**UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ**

**Colegio de Posgrados**

# Aprendizaje No Supervisado para la Identificación de Perfiles Educativos en el Ámbito de la Educación Superior en el Ecuador

**Proyecto de Titulación**

**Luis Alberto Baca Guerrero**

**Israel Pineda, Ph.D.**

**Director de Trabajo de Titulación**

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magister en Ciencia de Datos

Quito, 02 de diciembre de 2024

# UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
# COLEGIO DE POSGRADOS

## HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

### Aprendizaje no supervisado para la identificación de Perfiles Educativos en el ámbito de la Educación Superior en el Ecuador

**Luis Alberto Baca Guerrero**

| | |
|---|---|
| Nombre del Director del Programa: | Felipe Grijalva Arévalo |
| Título académico: | Ph.D. en Ingeniería Eléctrica |
| Director del programa de: | Ciencia de Datos |
| Nombre del Decano del colegio Académico: | Eduardo Alba |
| Título académico: | Doctor en Ciencias Matemáticas |
| Decano del Colegio: | Ciencias e Ingenierías |
| Nombre del Decano del Colegio de Posgrados: | Dario Niebieskikwiat |
| Título académico: | Doctor en Física |

**Quito, diciembre 2024**

# © DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

| | |
|---|---|
| Nombre del estudiante: | Luis Alberto Baca Guerrero |
| Código de estudiante: | 00338886 |
| C.I.: | 1721169496 |
| Lugar y fecha: | Quito, 02 de Diciembre de 2024. |

# ACLARACIÓN PARA PUBLICACIÓN

**Nota:** El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

# UNPUBLISHED DOCUMENT

**Note:** The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

# DEDICATORIA

A mi familia por todo su amor incondicional y su paciencia, este logro es el reflejo del esfuerzo compartido con cada uno de ustedes.

A mis padres y mi hermana, por guiarme y estar presentes en cada paso de mi vida.

A Kimberly, mi compañera de vida, que comparte el día a día conmigo en las buenos y malos momentos.

A mis hijos Luanna y Enzo, espero despertar siempre en ellos esa curiosidad por el aprendizaje constante.

# AGRADECIMIENTOS

A la Universidad San Francisco de Quito por facilitarme los recursos físicos y tecnológicos necesarios para realizar este trabajo de investigación. A mi tutor Israel Pineda, por su guía durante este proceso y por toda la retroalimentación a lo largo del desarrollo de este proyecto. Por último, quiero reiterar el agradecimiento a mi familia por el apoyo y motivación.

# RESUMEN

El uso de técnicas de aprendizaje no supervisado en el análisis de datos educativos está revolucionando la forma en que las instituciones de educación superior y los entes gubernamentales enfrentan los retos del sector. Este trabajo de investigación se centra en la implementación de dos algoritmos de clustering. K-modes, diseñado de forma específica para clusterizar datos categóricos o binarios; y clustering jerárquico con el método de enlace haming, que también esta diseñado para ser usado con datos categóricos, el propósito es identificar perfiles característicos en los estudiantes universitarios en el Ecuador. Se utiliza un conjunto de datos de acceso público que comprende características demográficas y académicas de los estudiantes matriculados en las instituciones de educación superior del Ecuador durante el año 2022, se identifican 43 perfiles característicos diferenciados. Estos perfiles revelan patrones que podrían estar relacionados con la concentración de la oferta académica de carreras/programas en ciertas provincias del Ecuador y las dinámicas de matrícula de estudiantes de acuerdo con sus preferecnias entre los diferentes campos de conocimiento y las diferentes modalidades de estudio.

El análisis destaca algunas problemáticas importantes que afectan a la educación superior, como las brechas existentes en el acceso a la educación superior, las dificultades para garantizar la permanencia estudiantil y la desconexión existente entre la formación académica y las demandas cambiantes del mercado laboral. Estos problemas, además de limitar las oportunidades de los estudiantes, frenan el desarrollo social y económico del país. En este contexto, el uso de algoritmos de clustering se presenta como una herramienta valiosa, identificando patrones en los datos que orienten la toma de decisiones eficaces. Los resultados de este trabajo de investigación sientan las bases para la formulación de nuevas políticas públicas sustentadas en el análisis de datos y orientadas a resolver o mitigar los problemas planteados.

En la implementación de este trabajo se define un pipeline con cinco fases, empezando por la carga y limpieza de datos hasta concluir con la aplicación de los algoritmos de clustering y la interpretación de los perfiles generados. Los resultados obtenidos son evaluados por métricas, como el coeficiente de Silhouette y el índice de Calinski-Harabasz.

La integración de tecnologías de análisis avanzado no solo permite comprender de mejor manera las dinámicas estudiantiles, sino que tambíen puede guiar desde las instituciones de educación superior, intervenciones específicas que fomenten modelos educativos más eficientes, inclusivos y orientados a una educación integral y de calidad.

En resumen, este trabajo de investigación presenta una propuesta innovadora enfocada en la aplicación de técnicas avanzadas de aprendizaje no supervisado, y en el análisis de las problemáticas actuales en el sector de la educación superior, estos hallazgos pueden ser utilizados para promover una educación superior más inclusiva y orientada hacia el desarrollo integral de los estudiantes.

**Palabras clave:** educación superior, desigualdad, universidad, aprendizaje automático, clustering, aprendizaje no supervisado, k-modes, clustering jerárquico

# ABSTRACT

The use of unsupervised learning techniques in educational data analysis is transforming how higher education institutions and governmental entities address challenges in the sector. This research focuses on the implementation of two clustering algorithms: K-modes, specifically designed for clustering categorical or binary data, and hierarchical clustering with the Hamming linkage method, also tailored for categorical data. The objective is to identify distinctive student profiles in Ecuadorian universities. The study utilizes a public available dataset comprising demographic and academic characteristics of students enrolled in Ecuadorian higher education institutions during 2022, identifying 43 differentiated characteristic profiles. These profiles reveal patterns that could be linked to the concentration of academic offerings in specific provinces of Ecuador and the enrollment dynamics of students according to their preferences across different fields of study and modalities.
The analysis highlights several critical issues affecting higher education, such as existing gaps in access to higher education, challenges in ensuring student retention, and the disconnection between academic training and the demands of the labor market. These issues not only limit students' opportunities but also hinder the country's social and economic development. In this context, the use of clustering algorithms emerges as a valuable tool, identifying patterns in data to guide effective decision-making. The results of this research lay the foundation for the formulation of new public policies supported by data analysis, aimed at solving or mitigating the identified problems.
The implementation of this study is structured into a five-phase pipeline, starting with data loading and cleaning and concluding with the application of clustering algorithms and the interpretation of the generated profiles. The obtained results are evaluated using metrics such as the Silhouette coefficient and the Calinski-Harabasz index.
The integration of advanced analytical technologies not only enhances the understanding of student dynamics but also enables higher education institutions to design targeted interventions that promote more efficient, inclusive, and quality-oriented educational models.
In summary, this research presents an innovative proposal focused on the application of advanced unsupervised learning techniques and the analysis of current challenges in the higher education sector. These findings can be used to foster a more inclusive higher education system oriented toward the comprehensive development of students.

**Keywords:** higher education, inequity, university, machine learning, clustering, unsupervised learning, k-modes, hierarchical clustering

# TABLA DE CONTENIDO

# ÍNDICE DE TABLAS

# ÍNDICE DE FIGURAS

# Aprendizaje No Supervisado para la Identificación de Perfiles Educativos en el Ámbito de la Educación Superior en el Ecuador

Luis Alberto Baca Guerrero , Maestria de Ciencia de Datos

***Abstract*—El uso de técnicas de aprendizaje no supervisado en el análisis de datos educativos está revolucionando la manera en que las instituciones de educación superior enfrentan desafíos estructurales. Este trabajo de investigación se enfoca en la aplicación de algoritmos de clustering, como *K-modes* y *hierarchical clustering*, para identificar perfiles educativos y sociales entre los estudiantes universitarios en el Ecuador. Se utiliza un conjunto de datos extenso que comprende características demográficas y académicas de los estudiantes que han sido matriculados en las instituciones de educación superior del Ecuador durante el año 2022, como resultado de este ejercicio se identifican 43 perfiles característicos diferenciados. Estos perfiles revelan patrones que podrían estar relacionados con la concentración de la oferta académica de carreras/programas en ciertas provincias del Ecuador, las dinámicas de matrícula de estudiantes de acuerdo con los diferentes campos de conocimiento y sus preferencias por las modalidades de estudio. Además, se destaca problemáticas importantes, como las brechas existentes en el acceso a la educación superior, las dificultades para garantizar la permanencia estudiantil y la desconexión entre las competencias adquiridas durante la formación académica y las demandas del mercado laboral. Los resultados obtenidos mediante la aplicación de algoritmos de clustering ofrecen una base fundamentada que pudiera respaldar acciones como, el diseño de políticas públicas que promuevan la equidad educativa y favorezcan la inclusión, permitiendo una mejor adaptación de los recursos a las necesidades de los estudiantes y las instituciones.***

***Index Terms*—educación superior, desigualdad, universidad, aprendizaje automático, *clustering*, aprendizaje no supervisado, *k-modes*, *hierarchical clustering*, deserción estudiantil.**

## I. Introducción

ESTE trabajo pretende abordar el reto de mejorar la toma de las decisiones de política pública relacionadas con la educación superior, mediante la identificación de perfiles característicos de los estudiantes, utilizando técnicas de aprendizaje no supervisado se busca analizar la situación específica de la educación superior en el Ecuador. El objetivo principal es agrupar a los estudiantes que han sido matriculados en las instituciones de educación superior durante el año 2022, en base a sus carácteristicas más representativas, formando grupos diferenciados que brinden una perspectiva mas profunda sobre la estructura y la conformación del sistema de educación superior, permitiendo la optimización de las políticas y estratégias planteadas desde los organismos gubernamentales correspondientes en busqueda de brindar una solución efectiva a problemas como las altas tasas de deserción estudiantil, la inaccesibilidad o el limitado acceso a la educación superior, la desconexión entre las competencias adquiridas por los estudiantes durante su formación académica y las demandas de perfiles laborales por parte del sector productivo, entre otros. Por ejemplo, la identificación de perfiles estudiantiles puede ser una herramienta clave para abordar el problema de la deserción estudiantil, que no solo representa una pérdida significativa de talento humano, sino que también implica un desperdicio de recursos para las instituciones y los gobiernos.

Asimismo, el análisis de perfiles permite avanzar en busqueda de la igualdad de oportunidades, al identificar a los estudiantes que enfrentan barreras sociales o culturales que dificultan su permanencia y éxito académico, las instituciones pueden generar políticas de apoyo que ayuden a reducir las brechas de desigualdad.

Por otro lado, también permite mejorar la conexión entre la academia y el sector productivo, al identificar los perfiles de los estudiantes, es posible adaptar los currículos y las estrategias educativas para alinearlos con las demandas actuales y futuras del sector productivo.

En este contexto, la capacidad de segmentar a los estudiantes en perfiles representativos, permitirá a las instituciones de educación superior y entes de gobierno, plantearse acciones de intervención más precisas y efectivas, que se adapten de forma adecuada a las características específicas de cada grupo.

El enfoque adoptado en este trabajo de investigación permite agrupar a los estudiantes en función de la similitud entre sus diversas caracteristicas como, su lugar de residencia, sexo, étnia, nivel de formación, situación socioeconómica o su modalidad de estudio. La ventaja de este enfoque es que permite identificar patrones en los datos sin necesidad de definir categorías previas, esto resulta especialmente útil en el contexto de la educación superior.

El conjunto de datos utilizado presenta información demográfica y académica de los estudiantes de todas las universidades del Ecuador durante el año 2022. Estos datos, se preprocesan y se utilizan para realizar un análisis exploratorio y obtener conclusiones generales

sobre la estructura del sistema de educación superior en el país, posteriormente, los datos se codifican para poder implementar técnicas de clustering, con el objetivo de identificar en estos grupos de estudiantes, algunos perfiles característicos que puedan ser utilizados para entender a profundidad la situación del sistema de educación superior. Tras la segmentación de los estudiantes en grupos homogéneos, se analizan las características predominantes de cada cluster, con el fin de proporcionar recomendaciones que permitan diseñar intervenciones específicas. Este trabajo de investigación no solo pretende ser una contribución técnica en el uso y aplicación de algoritmos de machine learning en el ámbito educativo, sino que también busca transformarse en una línea base para el sistema de educación superior presentando un análisis aplicado que permita tanto a las instituciones educativas como a los entes gubernamentales, tomar decisiones basadas en el análisis de datos.

En resumen, este trabajo de investigación propone un enfoque diferente para comprender la situación del sistema de educación superior mediante el análisis de perfiles característicos, que proporcione a las instituciones educativas y a los entes de gobierno una herramienta concreta para tomar decisiones de política pública.

Las conclusiones que se derivan de este trabajo de investigación brindan una alternativa para optimizar el uso de recursos y están enfocadas hacía una educación superior inclusiva y de calidad, que se conecte con las realidades y demandas del sector productivo.

Este proyecto se encuentra alojado en un repositorio público en la plataforma GitHub, por lo que en caso de necesitar replicarlo o usarlo como referencia, se puede conseguir toda la información en el siguiente link: https://github.com/LuisStatsOps/Clustering_educacion

## II. Problemática

La educación superior en Latinoamérica enfrenta diversos problemas estructurales que comprometen la capacidad que tienen sus sistemas de educación para satisfacer las demandas de los estudiantes y del sector productivo. Entre los desafíos más importantes se puede mencionar los altos índices de deserción estudiantil, la desigualdad de oportunidades de acceso a programas de estudio, y la desconexión entre la formación académica impartida por las instituciones de educación superior y el mercado laboral. Esta problemática limita las oportunidades de éxito de los estudiantes, incidiendo de forma negativa en el desarrollo de las naciones. En este contexto, la identificación de perfiles académicos y sociales mediante técnicas avanzadas de machine learning y clustering puede ofrecer soluciones efectivas para mejorar la calidad de la educación y optimizar la toma de decisiones en políticas educativas.

González & Espinoza [1], describen la deserción como el abandono de la carrera por parte del estudiante, ya sea de forma voluntaria o bajo presión, influido por factores internos o externos que pueden impactar positiva o negativamente su decisión. Según el Banco Mundial, la tasa de abandono estudiantil, principalmente en los primeros años de universidad en América Latina alcanza valores promedio entre 30% y 40%, dependiendo del país y el contexto socioeconómico, llegando a cifras de hasta el 60% en países como Bolivia. En Ecuador, la deserción en la educación superior varía entre un 20% y un 27% [2]. La deserción no solo representa una pérdida significativa de talento humano y oportunidades para los estudiantes, sino que también supone un desperdicio de recursos para las instituciones educativas y el Estado. Los factores económicos desempeñan un rol central en contextos como América Latina, donde las desigualdades socioeconómicas limitan el acceso y la permanencia de estudiantes provenientes de sectores vulnerables [3]. La necesidad de combinar los estudios con un empleo para sostener gastos educativos se encuentra estrechamente relacionada con horarios académicos inflexibles, lo que dificulta la conciliación entre ambas actividades [4]. Otro factor clave es la falta de orientación vocacional adecuada, que conduce a decisiones erróneas al momento de elegir una carrera. Los estudiantes que abandonan suelen reportar que la carrera no correspondía a sus intereses o habilidades, evidenciando la necesidad de reforzar los mecanismos de orientación desde etapas tempranas [5]. Asimismo, problemas académicos como la dificultad en cursar las materias o reprobarlas, están directamente vinculados a bajos niveles de preparación previa, y son acentuados por métodos de enseñanza poco efectivos o falta de apoyo institucional [4].

La inequidad en las oportunidades de acceso a la educación superior es otro reto que enfrenta la región. A pesar de los esfuerzos por democratizar el acceso universitario, aún existen brechas significativas en términos de ingreso y éxito académico. La falta de recursos económicos y la limitada infraestructura en áreas rurales impiden que muchos estudiantes accedan a la educación superior, y para quienes lo logran, su permanencia es incierta. Según Garcia-Penalvo[6], los estudiantes de áreas rurales enfrentan mayores dificultades de acceso, debido a la distancia a la que se encuentran las instituciones educativas y la falta de apoyo financiero, lo que incrementa el riesgo de deserción. Asimismo, el informe de la UNESCO[7] señala que las desigualdades socioeconómicas son un factor crucial que impacta el aprendizaje y la continuidad educativa, afectando desproporcionadamente a estudiantes en entornos de pobreza. Además, investigaciones como las de De Vries *et al.*[4] destacan que en América Latina las tasas de permanencia están marcadas por el contexto socioeconómico, siendo más bajas entre estudiantes de familias con menos recursos económicos.

Otro de los retos importantes es la desconexión entre la formación académica recibida por los estudiantes y las necesidades reales del sector productivo local. Según Sanahuja *et al.*[8], una de las principales causas de esta brecha radica en la limitada vinculación entre las universidades y los sectores productivos, lo que dificulta la adaptación de los contenidos académicos a las

demandas cambiantes del sector productivo. Asimismo, la globalización y la rápida evolución tecnológica han transformado las habilidades requeridas en el ámbito laboral, mientras que los currículos universitarios suelen ser rígidos y poco flexibles, dificultando su adaptación a los nuevos escenarios económicos [9]. Esto hace que los estudiantes culminen sus estudios sin adquirir las competencias actuales necesarias para satisfacer las demandas del sector productivo, lo que dificulta su inserción en el mundo laboral y les genera frustración.

La definición de estos perfiles académicos mediante el proceso de clustering beneficia a las universidades, permitiendoles adaptar sus políticas y estrategias hacia los grupos vulnerables; beneficia también a los estudiantes, quienes podrán recibir una formación de mejor calidad y a las empresas, que pueden utilizar estos perfiles para identificar talento de manera más eficiente. Otro beneficio sería la posibilidad de crear alianzas estratégicas entre el sector académico y el sector empresarial, para que mediante un trabajo conjunto aseguren que las competencias desarrolladas por los estudiantes sean las requeridas en el corto plazo en el mercado laboral.

La clave para implementar estas soluciones radica en contar con datos de calidad y detallados que permitan aprovechar el potencial de las técnicas de machine learning para obtener conclusiones relevantes que aporten y fortalezcan la toma de decisiones académicas.

## III. Trabajos relacionados

Varios estudios han explorado el uso de técnicas de clustering para identificar perfiles en estudiantes de educación superior.

Romero & Ventura[10], resaltan que los datos educativos presentan desafíos únicos, ya que suelen ser heterogéneos, jerárquicos, y ricos en semántica, lo que exige adaptaciones de los algoritmos tradicionales para manejarlos, y aborda el problema de la deserción mediante la utilización de algoritmos de clustering jerárquico como herramienta para definir grupos de estudiantes basados en sus características

Urbano[11], se enfoca en el análisis de datos cualitativos en el ámbito educativo, busca comprender los factores y contextos que afectan a los estudiantes, mediante un análisis detallado de variables. Examina técnicas como la codificación abierta y axial, y hace enfásis en la importancia de identificar patrones y categorías en los datos. Además, menciona que estas técnicas permiten explorar dinámicas institucionales y sociales en la educación superior, ofreciendo una herramienta valiosa para entender fenómenos como la deserción y la desigualdad en el acceso educativo. Malzer & Baum[12], propone una adaptación al algoritmo HDBSCAN, llamada HDBSCAN($\epsilon$), que combina características de HDBSCAN y DBSCAN para abordar desafíos en datasets con densidades variables. Se introduce un umbral adicional ($\epsilon$) que limita las divisiones en la jerarquía de clusters, evitando microclusters en áreas densas y preservando clusters relevantes en áreas menos densas.

Mojarad *et al.*[13], se enfoca en la identificación de grupos de estudiantes según características similares como conocimiento inicial, esfuerzo, consistencia y ritmo, se aplica PCA para reducción de dimensionalidad, *Mean-Shift* para determinar el número de clusters y *K-means* para la clusterización en un grupo de 628 estudiantes. Concluye que los clusters formados son un buen punto de partida para plantear intervenciones grupales efectivas.

Talib *et al.*[14], utiliza una combinación de *K-means* y SVM para analizar el comportamiento de estudiantes en educación superior, logra identificar tres grupos de rendimiento, demostrando que las etiquetas generadas por los clusters mejoraron la precisión de las predicciones realizadas por el modelo SVM en comparación con las calificaciones semestrales.

Huang[15], presenta el algoritmo k-modes, diseñado para superar las limitaciones de k-means en el análisis de datos categóricos. Este método reemplaza la media por la moda como medida central, adaptando la actualización de centroides al valor más frecuente dentro de cada cluster. Además, utiliza una medida de disimilitud específica para variables cualitativas, que evalúa las diferencias entre categorías, permitiendo la agrupación eficiente de este tipo de datos sin sacrificar escalabilidad.

Argiento[16] propone un enfoque de clustering basado en modelos para datos categóricos sin un orden natural, considera la distancia de Hamming para definir una familia de funciones de masa de probabilidad que modelan los datos. Los elementos de esta familia se consideran como núcleos de un modelo de mezcla finita con un número desconocido de componentes. La inferencia bayesiana conjugada se desarrolla para los parámetros del modelo de distribución de Hamming, y se implementa un sampler de Gibbs bloqueado transdimensional para proporcionar una inferencia bayesiana completa sobre el número de clusters, su estructura y los parámetros específicos de cada grupo. Los resultados muestran mejoras en la recuperación de clusters en comparación con enfoques existentes.

Estos estudios demuestran el creciente interés en aplicar algoritmos de machine learning para identificar perfiles educativos que permitan personalizar la enseñanza y mejorar la experiencia académica de los estudiantes.

## IV. Conjunto de datos

El conjunto de datos utilizado para este análisis y entrenamiento, incluye variables relevantes para explicar de mejor manera la estructura del sistema de educación superior en el Ecuador. Los datos han sido obtenidos a traves del portal de datos abiertos que maneja el INEC, con información que ha sido clasificada como Pública. Este conjunto de datos se alimenta de la información que cargan las universidades de forma anual en el Sistema Integral de Información de la Educación Superior (SIIES). El conjunto de datos esta conformado por variables con información demográfica y académica, de los estudiantes que han sido

matriculados en las instituciones de educación superior durante el año 2022. Se compone de 140859 registros agregados que representan a 783211 estudiantes, y cuenta con 23 variables de la cuales 21 variables son categóricas. A continuación, se detallan las variables que componen el conjunto de datos:

- `'AÑO'`: Año en el que se reporta la matrícula de cada estudiante.
- `'CODIGO_IES'`: Código único que identifica a la institución de educación superior que reportó la matrícula de cada estudiante.
- `'CODIGO_CARRERA'`: Código único que identifica al curso en el que se matriculó cada estudiante.
- `'TIPO_SEDE'`: Tipo de sede en la que se encuentra matriculado cada estudiante (MATRIZ o EXTENSIÓN).
- `'PROVINCIA_SEDE'`: Provincia donde se encuentra ubicada la institución de educación superior.
- `'CANTON_SEDE'`: Cantón en el cual se encuentra ubicada la institución de educación superior.
- `'SEXO'`: Género de cada estudiante (Hombre o Mujer).
- `'ETNIA'`: Grupo étnico al que pertenece cada estudiante.
- `'PUEBLOS_NACIONALIDAD'`: Pueblo o nacionalidad indígena al que pertenece el estudiante.
- `'DISCAPACIDAD'`: Indica si el estudiante presenta alguna discapacidad.
- `'PAIS_NACIONALIDAD'`: País de nacionalidad de cada estudiante.
- `'PAIS_RESIDENCIA'`: País de residencia de cada estudiante en el momento que se realizó el proceso de matrícula.
- `'PROVINCIA_RESIDENCIA'`: Provincia de residencia del estudiante en el momento que se realizó el proceso de matrícula.
- `'CANTON_RESIDENCIA'`: Cantón de residencia del estudiante en el momento que se realizó el proceso de matrícula.
- `'NOMBRE_IES'`: Nombre de la institución de educación superior que reporta la matrícula de cada estudiante.
- `'TIPO_FINANCIAMIENTO'`: Tipo de financiamiento de la institución de educación superior que reporta la matrícula de cada estudiante.
- `'NOMBRE_IES'`: Nombre del curso en el que se matriculó cada estudiante.
- `'CAMPO_AMPLIO'`: Campo amplio de conocimiento al que pertenece la carrera/programa que se encuentra cursando cada estudiante.
- `'CAMPO_ESPECIFICO'`: Campo específico de conocimiento al que pertenece la carrera/programa que se encuentra cursando cada estudiante.
- `'CAMPO_DETALLADO'`: Campo detallado de conocimiento al que pertenece la carrera/programa que se encuentra cursando cada estudiante.
- `'NIVEL_FORMACIÓN'`: Nivel de formación del curso en el que se encuentra matriculado cada estudiante.
- `'MODALIDAD'`: Modalidad de estudios del curso en el que se encuentra matriculado cada estudiante.
- `'tot'`: Variable que representa la frecuencia o número de veces que un registro aparece en el dataset. Es decir, aporta con información sobre el número de individuos que comparten todas las características de cada registro.

El conjunto de datos utilizado en este proyecto se encuentra disponible en el link: https://datosabiertos.gob.ec/dataset/base-de-datos-matricula-uep-2022.

## V. Metodología

Para abordar el análisis del conjunto de datos, se utilizan varias técnicas de preprocesamiento y visualización, y algoritmos de clusterización enfocados en facilitar la interpretación de los perfiles característicos. Se construye el código en un archivo de formato .ipynb generando tareas que se agrupan en diferentes fases. El pipeline de ejecución definido, inicia con la carga y preprocesamiento de los datos, la exploración y análisis de datos, la selección de variables de interés y codificación del conjunto de datos, la aplicación de modelos de clusterización y la identificación de los perfiles académicos generados.

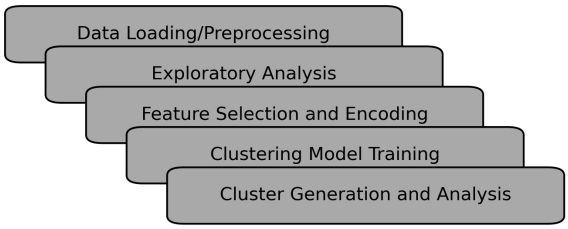


Figure 1. Workflow

### A. Carga y Preprocesamiento de datos

Los datos se encuentran en formato .xlsx, por lo que se realiza la carga al entorno de ejecución mediante el uso de la librería pandas y el motor openpyxl. El archivo cuenta con 399176 registros agregados correspondiente a los estudiantes matriculados en las instituciones de educación superior del Ecuador durante los años 2020, 2021 y 2022, sin embargo, para objeto de este estudio se procede a filtrar únicamente los registros del año 2022 (140859 registros), ya que los registros de los años 2020 y 2021 cuentan con una cantidad importante de estudiantes que han sido registrados también en el año 2022 debido a la duración de sus cursos de estudio.
El preprocesamiento garantiza la calidad de los resultados en los procesos de clustering, especialmente al manejar variables categóricas. Para el caso específico de variables categóricas, Kuhn y Johnson[17], resaltan la importancia de la utilización de técnicas de codificación de datos, como la creación de variables dummy o variables binarias, para

representar las categorías de manera que los algoritmos puedan calcular distancias relevantes entre observaciones. Además, los autores destacan que en el caso de valores cuantitativos la normalización o la estandarización de datos asegura que todos los predictores contribuyan equitativamente en algoritmos basados en métricas de distancia, como *Hierarchical Clustering o K-means*. Según Kuhn y Johnson[17], el impacto del preprocesamiento se refleja en la mejora de la calidad de los clusters formados, y plantean que los beneficios incluyen una mejor interpretación de los resultados, mayor robustez de los modelos y reducción del ruido en los datos, lo que aumenta la eficiencia de los algoritmos de aprendizaje no supervisado.

Después de realizar una exploración inicial de los datos y de verificar que no existen valores nulos o faltantes, se realiza un re etiquetado de algunas variables que cuentan con nombres extensos, con el objetivo de mejorar la calidad de las visualizaciones que son elaboradas posteriormente.

En cuanto al proceso de limpieza de datos, se identifica que en la variable `'CAMPO_AMPLIO'` existe una categoría `'NO_REGISTRA'` con 10 registros, en el caso de la variable `'PROVINCIA_RESIDENCIA'` también existe una categoría `'NO_REGISTRA'` con 810 registros y además existe una categoría `'ZONAS_NO_DELIMITADAS'` que cuenta con 10 registros, por lo que después de analizar los registros asociados a estas categorías se decide eliminarlos con el fin de que no generen mayor ruido en el proceso de clusterización, ya que estos pueden influir negativamente en la calidad del modelo y en los resultados del análisis. Además, debido a que hasta cierto punto, la categoría `'TERCER_NIVEL_TECNICO-TECNOLOGICO_SUPERIOR'` que se encuentra dentro de la variable `'NIVEL_FORMACIÓN'` es nuevo dentro del sistema de educación superior, se elimina esta categoría que cuentan con 1431 registros. Luego de todo este proceso de limpieza se obtiene un conjunto de datos de 138599 registros.

Además, para poder aplicar los algoritmos de clustering en este trabajo de investigación, es necesario transformar todas las variables categóricas a un formato numérico adecuado, debido a que esto facilita su procesamiento por los algoritmos de machine learning.

### *B. Exploración y análisis de datos (EDA)*

El EDA se centra en analizar las variables categóricas relacionadas con la forma en la que se distribuyen estudiantes matriculados y la oferta académica vigente, durante el año 2022, utilizando técnicas descriptivas y herramientas de visualización avanzadas que permiten observar tanto la composición de las categorías como las relaciones entre ellas.

El primer paso es calcular las distribuciones de frecuencias de las variables categóricas, este análisis permite obtener un resumen tabular que describe la cantidad de observaciones en cada categoría. La construcción de tablas de frecuencias es esencial para identificar posibles problemas de desbalanceo en los datos, como categorías con pocas observaciones o concentraciones excesivas en algunas clases. Este enfoque está respaldado por la literatura estadística, que enfatiza la importancia de explorar distribuciones iniciales como una base sólida para cualquier análisis subsecuente[18].

Para entender de mejor manera la interacción entre la oferta académica y los estudiantes, se emplean gráficos de tipo sunburst. Esta herramienta visual representa relaciones entre categorías principales y subcategorías mediante anillos concéntricos, donde las categorías más generales ocupan los niveles exteriores y las subcategorías se distribuyen hacia el interior. Los gráficos sunburst proporcionan una representación clara de las proporciones relativas entre los niveles jerárquicos, permitiendo identificar cómo se distribuyen los estudiantes en función de la modalidad de estudio, los niveles de formación y los campos de conocimiento. Esta técnica de visualización, reconocida por su capacidad de sintetizar estructuras complejas en formatos intuitivos[19], garantiza una visualización interactiva y flexible.

Se construye un treemap para visualizar la distribución de la oferta educativa por provincia. Este tipo de gráfico distribuye las categorías en bloques, cuyo tamaño es proporcional a la cantidad de observaciones asociadas, proporcionando una representación compacta y efectiva de los datos jerárquicos[20].

Como aporte complementario, se genera un grafo que busca analizar las relaciones entre la provincia de residencia de los estudiantes y los campos de conocimiento de la carrera/programa en la que se encuentran matriculados. La implementación del grafo se realiza usando la biblioteca NetworkX, que es reconocida por su capacidad para construir representaciones relacionales complejas[21]. Estas visualizaciones complementan los análisis previos, proporcionando una perspectiva relacional mas profunda y detallada sobre la situación del sistema de educación superior en el Ecuador.

En resumen, el EDA combina técnicas de visualización para analizar la estructura del dataset, identificar patrones clave y generar pistas sobre la distribución de los estudiantes y la oferta académica. Además, este enfoque metodológico se convierte en la base para la selección de variables y la modelización.

### *C. Selección de variables y codificación*

Se aplican técnicas de selección de variables para reducir la dimensionalidad del conjunto de datos, eliminando aquellas que aportan información redundante para los análisis de clustering. Se genera una matriz de correlación calculada mediante Cramér's V, una medida estadística que evalúa la fuerza de asociación entre variables categóricas. Cramér's V permite mediante el cálculo del chi cuadrado identificar relaciones significativas entre pares de variables, facilitando la detección de redundancias al no asumir relaciones lineales u ordinales entre las

categorías[22]. Se establece un umbral de correlación de 0.8 para considerar a dos variables como redundantes. Por ejemplo, variables como `'CAMPO_ESPECIFICO'` y `'CAMPO_DETALLADO'` fueron eliminadas al mostrar una alta correlación con `'CAMPO_AMPLIO'`, reteniendo esta variable como representativa del grupo. Este proceso se complementa con el conocimiento experto y el análisis del EDA que fue realizado previamente, permitiendo tomar decisiones fundamentadas en casos donde las correlaciones estadísticas no sean suficientes para justificar la exclusión de una variable. Como paso previo a la construcción de la matriz de Cramér's V se eliminan las variables numéricas `'AÑO'` que tenía únicamente el valor 2022, y `'tot'` que representaba la frecuencia de cada registro, lo que podría haber impactado negativamente en el proceso de clustering. También se elimina la variable `'PAIS_RESIDENCIA'` debido a que la mayoría de sus registros tienen el valor `'ECUADOR'`.

Las variables seleccionadas son transformadas mediante One-Hot Encoding, una técnica que convierte categorías en variables binarias, asignando un valor de 1 o 0 según la presencia o ausencia de cada categoría. Esta técnica es ampliamente utilizada en análisis no supervisado porque asegura que los algoritmos de clustering interpreten las categorías de manera adecuada, evitando relaciones implícitas de orden o distancia entre ellas[23]. Las variables son etiquetadas de forma descriptiva para preservar la trazabilidad con las categorías originales. La aplicación de One-Hot Encoding facilita la interpretación de los clusters generados, ya que estos pueden describirse en términos de las variables predominantes en cada grupo.

Con las variables codificadas, se recalcula la matriz de Cramér's V buscando identificar posibles correlaciones entre las nuevas variables binarias creadas. Como último paso se realiza un análisis de correlación de Pearson con las variables seleccionadas, aquí se eliminan aquellas variables que se encuentran correlacionadas de forma inversa y perfecta, para evitar redundancias. Por ejemplo, variables como 'TIPO_SEDE_EXTENSION' y 'NIVEL_FORMACIÓN_POSGRADO' fueron descartadas debido a su alta correlación con otras características que ya representaban la misma información. Esta preparación previa mejora el desempeño de los algoritmos de clustering.

### *D. Modelos de clustering*

En la fase de clustering se implementan los algoritmos *Hierarchical Clustering* y *K-Modes*. El objetivo principal es identificar la configuración más adecuada de parámetros y técnicas que permitan una agrupación óptima de los datos categóricos, maximizando la coherencia interna de los clusters y minimizando el ruido. En una fase preeliminar se probó también la implementación de HDBSCAN sin embargo no se obtuvo buenos resultados. Luego de la evaluación de los dos algoritmos, se concluye que el *Hierarchical Clustering* demuestra ser el que mejor se desempeña en términos de interpretabilidad y generación de los perfiles de caracterización.

***Hierarchical clustering***

Genera una representación jerárquica de los datos, permitiendo analizar la estructura de los grupos en diferentes niveles de granularidad. Este algoritmo genera un dendrograma que muestra cómo se forman los clusters mediante un enfoque aglomerativo o divisivo[24]. La capacidad para identificar subgrupos dentro de clusters principales lo hace útil para interpretar relaciones jerárquicas en los datos, lo que facilita la identificación de patrones específicos en subpoblaciones de los registros analizados.

Para configurar el modelo *Hierarchical clustering* se realizan pruebas con diferentes métricas de distancia como euclidean, bray-curtis, hamming y jaccard, debido a la naturaleza de los datos, además se considera métodos de enlace, como ward, complete, average, y single. Estas configuraciones han sido probadas para identificar la mejor estructura jerárquica de los datos, generando dendrogramas que permiten evaluar visualmente los niveles de agrupación en diferentes puntos de corte[24]. En este caso específico se decide usar el método average en conjunto con la métrica de enlace Hamming que esta diseñada de forma especial para datos binarios y categóricos. El dendrograma resultante permite definir un punto de corte óptimo con base en la distancia entre los clusters, asegurando una partición significativa, así como también una adecuada interpretación de los perfiles formados.

***K-modes***

Este algoritmo introducido por Huang[15], está diseñado para agrupar datos categóricos minimizando la disimilitud entre registros mediante una métrica basada en el número de discrepancias entre valores categóricos, es una variante de *K-means*, que emplea la moda para actualizar los centroides de los clusters. Es útil para identificar patrones en datos cualitativos como la modalidad de estudios o el campo de conocimiento. Sin embargo, su capacidad para capturar relaciones complejas es limitada debido a su estructura rígida de determinación de número de clusters.

La implementación de *K-modes* empieza determinando $k$ que es el número óptimo de clusters, para elegir el mejor $k$ se realiza una comparativa de varias métricas entre ellas el costo, siguiendo el criterio de la 'curva del codo'[25]. A continuación, con el valor de $k$ definido, se inicializan de los centroides probando con el método *Huang*, que selecciona centroides aleatorios, y luego con el método *Cao*, que los elige según la densidad de los datos[26]. Finalmente, los registros se asignan al cluster más cercano según la disimilitud, y los centroides se recalculan hasta que no haya cambios significativos. Este proceso genera clusters interpretables, permitiendo identificar perfiles útiles para la toma de decisiones.

Para evaluar la calidad de los clusters generados con estos dos algorítmos, se plantean tres métricas que tienen diferentes propósitos de evaluación.

- **Índice de Correlación Cophenética (CCC)**: Este índice que es específico para *Hierarchical Clustering*, mide la coherencia entre las distancias originales y las distancias representadas en el dendrograma. Un valor alto de CCC indica que el dendrograma refleja con precisión las relaciones entre los datos[27].
- **Coeficiente de Silhouette** : Esta métrica evalúa la separación entre los clusters y su cohesión interna, proporcionando un valor entre -1 y 1. Valores cercanos a 1 indican que los registros están bien agrupados dentro de sus clusters y alejados de otros grupo[28].
- **Índice de Calinski-Harabasz** : Este índice mide la proporción de la dispersión interna de los clusters frente a su dispersión externa, favoreciendo agrupaciones con alta cohesión interna y buena separación[29].

### *E. Evaluación de los perfiles académicos creados*

Para poder evaluar los perfiles acádemicos, se realiza la asignación de las etiquetas generadas mediante la aplicación de cada algoritmo de `clustering` a cada registro del conjunto de datos. Esta etapa es crucial para conectar los resultados del modelo con las características originales de los datos y establecer una base sólida para la interpretación de los perfiles generados James[30].
Posteriormente, se calculan las características promedio de cada cluster mediante funciones de agrupación. Esto permite identificar las categorías más representativas en términos de frecuencia relativa, en este punto toma especial importancia la codificación realizada previamente mediante one hot encoding ya que al tener cada columna en formato binario la interpretación de las características se vuelve sencilla. Los valores promedio cercanos a 1 indican que una categoría está altamente representada en el cluster correspondiente, además se calcula el tamaño de cada cluster, es decir, el número de registros que conforman cada grupo formado, esto aporta información clave sobre la representatividad de cada perfil identificado en el conjunto de datos. Este enfoque cuantitativo es ampliamente utilizado en el análisis de clusters, ya que facilita la interpretación de los datos categóricos al convertirlos en descripciones agregadas que son fáciles de comprender y comparar[31].
Se seleccionan aquellas características con valores promedio superiores a 0.5. Este criterio permite destacar las categorías más representativas dentro de cada grupo, excluyendo aquellas que no son relevantes o que presentan baja frecuencia. Además, se incluyen manualmente las características que presentaron un índice de correlación de Pearson de -1 luego de aplicar one-hot encoding; esto se debe a que, en el proceso previo de selección de características, se eliminaron las variables con las cuales tenían correlación inversa perfecta. De esta forma, si estas variables incluidas tienen valores promedio cercanos a 0, se puede concluir que en ese cluster existe una fuerte representación de la característica con la cual se encontraban perfectamente correlacionadas y que fue eliminada anteriormente. Así, se logra la interpretación de cada perfil académico basados en las características más frecuentes, proporcionando una visión completa de los patrones detectados en los datos. Esta metodología asegura que los clusters generados sean estadísticamente coherentes e interpretables Facilitando su uso en contextos prácticos, como el diseño de políticas educativas o la segmentación de estudiantes.
La selección de características con valores promedio superiores a un umbral específico es una práctica común en el análisis de clusters para identificar las variables más representativas de cada grupo. Este enfoque permite resaltar las categorías predominantes y facilita la interpretación de los perfiles generados[32]. Por otro lado, la inclusión manual de características con correlación inversa perfecta, eliminadas previamente, es una estrategia que busca mantener la integridad de la información y asegurar que las variables relevantes estén presentes en el análisis final. Esta práctica es esencial para garantizar que los clusters sean interpretables y útiles en aplicaciones prácticas, como la segmentación de estudiantes o el diseño de políticas educativas[30].

## VI. Resultados

En primera instancia se presentan los resultados del análisis exploratorio, estos han sido estructurados en dos segmentos, por un lado, está el análisis de la oferta académica que permite observar patrones significativos en las características de las carreras/programas que ofertan las instituciones de educación superior; y por otro lado tenemos a los estudiantes que son quienes demandan las carreras/programas ofertados. Estos resultados se resumen a continuación:

### *A. Oferta Académica de carreras/programas en Ecuador*

La oferta académica en el Ecuador al año 2022 es de 3674 carreras/programas, de estos el 54 % son ofertados por instituciones con tipo de financiamiento particular y el 46% por instituciones públicas. Además, se presenta una marcada concentración en carreras de pregrado, que representan aproximadamente el 68% del total, mientras que los programas de posgrado corresponden al 32%. Esto sugiere una predominancia de programas de formación de tercer nivel, en comparación con la educación de cuarto nivel. La modalidad presencial es la que predomina en la oferta, abarcando cerca del 74% de las carreras/programas. A continuación, se presenta un grafico de estilo sunburst que permite explorar de forma jerárquica los datos de la oferta académica en el Ecuador.

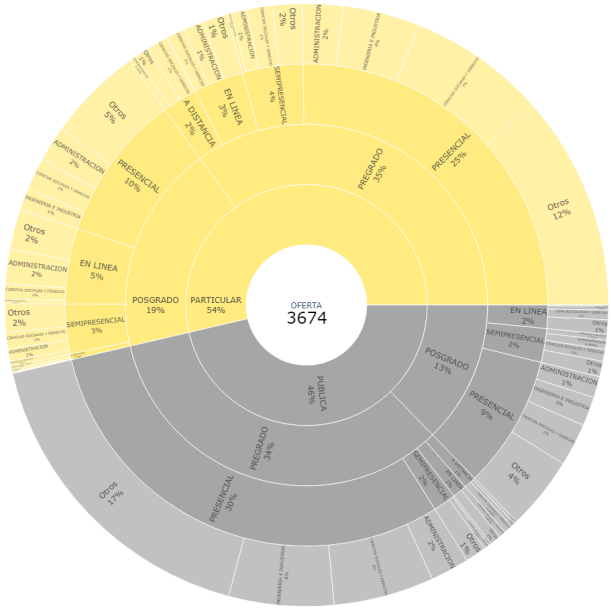


Figure 2. Oferta académica de Carreras/Programas en el Ecuador

En cuanto a los campos de conocimiento, la oferta académica esta distribuida mayoritariamente en los campos de Ciencias Sociales y Derecho (25%), Ingeniería e Industria (13%) y Administración (13%). Estos tres campos agrupan más de la mitad de las carreras/programas ofertados.

La distribución geográfica de la oferta destaca una fuerte concentración en las provincias de Guayas y Pichincha, que juntas representan aproximadamente el 53% del total de la oferta a nivel nacional. Si profundizamos a nivel cantonal, Quito y Guayaquil concentran la mayor parte de esta oferta, con 27% y 16% registros respectivamente. A continuación, se presenta un grafico de estilo treemap que permite entender la situación geográfica de la distribución de la oferta académica en el Ecuador.

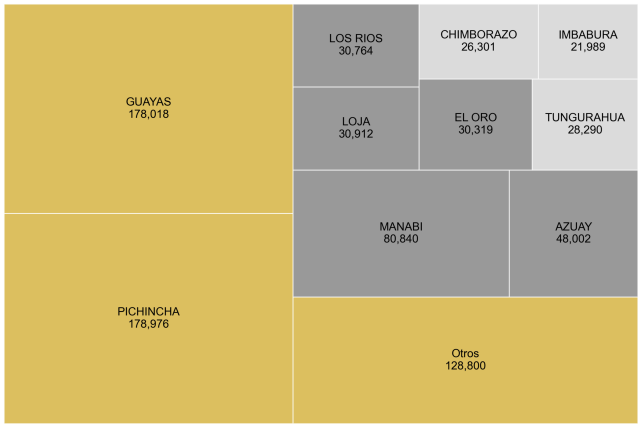


Figure 3. Distribución geográfica de la Oferta académica de Carreras/Programas en el Ecuador

*B. Distribución de estudiantes de educación superior en Ecuador*

La cantidad de estudiantes matriculados en Ecuador en el año 2022 es de 783211 estudiantes que se encuentran distribuidos de la siguiente manera; la mayoría de los estudiantes se encuentran matriculados en instituciones públicas 64% mientras que en instituciones con tipo de financiamiento particular se concentra el 36% de los estudiantes. Además, se presenta una marcada concentración de estudiantes en niveles de pregrado, que representan el 92% del total, mientras que los programas de posgrado corresponden al 8%.

La modalidad presencial es la que predomina en la distribución de estudiantes a nivel nacional, abarcando cerca del 72% de las carreras/programas. A continuación, se presenta un gráfico de estilo sunburst que permite explorar de forma jerárquica la distribución de los estudiantes en el sistema de educación superior en el Ecuador.

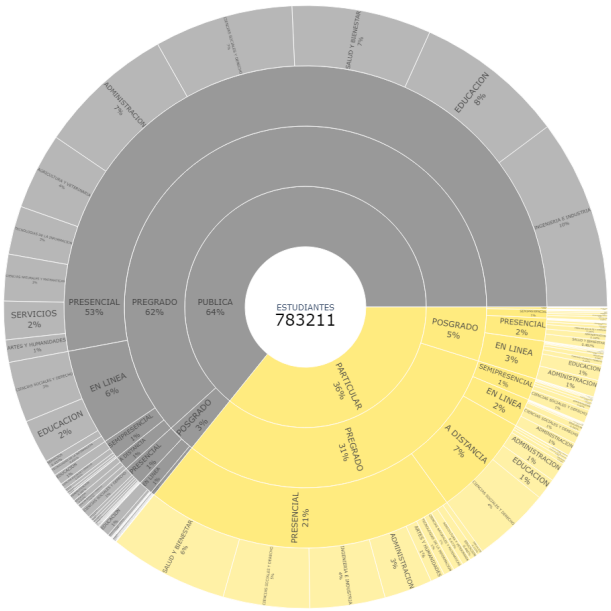


Figure 4. Distribución de estudiantes de educación superior en Ecuador

Desde la perspectiva de los campos de conocimiento, la oferta esta distribuida de forma predominante en las áreas de Ciencias Sociales y Derecho (23%), Educación (15%) e Ingeniería e Industria (14%). Estas tres áreas de conocimiento agrupan a más de la mitad de los estudiantes del sistema de educación nacional.

En cuanto a la distribución de estudiantes de acuerdo con sus características demográficas, se observa una ligera predominancia de estudiantes de genero Mujer (54%), y una mayor concentración de estudiantes pertenecientes al grupo étnico mestizo (61%). El resto de grupos étnicos, como, indígenas y montubios, tienen una menor representación.

### C. Correlaciones encontradas

El análisis de correlación de variables realizado después de aplicar *One Hot Encoding* reveló, de forma preeliminar, relaciones interesantes entre las variables seleccionadas para el ejercicio de clusterización. Por ejemplo, `'PROVINCIA_SEDE_LOJA'` muestra alta correlación con `'MODALIDAD_A DISTANCIA'` (0.733167), es decir, existe una relación importante entre las universidades con sede en la provincia de Loja y los estudiantes matriculados en carreras/programas de modalidad a distancia. `'PROVINCIA_SEDE_MANABI'` tiene una correlación importante con `'PROVINCIA_RESIDENCIA_MANABI'` (0.644846), esto muestra que los estudiantes que estudian en la provincia de Manabí tienden a residir en gran medida en la misma provincia. Algo parecido sucede con `'PROVINCIA_SEDE_GUAYAS'` que se encuentra correlacionada con `'PROVINCIA_RESIDENCIA_GUAYAS'` (0.512129) la siguiente tabla nos permite visualizar las correlaciones encontradas:

Table I
Correlación después de OHE. Se acorta el nombre de algunas variables

| Variable 1 | Variable 2 | Corr |
|---|---|---|
| PROV_SEDE_LOJA | MOD_A_DISTANCIA | 0.7331 |
| PROV_SEDE_MANABI | RESIDENCIA_MANABI | 0.6484 |
| PROV_SEDE_GUAYAS | RESIDENCIA_GUAYAS | 0.5121 |
| MOD_A_DISTANCIA | MOD_PRESENCIAL | -0.5064 |
| MOD_EN_LINEA | MOD_PRESENCIAL | -0.6392 |
| TIPO_SEDE_EXTENSION | TIPO_SEDE_MATRIZ | -1.0000 |
| TIPO_PARTICULAR | TIPO_PUBLICA | -1.0000 |
| NIVEL_POSGRADO | NIVEL_PREGRADO | -1.0000 |

### D. Modelos de clustering

Se probaron diversas configuraciones de parámetros en cada uno de los algoritmos buscando optimizar los resultados. Los dos métodos se ejecutaron sobre el mismo conjunto de datos, permitiendo una comparación directa en términos de eficiencia y calidad de los clusters formados. Es necesario mencionar que para calcular las métricas se ha utilizado una muestra representativa del conjunto de datos de 50000 registros, esto debido a que al calcular las métricas con todos los datos el costo computacional era demasiado elevado.

***K-modes***, Para este caso se define un $k$=20 clusters como resultado de la aplicación del método de la curva del codo y de analizar las métricas de costo, silhouette score y calinski harabasz score. A continuación se muestran los resultados obtenidos para los diferentes valores de $k$ ,10 y 20:

Table II
Resultados con K-modes

| Clusters | Costo | Silhouette | Calinski |
|---|---|---|---|
| 10 | 180398.0 | 0.086369 | 2032.408603 |
| 20 | 159135.0 | 0.109196 | 1459.076336 |

Se generan 20 clusters de diferentes tamaños, siendo el mayor el Cluster 0. con 43849 registros y el mas pequeño, el Cluster 19 con 475 registros

***Hierarchical Clustering***, permitió identificar una estructura jerárquica clara en los datos, lo que reveló cómo los grupos más amplios pueden subdividirse en subgrupos con características más específicas. Se probó varias configuraciones cambiando las métricas de distancia y los metodos de enlace, se define que la combinación de la metrica de distancia 'average' junto con el método de enlace 'hamming' es el que mejor se ajusta al conjunto de datos y brinda la mejor interpretación de los clusters construidos. A continuación se muestra el dendograma generado, en este dendograma se puede apreciar el punto de corte de los clustes, que ha sido definido en `0.10`.

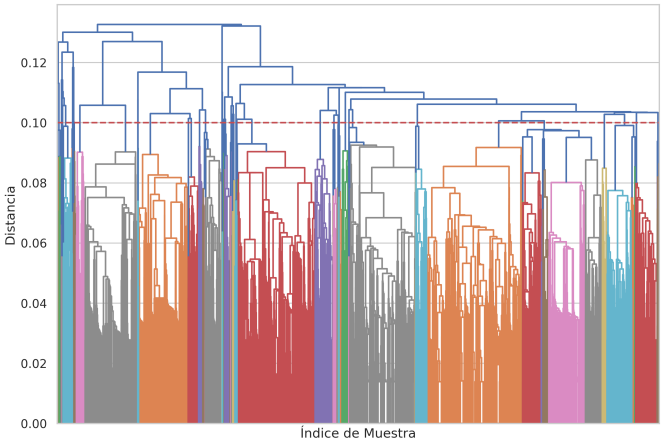


Figure 5. Dendograma de Hierarchical Clustering

Se generaron 43 clusters a una altura de corte de 0.10. Existen clusters de tamaño pequeño que al no aportar en mayor medida se pueden clasificar como ruido. Al igual que con k-modes, para calcular las métricas se ha utiliza un df_sample de 50000 registros, esto debido a que al calcular las métricas con todos los datos el costo computacional era demasiado elevado.

Table III
Resultados con Hierarchical Clustering

| Descripción | Valor |
|---|---|
| Clusters con altura de corte de 0.10 | 43 |
| Índice de Correlación Cophenética | 0.5891 |
| Silhouette Score | 0.1524 |
| Calinski-Harabasz Score | 662.7389 |

Luego de analizar las métricas de cada uno de los modelos aplicados y de revisar los perfiles caraterísticos generados, se define acoger los resultados del algoritmo de *Hierarchical Clustering*. A continuación se muestran las caracteristicas mas relevantes de los principales clusters generados con este algoritmo, esto permite una interpretación relativamente sencilla.

Table IV
Principales características del Cluster 38

| **SIZE** | **22141** |
|---|---|
| MODALIDAD_PRESENCIAL | 0.999 |
| NIVEL_FORMACIÓN_PREGRADO | 0.942 |
| TIPO_SEDE_MATRIZ | 0.908 |
| PROVINCIA_SEDE_GUAYAS | 0.884 |
| ETNIA_MESTIZO | 0.666 |
| PROVINCIA_RESIDENCIA_GUAYAS | 0.552 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.319 |

El Cluster 38 es el más grande y agrupa estudiantes mestizos que cursan estudios presenciales de pregrado en la provincia de Guayas.

Table V
Principales características del Cluster 12

| **SIZE** | **12108** |
|---|---|
| TIPO_SEDE_MATRIZ | 0.999 |
| MODALIDAD_A_DISTANCIA | 0.997 |
| NIVEL_FORMACIÓN_PREGRADO | 0.961 |
| PROVINCIA_SEDE_LOJA | 0.949 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.912 |
| ETNIA_MESTIZO | 0.658 |

El Cluster 12 incluye estudiantes mestizos que estudian carreras de pregrado en modalidad a distancia en instituciones de educación superior con financiamiento particular y que se ubican en la provincia de Loja.

Table VI
Principales características del Cluster 36

| **SIZE** | **15354** |
|---|---|
| PROVINCIA_SEDE_MANABI | 0.983 |
| NIVEL_FORMACIÓN_PREGRADO | 0.952 |
| MODALIDAD_PRESENCIAL | 0.843 |
| TIPO_SEDE_MATRIZ | 0.838 |
| PROVINCIA_RESIDENCIA_MANABI | 0.709 |
| ETNIA_MESTIZO | 0.566 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.035 |

El Cluster 36 agrupa a estudiantes meztisos de la provincia de manabí, que cursan estudios de pregrado en modalidad presencial, en instituciones públicas de educación superior de la provincia de Manabí.

### *E. Análisis*

La aplicación de técnicas de aprendizaje no supervisado y el análisis exploratorio han permitido obtener hallazgos significativos, tanto sobre la oferta académica, como sobre la distribución de estudiantes y los perfiles identificados en el sistema de educación superior del Ecuador.

El análisis de la oferta académica evidencia una predominancia de las carreras de pregrado (68%) sobre los programas de posgrado (32%). Además, se observa que la oferta académica se encuentra concentrada en provincias como Guayas y Pichincha, que juntas suman más del 50% de los programas académicos. Esto plantea una clara inequidad territorial, ya que las regiones periféricas tienen menos opciones de acceso a las carreras/programas.

En cuanto a la distribución de estudiantes, se observa que predomina la modalidad presencial (72%) y la alta concentración de estudiantes en instituciones públicas (64%) indica que aunque las instituciones privadas ofrecen un mayor número de programas, las públicas son las que abarcan la mayoría de la población estudiantil, este factor puede explicar en gran medida situaciones como el hacinamiento en las aulas o el mal estado de los ambientes de aprendizaje, ya que en ocasiones sobrepasan su capacidad óptima, comprometiendo los procesos de enseñanza-aprendizaje. Esta alta concentración de estudiantes en las instituciones públicas también puede estar influenciada por factores económicos y por las políticas de acceso a la educación pública. Sin embargo, esta disparidad sugiere la necesidad de evaluar el impacto de estas políticas para garantizar una distribución más equitativa de estudiantes.

Asimismo, el análisis de correlaciones entre variables destacó patrones relevantes. Por ejemplo, la alta correlación entre las instituciones de educación superior con sede en Loja y los estudiantes matriculados en modalidad a distancia, que se entiende por la fuerte apuesta de la Universidad Particular de Loja a ofertar carreras y programas en esta modalidad, captando estudiantes de diferentes partes del país e incluso estudiantes residentes en otros países como Estados Unidos. La alta correlación entre la ubicación de las instituciones y la residencia de los estudiantes en Manabí y Guayas sugiere que muchos estudiantes optan por estudiar cerca de sus lugares de residencia, lo que no es tan marcado en la provincia de Pichincha, esto podría deberse a las políticas públicas de asignación de cupos que en ocasiones implican que el estudiante para acceder a una plaza tenga que aceptar trasladarse a vivir en otra provincia. Este análisis de correlaciones si bien no indica causalidad directamente, si puede señalar ciertas áreas en las que tal vez es necesario intervenir: las bajas correlaciones en ciertas modalidades de estudio y ubicaciones geográficas indican posibles oportunidades para ampliar el acceso y diversificar las opciones educativas en regiones menos favorecidas.

El dendograma de representación del conjunto de datos Fig.5 aporta información importante, se puede identificar ciertas areas que se agrupan y se ajustan de forma natural, mientras existen otras areas en las que a simple vista se observa que los datos se encuentran solapados y que los clusters no se encuentran del todo diferenciados, este solapamiento es algo común cuando se trabaja con variables categóricas, especialmente cuando se cuenta con muchas categorías dentro de las variables. El valor obtenido para el silhouette score también respalda este análisis. Sin embargo, el objetivo de este trabajo es ir mas alla de las métricas y enfocar el análisis en las características que predominan en cada cluster formado y poder interpretarlas y relacionarlas con la realidad del sistema nacional de educación ecuatoriano.

El clustering jerárquico, configurado con la métrica de distancia de *Average* y método de enlace *Hamming*, generó

43 clusters con características distintivas Fig.5. Entre ellos, destacan:

'Cluster 38' (22141 registros)Tabla.IV: Esta compuesto por estudiantes predominantemente de modalidad presencial, con moderada representación en la provincia de Guayas, en su mayoría de nivel pregrado, pertenecientes al grupo étnico mestizo y al tener baja representatividad en la variable 'TIPO_FINANCIAMIENTO_PARTICULAR' (0.319) se puede deducir que este grupo de estudiantes pertenece en su mayoría a instituciones con financiamiento público. Este análisis es interesante debido a que en la provincia del Guayas existen mas instituciones de educación superior con financiamiento particular, sin embargo lo que nos muestra este cluster es que existe una mayor concentración de estudiantes matriculados en las instituciones publicas.

'Cluster 12' (12108 registros)Tabla.V: Identifica estudiantes en modalidad a distancia, con una concentración significativa en instituciones procedentes de la provincia de Loja y de tipo de financiamiento particular, orientados principalmente hacia las carreras de pregrado en el campo de conocimiento de las ciencias sociales y el derecho. Este perfil destaca la importancia de poder ofertar otras modalidades de educación como estrategia para ampliar el acceso en áreas con poca oferta académica. En este punto es importante entender que a pesar de que la modalidad de formación predominante de este cluster es, a distancia, muchas de estas carreras se están reformulando en base a una modalidad en línea que permite mayor interacción del estudiante con los docentes y los ambientes de aprendizaje, mejorando la calidad de los programas académicos.

'Cluster 36' (15354 registros)Tabla.VI: Agrupa estudiantes predominantemente de modalidad presencial, con alta representación en la provincia de Manabí, en su mayoría matriculados en carreras de pregrado, pertenecientes al grupo étnico mestizo y al tener baja representatividad en la variable 'TIPO_FINANCIAMIENTO_PARTICULAR'(0.035) se puede deducir que este grupo de estudiantes esta matricualdo en instituciones públicas. Este perfil podría reflejar una tendencia regional hacia la educación tradicional pública y presencial.

Con estos ejemplos se puede evidenciar como es que aportan estas técnicas de aprendizaje no supervisado a entender los datos. Este ejercicio es un primer avance para poner en perspectiva la realidad del país, sin embargo, como se mecionó anteriormente, a pesar de estos hallazgos las métricas de evaluación sugieren que la definición de los clusters aún se pudiera mejorar. El índice de correlación cophenética (0.5891) y el coeficiente de Silhouette (0.1524) indican una moderada representación y cierta superposición entre clusters. Estos resultados muestran la complejidad inherente de los datos categóricos y resaltan la necesidad de seguir optimizando los modelos.

## VII. Conclusiones

Desde el análisis exploratorio se puede concluir que la oferta de carreras/programas se concentra moyoritariamente en las instituciones privadas con el 54%, mientras que las públicas concentran el 46%. Sin embargo, al analizar la distribución de estudiantes en este mismo segmento se observa que las instituciones públicas son las que concentran a la mayoría de los estudiantes matriculados con el 64%, en contraste con el 36% que se encuentra matriculado en las instituciones privadas. Esto podría deberse a factores como el costo de los estudios o las políticas de acceso a la educación pública, este es un factor clave para entender la desigualdad en el acceso a la educación superior.

El análisis de los índices de evaluación aplicados al modelo de clustering jerárquico sugiere una representación razonable, aunque no óptima, de las relaciones entre las observaciones. El índice de correlación cophenética (0.5891) muestra que el dendrograma captura de forma moderada y aceptable las distancias originales entre los datos, generando una representación bastante aproximada de las similitudes y diferencias en las observaciones. Sin embargo, al analizar el Silhouette (0.1524) se concluye que los clusters no se encuentran del todo bien definidos ni separados entre sí, lo que sugiere que existe cierto grado de superposición entre los clusters o una estructura difusa en los datos.

Por otro lado, al analizar el índice de Calinski-Harabasz score (662.7389), que mide la relación entre la dispersión interna y externa de los clusters, se encuentra un valor alto. Esto implica que, aunque los clusters presentan cierta superposición o cercanía entre sí, internamente estos mantienen buena cohesión en relación con su separación respecto a otros grupos.

El análisis de los clusters revela perfiles distintivos que reflejan patrones relevantes en la modalidad de formación, la ubicación geográfica y el campo de estudio de los estudiantes. El Cluster 36 agrupa predominantemente a estudiantes en modalidad presencial, con una notable representación en la provincia de Manabí, lo que sugiere una concentración de este tipo de educación en dicha región. Por otro lado, el Cluster 12 caracteriza a estudiantes en modalidad de educación a distancia, con una fuerte presencia en la provincia de Loja y una orientación destacada hacia el campo de las ciencias sociales y el derecho. Este tipo de segmentación puede ser fundamental para la formulación de políticas educativas más adaptadas a las necesidades específicas de cada grupo, así como para optimizar la oferta académica en función de las particularidades regionales y disciplinares.

Los hallazgos evidencian desequilibrios importantes en la distribución geográfica de la oferta educativa y disparidades en el acceso de los estudiantes. La segmentación en clusters proporciona una herramienta valiosa para diseñar políticas públicas más inclusivas y focalizadas, como la implementación de programas que promuevan la educación en regiones con menor acceso o la

expansión de modalidades flexibles como la educación en línea. No obstante, el alto nivel de ruido identificado en los datos y la dificultad para lograr clusters bien definidos sugieren que sería beneficioso complementar estos análisis con enfoques mixtos que incluyan técnicas supervisadas y la integración de variables contextuales adicionales. Este enfoque permitiría capturar con mayor precisión las dinámicas educativas y sociales que influyen en los perfiles estudiantiles.

## Appendix A
## TABLAS CON LAS PRINCIPALES CARATERÍSTICAS DE LOS CLUSTERS FORMADO

Las siguientes tablas presentan los detalles de las principales características de los clusters generados mediante el algoritmo K-modes y el algoritmo hierarchical clustering, se destacan las categorías predominantes en cada cluster y el tamaño de los mismos. Esta información complementa la discusión presentada en la sección de resultados.

### *A. Clusters con Hierarchical Clustering*

Table VII
Principales características del Cluster 39

| **SIZE** | **18330** |
|---|---|
| MODALIDAD_PRESENCIAL | 0.999 |
| TIPO_SEDE_MATRIZ | 0.977 |
| NIVEL_FORMACIÓN_PREGRADO | 0.924 |
| ETNIA_MESTIZO | 0.808 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.453 |

Table VIII
Principales características del Cluster 15

| **SIZE** | **11512** |
|---|---|
| TIPO_SEDE_MATRIZ | 0.999 |
| PROVINCIA_SEDE_GUAYAS | 0.989 |
| MODALIDAD_EN_LINEA | 0.970 |
| NIVEL_FORMACIÓN_PREGRADO | 0.670 |
| ETNIA_MESTIZO | 0.566 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.455 |

Table IX
Principales características del Cluster 30

| **SIZE** | **4140** |
|---|---|
| MODALIDAD_PRESENCIAL | 0.998 |
| NIVEL_FORMACIÓN_PREGRADO | 0.948 |
| PROVINCIA_SEDE_IMBABURA | 0.842 |
| TIPO_SEDE_MATRIZ | 0.832 |
| ETNIA_MESTIZO | 0.642 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.210 |

Table X
Principales características del Cluster 41

| **SIZE** | **6609** |
|---|---|
| TIPO_SEDE_MATRIZ | 1.000 |
| NIVEL_FORMACIÓN_PREGRADO | 0.991 |
| PROVINCIA_SEDE_CHIMBORAZO | 0.852 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.001 |

Table XI
Principales características del Cluster 19

| **SIZE** | **4393** |
|---|---|
| MODALIDAD_EN_LINEA | 0.994 |
| TIPO_SEDE_MATRIZ | 0.990 |
| PROVINCIA_SEDE_PICHINCHA | 0.855 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.760 |
| ETNIA_MESTIZO | 0.736 |
| NIVEL_FORMACIÓN_PREGRADO | 0.311 |

### *B. K-Modes*

Table XII
Principales características del Cluster 0

| **SIZE** | **43849** |
|---|---|
| MODALIDAD_PRESENCIAL | 0.993 |
| NIVEL_FORMACIÓN_PREGRADO | 0.917 |
| TIPO_SEDE_MATRIZ | 0.917 |
| ETNIA_MESTIZO | 0.906 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.113 |

Table XIII
Principales características del Cluster 2

| **SIZE** | **11786** |
|---|---|
| MODALIDAD_A_DISTANCIA | 1.000 |
| TIPO_SEDE_MATRIZ | 0.999 |
| NIVEL_FORMACIÓN_PREGRADO | 0.966 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.921 |
| PROVINCIA_SEDE_LOJA | 0.877 |
| ETNIA_MESTIZO | 0.681 |

Table XIV
Principales características del Cluster 6

| **SIZE** | **7050** |
|---|---|
| TIPO_SEDE_MATRIZ | 0.999 |
| MODALIDAD_EN_LINEA | 0.935 |
| NIVEL_FORMACIÓN_PREGRADO | 0.865 |
| ETNIA_MESTIZO | 0.843 |
| PROVINCIA_SEDE_GUAYAS | 0.745 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.132 |

Table XV
Principales características del Cluster 8

| **SIZE** | **11083** |
|---|---|
| MODALIDAD_PRESENCIAL | 0.982 |
| PROVINCIA_SEDE_PICHINCHA | 0.956 |
| TIPO_SEDE_MATRIZ | 0.947 |
| NIVEL_FORMACIÓN_PREGRADO | 0.919 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.690 |

Table XVI
Principales características del Cluster 10

| **SIZE** | **10654** |
|---|---|
| PROVINCIA_SEDE_MANABI | 0.983 |
| PROVINCIA_RESIDENCIA_MANABI | 0.970 |
| NIVEL_FORMACIÓN_PREGRADO | 0.956 |
| MODALIDAD_PRESENCIAL | 0.948 |
| TIPO_SEDE_MATRIZ | 0.817 |
| TIPO_FINANCIAMIENTO_PARTICULAR | 0.031 |

## Agradecimiento

A la Universidad San Francisco de Quito por facilitarme los recursos físicos y tecnológicos necesarios para realizar este trabajo de investigación. A mi familia por el apoyo y la motivación constante brindados durante este año.

## References

[1] L. G. G. Fiegehen and O. E. Díaz, "Deserción en educación superior en américa latina y el caribe," *IESALC/UNESCO*, 2008.

[2] D. N. de Gestión de la Información (DNGI), *Estudio de Indicadores PND 2021-2025*. Quito, Ecuador: Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (Senescyt), 2021.

[3] S. Schwartzman, "Acceso y retrasos en la educación en américa latina," *Sistema de Información de Tendencias Educativas en América Latina*, 2004.

[4] J. F. R. M. W. De Vries, P. León Arenas and I. H. Saldaña, "¿desertores o decepcionados? distintas causas para abandonar los estudios universitarios," *Revista de la Educación Superior*, vol. XL, no. 4, pp. 29–49, 2011.

[5] M. G. Zúñiga, *Deserción estudiantil en el nivel superior. Causas y solución*. México: Trillas, 2006.

[6] D. G. F. J. García-Peñalvo and M. Johnson, "Addressing inequalities in access and achievement in higher education," *International Journal of Educational Development*, vol. 73, pp. 102–110, 2020.

[7] UNESCO, "Desigualdades socioeconómicas y aprendizaje," 2022. [Online]. Available: https://www.unesco.org/en/education-equity

[8] C. B.-A. A. Sanahuja and C. D. Angelis, "Prácticas inclusivas en el contexto escolar: una mirada sobre tres experiencias internacionales," *Revista Iberoamericana de Educación*, vol. 89, no. 1, pp. 17–37, 2022.

[9] C. E. para América Latina y el Caribe (CEPAL), "La educación técnico-profesional y el mercado laboral en américa latina y el caribe: Desafíos y oportunidades," *CEPAL*, 2013.

[10] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 12, no. 3, pp. 425–445, 2020.

[11] P. A. U. Gómez, "Análisis de datos cualitativos," *Revista Fedumar: Pedagogía y Educación*, vol. 3, no. 1, pp. 113–126, 2016.

[12] C. Malzer and M. Baum, "A hybrid approach to hierarchical density-based cluster selection," in *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, 2021, pp. 33–42.

[13] S. Mojarad, A. Essa, S. A. Mojarad, and R. S. Baker, "Data-driven learner profiling based on clustering student behaviors: Learning consistency, pace and effort," in *Proceedings of the International Conference on Intelligent Tutoring Systems (ITS)*. McGraw-Hill Education and University of Pennsylvania, 2018.

[14] N. I. M. Talib, N. A. A. Majid, and S. Sahran, "Identification of student behavioral patterns in higher education using k-means clustering and support vector machine," *Applied Sciences*, vol. 13, no. 3267, 2023.

[15] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[16] F.-M. E. Argiento R. and P. L., "Model-based clustering of categorical data based on the hamming distance," *arXiv preprint*, vol. arXiv:2212.04746, 2022.

[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.

[18] A. Agresti, *Categorical Data Analysis*, 3rd ed. Wiley, 2013.

[19] M. G. J. T. Stasko, R. Catrambone and K. McDonald, "An evaluation of space-filling information visualizations for depicting hierarchical structures," *International Journal of Human-Computer Studies*, pp. 663–694, 2000.

[20] B. Johnson and B. Shneiderman, "Treemaps: A space-filling approach to the visualization of hierarchical information structures," in *Proceedings of the 2nd International IEEE Visualization Conference (San Diego, Oct. 1991)*, 1991, pp. 284–291.

[21] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 2008, pp. 11–15.

[22] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[25] G. K. Y. Zhao and U. Fayyad, "Hierarchical clustering algorithms for document datasets," *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141–168, 2010.

[26] J. L. F. Cao and L. Bai, "A new initialization method for categorical data clustering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 223–10 228, 2009.

[27] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, vol. 11, no. 2, pp. 33–40, 1962.

[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987.

[29] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[31] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, *Handbook of Cluster Analysis*. Chapman and Hall/CRC, 2015.

[32] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Pearson, 2018.