistemas de Información Geográfica (GIS) y modelos de localización-asignación (p-media)

para minimizar tiempos de viaje y problemas de cobertura máxima (MCLP) para identificar hasta n ubicaciones que cubren la mayor demanda posible dentro de un umbral de acceso [42], junto con métricas de accesibilidad efectiva (E2SFCA) orientan la ubicación de hubs/centros de apoyo allí donde la brecha territorial y tecnológica es mayor.

En el plano macro-político, el International Institute for Educational Planning - UNESCO (IIPE) [43] sitúa la equidad, la inclusión y la gestión basada en evidencia como ejes de la planificación para la Agenda Educación 2030. Sus informes bienales subrayan que fortalecer capacidades estatales para planificar y gestionar con datos es condición para traducir políticas en resultados medibles de aprendizaje y cobertura, especialmente en contextos de desigualdad territorial. Esta línea pragmática se establece como un marco referencial para articular estándares de calidad y uso sistémico de evidencia en la toma de decisiones [44, 43]

Aterrizando este marco, el IIPE impulsa la integración de geodatos en la planificación (Data and Evidence | International Institute for Educational Planning, n.d.) para producir diagnósticos contextualizados y orientar micro-planificación (school mapping) a escala local. El principio operativo es “cruzar” registros del sistema educativo con capas espaciales (conectividad, densidad estudiantil, oferta de servicios) para alinear la oferta con necesidades reales, y hacer más eficiente la asignación de recursos. En el plano analítico, el IIPE ha formalizado el uso de Geographically Weighted Regression (GWR) para priorizar espacialmente políticas e intervenciones. El enfoque principal se basa en estimar cómo varía el efecto de los determinantes educativos según el territorio, GWR produce mapas o indicadores que señalan que estratégica académica (nueva sede, focos de acercamiento, transporte) tendrían mayor impacto esperado. La disponibilidad del paper técnico (Isochrone-based catchment areas for educational planning (2022)) y del código reproducible (GitHub - Iiepdev/GWR-in-Educational-Planning) permite escalar esta práctica y acoplarla a componentes operativos específicos. Desde este eje se articula la presente investigación al proponer un marco de análisis impulsado por datos para la identificación de perfiles estudiantiles y la definición estratégica de centros de apoyo híbridos en unidades administrativas locales LAU2 (cantones) con alto potencial de similitud demográfica, de modo que la política de acceso se traduzca en cobertura verificable, persistencia y titulación oportuna.

3. Materiales y métodos

Si bien la inteligencia artificial (IA) es la ciencia general que imita las habilidades humanas, el aprendizaje automático (Machine Learning o ML) es un subcampo específico de la IA enfocado en el desarrollo de algoritmos que permiten entrenar una máquina para aprender y tomar decisiones a partir de datos. Este paradigma se basa en la identificación de patrones y la construcción de modelos como predictivos o descriptivos que mejoran su rendimiento con la experiencia.

La relevancia del ML se enmarca en un contexto más amplio, relacionada con la generación de vastas cantidades de datos en muchos campos, que en cierta forma ha transformado el trabajo del estadístico frente a extraer patrones y tendencias importantes para entender “qué dicen los datos”. De la mano, los desafíos en las áreas de almacenamiento, organización y búsqueda de información han dado paso al campo de la “minería de datos”, estableciendo las bases conceptuales para el desarrollo del ML moderno.

Gracias a este conjunto de cambios y nuevas tecnologías informáticas, el aspecto iterativo del aprendizaje automático ha desarrollado una importancia crucial para su expansión y aplicación en diferentes entornos. A medida que los modelos se exponen a nuevos datos, estos pueden adaptarse de forma independiente a una variabilidad de escenarios, aprendiendo precisamente de estos datos, identificando patrones y tomando decisiones con mínima intervención humana.

Metodológicamente, el ML se clasifica según la naturaleza de los datos de entrada y el objetivo del modelo. Existen cuatro tipos principales de algoritmos de aprendizaje, entre los que podemos encontrar: supervisado, semi-supervisado, no supervisado y de refuerzo. Llegado a este punto, dependiendo del problema que queramos resolver, las preguntas que necesitemos responder y los datos a los que tengamos acceso, podremos enfocar nuestra aplicación en uno en particular.

- Aprendizaje Supervisado: se construye un modelo basado en datos etiquetados e identifica patrones que le permiten realizar predicciones precisas a partir de datos nuevos e inéditos,
- Aprendizaje No Supervisado: se construye un modelo basado en datos no etiquetados, cuyo objetivo principal es descubrir estructuras y patrones latentes o agrupar observaciones similares (clustering)
- Aprendizaje Semi-supervisado: Se sitúa entre los enfoques anteriores, construyendo un modelo basado en una combinación de datos etiquetados y no etiquetados para aprovechar la información estructural.
- Y el Aprendizaje por Refuerzo: Es un método de aprendizaje que se basa en la retroalimentación y un sistema de recompensas y “castigos” para acciones correctas e incorrectas, respectivamente. El objetivo central se basa en que el agente de aprendizaje reciba la máxima recompensa y, por lo tanto, mejore su rendimiento mediante la interacción con un entorno dinámico.

Si bien estos paradigmas de ML ofrecen soluciones distintas a problemas complejos, la elección del enfoque es una decisión estratégica dictada por el objetivo de la investigación y la naturaleza de los datos. En el contexto de este estudio, donde la meta es descubrir y segmentar perfiles socioeconómicos latentes sin una etiqueta de clasificación preexistente, el paradigma más adecuado es el Aprendizaje No Supervisado. Este enfoque permite que el modelo explore la estructura intrínseca de los datos, identificando patrones y similitudes que de otra manera permanecían ocultos. Concretamente, se empleará el análisis de conglomerados (Agrupamiento o Clustering), cuyo objetivo principal es agrupar

las observaciones (los individuos) de tal forma que los miembros dentro de un mismo grupo (cluster) sean más similares entre si (alta cohesión) que los miembros de otros grupos (alta separación). Al aplicar este paradigma, se busca generar los perfiles de los individuos de forma empírica, sentando las bases para una comprensión descriptiva y analítica de la población de la modalidad de estudios en Línea antes de evolucionar a una fase de clasificación.

En esta línea, a continuación en la Figura 2 se presenta una visión general de la arquitectura de este ecosistema analítico y predictivo constituido de manera modular permitiendo la integración de componentes que aporten al entorno de datos con el objetivo de consolidar una propuesta escalable. En esencia esta arquitectura se integra como una cadena de valor que transforma datos censales heterogéneos en (i) perfiles interpretables de la población y (ii) una capacidad de predicción y territorialización que permite escalar dichos perfiles sobre bases masivas y traducirlos en insumos cartográficos para la toma de decisiones.

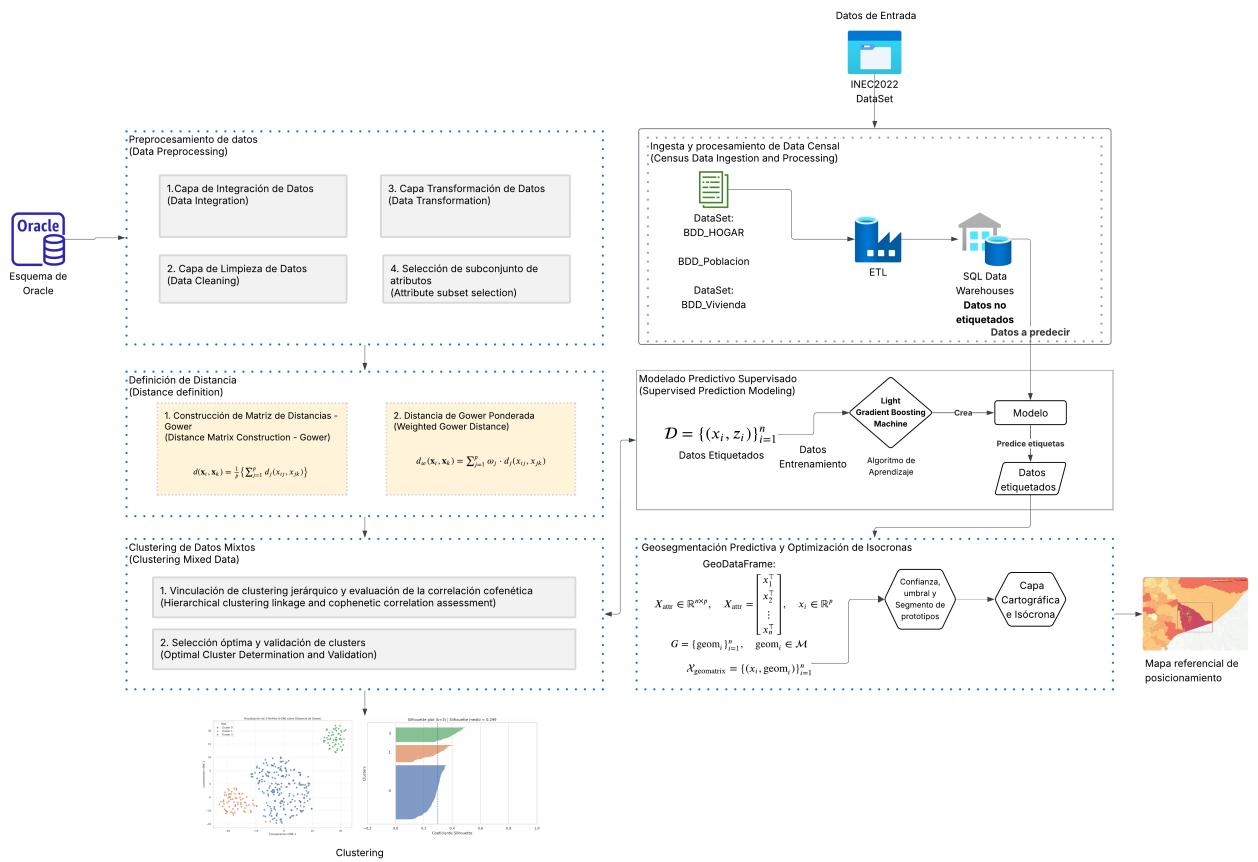


Figura 2: Arquitectura técnica general de la propuesta planteada constituida por 6 etapas

Esta arquitectura parte de dos pilares que constituyen el core del ecosistema. Por un lado, se cuenta con un conjunto de datos integrados y depurados (provenientes de repositorios institucionales de la UPS (Oracle)), y se realiza el trabajo metodológico de segmentación: limpieza, transformación, selección de atributos y definición de una medida de similitud adecuada para datos mixtos (Gower). Por otro lado, se dispone del universo censal INEC 2022, que se procesa mediante un esquema de ingesta

(ETL) hacia un repositorio tipo *data warehouse*, donde los registros se consideran **no etiquetados** en el sentido de que no poseen asignación previa a perfiles.

A partir de esta base, la propuesta articula dos etapas con objetivos distintos pero encadenados:

1. **Segmentación y construcción de etiquetas (clustering).** Se define la distancia de Gower ponderada como mecanismo central para medir disimilitud en presencia de variables categóricas, ordinales y numéricas. Con esta distancia se ejecuta un agrupamiento jerárquico aglomerativo (HCA) y se valida la partición mediante criterios internos. El resultado de esta etapa no es solo una partición, sino la identificación de estos perfiles (etiquetas) que representan una población objetivo basada en características determinadas.
2. **Escalamiento mediante aprendizaje supervisado y territorialización.** Las etiquetas generadas por el clustering se utilizan como variable objetivo para entrenar un modelo supervisado (LightGBM), cuyo propósito es aprender reglas de asignación y aplicarlas posteriormente sobre datos no etiquetados del INEC. La salida del modelo no se limita a la clase predicha, sino que incorpora un nivel de confianza, lo cual permite la aplicación de filtros operativos (umbral) para consolidar únicamente asignaciones optimas. Finalmente, estas predicciones se integran con geometrías cantonales para producir indicadores territoriales (densidad por perfil) y, en un segundo nivel, habilitar análisis operativos como isócronas para evaluar accesibilidad y cobertura.

En consecuencia, las secciones presentadas en este trabajo desarrollan cada bloque de este ecosistema, en el mismo orden lógico partiendo de una base contextual teórica. Primero, el **preprocesamiento y selección de variables 3.2**; luego, la **definición de la distancia y los pesos 3.3** y **agrupamiento jerárquico aglomerativo y criterios de enlace 3.4**; posteriormente, el **refinamiento interno 3.5** para mejorar coherencia interna; y finalmente, el paso a **aprendizaje supervisado 4.2** como resultados logrados, para replicar los perfiles a gran escala y traducirlos en productos territoriales (rankings cantonales, mapas temáticos coropléticos e isócronas 4.3). Esta estructura mantiene la coherencia entre el planteamiento conceptual y los resultados observados, de modo que el lector entienda no solo qué se implementó, sino por qué, cada decisión metodológica resulta clave para obtener una solución consistente, escalable y aplicable en un contexto real.

3.1. Base contextual analítica

3.1.1. Análisis de conglomerados

El análisis de conglomerados (clustering) es un proceso fundamental en el Aprendizaje No Supervisado cuyo objetivo es la partición de un conjunto de objetos de datos (u observaciones) en subconjuntos o grupos. Donde, cada subconjunto es un conglomerado, de tal manera que los objetos en un conglomerado son similares entre sí, pero diferentes de los objetos en otros conglomerados. El conjunto de conglomerados resultante de un análisis de conglomerados puede denominarse agrupamiento. En este contexto, diferentes métodos de agrupamiento pueden generar distintos agrupamientos sobre el mismo conjunto de datos [45], lo que subraya la necesidad de una validación rigurosa. Es crucial notar que este agrupamiento, no se realiza mediante intervención humana directa o reglas predefinidas, sino que es ejecutada por un algoritmo de clustering. Por consiguiente, el principal valor heurístico del clustering reside en su capacidad para facilitar el descubrimiento de grupo y estructuras

previamente desconocidas dentro de los datos, lo que es esencial para la exploración y la generación de información de valor que aporta conocimiento.

La efectividad de esta segmentación radica en el cumplimiento de requisitos básicos que determinan la viabilidad en entornos de datos reales y complejos. Fundamentalmente, los métodos de clustering deben ser altamente escalables; no es suficiente que funcionen en muestras pequeñas, si no que deben ser capaces de procesar bases de datos masivas, evitando resultados sesgados que surgen del simple muestreo. Complementariamente, se exige una capacidad para manejar diferentes tipos de atributos. Si bien muchos algoritmos están diseñados para datos puramente numéricos, las aplicaciones reales requieren agrupar datos mixtos, incluyendo atributos binarios, nominales (categóricos), ordinales o incluso tipos de datos más complejos.

En términos de detección de patrones, es crucial la capacidad de descubrir conglomerados con forma arbitraria, ya que muchos métodos basados en métricas estándar se orientan por formas esféricas, mientras que estructuras reales son frecuentemente irregulares y complejas. Esto se vincula con la necesidad de manejar datos ruidosos; los conjuntos de datos del mundo real contienen intrínsecamente valores atípicos (outliers) y/o datos erróneos, y los métodos de clustering deben ser robustos y no sensibles a estas imperfecciones para garantizar agrupaciones de alta calidad.

Finalmente, la usabilidad y la flexibilidad algorítmica imponen dos requisitos clave: primero, la necesidad de algoritmos que sean insensibles a parámetros de entrada difíciles de determinar (como número de clusters, K), evitando así la sobrecarga al usuario y la sensibilidad del resultado al conocimiento del dominio. Segundo, se necesitan algoritmos que permitan el agrupamiento incremental y que sean insensibles al orden de entrada, permitiendo incorporar actualizaciones de datos recientes en las estructuras existentes sin tener que re-calcular el clustering desde cero [45].

Para gestionar este proceso de partición de manera algorítmica y buscar la optimización de los requisitos mencionados, los algoritmos de clustering heurístico se pueden dividir en dos categorías principales: clustering particional, (ejemplo: K-medias o K-medoids) y jerárquico. En términos generales, los métodos de agrupamiento particional comienzan con una agrupación inicial de observaciones y actualizan iterativa-mente la agrupación hasta encontrar la "mejor" agrupación. Por otro lado, los métodos de agrupamiento jerárquico son de naturaleza más secuencial; construyen una estructura similar a un árbol de agrupamientos anidados (dendograma) mediante fusiones sucesivas de observaciones similares, de acuerdo con una métrica de disimilitud (distancia) definida. Se menciona los métodos basados en cuadrículas (Grid-based methods) que cuantifican el espacio de objetos en un número finito de celdas que forman una estructura de cuadriculas, con beneficios de tiempo de procesamiento. Para este estudio, en consideración del tipo de datos mixtos, las ventajas en cuanto a la versatilidad en medidas de similitud que se pueden implementar, y características deterministas, se opta por la aplicación del método de agrupamiento jerárquico.

3.1.2. Agrupamiento jerárquico

La agrupación jerárquica es otro paradigma de agrupación basado en disimilitudes, versátil y ampliamente utilizado. Aunque también se basa en disimilitudes, la agrupación jerárquica se diferencia de los métodos de agrupación tradicional, como K-means o K-medoides, en que normalmente no utiliza la noción de calcular las distancias a un prototipo central, ya sea un vector de medida centroide o un medoide, sino que construye una jerarquía de agrupaciones basada en las disimilitudes (distancias) entre las propias observaciones y los conjuntos de observaciones.

Por tanto, el agrupamiento jerárquico genera múltiples configuraciones de particiones, que van

desde un único cluster hasta tanto clusters como observaciones. Típicamente, los resultados de una agrupación jerárquica suelen presentarse en forma de una visualización de dendograma, que representa el conjunto de particiones exploradas durante el proceso, facilitando la selección de la configuración de agrupamiento más adecuada. Adicionalmente, el agrupamiento jerárquico presenta ciertos beneficios en comparación con los métodos de particionamiento. En primer lugar, se puede utilizar cualquier medida válida de distancia, por lo que no se limita a distancias euclidianas al cuadrado ni a la agrupación de datos puramente continuos. En segundo lugar, los algoritmos de agrupación jerárquica no requieren el propio conjunto de datos como entrada; basta con una matriz de distancias por pares. Esta característica permite procesar estructuras de datos más complejas, incluso aquellas donde la distancia se calcula mediante funciones no geométricas, como las que miden la conectividad de redes o la similitud de secuencias [46].

En términos generales, hay dos categorías de agrupamiento jerárquico:

- Aglomerativo: Partiendo de la base de la jerarquía, se empieza con cada observación en un grupo propio y se va fusionando sucesivamente pares de grupos a medida que se asciende en la jerarquía, hasta que todas las observaciones estén en un solo grupo. Este enfoque se conoce a veces como anidamiento aglomerativo. (AGNES; [47]) [46].
- Divisivo: Partiendo de la cima de la jerarquía, con todas las observaciones en un clúster, se dividen recursivamente los clústeres a medida que se desciende por la jerarquía, hasta que todas las observaciones se encuentren en un clúster propio. Este enfoque a veces se denominado análisis divisivo) (DIANA; [48]) [46].

No obstante, los enfoques divisivos de agrupamiento jerárquico, tales como DIANA, demandan un esfuerzo computacional considerablemente mayor, incluso al trabajar con volúmenes de datos de tamaño medio. Cabe señalar que también los métodos jerárquicos aglomerativos implican un consumo sustancial de recursos computacionales, y de memoria cuando se manejan grandes cantidades de observaciones.

Al aplicar el agrupamiento jerárquico aglomerativo, es necesario considerar tres aspectos fundamentales. Primero, es la métrica de distancia a emplear; segundo, el criterio de enlace (linkeo) para cuantificar las distancias entre los clusters fusionados a medida que el algoritmo asciende en la jerarquía; y tercero, el criterio para determinar en qué punto se debe dividir el dendrograma resultante con el fin de producir una partición final óptima.

3.1.3. Distancia de Gower

En el ámbito de la minería de datos y el ML, la medición precisa de la disimilitud entre observaciones es crucial. Sin embargo, las métricas tradicionales como la distancia euclídea están intrínsecamente diseñadas para operar únicamente sobre variables continuas. Su aplicación directa a conjuntos de datos con tipos mixtos (numéricos, categóricos) anula la validez geométrica del espacio y conduce a resultados sesgados. Para estudios de perfilamiento poblacional o socioeconómico, donde coexisten atributos como la Edad (continua) y el Género (categórico), la aplicación de métricas estándar es inapropiada. Para resolver este desafío, el presente estudio adopta una de las métricas creadas para conjuntos de datos mixtos, como la Distancia de Gower [49], siendo ampliamente utilizada en la medición de grado de disimilitud entre observaciones cuando existen características de tipo mixto. Donde un conjunto de datos de tipo mixto X con n observaciones tiene p características,

donde las primeras h características son continuas y las restantes, desde $h + 1$ hasta p , son categóricas. Por lo tanto, la distancia de Gower entre dos observaciones $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ip})$ y $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{jk}, \dots, x_{pk})$ del conjunto de datos X ($i, k \in \{1, 2, \dots, n\}$) es:

$$d(\mathbf{x}_i, \mathbf{x}_k) = \frac{1}{p} \left\{ \sum_{j=1}^p d_j(x_{ji}, x_{jk}) \right\} \quad (1)$$

donde $d_j(x_{ji}, x_{jk})$ se define de manera diferente para variables continuas y categóricas:

$$d_j(x_{ji}, x_{jk}) = \begin{cases} \frac{|x_{ji} - x_{jk}|}{R_j}, & \text{si } j \in \{1, 2, \dots, h\}, \\ \mathbb{I}(x_{ji} \neq x_{jk}), & \text{si } j \in \{h + 1, h + 2, \dots, p\}. \end{cases} \quad (2)$$

donde R_j es el rango para el valor de la j -ésima característica continua, con x_{ji} y x_{jk} siendo los valores de la j -ésima característica para las observaciones \mathbf{x}_i y \mathbf{x}_k separadamente. $\mathbb{I}(x_{ji} \neq x_{jk})$ es 1 si $x_{ji} \neq x_{jk}$ y 0 en caso contrario. La configuración por defecto en la distancia de Gower especifica pesos iguales para todas las variables, aunque un vector de importancia variable podría aplicarse a $d_j(\mathbf{x}_i, \mathbf{x}_k)$ como una opción, mismo que será abordado más adelante en este desarrollo [50].

3.1.4. Criterio de enlace

Por otro lado, el agrupamiento jerárquico aglomerativo descansa en dos nociones distintas de disimilitud. La primera es la medida de distancia d (por ejemplo, euclidiana, Manhattan o Gower), utilizada para cuantificar la separación entre pares de observaciones individuales del conjunto de datos. Dado que distintas métricas capturan diferentes geometrías y escalas de variación, la selección de esta puede producir soluciones de agrupamiento notablemente distintas; por ello, es recomendable ejecutar el procedimiento con varias medidas de distancia y contrastar los resultados.

La segunda noción aparece porque, en el enfoque aglomerativo, las observaciones se fusionan sucesivamente para formar conglomerados. En consecuencia, se requiere una forma de medir la disimilitud entre conjuntos de observaciones (clusters) a partir de las distancias por pares entre sus elementos. Esta idea conduce al criterio de enlace (o linkage), que define cómo se computa la “distancia” entre dos clusters. En cada iteración se combinan los dos conglomerados con menor disimilitud según dicho criterio de enlace; dicho criterio define cómo se calcula la distancia inter-cluster (mínimo, máximo, promedio, incremento de varianza, etc.). El criterio de enlace es precisamente la definición de “distancia más corta”, que diferencia a los otros enfoques aglomerativos. Nuevamente, la elección del criterio de enlace puede tener un impacto sustancial en el resultado de agrupación, por lo que se debe evaluar múltiples soluciones con diferentes combinaciones de medida de distancia y de criterio de enlace o lindeo.

Existen varios criterios de enlace de uso común que se describen a continuación:

Complete linkage: Define la disimilitud entre dos clusters como la distancia entre los dos elementos (uno en cada cluster) que están más alejados entre sí según la medida de distancia elegida d [46]:

$$D_{\text{complete}}(A, B) = \max_{a \in A, b \in B} d(a, b) \quad (3)$$

Single linkage: Define la disimilitud entre dos clusters como la distancia entre los dos elementos (uno en cada cluster) que están más cerca entre sí según la medida de distancia elegida d [46]:

$$D_{\text{single}}(A, B) = \min_{a \in A, b \in B} d(a, b) \quad (4)$$

Average linkage: Define la disimilitud entre dos clusters como la distancia promedio según la medida de distancia elegida d entre todos los pares de elementos (uno en cada cluster) [46]:

$$D_{\text{average}}(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (5)$$

Centroid linkage: Define la disimilitud entre dos clusters A y B como la distancia, según la medida de distancia elegida d , entre sus correspondientes vectores centroides μ_A y μ_B [46]. Matemáticamente, se expresa como :

$$D_{\text{centroid}}(A, B) = d(\mu_A, \mu_B) \quad (6)$$

Ward linkage: Se busca minimizar el aumento total en la suma de cuadrados dentro de los clusters, encontrando el par de clusters que, al fusionarse, produzca el menor incremento en la varianza total dentro de los clusters [46]:

$$\frac{|A||B|}{|A| + |B|} \|\mu_A - \mu_B\|^2 = \sum_{x \in A \cup B} \|x - \mu_{A \cup B}\|^2 - \sum_{x \in A} \|x - \mu_A\|^2 - \sum_{x \in B} \|x - \mu_B\|^2 \quad (7)$$

Los criterios de lindeo de Ward y de centroide difieren de los otros criterios de lindeo en que generalmente están diseñados para ser utilizados solo cuando las distancias iniciales entre pares de observaciones son **distancias euclidianas al cuadrado**.

3.1.5. Corte del Dendograma

En el agrupamiento jerárquico, la obtención de una partición final de los datos en clústeres disjuntos se logra mediante un corte horizontal del dendrograma a una altura específica. Este corte determina la asignación de las observaciones a los distintos clústeres. La altura a la que se realiza el corte influye directamente en el número de clústeres resultantes: a menor altura, mayor será el número de clústeres formados, llegando al extremo de tener un clúster por cada observación (es decir, n clústeres). Por el contrario, a mayor altura del corte, menor será el número de clústeres, considerando que puede llegar al caso límite de un único clúster que agrupa todas las observaciones, lo que indicaría la ausencia de una estructura de agrupamiento en los datos.

Una de las principales ventajas del agrupamiento jerárquico es su flexibilidad para determinar el número de clústeres sin necesidad de especificar K (el número de clústeres) a priori. En su lugar, se puede examinar el dendrograma resultante y seleccionar manualmente el nivel de corte que mejor se ajuste a la granularidad deseada, permitiendo así adaptar los resultados según las necesidades del análisis. No existe una regla universalmente aplicable para determinar la altura óptima de corte del árbol, pero es común seleccionar una altura en la región donde existe la mayor separación entre fusiones, es decir, donde existe un rango relativamente amplio de distancias en el que el número de grupos en la partición resultante no varía. Esto, por supuesto, depende en gran medida de la propia visualización. [46]

Sin embargo, ciertas combinaciones de medidas de disimilitud y criterios de enlace pueden generar dendrogramas con características indeseables. En particular, el enlace completo (complete linkage) suele presentar un rendimiento deficiente en presencia de valores atípicos, debido a su dependencia de las distancias máximas entre observaciones. Esto puede llevar a la formación de clústeres poco representativos, donde los valores atípicos distorsionan la estructura general de los grupos.

Por otro lado, el enlace simple (single linkage) tiende a producir un efecto conocido como "encadenamiento" en el dendrograma resultante. Esto ocurre porque el método se basa en las distancias mínimas entre observaciones, lo que genera que las observaciones se unan continuamente a grupos ya existentes, en lugar de fusionarse con otras observaciones para formar nuevos clústeres. Como consecuencia, las observaciones en extremos opuestos de un mismo clúster pueden ser bastante disímiles entre sí, lo que compromete la cohesión interna del grupo.

Estas limitaciones no son exclusivas de un criterio de enlace específico, sino que también pueden atribuirse a la naturaleza misma de la agrupación jerárquica. A diferencia de métodos como K-medias (K-means) o K-medoides, que optimizan objetivos globales para toda la partición, la agrupación jerárquica optimiza un criterio local en cada paso de fusión. Esto puede resultar en soluciones subóptimas desde una perspectiva global, especialmente en conjuntos de datos complejos o con estructuras no esféricas. [46]

3.2. Preprocesamiento y Selección de Variables

La presente investigación lleva a cabo un ejercicio de *clustering* de individuos que se encuentran cursando la educación superior en modalidad en línea a partir de información socio-demográfica, educativa y de acceso a recursos materiales y digitales. El propósito central es identificar perfiles relativamente homogéneos de estudiantes que permitan comprender mejor la heterogeneidad de la población y, para finalmente vincular dichos perfiles con individuos registrados a partir del Censo poblacional 2022, permitiendo realizar un mapeo de posibles nichos de población académica a la cual acercar la educación superior con enfoque en territorio.

Desde una perspectiva metodológica, el análisis se apoya en técnicas de aprendizaje no supervisado, en particular en métodos de agrupamiento (*clustering*) sobre una matriz de distancias construida a partir de variables mixtas.

Se han definido 6 hitos a desarrollar con esta investigación: (i) Definir y justificar el conjunto de variables y su codificación para el análisis de clustering. (ii) Construir una matriz de distancias basada en la distancia de Gower ponderada, acorde al tipo de variables. (iii) Aplicar un algoritmo de agrupamiento jerárquico aglomerativo (HCA) con enlace vinculado sobre la matriz de distancias. (iv) Seleccionar un número de clusters k con base en métricas internas de validación (índice de Silhouette, índice de Davies–Bouldin y Calinski–Harabasz), complementadas por criterios de interpretabilidad. (v) Refinar localmente la partición obtenida mediante Partitioning Around Medoids (PAM), actuando solo sobre individuos particulares (bajo silhouette). (vi) Caracterizar detalladamente los perfiles resultantes y agruparlos en macrosegmentos interpretables.

3.2.1. Fuente de información

La base de datos utilizada para este propósito contiene 375 observaciones (individuos) y un conjunto de 97 variables mixtas (numéricas, ordinales y categóricas nominales, incluidas dicotómicas).

Estas variables agrupan información en bloques temáticos que se refieren a: situación académica actual, composición y vínculos dentro del hogar, características socio-demográficas básicas, identidad cultural y *lingüística*, trayectoria educativa previa, estructura familiar (hijos), discapacidad en distintas dimensiones funcionales, localización geográfica actual y pasada, condiciones de actividad y ocupación laboral, características y materiales de la vivienda, acceso y uso de Tecnologías de la Información y Comunicación (TIC) y otros activos del hogar. Esta estructura modular permite, posteriormente, seleccionar subconjuntos de variables relevantes para distintos objetivos analíticos. Esta información proviene de los registros institucionales de la Universidad Politécnica Salesiana del Ecuador [12] a los cuales se dispone de acceso por medio de acuerdos de confidencialidad adecuados para su uso en rigor académico y de investigación. Estos registros son capturados durante los procesos de admisión de los estudiantes en las distintas modalidades de estudio que oferta la institución. El dataset de información fue consolidado los semestres académicos 2025-2026 y 2026-2026.

A efectos de claridad metodológica, las 97 variables se organizan en bloques conceptuales definidos en el Cuadro 2.

Cuadro 2: Categorización temática de las variables

Bloque	Categoría temática	N. variables
1	Situación académica actual y vínculo con el hogar	3
2	Identidad socio-demográfica básica	3
3	Identidad cultural y lingüística	3
4	Trayectoria educativa previa	2
5	Estado conyugal y estructura familiar	2
6	Discapacidad funcional (dificultades permanentes)	6
7	Localización geográfica actual (provincia, cantón, parroquia)	26
8	Condición de actividad y ocupación laboral	5
9	Residencia hace cinco años (provincia y cantón)	25
10	Ubicación geográfica de la vivienda actual	1
11	Condiciones físicas y tenencia de la vivienda	6
12	TIC, activos del hogar y uso de dispositivos	15
Total		97

Este dataset original es de alta dimensionalidad (97 variables) y cubre un espectro amplio del entorno del estudiante. Considerando esta dimensión de información, se lleva a cabo una fase de selección, reducción y transformación de variables antes de proceder a la aplicación del clustering.

3.2.2. Ingeniería de características

La aplicación o vinculación de las 97 variables no resulta metodológicamente razonable ni computacionalmente eficiente. Para ello, se realiza una fase previa de exclusión de variables, en consideración de:

- Varias variables ya captan variantes de una misma dimensión (por ejemplo, múltiples indicadores geográficos a nivel de cantón, tanto actual como hace 5 años; varias preguntas sobre el uso y disponibilidad de TIC).
- El interés central es caracterizar perfiles de estudiantes según condiciones personales, familiares, laborales, de discapacidad y de acceso a activos, más que describir en detalle la movilidad geográfica o la estructura de la vivienda.

- Un número excesivo de variables, muchas de ellas altamente correlacionadas o marginales respecto al objetivo, tiende a introducir ruido y dificulta la interpretación de los resultados.

En esta línea se evaluó el grado de asociación interna entre las variables candidatas (35) con el fin de identificar grupos de ítems altamente redundantes (especialmente dentro del bloque de Ocupación, TIC y activos), buscando reducir la dimensionalidad del conjunto de entrada al modelo de clustering y evitando incluir variables que aportan la misma información.

Dado que el conjunto preliminar incluye tanto variables categóricas nominales/dicotómicas como variables numéricas/ordinales, para este análisis se utilizaron dos medidas complementarias:

- El estadístico V de Cramer para pares de variables categóricas;
- El coeficiente de correlación de Spearman para pares de variables continuas u ordinales.

3.2.3. Asociación entre variables categóricas: V de Cramer

Para evaluar la fuerza de asociación entre variables categóricas nominales X y Y , se utilizó el estadístico V de Cramer, definido a partir del chi-cuadrado de independencia:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(r, c) - 1)}} \quad (8)$$

donde:

- χ^2 es el estadístico de la prueba de independencia en la tabla de contingencia $r \times c$ de X y Y ,
- n es el tamaño muestral,
- r y c son el número de categorías de X y Y , respectivamente.

Por construcción, $V \in [0, 1]$, donde valores cercanos a 0 indican asociación débil o nula, y valores cercanos a 1 indican una asociación más fuerte entre las dos variables.

Se calculó la matriz completa de V de Cramer para todas las variables categóricas nominales inicialmente disponibles en los bloques sociodemográfico, discapacidad, vivienda y TIC. El resultado se representó mediante un mapa de calor que permite visualizar patrones de asociación y detectar grupos de variables fuertemente correlacionadas (Figura 3).

En esta primera matriz, se puede identificar en particular una alta asociación dentro del bloque de TIC entre variables de "disponibilidad" y sus correspondientes variables de "uso" (ejemplo: teléfono celular, datos móviles, internet fijo, computadora, tablet), lo que sugiere una alta redundancia semántica. Por otro lado, se presentan asociaciones moderadas entre Q89InternetFijo, Q91Computadora y Q95Auto, reflejando un nivel común de material/tecnológico dentro del hogar.

A partir de esta matriz preliminar en bloques con las relaciones entre la disponibilidad y el uso fuertemente asociadas (V alto), se mantuvo solo la variable de disponibilidad (ejemplo: Q84Celular, Q89InternetFijo, Q91Computadora) y se descartaron las variables puramente de uso, para evitar duplicar la misma información o variables demasiado correlacionadas evitando multicolinealidad conceptual. Así también, se excluyeron variables con muy baja relevancia para el objetivo de segmentación

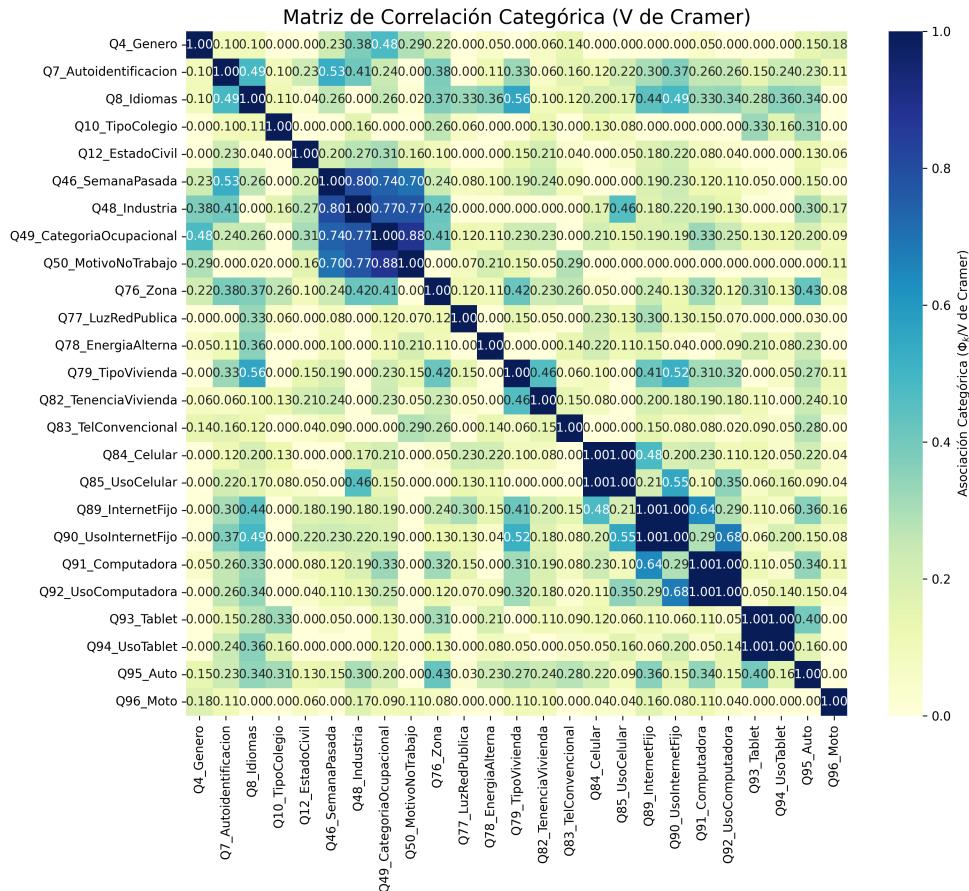


Figura 3: Matriz de asociación categórica (V de Cramer) entre las variables preliminares.

(por ejemplo, detalles de tenencia de teléfono convencional, tablet o motocicleta), priorizando una representación concisa del patrimonio.

Como paso previo al cálculo de distancias se aplicaron varias transformaciones, entre las que se destaca:

- **Recodificación ordinal:** La variable Q11_NivelInstrucción se recodificó en una escala ordinal que respeta el orden natural de los niveles educativos (por ejemplo, 0 = primaria incompleta, . . . , 4 = superior completa). Esta recodificación permitió tratar dicha variable de forma coherente con la distancia de Gower para ordinales.
- **Dicotomización:** A partir del conteo discreto de hijos, se derivó el constructor ordinal Q100. Esto se aplica con el propósito de captar el nivel de responsabilidad familiar del individuo en lugar de la cifra exacta. Se establecieron tres niveles (0: Ausencia, 1: Moderada, 2: Numerosa) para identificar los umbrales donde la carga familiar condiciona al individuo.
- **Síntesis de discapacidad:** El conjunto de seis variables sobre discapacidad o dificultades permanentes se consolidó en una variable de síntesis Q99_Tiene_Discapacidad (0 = no presenta ninguna discapacidad; 1 = presenta al menos una dificultad), en función de segmentar por condición de discapacidad sin entrar en detalles de los subtipos específicos.

- Homogeneización de Activos e Índice de Acceso Tecnológico:** Las variables de acceso Q84, Q89, Q91, Q95 y la condición laboral Q98 se estandarizaron a formato binario 1 = dispone/trabaja, 0 = ausencia. A partir de este bloque, se consolidó el indicador Q101_IAT_Acceso mediante la puntuación compuesta de la tenencia de celular, internet fijo y computadora. Este cálculo permite generar una escala de acceso digital de 0 a 3, cuyo propósito es definir con claridad nivel de acceso tecnológico del individuo. Lo que se busca con esto es que el modelo pueda inferir la brecha entre la exclusión total(0) y el acceso tecnológico total(3).
- Geografía:** Para este escenario inicial de identificación de perfiles, no se consideran las variables relacionadas a la procedencia y residencia con el objetivo de no sesgar los perfiles a regiones concretas, únicamente se captura el posicionamiento territorial rural o urbano.

Posterior a la depuración realizada, se re-calcularó la matriz de V de Cramer únicamente para las variables categóricas candidatas a permanecer en el modelo de clusterización. El nuevo mapa de calor muestra un conjunto más compacto de variables con asociaciones moderadas pero no tan altas entre sí (Figura 4).

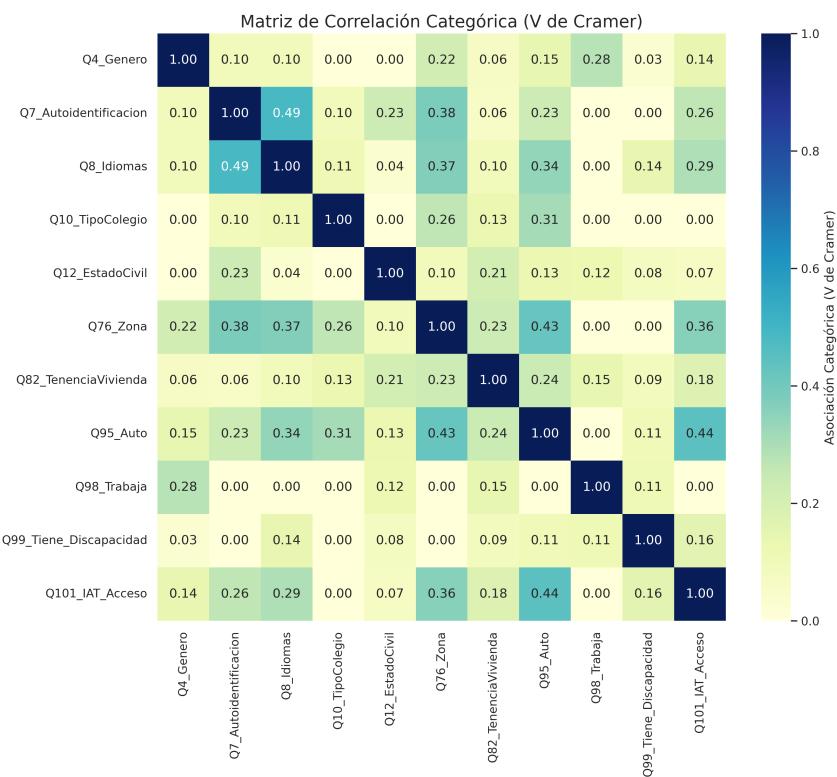


Figura 4: Matriz reducida de asociación categórica (V de Cramer) para las variables seleccionadas

3.2.4. Asociación de Variables Ordinales y Continuas: Coeficiente de correlación de rangos de Spearman

Se empleó el Coeficiente de Correlación de Spearman como una medida de asociación no paramétrica para evaluar la dependencia monótona entre variables ordinales y continuas. En este punto,

se descartó el uso del coeficiente de Pearson en favor de la capacidad de Spearman para detectar dependencias monótonas no lineales y en su inherente robustez ante distribuciones no paramétricas y la presencia de valores atípicos (*outliers*) [51].

En esta línea, el resultado de este cálculo expuesto en la Figura 5, evidencia que la asociación de 0.55 entre la edad Q5 y la carga familiar Q100 valida la estructura demográfica del conjunto de datos, mientras que la independencia observada del nivel de instrucción Q11 sugiere una segmentación donde el capital académico no es el motor primario en el crecimiento del hogar según los datos recabados en este estudio.

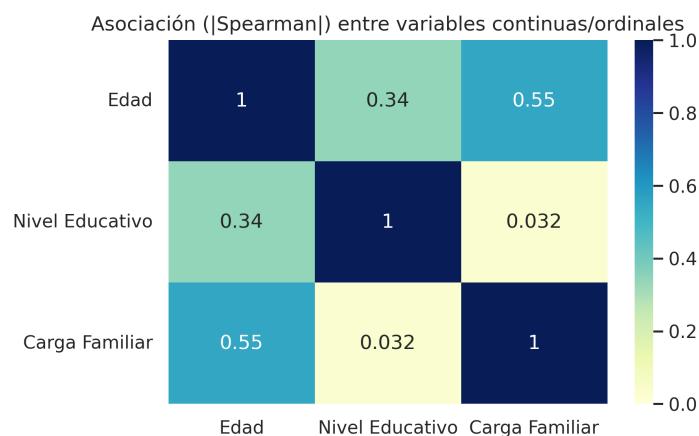


Figura 5: Matriz de coeficiente de Spearman para variables ordinales y continuas

Estas transformaciones permiten trabajar con un conjunto de 14 variables mezcla de numéricas, ordinales y categóricas (incluidas dicotómicas), preparadas para el cálculo de una medida de disimilitud, detallados en el Cuadro 3.

En resumen, la selección de variables se definió a priori con base en el objetivo del estudio y el conocimiento experto del dominio. Para evitar duplicidad de información y multicolinealidad conceptual, se evaluó la asociación entre variables categóricas y ordinales mediante la V de Cramér y la correlación de Spearman, consolidando o excluyendo variables altamente redundantes. Adicionalmente, se aplicó una recategorización de modalidades poco frecuentes para mejorar la estabilidad del análisis.

El Análisis de Componentes Principales (PCA) se descarta en este estudio debido a su incompatibilidad con la naturaleza predominantemente categórica de las variables. El PCA está diseñado para datos numéricos continuos, y su aplicación en variables categóricas requeriría codificaciones binarias o dummies, lo que introduciría varianza artificial y afectaría la interpretabilidad de los resultados. En su lugar, se priorizó un enfoque basado en la V de Cramér y la correlación de Spearman, que permiten cuantificar la fuerza de asociación entre variables sin alterar su naturaleza original.

Aunque se destaca una posible aplicación del Análisis Factorial de Datos Mixtos (FAMD) como una técnica recomendada para estudios con variables mixtas (categóricas y numéricas), en este caso, la justificación inicial permite cubrir y respaldar la selección de las 14 variables. Sin embargo, se reconoce que el FAMD podría ser una herramienta valiosa en futuros análisis para explorar la estructura subyacente de los datos de manera más profunda.

Según lo expuesto, se logra con estas 14 variables constituir el núcleo sólido de la segmentación, ya que capturan aspectos clave como la fase del ciclo de vida (edad), el capital educativo, la estruc-

Cuadro 3: Diccionario de las variables del dataset final

Índice	Variable	Tipo de dato	Descripción
0	Q4_Genero	Categoría nominal	Género declarado por la persona encuestada.
1	Q5_Edad	Numérica (entero)	Edad en años cumplidos.
2	Q7_Autoidentificacion	Categoría nominal	Autoidentificación étnica/cultural declarada.
3	Q8_Idiomas	Categoría nominal	Idioma(s) declarados por la persona.
4	Q10_TipoColegio	Categoría nominal	Tipo de institución educativa de procedencia (según clasificación del INEC).
5	Q11_NivelInstruccion_ord	Categoría ordinal (codificada numéricamente)	Máximo nivel de instrucción completado, codificado en orden creciente.
6	Q12_EstadoCivil	Categoría nominal	Estado civil declarado.
7	Q76_Zona	Categoría nominal	Zona de residencia (urbana/rural).
8	Q82_TenenciaVivienda	Categoría nominal	Régimen de tenencia de la vivienda (propia, arrendada, prestada, etc.).
9	Q95_Auto	Categoría nominal (binaria)	Tenencia de automóvil (sí/no).
10	Q98_Trabaja	Categoría nominal (binaria)	Condición laboral actual (trabaja: sí/no).
11	Q99_Tiene_Disapacidad	Categoría nominal (binaria)	Reporte de condición de discapacidad (sí/no).
12	Q100_Carga_Familiar_ord	Categoría ordinal (codificada numéricamente)	Nivel de carga familiar, codificado en orden creciente.
13	Q101_IAT_Acceso	Índice ordinal / numérico discreto	Índice de acceso tecnológico construido a partir de activos y/o conectividad (mayor valor implica mayor acceso).
Total de variables:			14

tura familiar (hijos), la inserción en el mercado laboral, la presencia de discapacidad, la localización territorial agregada y el capital tecnológico/material del hogar, que permite responder a lo planteado en el Objetivo secundario 1 de esta investigación (Sección 1.2.2).

3.3. Distancia de Gower ponderada para variables mixtas

El problema de fondo consiste en definir una función de distancia $d(i,j)$ entre pares de individuos i y j que sea coherente con la naturaleza de cada variable (numérica, ordinal, nominal, dicotómica), capaz de incorporar pesos para enfatizar dimensiones clave que resalte el enfoque del estudio, y computable en presencia de valores faltantes.

Para ello se utilizó la distancia de Gower ponderada como medida de disimilitud entre individuos.

Sobre la base de la formulación general de la distancia de Gower presentada anteriormente, este estudio hace uso de una versión *extendida y ponderada* que permite: (i) asignar una importancia relativa diferenciada a cada variable, y (ii) tratar de forma explícita las variables ordinales y dicotómicas, además de las continuas y categóricas nominales.

Asignación de pesos por Normalized Mutual Information (NMI) En la expresión original de Gower, todas las características contribuyen de forma equiponderada al promedio:

$$d(\mathbf{x}_i, \mathbf{x}_k) = \frac{1}{p} \sum_{j=1}^p d_j(x_{ji}, x_{jk}) \quad (9)$$

Lo cual equivale a asumir que todas las variables son igualmente relevantes para la medición de disimilitud. Sin embargo, en estudios de perfilamiento poblacional existen atributos que, por su naturaleza, se desea que influyan más en la distancia final (por ejemplo, nivel de estudio, condición laboral o discapacidad, en comparación con la tenencia de un automóvil). Para aplicar este concepto, se incorpora un vector de pesos específicos por variable $\omega = (\omega_1, \omega_2, \dots, \omega_p)$, que permite ajustar la contribución de cada variable en el cálculo de disimilitud. En lugar de definir los pesos de manera discrecional, por juicio de experto, o asumir la ponderación uniforme, en este estudio se adopta un criterio basado en la Información Mutua Normalizada (NMI, por sus siglas en inglés), adoptando el enfoque propuesto por [50].

La motivación principal de esta adopción se constituye por la determinación de variables con mayor capacidad de organizar la estructura multivariada en consideración con otras dimensiones del perfil, las cuales, debería ejercer cierto peso relativo sobre la noción de distancia (en concordancia con el objetivo del estudio a realizar).

Adicionalmente, se incorpora un indicador de comparabilidad δ_{ijk} que vale 1 si la característica j es observada y comparable para las observaciones \mathbf{x}_i y \mathbf{x}_k , y 0 en caso contrario (por ejemplo, en presencia de valores faltantes). En esta línea, la versión ponderada de la distancia de Gower se constituye por:

$$d(\mathbf{x}_i, \mathbf{x}_k) = \frac{\sum_{j=1}^p w_j \delta_{ijk} d_j(x_{ji}, x_{jk})}{\sum_{j=1}^p w_j \delta_{ijk}}. \quad (10)$$

La forma de $d_j(\cdot, \cdot)$ sigue siendo la descrita en la sección 3.1.3, diferenciando entre variables continuas y categóricas, pero ahora cada dimensión contribuye en proporción a su peso w_j .

La expresión (10) se reduce al caso no ponderado cuando $w_j = 1$ para todo j , y $\delta_{ijk} = 1$ en ausencia de valores faltantes.

En este contexto, el presente estudio en complemento con el análisis NMI, consideró para la definición del vector de pesos que ciertas dimensiones estructurales tengan mayor influencia en la distancia total. Bajo este criterio:

- La variable de *Q11_NivelInstruccion_ord* recibe un peso más alto en la ponderación, en coherencia con el objetivo del estudio y en consideración del enfoque de capital humano, que a diferencia de los activos materiales (como tener celular o un auto), el capital académico es un atributo estructural y de largo plazo. Esto evita que el algoritmo genere clústers superficiales basado en solo la posesión de objetos; y favorece la identificación de segmentos de la población académica que accede a la educación superior bajo una modalidad específica, permitiendo que el resto de variables maticen el perfil.

En consecuencia, la asignación de pesos w_j para cada variable se definió tomando como referencia (i) el resultado de aplicación del componente de cálculo basado en NMI, y (ii) un ajuste por criterio experto orientado a enfatizar dimensiones de carácter estructural alineados con el objetivo del estudio.

Por ende, los pesos asignados constituyen el resultado combinado del análisis NMI y de un ajuste sobre variables concretas. Este esquema de ponderación se formaliza en el cuadro 4 y se integra en el cálculo de la distancia de Gower, de manera que se determinan variables con mayor incidencia en la distancia total entre individuos, en coherencia con el objetivo del estudio.

Cuadro 4: Pesos w_j asignados por variable en la distancia de Gower

Peso	Variable(s)
2.5	Q11_NivelInstruccion_ord
2.0	Q101_IAT_Acceso
1.8	Q98_Trabaja
1.5	Q100_Carga_Familiar_ord
1.2	Q5_Edad
1.0	Q82_TenenciaVivienda
0.8	Q10_TipoColegio, Q95_Auto
0.5	Q4_Genero, Q12_EstadoCivil, Q76_Zona
0.4	Q7_Autoidentificacion, Q8_Idiomas
0.1	Q99_Tiene_Disapacidad

Los pesos obtenidos mediante NMI 4 muestran qué tan relacionada está cada variable con el nivel de instrucción, que se usó como variable ancla. En este sentido, un valor más alto de w_j indica que la variable X_j comparte más información con Q11_NivelInstruccion_ord; por tanto, tiende a diferenciar mejor perfiles asociados a trayectorias educativas y al eje de capital humano. Por el contrario, un peso bajo sugiere que la variable aporta poca información adicional para distinguir diferencias a lo largo de este eje, ya sea porque su relación con la educación es débil, porque varía poco en la muestra o porque sus categorías están muy dispersas.

Cabe señalar que, al utilizar Q11_NivelInstruccion_ord como variable ancla para el cálculo de NMI, los pesos estimados priorizan intencionalmente la discriminación asociada al eje educativo; por tanto, variables con baja NMI no se interpretan como irrelevantes en términos generales, sino que son menos informativas respecto a dicho eje.

Tratamiento de variables ordinales En la formulación original de Gower se distingue entre continuas y categóricas. Sin embargo, en muchos contextos aplicados, algunas variables categóricas poseen un orden natural en sus categorías (por ejemplo, nivel de instrucción). Estas características no son propiamente continuas, pero tampoco deberían tratarse como nominales puras, dado que ello ignoraría información valiosa sobre el orden. Sea Q11_NivelInstruccion_ord una variable ordinal con $L_j + 1$ categorías ordenadas, recodificadas internamente como $\tilde{x}_{ji} \in \{0, 1, \dots, L_j\}$ para la observación i . En apego a la literatura [46, 52], se procede a:

1. Mapear las categorías a una escala de rangos \tilde{x}_{ji} que respete el orden.
2. Normalizar estos rangos a la unidad:

$$z_{ji} = \frac{\tilde{x}_{ji}}{L_j}, \quad z_{ji} \in [0, 1]. \quad (11)$$

3. Aplicar la distancia absoluta normalizada, análoga al caso continuo:

$$d_j(x_{ji}, x_{jk}) = |z_{ji} - z_{jk}|. \quad (12)$$

De este modo, las diferencias entre niveles educativos más alejados (por ejemplo, educación básica frente a educación superior) se reflejan en valores de disimilitud mayores que aquellas entre niveles contiguos, sin aplicar las restricciones cuantitativas de una variable continua.

Tratamiento de variables dicotómicas En la base analizada existen múltiples variables dicotómicas (ejemplo, tener o no tener hijos, trabajar o no, presentar o no discapacidad, disponer o no de ciertos activos). Desde la perspectiva de Gower, estas pueden tratarse como categóricas binarias. En este trabajo se adopta un tratamiento *simétrico* de las dicotómicas, lo que implica:

$$d_j(x_{ji}, x_{jk}) = \begin{cases} 0, & \text{si } x_{ji} = x_{jk} \in \{0, 1\}, \\ 1, & \text{si } x_{ji} \neq x_{jk}. \end{cases} \quad (13)$$

Es decir, dos individuos son completamente similares en una variable dicotómica si comparten la misma categoría (tanto si esta es “sí” como “no”), y completamente disimilares si difieren. Esta elección es coherente con el objetivo de medir diferencias en la *configuración global* de atributos más que en la presencia aislada de una categoría específica.

Cabe señalar que, en aplicaciones donde se desea penalizar de forma diferencial ciertas coincidencias (por ejemplo, coincidencias en la presencia de una condición indeseable), es posible emplear formulaciones de Gower para variables binarias asimétricas. No obstante, en este estudio, la diferenciación sustantiva se realiza principalmente a través del esquema de pesos (??) y no mediante un tratamiento asimétrico de las dicotómicas.

Tratamiento de valores faltantes Un aspecto práctico relevante en la construcción de la matriz de disimilitudes y en los análisis complementarios (como el cálculo de la V de Cramer) es el tratamiento de los valores faltantes (NA). Una estrategia explícita para las variables categóricas y dicotómicas, consistente con el objetivo de conservar la información sobre patrones de no respuesta. Donde para ciertas variables categóricas y dicotómicas, los valores faltantes se recodificaron a una categoría adicional con código 99, interpretada como No respuesta / No aplica. En esta línea, la recodificación se aplicó tanto en los análisis de asociación entre variables categóricas (por ejemplo, tablas de contingencia y V de Cramer) como en la construcción final de la matriz de distancias de Gower.

Desde el punto de vista de Gower, esta decisión implica tratar la categoría 99 como una categoría nominal más dentro del dominio de la variable. Para una variable categórica j con dominio extendido

$$\mathcal{C}_j = \{c_{j1}, \dots, c_{jK_j}, 99\}, \quad (14)$$

la distancia parcial se define como en el caso nominal:

$$d_j(x_{ji}, x_{jk}) = \begin{cases} 0, & \text{si } x_{ji} = x_{jk}, \\ 1, & \text{si } x_{ji} \neq x_{jk}, \end{cases} \quad x_{ji}, x_{jk} \in \mathcal{C}_j. \quad (15)$$

En particular, esto implica dos consideraciones relevantes:

- Dos observaciones que comparten la categoría 99 en la misma variable son consideradas completamente similares en esa dimensión ($d_j = 0$), lo que refleja un patrón coincidente de no respuesta.
- Una observación con 99 y otra con una categoría válida (por ejemplo, “Usa datos móviles” o “No usa datos móviles” en una variable de conectividad) son consideradas completamente disímiles en esa dimensión ($d_j = 1$).

No obstante, es importante aclarar que esta formulación actuó principalmente como *mecanismo de robustecimiento* en la fase de exploración y diagnóstico, y no como componente activo en la fase final de clusterización. En concreto, el flujo metodológico fue el siguiente:

1. En la fase exploratoria se calcularon medidas de asociación entre variables categóricas (por ejemplo, V de Cramer) y se evaluó la colinealidad y redundancia entre variables, *incluyendo* las categorías 99 en las tablas de contingencia.
2. A partir de estos análisis, se llevó a cabo una depuración del conjunto de variables, eliminando aquellas con alta redundancia o fuerte colinealidad con otras, escasa variabilidad informativa, o patrones de no respuesta que las hacían poco útiles desde el punto de vista sustantivo.
3. Como consecuencia de esta depuración, las **columnas que contenían de forma relevante la categoría 99 quedaron excluidas del conjunto final de variables utilizadas para calcular la matriz de distancias de Gower y, por tanto, no participaron en la construcción de la matriz D ni en la clusterización final.**

En la práctica, tras la etapa de selección de variables basada en V de Cramer, colinealidad y criterio sustantivo, las variables donde la categoría 99 era operativamente relevante *no fueron retenidas* en el set final de variables de clustering.

Concretamente las variables no vinculadas fueron:

- Q48_Industria, que indica el detalle del campo desarrollo de las actividades laborales del individuo, con un 18.4 % de valores faltantes.
- Q49_Q49_CategoríaOcupacional, que indica el rol que ocupa el individuo en la actividad laboral principal, con un 18.4 % de valores faltantes.
- Q50_MotivoNoTrabajo, que indica la razón por las que no registra actividad laboral, con un 81.6 % de valores faltantes.
- Q85_UsoCelular, que indica la frecuencia de uso del celular, con un 14.9 % de valores faltantes.
- Q90_UsoInternetFijo, que indica la frecuencia de uso de internet fijo, con un 16.8 % de valores faltantes.
- Q92_UsoComputadora, que indica la frecuencia de uso de computador, con un 22.7 % de valores faltantes.

- Q94_UsoTablet, que indica la frecuencia de uso de tablet, con un 88.3 % de valores faltantes.

Se deduce que la mayoría de estas variables aportan detalles específicos que podrían ser útiles en estudios centrados en el mercado laboral o el uso de tecnología. Sin embargo, dado el alto porcentaje de valores faltantes y la falta de alineación con el enfoque de este estudio, se procede a excluirlas o limitar su uso a análisis secundarios.

Construcción de la matriz de distancias A partir de la ecuación (10) y de las definiciones anteriores para $d_j(\cdot, \cdot)$, se construyó la matriz de distancias

$$D = [d(\mathbf{x}_i, \mathbf{x}_k)]_{i,k=1}^n, \quad (16)$$

de dimensión $n \times n$ (en este caso, $n = 375$), simétrica y con ceros en la diagonal, que sintetiza la disimilitud entre todos los pares de individuos.

Esta matriz constituye el insumo único para las fases posteriores del análisis: (i) el *agrupamiento jerárquico aglomerativo* con distintos criterios de enlace (single, complete, average), (ii) el refinamiento local mediante *k-medoids (PAM)* sobre distancias precomputadas, (iii) la visualización de la estructura de proximidad mediante *t-SNE* con métrica precomputada, (iv) y el cálculo de métricas de validación internas (índice de Silhouette, índice de Davies–, índice de Calinski–Harabasz).

De esta forma, toda la metodología de clustering descansa sobre una definición coherente y flexible de disimilitud entre individuos, apropiada para datos de tipo mixto y ajustada a la relevancia sustantiva de las variables incluidas.

3.4. Agrupamiento jerárquico aglomerativo y criterios de enlace

Con la matriz de distancias D como entrada, se aplicó un **agrupamiento jerárquico aglomerativo (HCA)** que parte de considerar cada individuo como un cluster singleton y, en cada iteración, fusiona los dos clusters más próximos según una función de distancia entre clusters. Dados dos clusters C_a y C_b , se evaluaron tres criterios clásicos de enlace single linkage, complete linkage y average linkage.

Evaluación de los criterios de enlace Para evaluar qué criterio representaba mejor la estructura de distancias original, se calculó la **correlación cofenética** entre las distancias de Gower y las distancias inducidas por el dendrograma para cada método de enlace. La correlación cofenética (ρ_c) se calcula como el coeficiente de correlación de Pearson entre las distancias $d(i, j)$ originales y las distancias cofenéticas $d_c(i, j)$ derivadas del dendrograma:

$$\rho_c = \frac{\text{Cov}(d(i, j), d_c(i, j))}{\sigma_{d(i, j)} \cdot \sigma_{d_c(i, j)}}. \quad (17)$$

Los valores de la correlación cofenética para los métodos de enlace evaluados fueron los siguientes: **Single linkage:** 0.59; **Complete linkage:** 0.58; **Average linkage:** 0.65.

En este estudio se adoptó el método de agrupamiento jerárquico aglomerativo con criterio de enlace promedio (average linkage). La elección se fundamenta en su comportamiento más estable y menos sensible a observaciones atípicas en comparación con alternativas como enlace simple y enlace completo.

3.4.1. Selección del número de clusters k

Una vez definida la medida de disimilitud mediante la distancia de Gower ponderada y aplicado el agrupamiento jerárquico aglomerativo (HCA) con enlace promedio sobre la matriz de distancias D , el siguiente paso consiste en determinar el número apropiado de clusters k para cortar el dendrograma y obtener una partición útil de la población.

El dendrograma resultante del HCA sintetiza la estructura jerárquica de las fusiones sucesivas entre observaciones y grupos. En la Figura 6, se presenta el dendrograma obtenido a partir de la matriz de distancias de Gower, utilizando el criterio de enlace promedio. Cada fusión está representada por una unión horizontal cuya altura refleja la disimilitud (distancia) a la que se combinan los clusters implicados siendo un recurso de referencia para la definición del K óptimo.

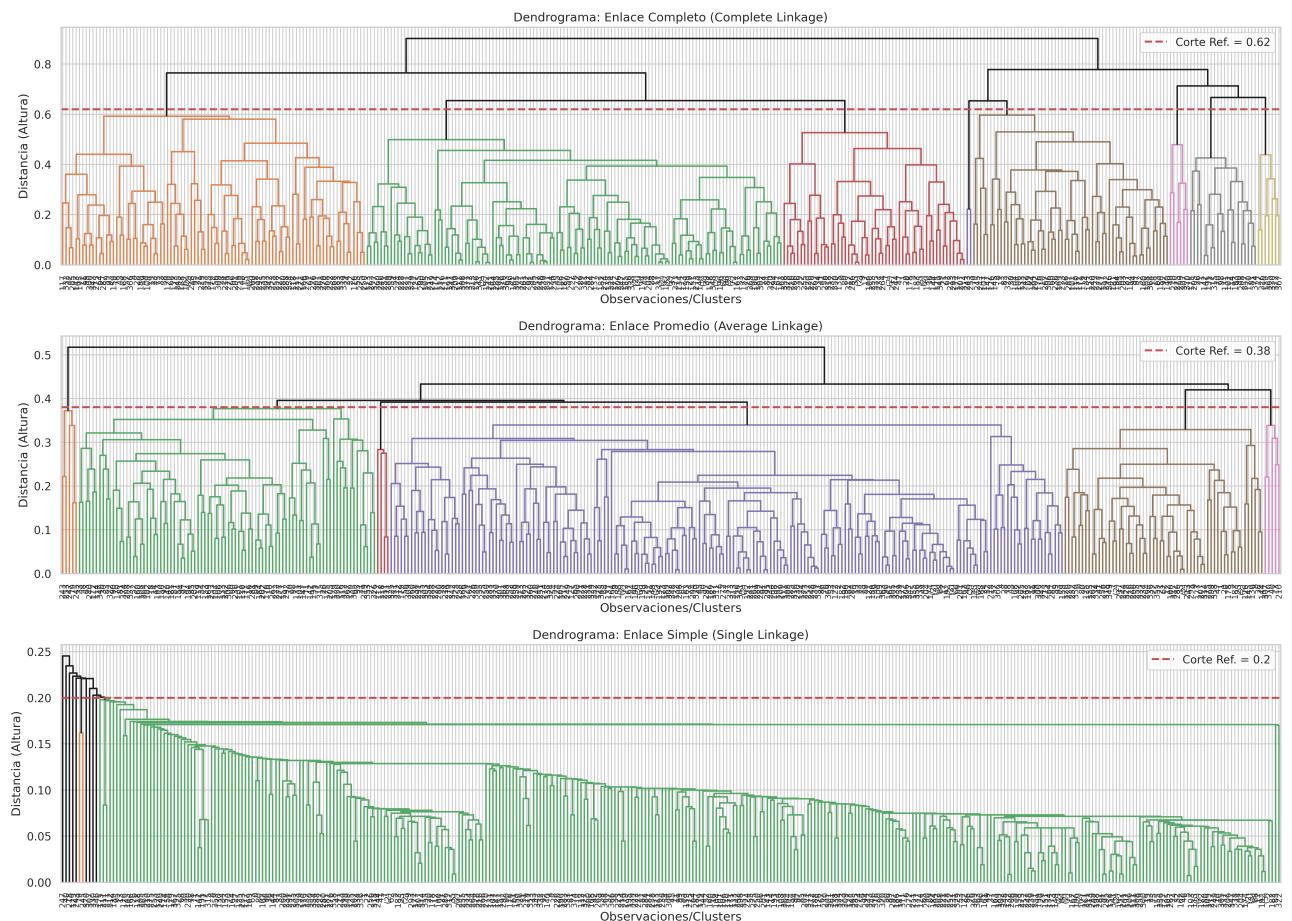


Figura 6: Dendrograma del agrupamiento jerárquico aglomerativo con enlace promedio sobre la matriz de distancias de Gower².

Desde un punto de vista cualitativo, el dendrograma muestra cómo se van formando los grupos a medida que se permite mayor ‘distancia’ entre individuos. Al usar el enlace promedio, se observa que la población tiende a agruparse inicialmente en pocos bloques grandes y estables. Esto es importante

²Versión en alta resolución disponible en: <https://github.com/coyolaf/TFMC0yola-UOC-2025/blob/main/DendrogramaHCA.png> (último acceso: 31/12/2025)

porque indica que existe una estructura principal de segmentación, sin depender de casos extremos o perfiles muy raros. Relizar un corte demasiado alto puede resultar en menos cantidad de clusters, excesivamente agregados; cortar demasiado bajo produce un número elevado de clusters pequeños, que pueden no aportar mucha información relevante al estudio.

Por otra parte, al trabajar con una distancia de Gower ponderada y con un conjunto amplio de variables categóricas y ordinales, es normal que aparezcan perfiles poco frecuentes o casos aislados que, en la práctica, se comporten como valores atípicos dentro del análisis. Cuando se utiliza el enlace completo, este tipo de casos suele tener un efecto desproporcionado, ya que se incrementa la distancia máxima entre grupos y puede llevar a una segmentación o clusters más fragmentados. En cambio, el enlace promedio atenua ese efecto al considerar el comportamiento general entre grupos, lo que permite generar una estructura de clústers más estable y representativa, manteniendo un equilibrio razonable entre claridad e interpretabilidad.

Para evitar una selección puramente visual de k , se complementó este análisis con la evaluación de diversos valores de k mediante *métricas internas de validación* del clustering. En particular, se consideraron valores en el rango $k \in \{2, 3, 4, \dots, 20\}$, y para cada partición se calculó, entre otros, el **índice de Silhouette** global. La Figura 7 muestra la evolución de dicho índice en función de k . Cada punto representa el valor medio del Silhouette para la partición correspondiente:



Figura 7: Índice de Silhouette global en función del número de clusters k .

El resultado de este análisis muestra que existe un patrón en relación a trabajar con pocos grupos, los cuales evidencian que existe una separación más consistente, pero a medida que k aumenta, las separaciones son mínimas. Es decir, al forzar demasiados grupos, el algoritmo termina dividiendo conjuntos que en realidad no están tan separados, lo que produce segmentos menos claros.

A partir de esta comparación, se identifica dos opciones concretas $k=2$ o $k=5$.

- $k = 2$, se presenta como una alternativa más sólida a nivel de estructura general, si tomamos de referencia este valor, estamos hablando de dos macroperfiles diferenciados, pero se considera que esta es una opción útil si se busca por ejemplo describir una brecha principal dentro de la población.

- $K = 5$, esta se presenta como una alternativa con mayor detalle y con coherencia ya que permite distinguir subperfiles dentro de esos macroperfiles presente en $k=2$. Esta opción emerge como candidato importante para el estudio, considerando que se requiere la identificación de segmentos más descriptivos, pero sin perder de vista las separaciones consistentes.

En complemento a la identificación objetiva del k óptimo, se incorpora al análisis de los índices de Davies–Bouldin (DBI) y Calinski–Harabasz (CH) en el rango $k \in \{2, \dots, 20\}$. Estos indicadores permiten evaluar el agrupamiento desde perspectivas distintas: el DBI penaliza soluciones con clústers poco separados o internamente dispersos (valores más bajos indican mejor desempeño), mientras que el CH favorece configuraciones con buena separación entre grupos y alta cohesión interna (valores más altos son preferibles) [53, 54]. En este sentido, DBI y CH permiten contrastar y reforzar el hallazgo del k óptimo sugerido por Silhouette, evitando que la decisión dependa de un único criterio. A partir de esta especificación, se observan los resultados de este análisis constituidos en la Tabla 5:

Cuadro 5: Índices de Davies–Bouldin (DBI) y Calinski–Harabasz (CH) para distintos valores de k .

k	DBI	CH
2	1.0237	1.9295
3	1.3418	8.7123
4	1.3185	11.9994
5	1.3185	15.4178
6	1.2556	13.9283
7	1.4648	11.8195
8	1.4207	10.5960
9	1.3338	9.3910
10	1.3104	13.3556
11	1.3062	16.1748
12	1.2219	15.0033

A partir de estos indicadores se observan dos patrones relevantes. Primero, si bien $k = 2$ presenta el menor valor de DBI, su índice CH es sustancialmente inferior al resto del rango, esto permite evidenciar que, aunque la partición en dos grupos maximiza la separación global, puede resultar demasiado agregada para capturar matrices dentro de la población. Segundo, para valores intermedios de k se aprecia un mejor balance entre ambas métricas: el CH alcanza sus valores más altos en torno a $k = 5$ y $k = 11$, mientras que el DBI se mantiene en niveles similares, pero una mejora marcada para $k = 12$.

Al evaluar $k \in \{2, \dots, 12\}$, se observa que los índices internos no señalan un único valor indiscutible. Por un lado, el DBI favorece soluciones más simples, con su mejor resultado en $k = 2$; por otro, el índice CH tiende a aumentar y alcanza su máximo en $k = 11$. Sin embargo, al priorizar una segmentación que sea fácil de explicar y útil para el estudio, $k = 5$ se presenta como una opción especialmente defendible, considerando que obtiene el segundo valor más alto de CH (15.4178) y, además, coincide con el máximo local observado en el análisis de Silhouette. Esto sugiere una separación adecuada entre los clusters sin caer en una partición excesivamente fina. En consecuencia, se adopta $k = 5$ como solución principal por su equilibrio entre calidad e interpretabilidad, manteniendo $k = 2$ como referencia de macroestructura y considerando $k \in \{11, 12\}$ como alternativas de mayor detalle.

3.4.2. Visualización con t-SNE

Con el fin de apoyar la interpretación visual de los clústeres, se elaboró una representación bidimensional utilizando *t-distributed Stochastic Neighbor Embedding* (t-SNE). Esta técnica se aplicó directamente sobre la matriz de distancias de Gower (con `metric="precomputed"`), utilizando una perplexidad intermedia y una semilla aleatoria fija para asegurar la reproducibilidad (Figura 8)

La proyección resultante representa a cada individuo como un punto en un plano. Los ejes mostrados como *Componente t-SNE 1* y *Componente t-SNE 2* no corresponden a variables observadas ni a dimensiones interpretables en sentido directo (como ocurre en PCA); más bien, son coordenadas latentes que t-SNE construye para ubicar los puntos de manera que, en lo posible, quienes son similares según Gower queden cercanos en el plano. En consecuencia, la utilidad principal de estas dos dimensiones es facilitar una lectura visual de la estructura del agrupamiento: la cohesión interna de cada clúster, la separación relativa entre clústeres y la existencia de zonas de transición. Por ejemplo, se aprecia un conglomerado amplio hacia la derecha (clúster 0) y otros grupos concentrados hacia la izquierda, con un clúster particularmente compacto en la parte inferior izquierda (clúster 3). Asimismo, la presencia de puntos cercanos en zonas intermedias sugiere áreas de frontera donde los perfiles comparten características entre clústeres.

Finalmente, los puntos se colorearon según el clúster asignado (solución de $k = 5$), lo que permite identificar agrupamientos relativamente compactos y reconocer visualmente dónde la separación es más clara y dónde existe mayor continuidad entre perfiles.

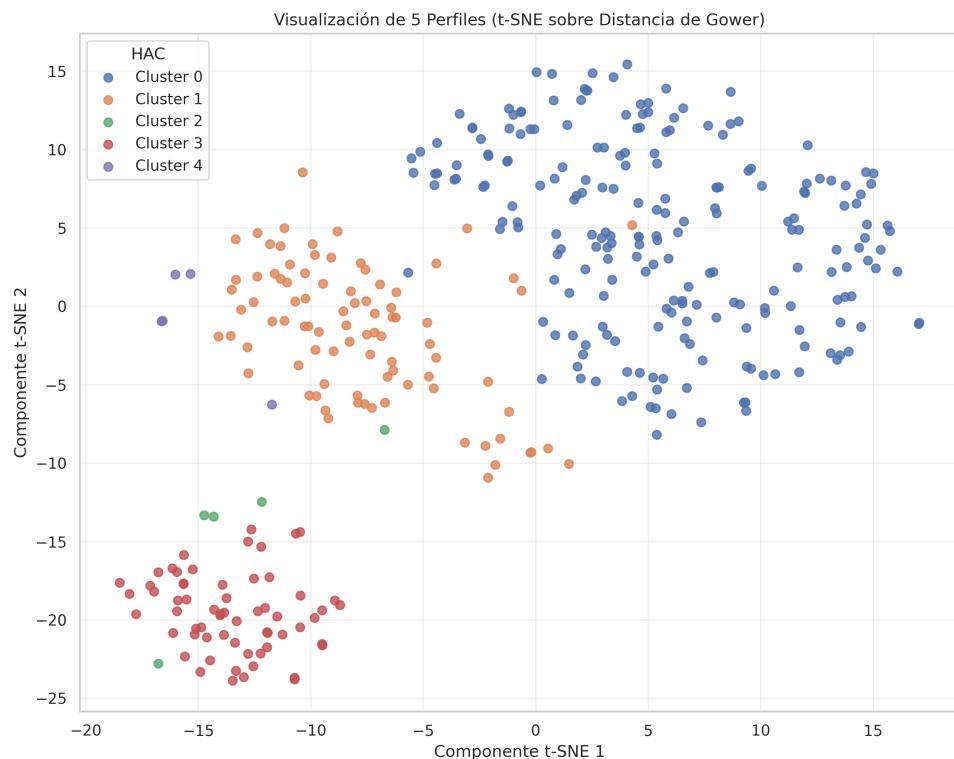


Figura 8: t-SNE sobre matriz de distancias de Gower coloreado por clusters HCA

3.4.3. Evaluación: índice de Silhouette

En esta etapa, la calidad interna de los clusters obtenidos con el modelo jerárquico (*average linkage*, $k = 5$) se evaluó mediante el índice de Silhouette. Este indicador permite valorar, de forma resumida, qué tan coherentes son los grupos formados comparando la cercanía de cada individuo con su propio clúster frente a su cercanía con el clúster más próximo. Para cada observación i , el coeficiente de Silhouette $s(i)$ se define como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (18)$$

donde $a(i)$ es la distancia media de i a los elementos de su propio cluster y $b(i)$ es la mínima distancia media de i a los elementos de cualquier otro cluster. Es así entonces que, $s(i) \in [-1, 1]$; valores cercanos a 1 indican una asignación más coherente, y valores cercanos a 0 corresponden a observaciones en la frontera entre dos clusters y valores negativos sugieren que el individuo podría estar mejor ubicado en otro cluster.

Para la solución con $k = 5$ se obtuvo un *Silhouette global* de $S_{\text{global}} \approx 0,254$. Este valor indica que, en promedio, los individuos tienden a estar más próximos a su propio cluster que al conglomerado más cercano, aunque con márgenes moderados. En un contexto sociodemográfico, donde los perfiles suelen presentar solapamiento entre características, este comportamiento es esperable y da cuenta de una estructura aprovechable más que grupos completamente separados. En consecuencia, el índice se utiliza como criterio comparativo entre configuraciones (valores de k , esquemas de enlace y ponderaciones) [55].

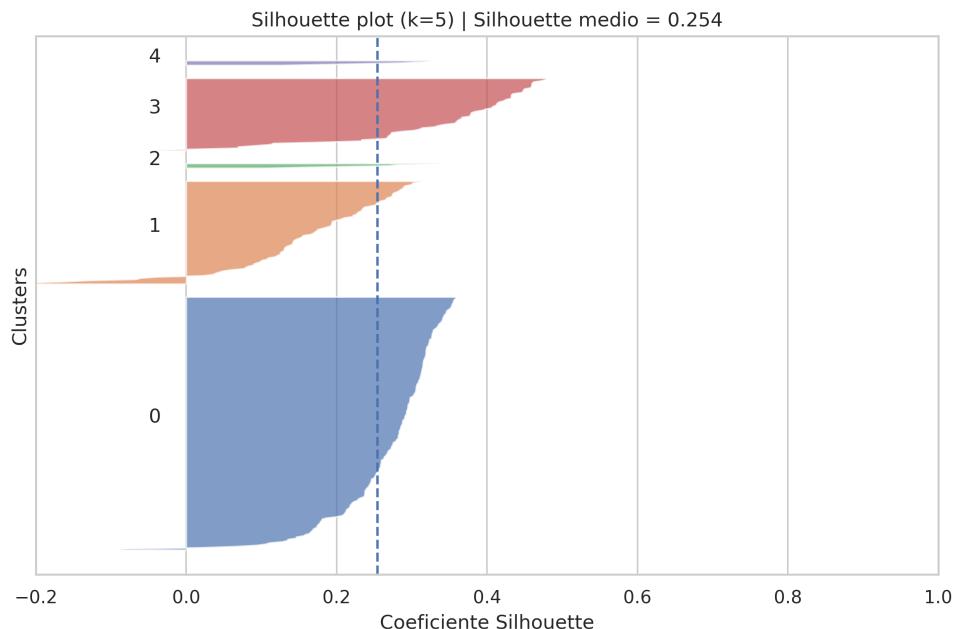


Figura 9: Gráfico de Silhouette para la solución jerárquica de $k = 5$ clusters. La línea vertical indica el Silhouette medio global.

Nota: “cerca de 0” corresponde a observaciones con Silhouette cercano a cero, interpretadas como casos en zona de transición entre clústers (según el umbral definido en el cálculo).

Cuadro 6: Resumen del coeficiente de Silhouette por clúster ($k = 5$).

Clúster	<i>n</i>	Media	Mediana	Mín	Máx	P25	P75	Prop. $s < 0$	Prop. $0 \leq s < 0,05$
3	62	0,342	0,378	-0,041	0,478	0,276	0,432	0,016	0,016
0	215	0,270	0,288	-0,091	0,359	0,242	0,316	0,009	0,005
4	5	0,251	0,285	0,126	0,325	0,206	0,314	0,000	0,000
2	5	0,236	0,269	0,107	0,346	0,178	0,282	0,000	0,000
1	88	0,157	0,170	-0,258	0,315	0,108	0,239	0,080	0,057

Nota: $s < 0$ identifica observaciones potencialmente mejor asignadas a otro clúster; $0 \leq s < 0,05$ corresponde a casos en zona de transición (frontera) entre clústers.

En la evaluación por clúster y tomando de referencia el resumen del coeficiente de Silhouette por clúster (Cuadro 6) podemos detallar los siguientes hallazgos:

- El **Clúster 3** destaca como el grupo más consistente. Con $n = 62$, presenta la mayor media (0,342) y mediana (0,378) del coeficiente de Silhouette, además de un percentil 75 elevado ($P_{75} = 0,432$), lo que sugiere una buena cohesión interna y una separación relativamente clara frente a otros clusters. Adicionalmente, la proporción de casos con Silhouette negativo ($s < 0$) y la proporción de observaciones en frontera ($0 \leq s < 0,05$) son bajas (ambas de 1,6%). En conjunto, estos resultados indican que se trata del clúster más compacto y mejor definido, por lo que resulta especialmente adecuado para perfilar e interpretar con mayor confianza.
- El **Clúster 0**, por otra parte, es el grupo más grande ($n = 215$, equivalente al 57% de la muestra) y evidencia un comportamiento estable. Su media (0,270) y mediana (0,288) se mantienen en niveles moderados, y registra proporciones muy reducidas tanto de asignaciones potencialmente que no corresponden ($s < 0$: 0,9%) como de casos en frontera ($0 \leq s < 0,05$: 0,5%). En resumen, este clúster puede interpretarse como el *perfil dominante* de la población analizada: no es el más separado, pero sí el más robusto por volumen y coherencia interna.
- El punto más sensible de la clusterización se concentra en el **Clúster 1** ($n = 88$, correspondiente al 23% de la muestra). Este grupo presenta el desempeño más débil: la media del coeficiente de Silhouette es la más baja (0,157), el mínimo alcanza valores negativos (-0,258) y se observan las mayores cantidades de casos negativos (7,95%) y de frontera (5,68%). En la práctica, esto indica que aquí se ubica la *zona de transición* del modelo, reflejando perfiles "particulares" que comparten características con otros clústers.
- Finalmente, se identifican dos **microclústers** (Clúster 2 y Clúster 4), cada uno con $n = 5$ (aproximadamente 1,3% de la muestra). Ambos presentan valores positivos de Silhouette y no registran casos negativos ni observaciones en frontera, lo que sugiere coherencia interna; sin embargo, su tamaño es demasiado reducido como para sostenerlos como grupos, posiblemente casos atípicos o corresponde a información sesgada durante la recolección.

3.5. Refinamiento interno mediante PAM

Con el objetivo de mejorar la coherencia interna de la partición generada por el agrupamiento jerárquico (HCA), se implementó un refinamiento posterior basado en *Partitioning Around Medoids* (PAM). La motivación es sencilla: el HCA produce una asignación inicial de clusters, pero en presencia de perfiles solapados es común que una fracción pequeña de observaciones quede ubicada en

zonas de transición entre grupos. Estas observaciones suelen reflejarse en valores bajos del coeficiente de Silhouette, indicando que no están claramente más próximas a su clúster asignado que a los clústers alternativos cercanos.

El procedimiento aplicado fue el siguiente:

- Primero, se calculó el coeficiente de Silhouette $s(i)$ para cada individuo utilizando directamente la matriz de distancias de Gower (métrica precomputada). A partir de estos valores, se identificaron como *puntos problemáticos* aquellas observaciones con $s(i) < 0,1$, umbral que permite aislar casos con asignación débil sin intervenir la estructura completa del agrupamiento. El resto de observaciones ($s(i) \geq 0,1$) se mantuvieron como base estable.
- En segundo lugar, para cada clúster del HCA se estimó un representante robusto mediante PAM, fijando $k = 1$ (un solo medoid por clúster). En este contexto, el *medoid* corresponde al individuo más central del grupo, es decir, aquel con menor disimilitud promedio respecto a los demás miembros del clúster. A diferencia de un centroide, el medoid es una observación real del conjunto de datos, lo cual es especialmente adecuado cuando se trabaja con distancias no euclidianas y variables mixtas (como en Gower).
- Una vez obtenidos los medoides de cada clúster, cada punto problemático se reasignó al clúster cuyo medoid resultó más cercano según la distancia de Gower. Este paso actúa como un ajuste local, es decir, no se reoptimiza toda la partición, pero sí corrige asignaciones mal realizadas, apoyándose en los individuos más representativos de cada grupo. Con ello se obtuvo una solución refinada, manteniendo $k = 5$ y preservando la estructura global del HCA.

Tomando de base este procedimiento, los resultados muestran una mejora en la calidad interna: el Silhouette global aumentó de 0,254 (solución original) a 0,261 (solución refinada), con un total de 13 observaciones reasignadas (Cuadro 7). Este cambio, se lo puede considerar como moderado, pero es consistente con la finalidad de reducir ambigüedad en la frontera entre clústers sin alterar drásticamente toda la segmentación.

Cuadro 7: Matriz de contingencia entre la partición HCA original y la partición refinada con PAM ($k = 5$)

HCA	Partición refinada					Total
	0	1	2	3	4	
0	213	2	0	0	0	215
1	6	78	1	0	3	88
2	0	0	5	0	0	5
3	0	0	1	61	0	62
4	0	0	0	0	5	5
Total	219	80	7	61	8	375

Asimismo, en el cuadro 8 se observa que el clúster 3 mejora de forma notable en cohesión (media $\approx 0,363$), mientras que los clusters pequeños tienden a mantener valores más bajos debido a su tamaño reducido y a la mayor sensibilidad a perfiles atípicos.

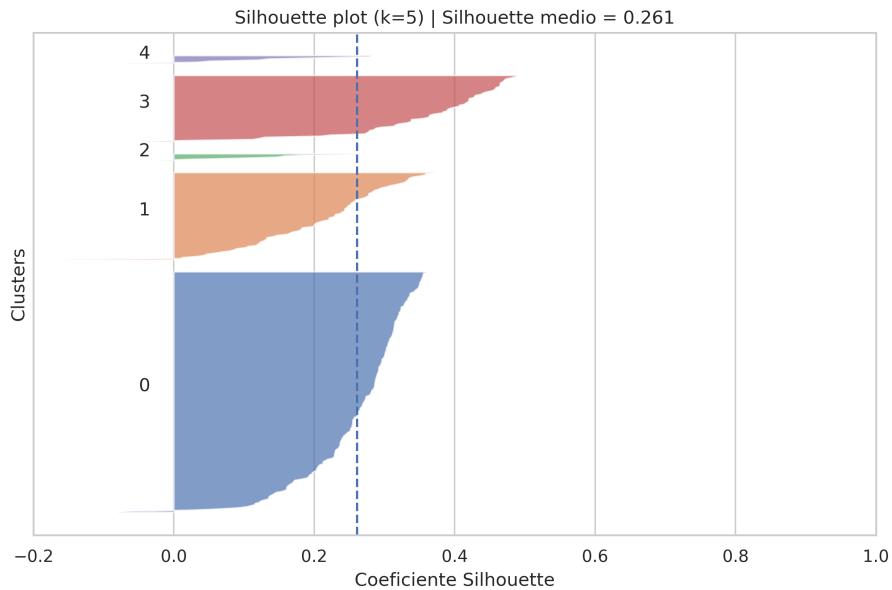


Figura 10: Gráfico de Silhouette con refinamiento PAM para la solución jerárquica de $k = 5$

Cuadro 8: Resumen del coeficiente de Silhouette por clúster ($k = 5$). Silhouette global: 0.2609.

Clúster	<i>n</i>	Media	Mediana	Mín	Máx	P25	P75	Prop. $s < 0$	Prop. $0 \leq s < 0,05$
3	61	0.363	0.392	-0.044	0.488	0.315	0.444	0.016	0.016
0	219	0.261	0.281	-0.082	0.361	0.230	0.312	0.009	0.005
1	80	0.208	0.229	-0.175	0.374	0.132	0.278	0.013	0.063
4	8	0.129	0.128	-0.075	0.281	0.047	0.225	0.125	0.125
2	7	0.123	0.148	-0.024	0.277	0.056	0.173	0.143	0.143

3.6. Limpieza de casos fronterizos

Con el fin de evaluar el efecto de observaciones con asignación débil sobre la calidad interna del agrupamiento, se realizó un experimento de *depuración* posterior a la partición refinada (PAM+HCA).

El criterio utilizado fue el coeficiente de Silhouette individual(*sil_hac_refined*), calculado sobre la matriz de distancias de Gower (métrica precomputada).

Dado que valores bajos de Silhouette indican proximidad entre el clúster asignado y clusters cercanos, se asumió que observaciones con $s(i) \leq 0,1$ presentan una *inconsistencia estructural* (casos en frontera o con asignación ambigua) y pueden deteriorar el indicador global.

A partir del conjunto completo ($n = 375$), se filtraron únicamente las observaciones con *sil_hac_refined* $> 0,1$, construyendo una muestra refinada:

$$\mathcal{D}^* = \{i : s(i) > 0,1\}. \quad (19)$$

Este filtro identificó 24 registros a no ser considerados, obteniendo 351 observaciones. Para garantizar coherencia en el cálculo posterior, se actualizó una submatriz de distancias de Gower (*gower_limpia*) restringida a los índices preliminares, y se conservaron las etiquetas de clúster correspondientes para el cómputo de Silhouette global en la nueva muestra depurada.

Tras la depuración, el Silhouette global aumentó de 0,261 (solución refinada PAM+HCA) a 0,2743

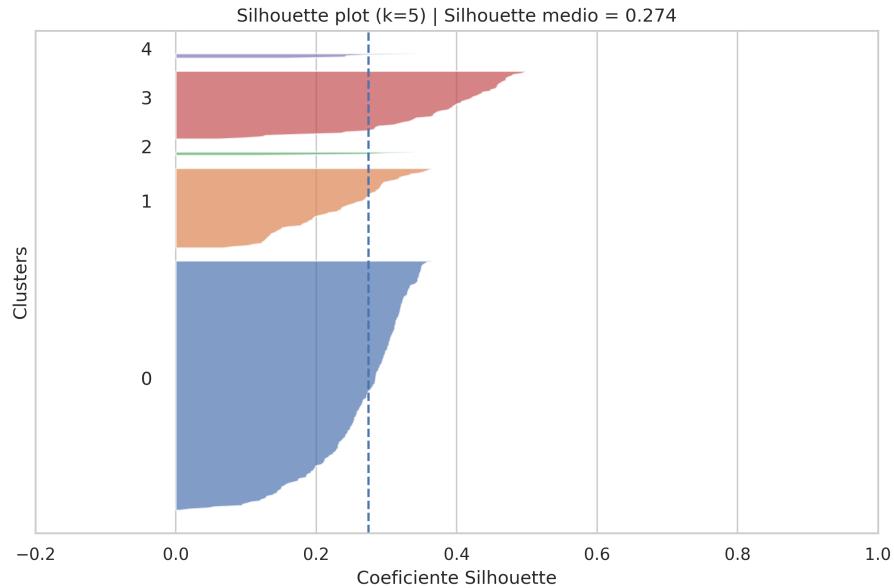


Figura 11: Gráfico de Silhouette con refinamiento de casos fronterizos para la solución jerárquica de $k = 5$

Cuadro 9: Resumen del coeficiente de Silhouette por clúster ($k = 5$). Silhouette global: 0.2743.

Clúster	<i>n</i>	Media	Mediana	Mín	Máx	P25	P75	Prop. $s < 0$	Prop. $0 \leq s < 0,05$
3	59	0.373	0.398	0.061	0.497	0.327	0.456	0.000	0.000
0	214	0.263	0.281	0.006	0.365	0.231	0.313	0.000	0.014
2	4	0.259	0.289	0.114	0.344	0.234	0.314	0.000	0.000
4	5	0.254	0.243	0.158	0.350	0.240	0.278	0.000	0.000
1	69	0.228	0.236	0.067	0.364	0.151	0.290	0.000	0.000

(muestra refinada casos fronterizos), evidenciando que una fracción pequeña de observaciones con asignación ambigua contribuía a reducir la coherencia interna promedio.

En primer lugar, se observa que **no existen asignaciones claramente inconsistentes**, ya que en todos los clústers la proporción de casos con Silhouette negativo es nula ($\Pr[s(i) < 0] = 0$). Esto implica que, para los registros filtrados, cada individuo se encuentra sistemáticamente más próximo a su propio clúster que a cualquier cluster cercano. Adicionalmente, tras el ajuste de casos en zona de frontera la presencia de estos es prácticamente inexistente: únicamente el Clúster 0 presenta una proporción reducida de observaciones con $0 \leq s(i) < 0,05$ (1,4 %), lo que es coherente con su condición de grupo mayoritario y, por tanto, con mayor diversidad interna.

En este escenario, el **Clúster 3** destaca como el segmento más nítido y mejor separado (media = 0,373, mediana = 0,398, $P_{75} = 0,456$), mientras que el **Clúster 0** conserva estabilidad y volumen (media = 0,263, $n = 214$). Por su parte, el **Clúster 1** mantiene valores más moderados (media = 0,228), lo que sugiere un perfil relativamente más difuso o de transición, aun cuando su asignación global resulta coherente tras la limpieza. Finalmente, los clústers 2 y 4 presentan tamaños muy reducidos ($n = 4$ y $n = 5$), por lo que se interpretan como perfiles anómalos o de baja implicación.

3.7. Consolidación de clústers

Con el fin de consolidar una segmentación más interpretable y con mayor consistencia interna, se realizó una etapa adicional de depuración sobre la muestra refinada. En esta etapa se seleccionaron únicamente los clústers con suficiente tamaño y relevancia analítica para el objetivo del estudio, definiéndose como *clústers válidos* aquellos identificados como 0, 1 y 3. En consecuencia, se excluyeron los clústers 2 y 4, previamente caracterizados como perfiles minoritarios de baja frecuencia, con el propósito de evitar que segmentos muy pequeños condicionen la interpretación global o introduzcan inestabilidad en el análisis posterior.

La distribución resultante en la muestra final evidencia una composición dominada por el Clúster 0 (62.57 %), seguido del Clúster 1 (20.18 %) y el Clúster 3 (17.25 %). A partir de esta submuestra se actualizó la matriz de distancias de Gower restringida a los individuos candidatos y se recalcularó el índice de Silhouette global para evaluar el efecto de este ajuste.

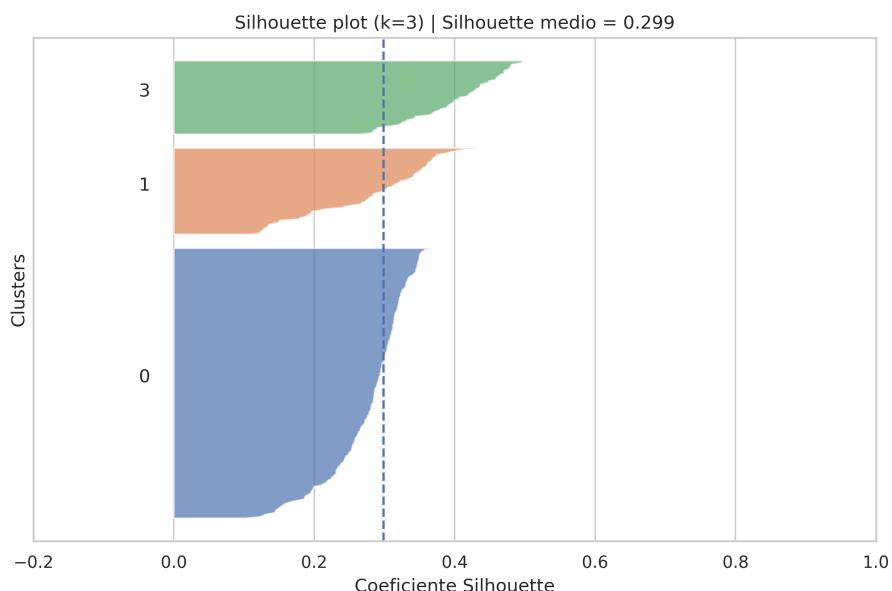


Figura 12: Gráfico de Silhouette con refinamiento de clusters para la solución jerárquica de $k = 5$

Cuadro 10: Resumen del coeficiente de Silhouette por clúster ($k = 5$). Silhouette global: 0.2989.

Clúster	<i>n</i>	Media	Mediana	Mín	Máx	P25	P75	Prop. $s < 0$	Prop. $0 \leq s < 0,05$
3	59	0.401	0.407	0.267	0.497	0.354	0.458	0.000	0.000
0	214	0.278	0.290	0.101	0.365	0.252	0.315	0.000	0.000
1	69	0.277	0.297	0.110	0.436	0.195	0.352	0.000	0.000

Como efecto de esta consolidación, el índice de Silhouette global se incrementó de $S_{\text{global}} \approx 0,274$ a $S_{\text{global}} \approx 0,299$, reflejando una mayor coherencia interna y una separación más clara entre los clústers retenidos. Además, no se registran casos con $s(i) < 0$ ni observaciones en frontera ($0 \leq s(i) < 0,05$).

La exclusión de clústers minoritarios y la focalización en los clusters 0, 1 y 3 permite obtener una segmentación final más robusta y defendible para el análisis del estudio, conservando los perfiles sobresalientes de la población y reduciendo el efecto de perfiles atípicos o de baja representatividad.

Por último, la Figura 13, muestra la proyección t-SNE de la muestra final (distancia de Gower) con tres clusters. Se observa una separación visual clara entre los grupos, donde el Cluster 3 (verde) aparece como un bloque compacto y bien aislado en la zona superior derecha, el Cluster 1 (naranja) se concentra en la parte inferior izquierda también con alta cohesión, mientras que el Cluster 0 (azul) ocupa una región central más amplia y dispersa, consistente con su mayor tamaño y heterogeneidad interna.

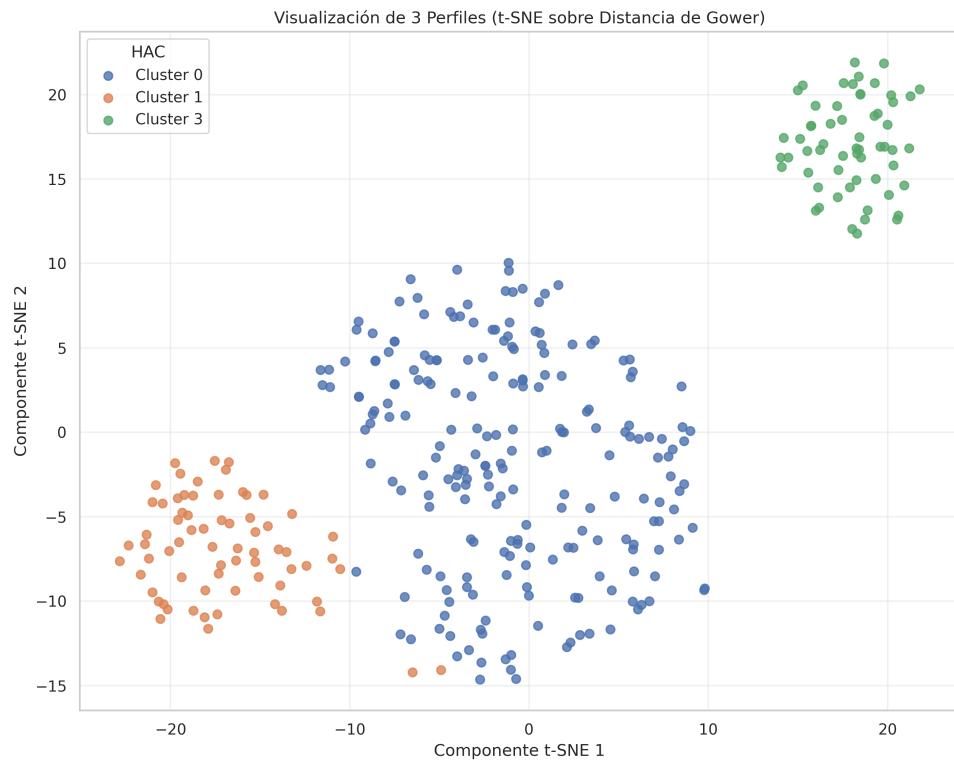


Figura 13: Proyección t-SNE de la segmentación final

4. Experimentos y resultados

En esta sección se presentan los resultados principales de la segmentación obtenida mediante un esquema de agrupamiento jerárquico aglomerativo aplicado sobre la distancia de Gower ponderada. A partir de esta partición inicial, se incorporó un refinamiento local mediante *Partitioning Around Medoids* (PAM), con el propósito de mejorar la coherencia interna de los grupos sin modificar su estructura general. Con base en este proceso, la solución de referencia se establece en $k = 3$ clusters y se construye a partir de variables sociodemográficas, familiares, laborales y de activos, descritas en la Sección 3.2.1.

En contexto, la Figura 14 resume la distribución porcentual final de los individuos por cluster. Se puede identificar que el **Cluster 0** concentra el 62.6% de la muestra, mientras que el **Cluster 1** agrupa el 20.2% y el **Cluster 3** el 17.3%. En términos prácticos, esto evidencia la presencia de un perfil mayoritario que describe a la población en general, acompañado de dos perfiles adicionales que introducen ciertas características relevantes sin caer en una segmentación demasiado detallada.

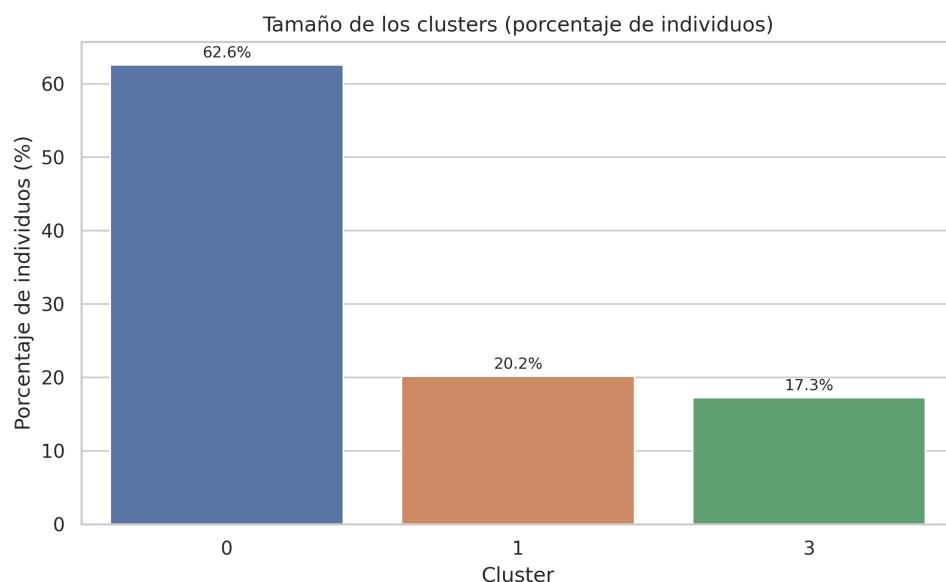


Figura 14: Tamaño de los clusters-porcentaje de individuos

4.1. Patrones identificados

Una vez definida la partición final con $k = 3$ clusters, el siguiente paso consiste en traducir los grupos estadísticos en perfiles comprensibles. Para ello se realizó un perfilado descriptivo por variable, cuyo objetivo es identificar rasgos caracterizan a cada cluster a partir de las variables empleadas en la segmentación (sociodemográficas, familiares, laborales y de activos). Dado que el agrupamiento se construyó sobre distancia de Gower ponderada, los individuos fueron asignados en función de su similitud global, considerando variables mixtas; sin embargo para comunicar los resultados de manera clara, es necesario resumir, por separado, el comportamiento de las variables categóricas y las numéricas dentro de cada grupo.

En este marco, para las variables categóricas y ordinales se calcularon dos estadísticos sencillos: (i) la modalidad dominante (*top_cat*), definida como la categoría más frecuente dentro del cluster, y (ii) su proporción de dominancia (*top_pct*), definida como el porcentaje de individuos del cluster que presentan dicha modalidad. Esta información se la detalla en el cuadro 11. En términos prácticos, *top_cat* responde a la pregunta “¿cuál es el valor más común en este grupo?”, mientras que *top_pct* responde a “¿qué tan homogéneo es el grupo respecto a esa característica?”. Valores altos de *top_pct* indican rasgos ampliamente compartidos y, por tanto, perfiles más consistentes en esa dimensión; en cambio, valores moderados sugieren mayor heterogeneidad interna o la coexistencia de subpatrones dentro del mismo cluster.

Cuadro 11: Modalidad dominante (*top_cat*) y proporción (*top_pct*) por variable y cluster.

Variable	Cluster 0	Cluster 1	Cluster 3
Q100_Carga_Familiar_ord	0 (62.15 %)	0 (60.87 %)	0 (77.97 %)
Q101_IAT_Acceso	3 (60.75 %)	3 (94.20 %)	3 (64.41 %)
Q10_TipoColegio	1 (73.83 %)	1 (49.28 %)	1 (67.80 %)
Q12_EstadoCivil	6 (62.62 %)	6 (71.01 %)	6 (84.75 %)
Q4_Genero	2 (64.95 %)	1 (53.62 %)	2 (89.83 %)
Q76_Zona	1 (55.61 %)	1 (78.26 %)	1 (61.02 %)
Q7_Autoidentificacion	6 (62.15 %)	6 (82.61 %)	6 (71.19 %)
Q82_TenenciaVivienda	4 (40.19 %)	1 (36.23 %)	4 (38.98 %)
Q8_Idiomas	2 (74.77 %)	2 (65.22 %)	2 (79.66 %)
Q95_Auto	0 (79.44 %)	0 (50.72 %)	0 (74.58 %)
Q98_Trabaja	1 (100 %)	1 (100 %)	0 (100 %)
Q99_Tiene_Discapacidad	1 (52.80 %)	0 (59.42 %)	0 (62.71 %)

Para facilitar la interpretación, los códigos numéricos reportados (modalidad) se traducen mediante la tabla de referencia de variables y codificación (cuadro 12).

Cuadro 12: Diccionario de variables y esquema de codificación

Variable	Categorías (resumen)
Q4_Genero	1: hombre; 2: mujer
Q5_Edad	numérica (rango observado: 17–68)
Q10_TipoColegio	1: fiscal (del Estado); 2: particular (privado); 3: fiscomisional; 4: municipal
Q11_NivelInstruccion_ord	0: bachillerato; 1: postbachillerato (no superior); 2: técnico/tecnológico superior; 3: educación superior; 4: maestría/especialización; 5: PhD/doctadorado
Q12_EstadoCivil	1: unida/o; 2: separada/o; 3: divorciada/o; 4: viuda/o; 5: casada/o; 6: soltera/o
Q76_Zona	1: área urbana; 2: área rural
Q7_Autoidentificacion	1: indígena; 2: afrodescendiente; 3: negra/o; 4: mulata/o; 5: montubia/o; 6: mestiza/o; 7: blanca/o
Q82_TenenciaVivienda	1: propia pagada; 2: propia pagando; 3: propia (donada/heredada/posesión); 4: arrendada/anticrésis; 5: prestada o cedida; 6: por servicios
Q8_Idiomas	1: sólo idioma indígena; 2: sólo castellano/español; 3: sólo idioma extranjero; 4: sólo lengua de señas ecuatoriana; 5: no habla/no se comunica; 6: indígena y castellano; 7: castellano e idioma extranjero; 8: indígena, castellano e idioma extranjero; 9: otras combinaciones
Q95_Auto	0: no tiene; 1: sí tiene
Q98_Trabaja	0: no tiene trabajo; 1: sí tiene trabajo
Q99_Tiene_Discapacidad	0: no tiene; 1: sí tiene
Q100_Carga_Familiar_ord	0: baja; 1: media; 2+: alta
Q101_IAT_Acceso	0–1: baja; 2: media; 3: alta

Por su parte, para las variables numéricas (por ejemplo, edad) y para aquellas ordinales codificadas numéricamente que se analizan como escala (por ejemplo, nivel de instrucción), se reportaron estadísticos descriptivos de tendencia central y dispersión (media, mediana, mínimo y máximo). Este resumen permite identificar diferencias de nivel entre clusters y complementar la lectura de las modalidades dominantes con magnitudes comparables (cuadro 13).

Cuadro 13: Estadísticos descriptivos por cluster para variables numéricas/codificadas.

Variable	Cluster	n	Media	Mediana	Mín	Máx
Q5_Edad	0	214	27.24	26	17	58
	1	69	32.91	30	21	68
	3	59	22.53	21	17	42
Q11_NivelInstruccion_ord	0	214	0.02	0	0	1
	1	69	2.67	3	1	4
	3	59	0.12	0	0	2

Lo que se busca con esta información es identificar rápidamente rasgos estructurales que diferencian a cada cluster (por ejemplo, trabajo, acceso tecnológico, zona); ayudar a distinguir variables que no segmentan (cuando la modalidad dominante es similar en todos los clusters); y facilitar la construcción de una narrativa comparativa entre perfiles, manteniendo coherencia con la naturaleza categórica/ordinal de la mayor parte del conjunto de datos.

4.1.1. Lectura Ejecutiva de los Clusters - Patrones identificados

En esta línea se traducen los resultados de los tres perfiles diferenciados, empleando las etiquetas de las variables:

Cluster 0 (62.6 %) — “Jóvenes trabajadores con Educación Media (Bachillerato)” Este grupo concentra la mayor parte de la población y marca el patrón dominante del estudio. En promedio tienen 27 años (mediana 26) y su nivel educativo se ubica casi totalmente en bachillerato o en la enseñanza media. La mayoría reporta carga familiar baja (62 %). Predominan los residentes en zona urbana (56 %) y el estado civil más frecuente es soltero/a (63 %).

En condiciones de vida, destaca que su situación de vivienda más común es arrendada/anticresis (40 %), lo cual sugiere una etapa de independencia parcial o movilidad residencial. En activos, la mayoría no tiene auto (79 %), y en conectividad el perfil es fuerte: predomina el acceso tecnológico alto (61 %). En empleo, el rasgo es contundente: trabajan (100 %). Además, se observa predominio de mujer (\approx 65 %), y una identidad y lengua predominantes comunes en la muestra constituida por mestiza/o que solo hablan castellano/español. La particularidad de este segmento es que la mayoría de individuos (53 %) reportaron que presentan algún tipo de discapacidad. Cabe indicar que este estudio no diferencia entre tipos y grados de discapacidad, lo que podría abrir espacio para futuras líneas de investigación aplicadas en esta línea.

Podemos identificar como conclusiones más relevantes que: (i) Es el segmento mayoritario: define el “perfil típico” de la muestra, (ii) Perfil de jóvenes trabajadores con bachillerato (educación media), conectividad alta, pero con bajo poder adquisitivo, (iii) Predomina el arriendo (alquiler), lo que sugiere una situación residencial más flexible o no estable.

Cluster 1 (20.2 %) — “Adultos con educación superior, conectividad alta y empleo” Este es el grupo con mayor diferenciación por capital académico. En promedio tienen 33 años (mediana 30), con un rango que alcanza hasta 68 años y el nivel educativo típico es educación superior universitaria (mediana 3), con presencia relevante de maestría/especialización (máximo 4), reflejando un segmento donde la formación universitaria y de posgrado es considerablemente más frecuente que en los otros clusters.

En lo social, predomina el estado civil soltero/a (71 %) y la residencia urbana es muy marcada (78 % urbano). En conectividad, la señal es muy fuerte: el acceso tecnológico alto alcanza 94 %, el valor más alto entre los tres clusters. También es un grupo con mayor probabilidad relativa de poder adquisitivo: “no tiene auto” es 51 %, lo que implica casi la mitad sí tiene auto, superior al resto de clusters; en vivienda predomina “propia y totalmente pagada” (36 %). En empleo, el cluster mantiene un patrón marcado donde trabajan el 100 % de la muestra. Este cluster en contraste con el segmento 0, predominan con mínima diferencia los hombres (54 %) que solo hablan castellano/pañol, con una autoidentificación como mestiza/o, sin discapacidad reportada (59 %).

Podemos identificar como conclusiones más relevantes que: (i) Es el cluster con mayor capital académico: concentra educación superior y posgrado, (ii) Tiene la conectividad tecnológica más alta del estudio (IAT alto - 94 %), lo que lo vuelve el perfil más “conectado”, (iii) Muestra señales de mayor poder adquisitivo (más vivienda propia y más auto relativo).

Cluster 3 (17.3 %) — “Jóvenes sin trabajo, urbanos y con conectividad alta” El Cluster 3 representa un perfil claramente diferenciado por la dimensión laboral y por su edad. Este grupo concentra a los más jóvenes en promedio de edad 22.5 años (mediana 21) y un máximo de 42 años, y su educación se concentra en bachillerato o educación media (mediana 0), con algunos casos que alcanzan ciclo postbachillerato o técnica (máximo 2).

En lo social, se observa una fuerte homogeneidad: soltero/a en 85 % coherente con la edad y urbano en 61 %. En vivienda predomina arrendada/anticresis (39 %). En activos y conectividad, el patrón es similar al cluster 0: predominio de no tener auto (75 %) y acceso tecnológico alto (64 %), lo que representa un punto relevante para estrategias de intervención o comunicación basadas en entornos digitales. La diferencia decisiva está en empleo, aquí predominan individuos que no tiene trabajo (100 %). Un rasgo adicional que destaca es la composición por género, la modalidad dominante es mujer con un peso muy alto (\approx 90 %), lo que indica un perfil fuertemente feminizado dentro del segmento juvenil sin empleo. También es el grupo con mayor concentración de carga familiar baja (78 %).

Podemos identificar como conclusiones más relevantes que: (i) Es el perfil que representa una etapa de transición: jóvenes, bachilleres, sin empleo, (ii) Aunque no trabajan, mantienen acceso tecnológico alto, lo cual es una oportunidad para estrategias digitales, (iii) Es un segmento fuertemente feminizado (predominio mujer), responsabilidad familiar baja y mayoritariamente urbano.

Por último, podemos indicar que en cuanto a las variables de autoidentificación e idiomas, al repetirse de forma similar en los tres clusters, aportan contexto descriptivo pero no funcionan como variables de segmentación. En contraste, la variable tiene discapacidad presenta una diferencia leve (con mayor presencia de “sí tiene” en el Cluster 0). Esta variable debe analizarse con mayor detalle y abordarse principalmente desde un enfoque de inclusión y accesibilidad, ya que aporta un matiz relevante para comprender necesidades potenciales, pero no es apropiado utilizarla para etiquetar perfiles.

En conjunto, los perfiles obtenidos son coherentes con las dimensiones priorizadas en el vector

de pesos, donde el capital humano, inserción laboral y acceso tecnológico eran los más valorados. De hecho, la separación más nítida se observa en la condición de empleo, lo cual es consistente con el peso asignado a Q98_Trabaja (1.8) y con la partición final, donde se distinguen segmentos de “trabaja” frente a “no trabaja”. Asimismo, el hecho de que variables con pesos bajos no determinen la estructura de los grupos refuerza la interpretabilidad del modelo y reduce el riesgo de sesgos, en este caso la autoidentificación e idiomas permanecen como rasgos comunes y, aunque discapacidad muestra una variación moderada, su bajo peso evita que condicione la partición, permitiendo que se incorpore como una lectura complementaria o perfilado descriptivo que aporta una matiz.

4.2. Aprendizaje supervisado: predicción de perfiles a partir de etiquetas de cluster (LightGBM)

Concluida la etapa de segmentación no supervisada, el análisis se extendió hacia un enfoque de aprendizaje supervisado con el objetivo de predecir automáticamente el perfil (cluster) de nuevos individuos. En esta fase, las etiquetas obtenidas por el agrupamiento jerárquico refinado (HAC + PAM) se utilizan como “marco de referencia” para entrenar un modelo que aprenda los patrones que caracterizan a cada perfil. Esta estrategia permite pasar de una segmentación válida para la muestra a un esquema reproducible de clasificación a gran escala, manteniendo consistencia con los perfiles identificados.

Preparación de etiquetas y consistencia de clases Dado que en la consolidación final se trabajó con los clusters 0,1,3 se realizó una recodificación para disponer de clases consecutivas y compatibles con un modelo multiclas. En particular, se aplicó el mapeo: $\{0 \rightarrow 0, 1 \rightarrow 1, 3 \rightarrow 2\}$ de modo que la variable objetivo quede definida como $y \in \{0, 1, 2\}$ (equivalente a $\{C_0, C_1, C_3\}$). Adicionalmente, se almacenó el listado de variables usadas como predictores (`feature_columns.pkl`) para asegurar trazabilidad y evitar inconsistencias al aplicar el modelo fuera del entorno de entrenamiento.

Partición de entrenamiento y manejo de desbalance Para evaluar la capacidad de generalización del clasificador, se dividieron los datos en conjuntos de entrenamiento y prueba mediante una partición estratificada (80 %–20 %), preservando la proporción relativa de cada clase. Dado que el tamaño de los grupos es heterogéneo (un perfil mayoritario y dos minoritarios), se incorporó un esquema de ponderación por clase durante el entrenamiento (asignando mayor peso al grupo minoritario). Este enfoque compensa el desbalance sin introducir registros sintéticos; de hecho, alternativas como SMOTE se descartaron en esta instancia debido al tamaño reducido de la muestra y a la sensibilidad que puede generar al replicar patrones en variables mixtas.

Entrenamiento con LightGBM (multiclas) Aunque inicialmente se consideró el uso de Random Forest por su reconocida estabilidad y sencillez, la balanza se inclinó hacia LightGBM debido a su capacidad superior para gestionar el gran volumen de datos del Censo y su habilidad para detectar patrones complejos que modelos más tradicionales no consideran. Esta elección no solo permitió manejar de forma más eficiente el desbalanceo de los perfiles, sino que también garantizó un despliegue ágil, preciso y totalmente escalable a nivel nacional. Para fortalecer esta decisión; al aprovechar su

especialización en variables categóricas, pudimos confirmar que los resultados obtenidos son consistentes y sólidos, asegurando que la segmentación final no sea fruto de un algoritmo específico, sino de una realidad latente en los datos.

LightGBM Light Gradient Boosting Machine es un framework de código abierto de alto rendimiento desarrollado por Microsoft que emplea un método potenciado de gradiente para el aprendizaje automático. Está diseñado específicamente para manejar grandes volúmenes de datos y tener un buen rendimiento en términos de velocidad y uso de memoria. Concretamente, el entrenamiento continuo de LightGBM utiliza una técnica llamada potenciación de gradientes, que combina múltiples aprendices "débiles" (normalmente árboles de decisión), para crear un modelo predictivo sólido [56].

La elección de *Light Gradient Boosting Machine* (LightGBM) como algoritmo de clasificación se fundamenta en razones tanto metodológicas como prácticas, alineadas con el objetivo central del estudio que busca aplicar un modelo capaz de identificar automáticamente el perfil al que pertenece cada individuo. Este modelo, entrenado con los datos previamente segmentados, permite clasificar grandes volúmenes de información provenientes de registros administrativos o encuestas extensas de manera consistente y eficiente.

- En primer lugar, LightGBM es especialmente adecuado para procesar información masiva debido a su eficiencia computacional y su capacidad para trabajar con grandes volúmenes de datos sin perder rendimiento. Esto resulta fundamental al considerar la aplicación del modelo sobre archivos extensos como el *dataset* del INEC, donde tanto el tiempo de entrenamiento como la velocidad de clasificación se convierten en factores prácticos determinantes.
- En segundo lugar, LightGBM maneja de forma robusta los **valores faltantes o datos incompletos**. A diferencia de otros métodos que requieren procesos de imputación de información ausente, este algoritmo puede trabajar directamente con datos faltantes, aprendiendo cómo procesarlos durante el entrenamiento. Cabe considerar que cuando no hay valores NaN en el conjunto de datos de entrenamiento (como en este caso de estudio), la predicción de las filas con NaN serán los valores de las hojas más a la izquierda de los árboles [57]. Esta característica es particularmente valiosa al trabajar con datos de encuestas o registros administrativos, donde es común encontrar información omitida, preguntas sin respuesta o saltos en los cuestionarios. En consecuencia, el modelo resulta más tolerante a inconsistencias en los datos nuevos, siempre que se mantenga la compatibilidad en la estructura y codificación de las variables.
- En tercer lugar, LightGBM tiene la capacidad de detectar **relaciones complejas e interacciones** entre variables sin necesidad de asumir patrones predefinidos. Esto significa que el modelo puede identificar, por ejemplo, cómo la combinación de edad y nivel educativo influye conjuntamente en la pertenencia a un perfil, sin que sea necesario especificar esta relación de antemano. Esta flexibilidad permite utilizar tanto variables numéricas ordenadas (como nivel de instrucción) como variables categóricas (como tipo de colegio) manteniendo resultados estables y coherentes.
- Finalmente, una ventaja adicional es que LightGBM permite identificar qué variables son más importantes para la clasificación mediante métricas interpretables (como la *importancia por ganancia*). Esto posibilita verificar que el modelo está efectivamente basando sus decisiones en las dimensiones relevantes del estudio (formación académica, situación laboral, acceso tecnológico) y no en variables secundarias. Esta transparencia refuerza la confiabilidad del proceso.

y facilita explicar el fundamento de las predicciones cuando el modelo se aplica sobre nuevos registros.

El modelo supervisado se entrenó con LightGBM de Python bajo un objetivo multiclase, empleando un esquema de boosting tipo gbdt. Los hiperparámetros se configuraron con un criterio conservador para reducir sobreajuste: fracción de variables por iteración (`feature_fraction=0.7`), mínimo de observaciones por hoja (`min_data_in_leaf=8`) y complejidad moderada (`num_leaves=15`), junto con una tasa de aprendizaje gradual (`learning_rate=0.05`) y un número suficiente de iteraciones (`num_boost_round=200`). El propósito no fue maximizar una métrica a cualquier costo, sino replicar de forma estable la lógica de asignación de perfiles derivada del clustering.

Se comparó el modelo base (parámetros definidos anteriormente) frente a una configuración seleccionada mediante búsqueda en rejilla con validación cruzada (Hiperparametros - GridSearch). En el conjunto de prueba ($n = 69$), el modelo base alcanzó accuracy = 0,72 y macro-F1 = 0,72, mientras que el modelo ajustado por GridSearch obtuvo accuracy = 0,70 y macro-F1 = 0,69. La diferencia también se observa en el desempeño del perfil minoritario (Clase 1), donde el modelo base logró $F1 = 0,39$ frente a $F1 = 0,32$ con GridSearch.

Evaluación del modelo: desempeño y matriz de confusión En el conjunto de prueba, el clasificador obtuvo una **exactitud global de 0.72**. El rendimiento por clase muestra un patrón relevante: el perfil correspondiente a la clase 2 (equivalente al cluster 3 original) se predice con alta precisión, mientras que la mayor confusión ocurre entre las clases 0 y 1. Esto se refleja en la matriz de confusión (Figura 15), donde se evidencia que ciertos casos del Cluster 0 se asigna como Cluster 1, y viceversa. Este comportamiento es consistente con el hallazgo previo donde ambos grupos registraban rasgos estructurales similares (por ejemplo, inserción laboral y contexto urbano) y se diferencian principalmente por edad y niveles de estudios, por lo que existe una zona natural de transición entre ellos.

Importancia de variables por ganancia Para comprender qué factores explican la asignación supervisada de perfiles, se analizó la importancia de variables por ganancia (*gain*). Los resultados muestran que Q98_Trabaja y Q5_Edad son los predictores más influyentes, seguidos por Q101_IAT_Acceso (Figura 16). En un segundo nivel aparecen variables asociadas a activos y contexto (ejemplo: tenencia de auto, género, carga familiar). Este resultado se relaciona directamente con el marco metodológico, identificando que la condición laboral y edad son separadores principales, mientras que el acceso tecnológico sirve como un componente habilitante que refuerza la segmentación socio-tecnológica.

Estabilidad por validación cruzada Para verificar que el desempeño no dependa de una sola partición entrenamiento/prueba, se ejecutó validación cruzada estratificada de cinco pliegues [58]. Los resultados muestran una **exactitud promedio de 0.7867** (± 0.0370) y un **Kappa de Cohen de 0.5899**, lo que sugiere un acuerdo moderado por encima del azar y, sobre todo, una estabilidad razonable del modelo entre particiones. En términos generales, este resultado respalda la idea de que el clasificador captura patrones consistentes y puede usarse como mecanismo reproducible de asignación de perfiles.

Clasificación masiva y control de consistencia de variables Finalmente, el modelo entrenado se aplicó a un archivo masivo (csv to parquet) para generar predicciones a gran escala. Antes

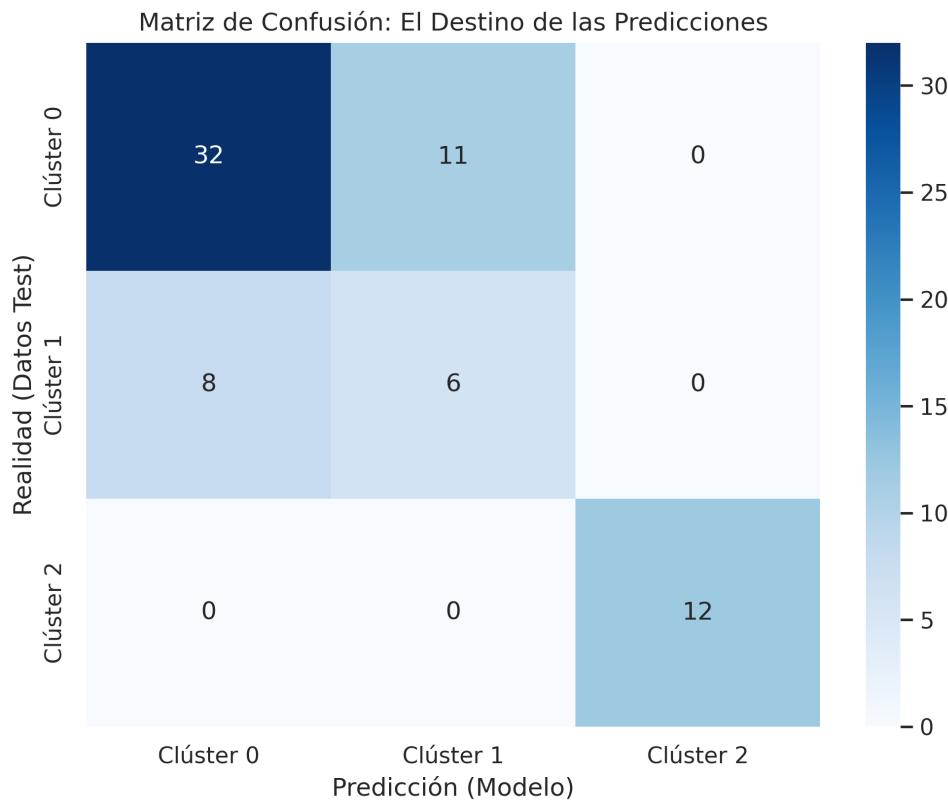


Figura 15: Matriz de confusión del resultado de clasificación de LightGBM

de predecir, se verificó la compatibilidad entre el dataset de entrenamiento y el dataset objetivo: (i) presencia completa de columnas esperadas, (ii) coincidencia de tipos de datos y (iii) diagnóstico de valores faltantes. Este paso es crítico, ya que la clasificación supervisada solo es confiable si las variables de entrada se encuentran armonizadas y codificadas bajo el mismo esquema.

Como salida operativa, el modelo produce probabilidades por clase (`Prob_C0`, `Prob_C1`, `Prob_C3`) y una etiqueta final `CPRED` definida por la clase con mayor probabilidad. Adicionalmente, se derivó una medida simple de confianza (`CPRED_Conf`) como el máximo de las probabilidades, lo que permite priorizar el análisis de casos con predicción más segura y, si se desea, establecer umbrales de calidad, relevante para la asignación del indicador territorial por cantón.

Finalmente, el modelo entrenado se aplicó a un volumen masivo de datos del INEC con el objetivo de generar predicciones a gran escala. Este paso implicó un reto previo de integración y estandarización, debido a que la información censal se encontraba distribuida en tres archivos fuente: *Vivienda* (6 611 555 registros, 32 variables), *Hogar* (5 193 548 registros, 44 variables) y *Población* (16 938 986 registros, 93 variables). Dado el tamaño y la heterogeneidad de estas fuentes, la solución no fue operar directamente sobre archivos planos, sino consolidar una arquitectura de datos que permita trazabilidad, eficiencia y actualización.

En una primera fase, se realizó el volcado de los archivos `.csv` hacia una base de datos PostgreSQL. A partir de esta carga, se aplicaron procesos SQL para transformar las tablas originales en una estructura dimensional más manejable y consistente para análisis, reduciendo redundancia y facilitando uniones (*joins*) reproducibles. Como resultado, se generaron tres componentes principales:

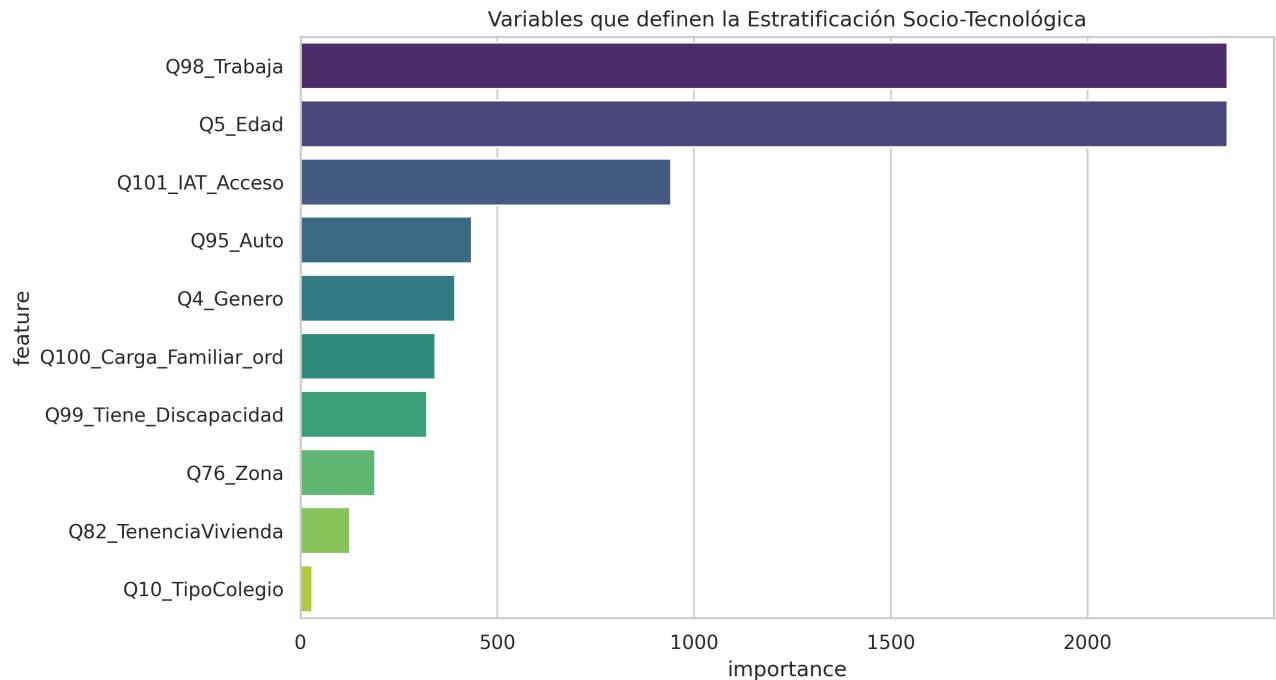


Figura 16: Importancia de variables por ganancia en el modelo LightGBM

`dim_hogar` (1.8 GB), `dim_vivienda` (965 MB) y `fact_poblacion` (4.3 GB). Esta organización no solo optimiza el procesamiento en una primera ejecución, sino que también permite mantener un esquema estable frente a futuras actualizaciones del Censo, al conservar una lógica de integración y etiquetado persistente en el tiempo.

Para operacionalizar la limpieza y homologación de variables (tratamiento de nulos, recodificaciones, mapeo de rangos y estandarización de campos), se consolida un dataset del INEC para clasificación, para esto, se implementaron transformaciones ETL en *Spoon* (Pentaho Data Integration). La Figura 17 ilustra el flujo de procesamiento utilizado como parte de esta etapa.

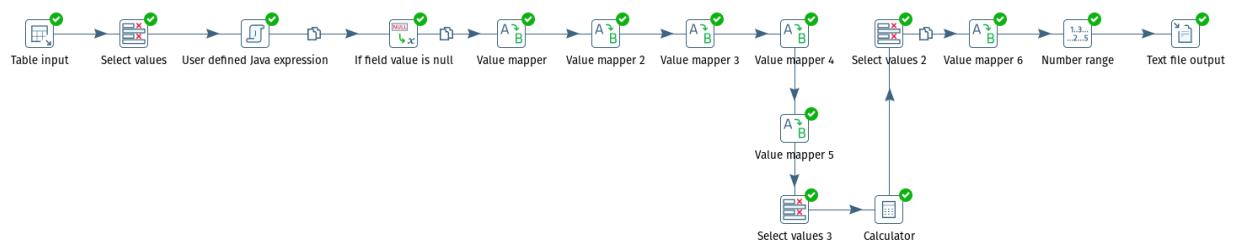


Figura 17: Flujo ETL en *Spoon* para homologación, recodificación y control de calidad previo a la consolidación del dataset INEC.

Posteriormente, el dataset del INEC requerido se exportó como .csv y se procesó en *Azure Data Factory* (ADF) para convertirla a .parquet, un formato más eficiente para lectura y scoring masivo. Sobre este dataset se aplicó un filtro de elegibilidad (edad mayor a 15 y menor a 70 años, y nivel educativo desde bachillerato en adelante), obteniendo un conjunto final de 8 123 393 registros a partir del universo original de 16 938 986 personas.

Antes de ejecutar la predicción, se verificó la compatibilidad entre el dataset de entrenamiento (UPS) y el dataset objetivo (INEC), considerando que cumpla con, (i) presencia completa de columnas esperadas, (ii) coincidencia de tipos de datos y (iii) diagnóstico de valores faltantes. Este control es crítico, ya que la clasificación supervisada solo es confiable si las variables de entrada mantienen el mismo esquema de codificación y estructura que el utilizado durante el entrenamiento.

Como salida operativa, el modelo genera probabilidades por clase (Prob_C0, Prob_C1, Prob_C3) y una etiqueta final CPRED, definida por la clase con mayor probabilidad. Adicionalmente, se definió una medida de confianza (CPRED_Conf) como el máximo de dichas probabilidades, lo que permite (i) priorizar análisis sobre predicciones más seguras, (ii) establecer umbrales de calidad y (iii) controlar la robustez del indicador territorial agregado por cantón.

4.3. Indicador territorial por cantón

En primer lugar, se generó un identificador único de cantón (Canton_ID) concatenando los códigos administrativos i01 (provincia) y i02 (cantón), (del dataset INEC) estandarizados con relleno a dos dígitos para mantener consistencia:

$$\text{Canton_ID} = \text{zfill}_2(\text{i01}) \parallel \text{zfill}_2(\text{i02}).$$

Este identificador se utilizó posteriormente como clave de enlace con la cartografía oficial cantonal.

Filtrado por umbral de confianza Dado que la predicción supervisada produce, además de una etiqueta final, una medida de confianza (CPRED_Conf) asociada a la probabilidad máxima predicha, se definió un umbral de corte $\tau = 0,65$ para retener únicamente casos con predicción consistente. Se definió el indicador binario:

$$\text{IDP} = \mathbb{I}(\text{CPRED_Conf} \geq \tau),$$

donde $\mathbb{I}(\cdot)$ denota la función indicadora. Este filtrado cumple un objetivo metodológico de reducir ruido en la agregación territorial y priorizar registros donde el modelo muestra mayor certeza.

Estimación de densidades cantonales por perfil (C0, C1 y C3) Con los registros filtrados (IDP=True), se calcularon conteos por cantón y perfil predicho ($\text{CPRED_NAME} \in \{\text{C0}, \text{C1}, \text{C3}\}$). Para asegurar comparabilidad entre cantones de distinto tamaño muestral, los conteos se transformaron en proporciones dividiendo por el total de registros por cantón. Para cada cantón c y perfil k se definió:

$$p_{c,k} = \frac{N_{c,k}}{N_c},$$

donde $N_{c,k}$ es el número de individuos del cantón c clasificados como perfil k (con alta confianza), y N_c es el total de individuos observados en el cantón. En consecuencia, cada cantón queda carac-

terizado por la distribución $(p_{c,C0}, p_{c,C1}, p_{c,C3})$, lo cual permite identificar patrones de predominio relativo por perfil.

Definición del puntaje cantonal y normalización Min–Max Para fines de priorización y visualización, se definió un puntaje cantonal asociado a un perfil de interés (por ejemplo, C3). En este caso, el *score* se definió como la proporción cantonal del perfil seleccionado:

$$\text{Similitud_Score}(c) = p_{c,k^*},$$

donde $k^* \in \{C0, C1, C3\}$ representa el perfil focal. Aunque en el ejemplo se ilustró el caso $k^* = C3$, el procedimiento se replica de forma análoga para C0 y C1 cuando se requiere evaluar concentraciones territoriales alternativas.

Dado que las magnitudes del puntaje pueden variar entre cantones y perfiles, se aplicó una normalización Min–Max para llevar el indicador al intervalo [0, 1]:

$$\text{Score_Normalizado}(c) = \frac{\text{Similitud_Score}(c) - \min(\text{Similitud_Score})}{\max(\text{Similitud_Score}) - \min(\text{Similitud_Score})}.$$

Este paso facilita la construcción de rankings comparables y mapas coropléticos interpretables.

Integración geográfica y visualización cartográfica (choropleth) Para la visualización espacial, el indicador cantonal se integró con una capa cartográfica oficial de límites cantonales (.shp). Se estandarizó el código cantonal de la cartografía (DPA_CANTON) a cuatro dígitos y se unió con Canton_ID bajo el mismo formato.

A partir de esta unión se generó un mapa coroplético donde el color de cada cantón representa el Score_Normalizado. Adicionalmente, se incorporó un *tooltip* con el nombre del cantón y el puntaje real del perfil visualizado, facilitando una lectura exploratoria e interactiva de los resultados (Figura 18).

Ranking cantonal por densidad de perfil Como complemento al mapa, se construyó un ranking cantonal ordenando los cantones por la densidad del perfil focal p_{c,k^*} . En el ejemplo ilustrativo (Cuadro 14) se presenta el Top 10 para el perfil C3, identificando cantones con alta concentración relativa del perfil en la muestra predicha con alta confianza. Este mismo procedimiento se extiende de forma equivalente para los perfiles C0 y C1, permitiendo generar rankings diferenciados y contrastar la distribución territorial entre perfiles.

Para ello se definió una función de consulta a la API de OpenRouteService (driving-car), que solicita polígonos de alcance para distintos horizontes temporales y los incorpora como capas geográficas sobre el mapa base. Esta visualización permite aproximar el radio efectivo de conectividad territorial y apoyar la priorización e identificación de nodos estratégicos.

Selección de nodos estratégicos y generación de capas A modo de ejemplo, se seleccionaron nodos estratégicos basados en cantones priorizados por densidad del perfil focal (por ejemplo, Top 10 del perfil C3). Para cada nodo se incorporó un marcador geográfico y se superpusieron las isócronas de 10, 30 y 60 minutos. Este enfoque permite dar respuesta a (*qué territorios quedan cubiertos*

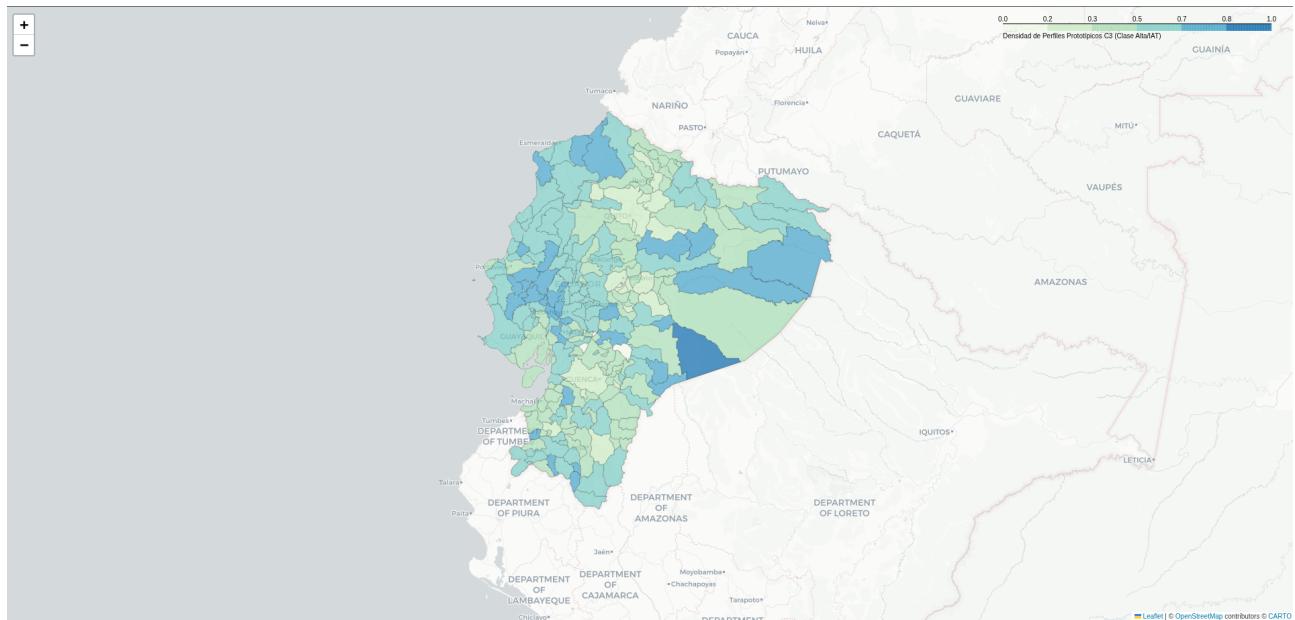


Figura 18: Identificación de cantones prioritarios mediante mapa coroplético

Cuadro 14: Top 10 cantones con mayor densidad de perfil C3

Ranking	Código Cantón	Cantón	Proporción C3	Score Normalizado
1	1409	Taisha	0.8167	1.0000
2	1410	Logroño	0.7102	0.8120
3	1604	Arajuno	0.7049	0.8026
4	1316	24 de Mayo	0.7024	0.7982
5	1407	Huamboya	0.6990	0.7922
6	0905	Colimes	0.6950	0.7851
7	1112	Sozoranga	0.6897	0.7757
8	1412	Tiwintza	0.6865	0.7700
9	1106	Espíndola	0.6864	0.7699
10	0803	Muisne	0.6845	0.7665

(dentro de ventanas de accesibilidad). El resultado se exportó como un mapa interactivo en formato HTML para exploración y consulta.

La Figura 19 presenta la territorialización del perfil de ejemplo C3 a partir de las predicciones del modelo sobre la base INEC, agregadas a nivel cantonal y filtradas por un umbral de confianza. Los tonos más intensos señalan cantones con mayor densidad relativa del perfil.

Sobre esta superficie se ubican los 10 cantones con mayor concentración (nodos priorizados), y se incorporan isócronas de accesibilidad para estimar el alcance desde cada uno de estos puntos propuestos, en consideración de vías, rutas o caminos, que los conectan.

El resultado integra tres capas de decisión: dónde se concentra el perfil, qué cantones priorizar y qué cobertura territorial puede lograrse bajo ventanas de tiempo de desplazamiento (10, 30 y 60 minutos).

Este mismo procedimiento se replica para los perfiles restantes, permitiendo comparar patrones y

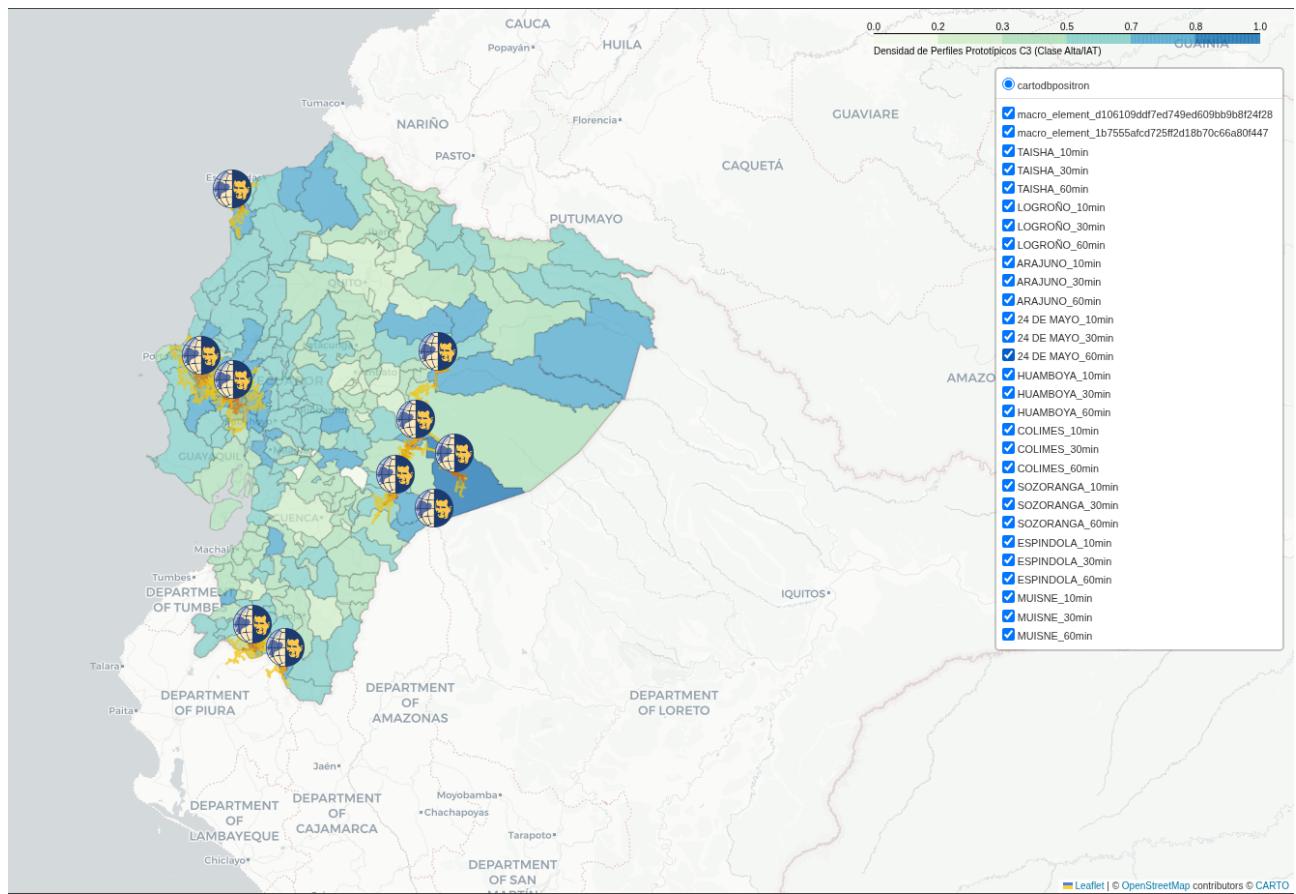


Figura 19: Isócronas de acceso en tiempo 10, 30 y 60 minutos sobre mapa de densidad por perfil

diseñar estrategias diferenciadas por segmento.

Con los resultados alcanzados, se confirma la aplicación de técnicas de similitud y segmentación para apoyar la planificación territorial educativa cuando el objetivo no es únicamente dimensionar población, sino identificar territorios con perfiles comparables de demanda potencial y condiciones de acceso. A partir de una tipología derivada mediante clustering sobre distancia de Gower ponderada y refinada con PAM, se obtuvo una segmentación interpretable que prioriza dimensiones estructurales (capital humano, inserción laboral y acceso tecnológico). Posteriormente, estas etiquetas se escalaron al Censo INEC 2022 mediante aprendizaje supervisado con LightGBM, permitiendo la clasificación masiva de millones de registros con control de consistencia de variables y un umbral de confianza para reducir asignaciones ambiguas. Con ello, el resultado final ya no es un “análisis de datos”, sino un producto de decisión constituido por indicadores de densidad por perfil a nivel cantón (LAU2), rankings territoriales y cartografía que permite priorizar dónde intervenir. Al incorporar isócronas, se completa el salto desde “dónde están los perfiles” hacia “qué tan accesible es atenderlos” considerando diversidad cultural y territorial, lo que traduce la evidencia en lenguaje accionable para planificadores y gestores educativos. En suma, el trabajo no se queda en el cálculo ya que entrega una arquitectura replicable y escalable que conecta la composición sociocultural del país, segmentación, predicción masiva y planificación territorial, con un sentido social.

5. Conclusiones y Trabajos Futuros

El aporte principal de esta investigación fue diseñar una arquitectura de ciencia de datos para datos mixtos que además es escalable. Se comprobó que es viable escalar de una muestra institucional (UPS) a una aplicación masiva en INEC 2022, siempre que se mantenga continuidad en variables homologadas, codificación consistente, medida de similitud adecuada y un modelo supervisado que replique etiquetas a gran escala.

Para lograr perfiles auténticos, fue fundamental respetar la naturaleza diversa de la información, que mezcla niveles educativos con activos del hogar y situaciones laborales. El uso de la distancia de Gower y medidas de asociación específicas permitió que el modelo no forzara los datos a moldes rígidos, sino que aprendiera de sus diferencias reales. De esta manera, evitamos las simplificaciones excesivas y logramos una segmentación que refleja fielmente las brechas sociales y tecnológicas que buscamos describir, garantizando que el análisis tenga un sustento sólido y no solo estadístico.

El proceso de refinamiento fue otro pilar clave para la estabilidad del estudio. Se aplican técnicas de limpieza y ajuste para separar el "ruido" de la estructura real. Al depurar las observaciones que se encontraban en los límites de cada perfil, la claridad de los resultados se mejora. Esto permite obtener grupos más definidos y fáciles de interpretar, asegurando que cada perfil identificado represente a un segmento poblacional con características claras y diferenciables.

Finalmente, la transición hacia el aprendizaje supervisado con LightGBM permitió que el estudio cobrara vida en el territorio. Al enseñarle al modelo a identificar estos perfiles en el universo del Censo, se logra mapear la realidad nacional con un enfoque de confianza y seguridad. Al integrar estos datos con análisis de rutas e isócronas, el resultado deja de ser un gráfico estático y se convierte en una herramienta de planificación. Ahora es posible identificar no solo quiénes necesitan intervención, sino dónde se encuentran y qué tan factible es llegar a ellos, cerrando así el ciclo entre la investigación y la acción requerida.

El reto más importante es fortalecer la base estadística del modelo. Si bien el sistema es técnicamente robusto, incrementar el tamaño de la muestra inicial permitiría que los perfiles sean aún más precisos y estables, especialmente para identificar a grupos minoritarios con mayor claridad. Una muestra más amplia reduciría cualquier fragilidad analítica y permitiría que el clasificador aprenda fronteras más nítidas entre los ciudadanos, elevando la calidad de las predicciones a nivel nacional.

Asimismo, es fundamental que este modelo no se vea como algo estático, sino como un ecosistema que requiere mantenimiento. Se recomienda crear una infraestructura de datos sólida (dominios de datos) que permita actualizar y comparar los resultados conforme cambie la realidad del país o lleguen nuevas fuentes de información. Finalmente, contrastar estos hallazgos con indicadores externos como el acceso real a servicios, discapacidad, o datos de pobreza, que será el paso definitivo para validar el impacto social de la herramienta y asegurar que sea un recurso confiable para la toma de decisiones estratégicas.

Listado de Acrónimos

Análisis Geoespacial Técnica de ciencia de datos que utiliza información geográfica (ejemplo: coordenadas, shapefiles) para analizar patrones espaciales, como la accesibilidad a centros de apoyo mediante isocronas. [5](#)

Cantones Unidades administrativas locales LAU2, en Ecuador equivalentes a los cantones (221 en total), definida por el Clasificador Geográfico Estadístico (DPA) del INEC, utilizada como base para correlacionar datos demográficos y mapear centros de apoyo. [6](#)

Choropleth Mapa temático que utiliza colores para representar valores cuantitativos (ejemplo: scores de similitud) por unidad geográfica, como cantones, en la visualización geoespacial del proyecto. [6](#)

Clustering No Supervisado Técnica de aprendizaje automático, como el algoritmo K-means, que agrupa datos en clusters basados en similitud (ejemplo: edad, educación), utilizada para segmentar perfiles estudiantiles. [5](#)

GraphHopper API Herramienta de código abierto que genera isocronas para análisis de accesibilidad, empleada para mapear áreas cubiertas por centros de apoyo en un tiempo dado (ejemplo: 1 hora). [5](#)

Isocrona Representación geográfica de áreas accesibles desde un punto en un tiempo determinado (ejemplo: 30-60 minutos en auto), utilizada para priorizar ubicaciones de centros de apoyo. [6](#)

Shapefiles DPA Archivos geográficos proporcionados por el INEC que representan los límites cantonales (DPA_CANTON), esenciales para el merge con scores de similitud y visualización choropleth. [6](#)

Bibliografía

- [1] Rodrigo Fernández, Carmen Pagés, Miguel Székely, and Isabel Acevedo. Education inequalities in latin america and the caribbean. *Oxford Open Economics*, 4:i55–76, feb 2025. doi: 10.1093/ooec/odae013. URL <https://dx.doi.org/10.1093/ooec/odae013>. [cited 2025 Oct 3].
- [2] Leonardo Herskovic and Jorge Silva. The rural-urban divide in transitions to higher education in chile. *Journal of International and Comparative Education (JICE)*, 13(1):17–33, apr 2024. URL <https://ejournal.um.edu.my/index.php/JICE/article/view/51672>. [cited 2025 Oct 4].
- [3] Andrés Guzmán Rincón, Sandra Barragán, and Fernanda Cala-Vitery. Rural higher education in colombia: An analysis of public policy evolution. *Latin American Policy*, 14(2):252–266, jun 2023. URL <https://doi.org/10.1111/lamp.12294>. [cited 2025 Oct 4].
- [4] Sergio Ziegler, Juan Arias Segura, Marcelo Bosio, and Karina Camacho. *Conectividad rural en América Latina y el Caribe. Un puente al desarrollo sostenible en tiempos de pandemia*. Instituto Interamericano de Cooperación para la Agricultura (IICA), 2020. URL <https://hdl.handle.net/11324/12896>. [cited 2025 Oct 3].
- [5] Eduardo A. Ortiz, Ximena Dueñas, Cecilia Giambruno, and Ángela López. The state of education in latin america and the caribbean: Learning assessments, sep 2024. [cited 2025 Oct 3].
- [6] John D. Meyer and Amanda C. Barefield. Infrastructure and administrative support for online programs. *Online Journal of Distance Learning Administration*, 13(3), 2010. URL https://www.westga.edu/~distance/ojdl/Fall133/meyer_barfield133.html. [cited 2025 Oct 8].
- [7] Ana C. Useche, Álvaro H. Galvis, Frida Díaz-Barriga Arceo, Andrés E. Patiño Rivera, and Claudia Muñoz-Reyes. Reflexive pedagogy at the heart of educational digital transformation in latin american higher education institutions. *International Journal of Educational Technology in Higher Education*, 19(1):1–15, dec 2022. URL <https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-022-00365-3>. [cited 2025 Oct 4].
- [8] Universidad Técnica Particular de Loja. Resolución rectoral - procuraduría universitaria utpl rct_rr_1_2024_v1: Resolución rectoral de reestructuración de la red de centros de apoyo. Resolución Rectoral emitida por el Rector Ph.D. Santiago Acosta Aide, febrero 2024. URL <https://www.utpl.edu.ec/>. Loja, Ecuador.

- [9] Claudio Rama. La virtualización universitaria en américa latina. *RUSC Universities and Knowledge Society Journal*, 11(3):33–43, jul 2014. URL <https://link.springer.com/articles/10.7238/rusc.v11i3.1729>. [cited 2025 Oct 4].
- [10] Juan Alcides Cárdenas Tapia. Universidad divergente: El caso de la universidad politécnica salesiana como un nuevo modelo de organización y gestión. *Universidad divergente: El caso de la Universidad Politécnica Salesiana como un nuevo modelo de organización y gestión*, 2 2025. doi: 10.17163/ABYAUPS.107.
- [11] Consejo de Educación Superior. Reglamento para carreras y programas académicos en modalidades en línea, a distancia y semipresencial o de convergencia de medios, 2015. URL <https://www.ces.gob.ec>. Resolución RPC-SO-42-No.559-2015 (Art. 13).
- [12] Universidad Politécnica Salesiana. Ups en cifras 2024. Informe institucional, Universidad Politécnica Salesiana, Cuenca, Ecuador, 2024. URL <https://www.ups.edu.ec/documents/20121/262148/2024+UPS+en+cifras.pdf>.
- [13] Instituto Nacional de Estadística y Censos. Censo ecuador 2022: Resultados, 2022. URL <https://www.censoecuador.gob.ec/resultados-censo/>. Consultado: 2025-11-02.
- [14] Peter Chapman. *CRISP-DM 1.0: Step-by-step data mining guide*, 2000.
- [15] J. Luis, C. Ramos, J. C. Sedraz Silva, R. L. Rodrigues, P. Letícia, and S. De Oliveira. Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1092–1101, nov 2020. URL <https://sol.sbc.org.br/index.php/sbie/article/view/12865>. [cited 2025 Oct 7].
- [16] *Partitioning Around Medoids (Program PAM)*, chapter 2, pages 68–125. John Wiley Sons, Ltd, 1990. ISBN 9780470316801. doi: <https://doi.org/10.1002/9780470316801.ch2>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2>.
- [17] Alexandre Gagnon, Gabriel Vargas Mesa, and UNESCO IIEP. *Isochrone-based catchment areas for educational planning*. UNESCO IIEP, 2023. [cited 2025 Oct 8].
- [18] Rosa Castro-Zarzur, Ricardo Espinoza, and Miguel Sarzosa. Unintended consequences of free college: Self-selection into the teaching profession. *Economics of Education Review*, 89:102260, 8 2022. ISSN 0272-7757. doi: 10.1016/J.ECONEDUREV.2022.102260. URL <https://www.sciencedirect.com/science/article/abs/pii/S0272775722000371?via%3Dihub>.
- [19] Oscar Espinoza, Bruno Corradi, Luis González, Luis Sandoval, Noel McGinn, Karina Maldonado, and Yahira Larrondo. The effects of free tuition on the persistence of university students in chile. *International Journal of Educational Development*, 101, 9 2023. ISSN 07380593. doi: 10.1016/J.IJEDUDEV.2023.102838.
- [20] Claudio Ruff, Alexis Matheu, Marcelo Ruiz, Paola Juica, and María Teresa Gómez Marcos. Cost-free education as a new variable of mixed financing policies in chilean higher education and its impact on student trajectory and social mobility. *Heliyon*, 9:e17415, 7 2023. ISSN 24058440. doi: 10.1016/j.heliyon.2023.e17415.

- [21] Maria Marta Ferreyra, Carlos Garriga, Juan David Martin-Ocampo, and Angelica Maria Sanchez-Diaz. The limited impact of free college policies. *European Economic Review*, 168: 104800, 9 2024. ISSN 0014-2921. doi: 10.1016/J.EUROECOREV.2024.104800.
- [22] Flavio Carvalhaes, Adriano S. Senkevics, and Carlos A. Costa Ribeiro. The intersection of family income, race, and academic performance in access to higher education in Brazil. *Higher Education*, 86:591–616, 9 2023. ISSN 1573174X. doi: 10.1007/S10734-022-00916-7.
- [23] Tatiane Pelegrini, Paola Liziane Silva Braga, Gustavo Saraiva Frio, Marco Túlio Aniceto França, Tatiane Pelegrini, Paola Liziane Silva Braga, Gustavo Saraiva Frio, and Marco Túlio Aniceto França. Are there performance differentials between quota and non-quota Brazilian students? *Journal of Economics, Race, and Policy*, 5:41–53, 3 2022. ISSN 2520-8411. doi: 10.1007/S41996-021-00080-7. URL https://EconPapers.repec.org/RePEc:spr:joerap:v:5:y:2022:i:1:d:10.1007_s41996-021-00080-7.
- [24] Flavio Carvalhaes, Marcelo Medeiros, and Clarissa Tagliari Santos. Higher education expansion and diversification: Privatization, distance learning, and market concentration in Brazil, 2002–2016. *Higher Education Policy*, 36:578–598, 9 2023. ISSN 17403863. doi: 10.1057/S41307-022-00275-Z/METRICS.
- [25] Sonnia Valeria Zapatier Castro, Delia Dolores Noriega Verdugo, Ruth María Farías Lema, Ruth Rubí Peña Holguín, and Juan Diego Valenzuela Cobos. Quality of educational service in public universities in Ecuador: a sustainable and equitable education approach. *Frontiers in Education*, 10:1595257, 9 2025. ISSN 2504284X. doi: 10.3389/FEDUC.2025.1595257/BIBTEX.
- [26] Paloma Sepúlveda-Parrini, Pilar Pineda-Herrero, Paloma Valdivia-Vizarreta, and Sara Rodríguez-Pérez. Examining the quality of online higher education in Chile from the perspective of equity. *Quality in Higher Education*, 30:393–409, 2024. ISSN 14701081. doi: 10.1080/13538322.2023.2270398.
- [27] Juan Quemada Vives, Aldo Gordillo, Enrique Barra Arias, Juan Carlos Torres-Díaz, Diana Rivera-Rogel, Ana María Beltrán-Flandoli, and Lucy Andrade-Vargas. Effects of covid-19 on the perception of virtual education in university students in Ecuador; technical and methodological principles at the Universidad Técnica Particular de Loja. *Sustainability* 2022, Vol. 14, Page 3204, 14:3204, 3 2022. ISSN 2071-1050. doi: 10.3390/SU14063204. URL <https://www.mdpi.com/2071-1050/14/6/3204>.
- [28] David Bañeres, M Elena Rodríguez-González, Ana-Elena Guerrero-Roldán, and Pau Cortadas. An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, 20(1):3, 2023.
- [29] Denise Stanley and Yamell Rocio Montero Fortunato. The efficacy of online higher education in Latin America: A systematic literature review. *Revista Iberoamericana de Tecnologías del Aprendizaje*, 17:262–269, 8 2022. ISSN 19328540. doi: 10.1109/RITA.2022.3191299.
- [30] Ana Ivenicki. Digital learning and higher education in Brazil: A multicultural analysis. *Journal of Comparative and International Higher Education*, 16(2):127–135, 2024.

- [31] Alejandro Peña-Ayala. Learning analytics: fundaments, applications, and trends. *A view of the current state of the art to enhance e-learning*, 2017.
- [32] Cristobal Romero and Sebastian Ventura. Educational data mining and learning analytics an updated survey. *Wiley interdisciplinary reviews Data mining and knowledge discovery*, 10(3):e1355, 2020.
- [33] Dirk Ifenthaler and Jane Yin-Kim Yau. Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*, 68(4):1961–1990, 2020.
- [34] Sharon Slade and Paul Prinsloo. Learning analytics ethical issues and dilemmas. *American Behavioral Scientist*, 57(10):1510–1529, 2013.
- [35] Lei Yang, Li Feng, Longqing Zhang, and Liwei Tian. Predicting freshmen enrollment based on machine learning. *The Journal of Supercomputing*, 77:11853 – 11865, 2021. URL <https://api.semanticscholar.org/CorpusID:233625118>.
- [36] Bryan Mann, Wei Li, and Kevin Besnoy. Digital divides: K-12 student profiles and online learning. *education policy analysis archives*, 29, 09 2021. doi: 10.14507/epaa.29.6351.
- [37] Daniel A Gutierrez Pachas, Germain Garcia Zanabria, Ernesto Cuadros Vargas, Guillermo Camara Chavez, and Erick Gomez Nieto. Supporting decision making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the latin american context. *Education Sciences*, 13(2):154, 2023.
- [38] João Gabriel Corrêa Krüger, Alceu de Souza Britto Jr, and Jean Paul Barddal. An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, 233:120933, 2023.
- [39] Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- [40] Llatina i el Carib. *Panorama Social de América Latina*. Cepal, 2002.
- [41] Kingsley Okoye, Haruna Hussein, Arturo Arrona-Palacios, Héctor Nahún Quintero, Luis Omar Peña Ortega, Angela Lopez Sanchez, Elena Arias Ortiz, Jose Escamilla, and Samira Hosseini. Impact of digital technologies upon teaching and learning in higher education in latin america: an outlook on the reach, barriers, and bottlenecks. *Education and information technologies*, 28(2):2291–2360, 2023.
- [42] Wei Luo and Yi Qi. An enhanced two-step floating catchment area (e2sfca) method for measuring spatial accessibility to primary care physicians. *Health & place*, 15(4):1100–1107, 2009.
- [43] UNESCO-IIEP. International Institute for Educational Planning. <https://www.iiep.unesco.org/id>, 2025. Accedido: 01-dic-2025.

- [44] John M.. Mendelsohn. Education planning and management, and the use of geographical information systems. page 78, 1996. URL <https://unesdoc.unesco.org/ark:/48223/pf0000105758>.
- [45] Jiawei Ha, Micheline Kambe, and Jian Pe. Data mining: Concepts and techniques. *Data Mining: Concepts and Techniques*, pages 1–703, 1 2011. doi: 10.1016/C2009-0-61819-5. URL <https://www.oreilly.com/library/view/data-mining-concepts/9780123814791/>.
- [46] Keefe Murphy, Sonsoles López-Pernas, and Saqr. *Dissimilarity-Based Cluster Analysis of Educational Data: A Comparative Tutorial Using R*, pages 231–283. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-54464-4_8. URL https://doi.org/10.1007/978-3-031-54464-4_8.
- [47] Agglomerative Nesting (Program AGNES), chapter 5, pages 199–252. 1990. ISBN 9780470316801. doi: <https://doi.org/10.1002/9780470316801.ch5>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch5>.
- [48] Divisive Analysis (Program DIANA), chapter 6, pages 253–279. 1990. ISBN 9780470316801. doi: <https://doi.org/10.1002/9780470316801.ch6>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch6>.
- [49] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857, 1971. ISSN 0006341X. doi: 10.2307/2528823.
- [50] Pinyan Liu, Han Yuan, Yilin Ning, Bibhas Chakraborty, Nan Liu, and Marco Aurélio Peres. A modified and weighted gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, 24(1):305, 2024.
- [51] Joost CF De Winter, Samuel D Gosling, and Jeff Potter. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 21(3):273, 2016.
- [52] James E. Corter. Clustering approaches. In Robert J Tierney, Fazal Rizvi, and Kadriye Erçikan, editors, *International Encyclopedia of Education (Fourth Edition)*, pages 627–636. Elsevier, Oxford, fourth edition edition, 2023. ISBN 978-0-12-818629-9. doi: <https://doi.org/10.1016/B978-0-12-818630-5.10072-7>. URL <https://www.sciencedirect.com/science/article/pii/B9780128186305100727>.
- [53] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.
- [54] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.
- [55] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

- ISSN 0377-0427. doi: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). URL <https://www.sciencedirect.com/science/article/pii/0377042787901257>.
- [56] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. 12 2017.
 - [57] Andry W. Marques. How lgbm deals with missing values, 2024. URL <https://medium.com/@andrywmarques/how-lgbm-deals-with-missing-values-bd361636357f>.
 - [58] GeeksforGeeks. Cross validation and hyperparameter tuning of lightgbm model, 2024. URL <https://www.geeksforgeeks.org/machine-learning/cross-validation-and-hyperparameter-tuning-of-lightgbm-model/>.

6. Anexos

Anexo 1: Composición de las Bases de Datos del INEC

Cuadro 15: Composición de las Bases de Datos del INEC (Vivienda, Hogar, Mortalidad, Emigración y Población)

hhDataset	Variables Clave Relevantes	Dataset	Variables Clave Relevantes
Vivienda	V09/V10: Agua por tubería/fuente V11: Servicio higiénico V12/V13: Electricidad V14: Eliminación basura V15: Cuartos DEF_HAB: Déficit habitacional	Hogar	H01-H04: Dormitorios/cocina/servicio higiénico/ducha H05: Combustible cocinar H06: Agua potable H07: Separación basura H08: Mascotas H09: Tenencia vivienda H10: Servicios/TIC H11/H12: Fallecidos/emigrantes H13: Total personas/hombres/mujeres
Mortalidad	M02: Mes/año fallecimiento M03: Edad al fallecer M04: Sexo M05: Muerte materna M06: Causa	Emigración	E01: Año salida E02: Sexo E03: Edad al salir E04: País destino
Población	P03: Edad P04: Fecha nacimiento P07: Discapacidades P08/P09: Lugar nacimiento/residencia P10: Idiomas P11: Autoidentificación étnica P12: Pueblo indígena P15-P20: Educación P21: Uso TIC P22-P30: Trabajo/ocupación P31: Estado conyugal P32-P35: Fertilidad P36/P37: Género/orientación sexual		

Fuente: <https://www.censoecuador.gob.ec>