UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE NEGOCIOS DIGITALES

APLICACIÓN DEL ALGORITMO K-PROTOTYPES PARA LA SEGMENTACIÓN DE COMPAÑÍAS ECUATORIANAS SUPERVISADAS POR LA SUPERINTENDENCIA DE COMPAÑÍAS, VALORES Y SEGUROS (SUPERSCIAS) 2022- 2024

Trabajo de titulación previo a la obtención del
Título de licenciado en Negocios Digitales.

Autores: Mateo Francisco Narváez Rea

Tutor: Ing. Mario Javier Caiza Simbaña, M. Sc.

Quito – Ecuador
2026

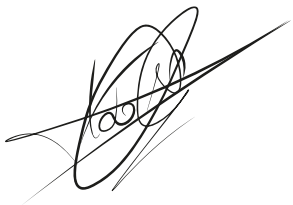
CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, Mateo Francisco Narváez Rea con documento de identificación N° 1726586009 manifiesto que:

Soy el autor y responsable del presente trabajo, declaro que he utilizado herramientas de inteligencia artificial solo para efectos de mejorar la redacción y puntuación del trabajo, lo cual consta en la citas y referencias; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 5 de enero del año 2026

Atentamente,



Firma
Mateo Francisco Narváez Rea
1726586009

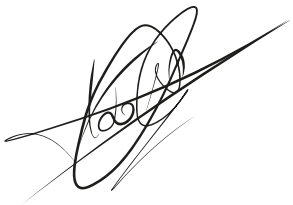
CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Yo, Mateo Francisco Narváez Rea con documento de identificación No 1726586009 , expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor del proyecto de investigación: **APLICACIÓN DEL ALGORITMO K-PROTOTYPES PARA LA SEGMENTACIÓN DE COMPAÑÍAS ECUATORIANAS SUPERVISADAS POR LA SUPERINTENDENCIA DE COMPAÑÍAS, VALORES Y SEGUROS**, el cual ha sido desarrollado para optar por el título de: Licenciado en Negocios Digitales, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 5 de Enero del año 2026

Atentamente,



Firma
Mateo Francisco Narváez Rea
1726586009

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Mario Javier Caiza Simbaña con documento de identificación N° 1727586347, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: "**APLICACIÓN DEL ALGORITMO K-PROTOTYPES PARA LA SEGMENTACIÓN DE COMPAÑÍAS ECUATORIANAS SUPERVISADAS POR LA SUPERINTENDENCIA DE COMPAÑÍAS, VALORES Y SEGUROS**", realizado por Mateo Francisco Narváez Rea con documento de identificación N° 1726586009, obteniendo como resultado final el trabajo de titulación bajo la opción proyecto de investigación que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 5 de Enero del año 2026

Atentamente,

Firma
Mario Javier Caiza Simbaña
1727586347

Agradecimiento

Para empezar, quiero dar gracias a Dios por darme la vida y permitirme vivir todo lo que he vivido. Agradezco profundamente a mis padres, quienes siempre han estado conmigo desde el principio; son quienes me vieron crecer, ganar y también perder, pero, sobre todo, aprender. En cada etapa de mi vida han estado a mi lado, guiándome y acompañándome en cada paso que he dado. A mis hermanos, gracias por estar siempre pendientes de mí, por brindarme ánimo y consejos cuando la situación lo requería. Agradezco a la carrera y, de manera especial, a la profesora Pao, quien desde el inicio me aconsejó y me guio en cada momento que lo necesité. Asimismo, agradezco a la profesora Marce, quien estuvo pendiente de mí y de todo el proceso, no solo académico, sino también personal; sus consejos y conversaciones me permitieron aprender y crecer significativamente.

Un agradecimiento especial a mi tutor de proyecto, el profesor Javier, quien no solo me guio durante el proceso de titulación, sino que también me orientó en el ámbito laboral, enseñándome aspectos fundamentales sobre cómo desenvolverme en ese entorno y, sobre todo, recordándome que lo importante no es solo saber, sino saber aprender. Agradezco a todos y cada uno de los profesores y amigos que he conocido a lo largo de la carrera; cada uno de ellos ha aportado algo especial en mí, y gracias a ello he podido crecer tanto a nivel académico como personal.

De igual manera, agradezco profundamente a mis amigos del trabajo: Fer, Jaquy y Francisco, quienes no solo me guiaron en el ámbito laboral, sino también en el personal. Sus consejos, conversaciones y enseñanzas hicieron posible que pueda desenvolverme y crecer no solo como profesional, sino también como ser humano.

Finalmente, un agradecimiento especial a todas las personas que he conocido, ya que, de una u otra forma, han aportado en mi vida y han hecho posible que hoy sea quien soy.

Índice de Contenidos

Índice de tablas

Índice de figuras

Resumen

La presente investigación se enfoca en el análisis y segmentación de las empresas ecuatorianas registradas en la Superintendencia de Compañías, Valores y Seguros (SuperCias), utilizando variables financieras, sectoriales y geográficas correspondientes al año 2024. Teniendo en cuenta la importancia de la SuperCias en la regulación y transparencia, se ha identificado que la complejidad y heterogeneidad de la información reportada en los balances (datos categóricos y numéricos) dificultan la detección de patrones mediante métodos estadísticos tradicionales. Esta limitación restringe la capacidad para orientar decisiones basadas en evidencia, generando lo que genera un análisis incompleto.

Utilizando técnicas de aprendizaje no supervisado, específicamente el algoritmo K-Prototypes en conjunto con el marco de trabajo *Team Data Science Process* (TDSP), se busca identificar agrupamientos naturales que capturen tanto los datos categóricos como numéricos que se encuentran en la base. La implementación de este modelo permite procesar tanto datos financieros cuantitativos de las empresas como datos cualitativos. Los resultados de la segmentación buscan mostrar clusters empresariales diferenciados, destacando grupos con características particulares entre si. Estas estrategias están diseñadas para optimizar el análisis de las bases de datos con datos mixtos de los mismos.

Palabras clave: K-Prototypes, Segmentación empresarial, Transformación digital, Aprendizaje automático.

Abstract

This research focuses on the analysis and segmentation of Ecuadorian companies registered with the Superintendence of Companies, Securities, and Insurance (SuperCias), utilizing financial, sectoral, and geographical variables from the fiscal year 2024. Despite the significance of SuperCias in regulating and stabilizing the national corporate environment, it has been identified that the complexity and heterogeneity of the reported information hinder the detection of performance patterns through traditional statistical methods. This limitation restricts the ability to guide evidence-based decisions, resulting in an incomplete analysis of the sector.

By employing unsupervised learning techniques, specifically the K-Prototypes algorithm within the Team Data Science Process (TDSP) framework, this study aims to identify natural groupings that capture both the categorical and numerical data present in the dataset. The implementation of this model allows for the simultaneous processing of quantitative financial metrics and qualitative corporate data. The segmentation results intend to reveal distinct business clusters, highlighting specific characteristics within each group. These strategies are designed to optimize the analysis of mixed-type datasets, providing a more robust understanding of the corporate landscape.

Keywords: K-Prototypes, Business segmentation, Digital transformation, Machine learning, Mixed data.

1. CAPÍTULO I: CONTEXTUALIZACIÓN

1.1 Contextualización

La Superintendencia de Compañías, Valores y Seguros (SuperCias) constituye el pilar fundamental para la supervisión y regulación del ecosistema corporativo en Ecuador, velando por la transparencia y la estabilidad normativa de las entidades bajo su control (Ministerio de Telecomunicaciones y de la Sociedad de la Información, 2026). En el marco de la Agenda de Transformación Digital del Ecuador 2022-2025, se ha establecido una hoja de ruta para modernizar el sistema productivo mediante el uso de tecnologías emergentes y la toma de decisiones basadas en evidencia técnica(Meza Fabián Íñiguez Mauricio Becerra Oswaldo Rivera Johanna Vera Juan Carlos Chiluiza Jorge Ortega Rocío Malla Franklin Simbaña Adriana Valverde Sheldon López Marcelo Sotaminga & Chang Calvache David Hurtado Vicente Palacios, 2022) (Ministerio de Telecomunicaciones y de la Sociedad de la Información, 2022). Durante el año fiscal 2024, las bases de datos que almacena la SuperCias contiene un acumulado número de variables contables, sectoriales y geográficas que permiten realizar análisis profundos sobre el desempeño productivo nacional (*www.supercias.gob.ec*, s/f)

1.2 Planteamiento del Problema

A pesar de la disponibilidad de información detallada, el análisis de las empresas ecuatorianas enfrenta un obstáculo crítico: la heterogeneidad de los datos. Las bases de datos combinan variables cuantitativas con atributos cualitativos, y los métodos estadísticos tradicionales resultan limitados para capturar esta naturaleza multidimensional y detectar agrupaciones naturales en estructuras tan diversas. Como consecuencia, patrones subyacentes

que explican las diferencias de desempeño suelen quedar ocultos bajo análisis superficiales, limitando la detección de riesgos o éxitos invisibles a simple vista (Ping et al., 2025)

1.3 Justificación

Este proyecto se justifica al demostrar cómo el uso del machine learning y analítica de datos puede transformar grandes volúmenes de información en conocimiento estratégico. La aplicación del algoritmo **K-Prototypes** es esencial, ya que permite procesar simultáneamente variables numéricas y categóricas para formar agrupaciones de acuerdo con las características en común, siendo considerada una de las técnicas más robustas para este objetivo (Huang, 1998). Este enfoque se alinea directamente con los ejes estratégicos de la agenda nacional, que buscan utilizar la analítica para fortalecer el tejido empresarial.

1.4 Alcance

El estudio se delimita al análisis de las compañías bajo control de la SuperCias que reportaron información financiera en el año 2024. El alcance técnico comprende la implementación del marco de trabajo **Team Data Science Process (TDSP)**, abarcando desde la adquisición de datos hasta la validación de clusters (What is TDSP?, s/f).

1.5 Objetivos

1.5.1 Objetivo general

Aplicar un modelo de clustering basado en el algoritmo K-Prototypes sobre las bases de datos empresariales de los años 2022, 2023 y 2024 de la Superintendencia de Compañías, Valores y Seguros (SuperCias), con el fin de identificar patrones ocultos en el comportamiento financiero y estructural de las empresas ecuatorianas, que permitan detectar oportunidades de mercado, riesgos emergentes y señales de crecimiento o estabilidad que no son evidentes mediante métodos tradicionales de análisis estadístico..

1.5.2 Objetivos específicos

- Analizar las variables contables, financieras y estructurales contenidas en las bases empresariales 2022–2024 de la SuperCias, con el propósito de seleccionar los indicadores más relevantes para el proceso de segmentación.
- Depurar, integrar y normalizar la información proveniente del periodo de análisis, garantizando la coherencia temporal y la calidad de los datos para su posterior tratamiento mediante técnicas de minería de datos.
- Implementar el algoritmo K-Prototypes para generar agrupamientos homogéneos de empresas, considerando simultáneamente variables numéricas y categóricas que reflejen la estructura y desempeño financiero de cada entidad.
- Evaluar y validar los resultados del modelo de clustering, identificando las variables con mayor influencia en la conformación de los grupos y analizando las diferencias estructurales y financieras entre ellos.
- Interpretar los patrones descubiertos desde una perspectiva de transformación digital y desarrollo económico, proponiendo lineamientos que contribuyan a la toma de decisiones estratégicas, la formulación de políticas diferenciadas y la comprensión del tejido empresarial ecuatoriano.

CAPITULO II: MARCO TEORICO REFERENCIAL

2.1 La Superintendencia de Compañías, Valores y Seguros (SuperCias)

2.1.1 Rol Institucional y Marco Regulatorio

La Superintendencia de Compañías, Valores y Seguros (SuperCias) es un organismo nacional que se encarga de vigilar, controlar y auditar las actividades de las empresas mercantiles en el Ecuador. Su función principal es asegurar que todo se trabaje bajo el marco de la ley para así promover una transparencia en el mercado de los valores. Todas las entidades que están bajo el control de este organismo están obligadas a presentar de forma anual sus estados financieros, lo cual hace que este organismo tenga uno de los repositorios con información corporativa más grandes del país. (*www.supercias.gob.ec*, s/f)

2.1.2 Base de Datos de la SuperCias como activo estratégico

Para la presente investigación, la base de datos del ejercicio fiscal 2024 alojada en la página web de la SuperCias, se considera un "activo de datos mixtos" ya que cuenta con información cuantitativa como lo son los datos financieros de los balances de estas, además de contar con información cualitativa como la actividad económica.

2.1.2 Estadistica descriptiva presentada por la Superintendencia de Compañias, Valores y Seguros (SuperCias)

La SuperCías difunde, mediante sus medios oficiales, estadísticas sobre el comportamiento del sector empresarial ecuatoriano. Dentro de la información disponible se encuentran el Ranking de Compañías (ordenado por ingresos y otros indicadores), además de reportes societarios que incluyen la constitución de compañías por año, el tipo societario, la distribución por provincia y su evolución en el tiempo. Adicionalmente, la base institucional incorpora información que nos permite calcular índices derivados de los estados financieros (por ejemplo, indicadores de liquidez, endeudamiento o rentabilidad), lo que amplía el alcance del análisis cuantitativo. Sin embargo, al evaluar estas variables de forma separada, se reduce la capacidad de identificar relaciones entre el comportamiento financiero y las características

cualitativas, por lo que el uso de técnicas de Machine Learning permite complementar el enfoque descriptivo y obtener segmentaciones más integrales y explicativas.

2.2 Transformación Digital

2.2.1 Marco Estratégico del Ecuador 2022 - 2025

El País está atravesando un proceso de modernización institucional que está siendo orientado por la Agenda de Transformación Digital del Ecuador 2022 – 2025, la cual fue diseñada por el Ministerio de Telecomunicaciones y de la Sociedad de la Información (MINTEL).

Lo que busca este marco es el fomentar el uso de tecnologías emergentes como la Big Data yl a Inteligencia Artificial, para así lograr una mejora en la competitividad de las empresas y en la eficiencia de la Gobernanza Digital (Meza Fabián Íñiguez Mauricio Becerra Oswaldo Rivera Johanna Vera Juan Carlos Chiluiza Jorge Ortega Rocío Malla Franklin Simbaña Adriana Valverde Sheldon López Marcelo Sotaminga & Chang Calvache David Hurtado Vicente Palacios, 2022).

La Agenda de Transformación Digital del Ecuador 2022–2025 define los principales ejes estratégicos para orientar el proceso de digitalización del país, abordando aspectos institucionales, productivos, tecnológicos y sociales. Estos ejes constituyen un marco de referencia para el análisis de la adopción de tecnologías y el uso de analítica de datos en el contexto empresarial ecuatoriano.

Tabla 1 Ejes de la Agenda de Transformación Digital del Ecuador 2022–2025

EJE ESTRATÉGICO	DESCRIPCIÓN
INFRAESTRUCTURA DIGITAL	Busca el fortalecimiento de la infraestructura tecnológica en el país, teniendo así una mejor conectividad y mejorando la calidad de los servicios para un mejor desarrollo digital en el Pais.
CULTURA E INCLUSIÓN DIGITAL	Busca mejorar las competencias digitales de la población mediante la educación, la salud e impulsando una cultural digital.
ECONOMIA DIGITAL	Se enfoca en la transformación digital de las empresas, en especial la de las MIPYMES, mejorando el comercio electrónico, la innovación y la competitividad
TECNOLOGIAS EMERGENTES PARA EL DESARROLLO SOSTENIBLE	Busca la adopción de tecnologías emergentes como lo son la Inteligencia Artificial, big data, blockchain, para aplicarlas en el ámbito administrativo e industrial.
GOBIERNO DIGITAL	Dirigido a la modernización del estado mediante la digitalización y simplificación de tramites
INTEROPERABILIDAD Y TRATAMIENTO DE DATOS	Busca tener un intercambio de información de forma segura y eficiente entre sistemas.
SEGURIDAD DIGITAL Y CONFIANZA	Garantiza la seguridad de la información, la protección de datos fortaleciendo así un sentimiento de confianza digital.

Fuente: Elaborado a partir de la *Agenda de Transformación Digital del Ecuador 2022–2025* (Meza Fabián Íñiguez Mauricio Becerra Oswaldo Rivera Johanna Vera Juan Carlos Chiluiza Jorge Ortega Rocío Malla Franklin Simbaña Adriana Valverde Sheldon López Marcelo Sotaminga & Chang Calvache David Hurtado Vicente Palacios, 2022)

La aplicación de analítica avanzada sobre la base del ejercicio fiscal 2024 que gobierna la SuperCias, nos da un accionar positivo al transformar los datos administrativos y financieros e conocimientos accionables, alineándose así con el objetivo de modernizar la estructura

mediante una evidencia técnica. (Meza Fabián Íñiguez Mauricio Becerra Oswaldo Rivera Johanna Vera Juan Carlos Chiluiza Jorge Ortega Rocío Malla Franklin Simbaña Adriana Valverde Sheldon López Marcelo Sotaminga & Chang Calvache David Hurtado Vicente Palacios, 2022)

2.3 Machine Learning y Minería de Datos

2.3.1 Fundamentos de Machine Learning y Minería de Datos

El Aprendizaje Automático (*Machine Learning*, ML) es una rama fundamental de la inteligencia artificial que se centra en el desarrollo de modelos y algoritmos capaces de aprender a partir de los datos, identificar patrones complejos y generar conocimiento sin intervención humana directa en cada etapa del proceso. Su objetivo principal es permitir que los sistemas mejoren su desempeño conforme se incrementa la cantidad de información disponible, facilitando la automatización del análisis y la toma de decisiones basada en evidencia (Alpaydin, 2020)

Según (Knox, 2018), el propósito central de la aplicación de Machine Learning dentro del análisis de datos es la extracción de la máxima cantidad de información relevante posible a partir de una base de datos, con el fin de generar un producto analítico útil, preciso y aplicable en contextos reales.

2.3.2 Algoritmo K-means

El Algoritmo de K-means es una de las técnicas de agrupamiento de datos mas utilizadas y desarrolladas cuando el número de clusters es conocido (Ayaquica-Martínez et al., 2006).

La popularidad del método se centra en la simplicidad y eficiencia; sin embargo, la formulación de esta se limita solo al uso exclusivo de las variables numéricas que se encuentran en las bases por lo que no lo combina con las variables categóricas de las mismas por lo cual esta generación de cluster carece de una descripción semántica (Ayaquica-Martínez et al., 2006).

El algoritmo de K-means es un algoritmo de agrupación iterativa basada en centroides, el cual divide los datos en grupos similares en función a sus centroides. Los centroides de este modelo es la media o mediana de todos los datos de cada cluster o agrupación. (¿Qué es la agrupación en clústeres k-means? | IBM, s/f).

2.3.3 Algoritmo K-modes

El Algoritmo de K-modes nace a partir de la limitación del algoritmo de k-means puesto que este algoritmo solo analiza valores numéricos, el objetivo de k-modes es el agrupamiento de datos cuyas variables sean categóricas permitiendo así analizar y agrupar los datos cualitativos de los mismos (Li & Li, 2015)

2.3.4 Algoritmo K-prototypes

El Algoritmo de K-prototypes está diseñado para un análisis y agrupamiento mixto de datos ya que contempla tanto los datos cualitativos como cuantitativos de las bases. Este algoritmo nace a raíz de la limitación de la aplicabilidad de los algoritmos de k-means y k-modes en el análisis y clustering de bases con información numérica y categórica de forma simultánea. Por lo que, (Li & Li, 2015)destaca el aporte realizado por (Huang, 1998) quien también propuso este algoritmo como una vía útil para el clustering de datos híbridos.

CAPITULO III: METODOLOGÍA

3.1 Metodología investigación

La investigación tendrá un enfoque aplicado, con alcance descriptivo–exploratorio, combinando revisión bibliográfica y análisis empírico basado en datos reales. Para la revisión bibliográfica se consultarán artículos académicos, informes oficiales, publicaciones digitales especializadas, estadísticas públicas y documentación disponible en fuentes web confiables. No se utilizarán libros físicos, pero sí fuentes secundarias digitales verificables.

Para el desarrollo de este trabajo se utilizará información de la Base de Empresas Ecuador 2024, publicada por la Superintendencia de Compañías, Valores y Seguros (SuperCias), la cual contiene variables cuantitativas y cualitativas de empresas reales del país. Esta base será trabajada previamente mediante procesos de limpieza y preparación de los datos, con el fin de contar con información consistente y adecuada para el análisis.

Una vez lista la base de datos, se aplicará el algoritmo K-Prototypes como técnica de clustering no supervisado, ya que permite trabajar con información numérica y categórica al mismo tiempo, sin necesidad de definir categorías. De esta forma, se podrán identificar grupos de empresas con características similares según su tamaño económico y su estructura financiera, permitiendo una segmentación que represente de mejor manera la realidad del tejido empresarial ecuatoriano.

3.2 Metodología Analítica de Datos

El desarrollo del presente proyecto se basa en la aplicación de la metodología TDSP (Team Data Science Process) propuesta por Microsoft, un marco de trabajo ágil, estructurado y colaborativo diseñado para proyectos de ciencia de datos y analítica avanzada. Este enfoque se adoptó por su capacidad para integrar la gestión de datos, el modelado predictivo y la entrega

de resultados analíticos de forma iterativa y controlada, garantizando la calidad técnica, la trazabilidad y la reproducibilidad del proceso. (Provost & Fawcett, 2013; *What is TDSP?*, s/f)

3.2.1 Comprensión del Negocio

Se establecerá como objetivo central entender cómo se organiza la estructura empresarial del país y determinar cuál será el verdadero driver de heterogeneidad: si lo será el sector económico o si lo será el tamaño financiero y la estructura contable. La finalidad será identificar segmentos reales de empresas que puedan servir para toma de decisiones, priorización estratégica, riesgo y política pública. (Hair, 2019; Porter, 2007)

3.2.2 Adquisición y Comprensión de Datos

Se recopilará y documentará la base empresarial con información contable, financiera y categórica. En esta etapa se evalúa la calidad de la data, la presencia de valores faltantes y la variedad de campos disponibles. Se revisarán los diccionarios de variables para asegurar la coherencia semántica y se validará que existan datos suficientes para análisis no supervisado. (Han et al., 2011; Provost & Fawcett, 2013)

3.2.3 Análisis exploratorio

En esta etapa se realizará una revisión general de las principales variables financieras y estructurales de la base de datos, con el objetivo de entender cómo se comportan los datos antes de aplicar el modelo de clustering. Este análisis permitirá identificar la presencia de valores muy altos o muy bajos, así como diferencias marcadas entre empresas, lo que puede evidenciar que existen compañías de tamaños y estructuras muy distintas dentro de la base.

A partir de este análisis se anticipará si la población empresarial estará compuesta por empresas de escalas muy diferentes, lo cual justificará la necesidad de segmentación. Esta estadística descriptiva servirá además como referencia comparativa para interpretar posteriormente los clusters.(James et al., 2013)

3.2.4 Preparación de Datos

Se realizarán labores de limpieza, corrección de tipos, derivación de indicadores adicionales y normalización. Se seleccionarán variables estratégicas de estructura y desempeño y se excluirán columnas que no aporten poder explicativo o que introduzcan granularidad innecesaria. Se aplicarán transformaciones para controlar el efecto de valores extremos y se estandarizarán los valores numéricos para permitir un modelado consistente. (García et al., 2016; Han et al., 2011)

3.2.5 Modelado

Se aplicará un algoritmo apto para segmentar datos mixtos (numéricos y categóricos), como KPrototypes. Previo a definir el número final de clusters se evaluarán varios posibles valores mediante el método del codo para identificar el punto óptimo de separación. Luego de elegir el número de clusters, se ejecutará el modelo definitivo y se generarán las etiquetas correspondientes para cada empresa. (Huang, 1998)

3.2.6 Evaluación e Interpretación

Se compararán los resultados de clustering frente a la estadística descriptiva inicial para verificar coherencia. Si los clusters agrupan empresas según patrones financieros y no meramente por sector, se confirmará que la heterogeneidad observada en la descripción inicial se deberá principalmente a diferencias en escala y estructura económica. Se caracterizarán los clusters a través de perfiles descriptivos (estadísticas internas y predominancias categóricas) para extraer conclusiones empresariales. (Everitt, 2011; Rendón et al., 2011).

CAPÍTULO IV: RESULTADOS

4.1 Comprensión del Negocio

Parte de la metodología aplica la cual es la TDSP, como primera fase se realizó la etapa de comprensión del negocio, con el fin de entender la importancia del análisis de los datos empresariales dentro del marco regulatorio. Para ello se tomó como referencia la información que proporciona la SuperCias en su página web, la misma que está encargada de supervisar y a las empresas mercantiles del país.

Este análisis se enfocó en diversos aspectos que se encuentran relacionados con el rol de la SuperCias, las características de los procesos que se contemplan y la naturaleza de la información que se encuentra en la página. Este análisis fue fundamental para comprender el alcance, las limitaciones y las oportunidades del análisis de segmentación empresarial que se propuso.

La Tabla 2 resume en si los principales aspectos institucionales y de negocio considerados, además de su implicación directa en el análisis de datos y la selección del algoritmo de clustering utilizada en la investigación.

Tabla 2 Aspectos Institucionales y de Negocio considerados en el análisis de datos empresariales (SuperCias)

Aspecto	Descripción	Implicación para el análisis
Rol Institucional	La SuperCias es un organismo de control, encargado de vigilar y auditar las empresas mercantiles del Ecuador.	Por ende, garantiza que la naturaleza de la información sea de una fuente oficial y regulada.
Obligación de Reporte	Las compañías que se encuentra bajo el control de este organismo tienen que presentar la información financiera de forma anual.	Permite trabajar con información estandarizada y comparable.

Aspecto	Descripción	Implicación para el análisis
Cobertura Empresarial	La actividad económica de las compañías es de amplia variedad, además del tamaño de estas y sus estructuras financieras.	Justifica la necesidad de aplicar técnicas de segmentación no supervisadas.
Naturaleza de los datos	La base contiene variables cuantitativas (balances y estado de resultados) y cualitativas (actividad económica).	Justifica el uso del algoritmo K-Prototypes para el análisis.
Periocidad	Información correspondiente al año fiscal 2024.	Permite un análisis transversal.
Enfoque Regulatorio	Los estados y balances presentados siguen normas contables y regulatorias nacionales.	Reduce el riesgo de inconsistencias en las bases.
Escala Empresarial	Presencia de micro, pequeñas, medianas y grandes empresas	Prevé una alta heterogeneidad en los datos y la existencia de los clusters diferenciados.

Fuente: Elaboración propia.

El análisis de estos aspectos permitió identificar que la base que maneja la SuperCias es un activo estratégico de datos mixtos (cualitativos – cuantitativos) lo cual genera una alta heterogeneidad empresarial, por lo que refuerza la necesidad de aplicar técnicas de segmentación no supervisadas enfocadas a encontrar patrones ocultos que van más allá de las clasificaciones tradicionales como lo sería el sector económico.

4.2 Adquisición y comprensión de datos

La información utilizada en la presente investigación fue obtenida directamente desde la página oficial de la Superintendencia de Compañías, Valores y Seguros, a través de su módulo de consulta y en la sección de descarga de estados financieros por ramo. (*www.supercias.gob.ec*, s/f)

El proceso de adquisición de la data se accedió a la sección de “Información msobre estados financieros por ramo”, el cual permite seleccionar el año fiscal a interés y descargar los archivos correspondientes a las empresas que están bajo el control del organismo de la SuperCias. Para esta investigación, se selecciono exclusivamente el ejercicio fiscal correspondiente al año 2024, con el fin de realizar un análisis transversal.

Ilustración 1 Interfaz de descarga de estados financieros por ramo – Superintendencia de Compañías, Valores y Seguros.

Tal como se muestra en la Figura 4.2, la pagina permite filtrar la información por año fiscal y una vez validado el proceso de seguridad, se descarga la base con los estados financieros reportados por las empresas. Este mecanismo nos asegura que los datos utilizdos vienen de una fuente oficial, publica y regulada, lo cual refuerza la validez del análisis.

La base descargada contiene información estructurada a nivel empresarial, inluyendo variables financieras provenientes del balance general y del estado de resultados, así como también contiene variables categóricas relacionadas con la actividad de la empresa y características societarias. Por lo que la base de datos se caracteriza por tener variables mixtas.

Previo al análisis, se realizó una revisión de la estructura de la base con el fin de comprender el tipo de variables disponibles, la presencia de datos faltantes y la cobertura

empresarial. En lo que se concluyó que la información descargada es adecuada para la aplicación del algoritmo K-Prototypes para la segmentación empresarial.

4.3 Preparación de Datos

Previo a la aplicación del modelo de segmentación, se llevó a cabo un proceso estructurado de preparación de datos con el fin de garantizar la calidad, consistencia y comparabilidad de la información utilizada.

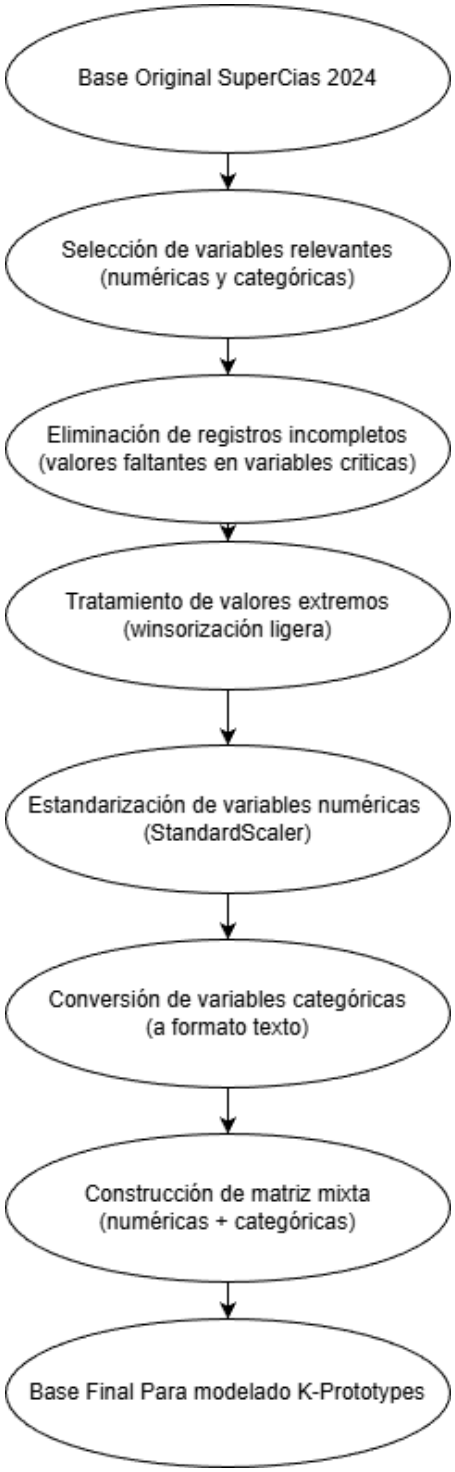


Ilustración 2 Proceso de preparación de datos para el modelado de segmentación empresarial.

Fuente: Elaboración propia.

Se partió sobre la base original del año fiscal 2024 descargada desde la pagina web de la Superintendencia de Compañías, Valores y Seguros, posteriormente se realizó un análisis y selección de variables relevantes, diferenciando así las variables numéricas de características

financieras de las variables categóricas que están asociadas a la estructura de la empresa. Esta selección permitió reducir el ruido logrando así enfocar nuestro análisis en un campo con potencial explicativo para la segmentación.

Posteriormente, se procedió con la eliminación de registros con valores faltantes en variables críticas, asegurándonos así que todas las observaciones utilizadas tendrán información completa en las dimensiones seleccionadas. Esta decisión se dio para evitar sesgos en el calculo de las distancias y en la asignación de clusters.

Con el objetivo de reducir el impacto de valores extremos sin distorsionar la distribución general de los datos, se aplicó una winsorizacion ligera sobre las variables numéricas, recortando así los percentiles extremos. Este tratamiento permitió disminuir la influencia de los outliers severos, manteniendo así la estructura relativa de la información financiera.

A continuación, las variables numéricas fueron estandarizadas mediante escalamientos z.score, con el objetivo de homogeneizar las escalas y evitar que las variables con magnitudes superiores dominen el proceso de agrupamiento. A la par, las variables categóricas fueron transformadas a texto, garantizando así su correcta interpretación durante el modelado.

Finalmente, se construyo una matriz de datos mixtos, integrando así las variables numéricas estandarizadas y las variables categóricas, siendo esta matriz la entrada para el algoritmo K-Prototypes. Este conjunto de datos preparado constituye la base final sobre la que se realizo el proceso de segmentación empresarial.

4.4 Modelado

4.4.1 Preparación y limpieza de los datos

Previo a la aplicación del algoritmo de segmentación, se llevó a cabo un proceso sistemático de preparación y limpieza de los datos, con el objetivo de garantizar la calidad, consistencia y comparabilidad de la información utilizada en el modelo. Esta etapa resulta

fundamental en técnicas de aprendizaje no supervisado, ya que los algoritmos de clustering son altamente sensibles a valores atípicos, escalas heterogéneas y datos incompletos.

4.4.1.1 Selección de variables

En primer lugar, se realizó una selección explícita de las variables a incluir en el análisis, diferenciando entre variables numéricas y categóricas. Las variables numéricas corresponden a indicadores financieros y estructurales directamente reportados por las empresas, tales como número de empleados, activos corrientes, pasivos corrientes, patrimonio, ventas operativas, resultado neto del ejercicio y liquidez corriente. Estas variables permiten capturar la escala operativa, la estructura financiera y la capacidad de pago de las compañías analizadas.

```
# SELECCIÓN DE VARIABLES PARA EL ANÁLISIS
# ------------------------------------------------

# Definición de variables numéricas.
# Se incluyen indicadores financieros y estructurales directamente reportados
# por las compañías
num_cols = [
    'Cant. Empleados',           # Tamaño operativo de la empresa
    'TOTALACTIVOCORRIENTE',      # Activos de corto plazo
    'TOTALPASIVOCORRIENTE',      # Pasivos de corto plazo
    'TOTALPATRIMONIO',           # Patrimonio total
    'VENTASOPERATIVAS',          # Nivel de ingresos operativos
    'UTILIDADPERDIDANETA',       # Resultado neto del ejercicio
    'LIQUIDEZ_CORRIENTE'         # Capacidad de pago a corto plazo

]

# Definición de variables categóricas.
# Se consideran únicamente atributos cualitativos propios de la base original
cat_cols = [
    'Tipo Compañia',             # Tipo legal de la empresa
    'Región'                     # Ubicación geográfica

]
```

Ilustración 3 Selección de Variables Numéricas y Categóricas

Fuente: Elaboración propia.

4.4.1.2 Tratamiento de valores faltantes

Una vez seleccionadas las variables, se procedió al tratamiento de valores faltantes. Dado que las variables seleccionadas son críticas para el análisis de segmentación, se optó por eliminar las observaciones que presentaban valores nulos en cualquiera de ellas. Esta decisión garantiza que el algoritmo de clustering opere únicamente sobre observaciones completas, evitando distorsiones en el cálculo de distancias y en la asignación de clústeres.

Si bien este enfoque reduce marginalmente el tamaño de la muestra, asegura una mayor consistencia en los resultados y evita la introducción de supuestos adicionales asociados a métodos de imputación.

```
# ------------------------------------------------
# LIMPIEZA DE DATOS: TRATAMIENTO DE VALORES FALTANTES
# ------------------------------------------------

# Elimina las filas que contienen valores nulos (NA) en cualquiera de las
# variables seleccionadas como críticas para el análisis.
# Esto garantiza que el algoritmo de clustering trabaje únicamente con
# observaciones completas, evitando distorsiones en el cálculo de distancias.
data = data.dropna(subset=num_cols + cat_cols).reset_index(drop=True)
```

Ilustración 4 Tratamiento de Valores Faltantes

Fuente: Elaboración propia.

4.4.1.3 Tratamiento de valores extremos

Posteriormente, se manejó el tratamiento de valores extremos mediante un proceso de winsorización ligera aplicado a todas las variables numéricas. Este procedimiento trata en recortar los valores ubicados en los percentiles extremos de la distribución, específicamente el 1% inferior y el 1% superior, reduciendo así la influencia de outliers sin eliminar observaciones del conjunto de datos.

La winsorización suele ocuparse mayormente en contextos financieros, donde las distribuciones suelen presentar asimetrías pronunciadas y valores extremos que pueden afectar de forma desproporcionada el cálculo. Este enfoque permite garantizar la estructura general de los datos, al tiempo que mejora la estabilidad y robustez del proceso de agrupamiento.

```python
# ------------------------------------------------
# TRATAMIENTO DE VALORES EXTREMOS (WINSORIZACIÓN)
# ------------------------------------------------

# Definición de una función de winsorización ligera.
# Esta función recorta los valores de una variable numérica a los percentiles
# inferior (p) y superior (1-p), reduciendo la influencia de outliers extremos
# sin eliminar observaciones del conjunto de datos.
def winsorize_series(s, p=0.01):
    # Calcula el percentil inferior (p1) y superior (p99) de la serie
    lo, hi = s.quantile(p), s.quantile(1 - p)
    # Recorta los valores que se encuentren fuera de estos límites
    return s.clip(lower=lo, upper=hi)

# Aplicación de la winsorización a todas las variables numéricas seleccionadas.
# Se utiliza p = 0.01, lo que implica un recorte al 1% inferior y 1% superior.
# Este procedimiento es especialmente útil en datos financieros, donde suelen
# existir distribuciones asimétricas y valores extremos.
for c in num_cols:
    data[c] = winsorize_series(data[c], p=0.01)
```

Ilustración 5 Tratamiento de Valores Extremos

Fuente: Elaboración propia.

4.4.1.4 Escalado de variables numéricas

Tomando en cuenta que las variables numéricas tienen distintas magnitudes y unidades de medida, se realizó un proceso de estandarización de los datos para que todas las variables se encuentren en una misma escala. Con este ajuste se evita que aquellas variables con valores más grandes tengan mayor peso en el cálculo de distancias del algoritmo.

Este paso es importante dentro del proceso de clustering, ya que permite que todas las variables aporten de forma equilibrada a la segmentación, independientemente de la escala en la que se encuentren originalmente.

```
# ------------------------------------------------
# ESCALADO DE VARIABLES NUMÉRICAS
# ------------------------------------------------

# Importa el escalador estándar de scikit-learn.
# StandardScaler transforma las variables numéricas para que tengan
# media cero y desviación estándar uno.
from sklearn.preprocessing import StandardScaler

# Inicializa el objeto escalador.
scaler = StandardScaler()

# Aplica el escalado a las variables numéricas seleccionadas.
# Este paso es fundamental en algoritmos de clustering, ya que evita
# que variables con mayor magnitud dominen el cálculo de distancias.
X_num = scaler.fit_transform(data[num_cols])
```

Ilustración 6 Escalado de Variables Numéricas

Fuente: Elaboración propia.

4.4.1.5 Construcción de la matriz mixta

Finalmente, se construyó la matriz de datos mixta requerida por el algoritmo K-Prototypes. Las variables numéricas, previamente escaladas, se combinaron con las variables categóricas transformadas a formato texto, preservando su naturaleza cualitativa. En la matriz resultante, las variables numéricas se ubicaron en las primeras columnas, seguidas de las variables categóricas, permitiendo identificar de forma explícita los índices correspondientes a cada tipo de variable.

Esta estructura permitió al algoritmo optimizar simultáneamente la distancia euclidiana para las variables numéricas y la disimilitud categórica para las variables cualitativas, asegurando un tratamiento adecuado de la información mixta presente en el universo empresarial analizado.

```
# ------------------------------------------------
# CONSTRUCCIÓN DE MATRIZ MIXTA PARA K-PROTOTYPES
# ------------------------------------------------

# Convierte las variables categóricas a texto (string).
# Esto asegura que K-Prototypes las trate como categorías y no como valores numéricos.
X_cat = data[cat_cols].astype(str).to_numpy()

# Concatena las variables numéricas ya escaladas (X_num) con las variables categóricas (X_cat),
# formando una única matriz mixta que contiene ambos tipos de datos.
# Nota: en esta matriz, primero están todas las numéricas y luego todas las categóricas.
X_mixed = np.concatenate([X_num, X_cat], axis=1)
```

Ilustración 7 Construcción de Matriz Mixta Para K-Prototypes

Fuente: Elaboración propia.

4.4.2 Selección del número óptimo de clústeres: Método del codo (Elbow Method)

Con el fin de determinar el número óptimo de clústeres (*k*) para el algoritmo K-Prototypes, se aplicó el método del codo (*Elbow Method*). Este procedimiento consiste en ajustar el modelo para distintos valores de *k* y analizar la evolución del costo intra-clúster. En la medida en que *k* aumenta, el costo total tiende a disminuir debido a que los grupos se vuelven más específicos; sin embargo, llega un punto a partir del cual la mejora marginal se reduce significativamente. Este punto de cambio se interpreta como el “codo” y suele tomarse como un equilibrio adecuado entre la complejidad del modelo y la facilidad para interpretar los resultados.

En el caso del algoritmo K-Prototypes, el indicador *cost* representa el nivel de diferencia total dentro de los grupos formados, considerando al mismo tiempo dos aspectos: por un lado, la distancia entre las variables numéricas previamente estandarizadas y, por otro, las diferencias entre las variables categóricas.

Por esta razón, aunque el procedimiento es similar al método del codo utilizado en K-Means, el valor del costo en K-Prototypes refleja un criterio mixto, acorde a la naturaleza diversa de los datos empresariales que combinan información financiera y estructural.

Dado que el universo empresarial analizado contiene decenas de miles de observaciones, y con el objetivo de optimizar el tiempo computacional sin perder

representatividad, el cálculo del codo se realizó sobre una muestra aleatoria de tamaño **SAMPLE_N = 1000**, seleccionada sin reemplazo y fijando una semilla (random_state = 42) para garantizar la reproducibilidad. Esta decisión metodológica es válida en bases grandes, ya que permite capturar la forma general de la curva de costo manteniendo condiciones controladas de replicación.

En consecuencia, se seleccionó **k = 4** como el número óptimo de clústeres para el entrenamiento final del modelo, ya que representa un balance adecuado entre capacidad de segmentación, estabilidad de la solución e interpretabilidad de los grupos obtenidos.

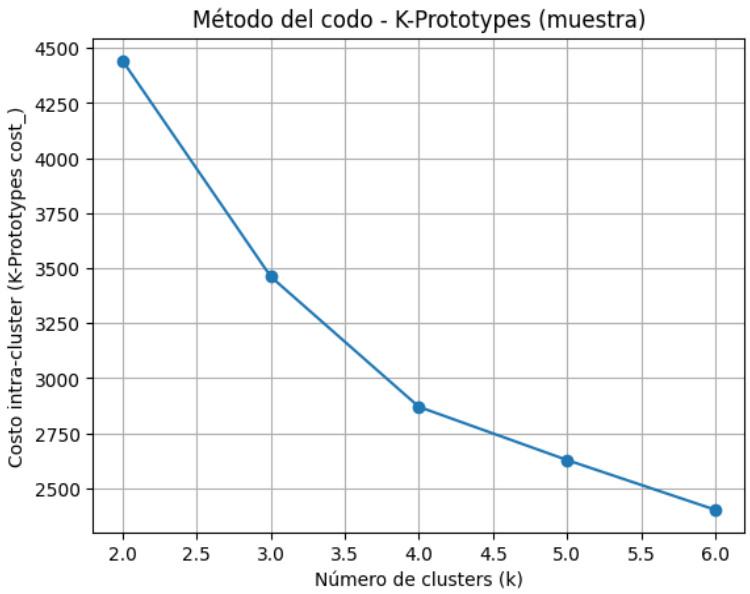


Ilustración 8 Método del Codo

Fuente: Elaboración propia.

4.4.3 Entrenamiento del modelo final K-Prototypes y asignación de clústeres

Una vez determinado el número óptimo de clústeres mediante el método del codo, se procedió al entrenamiento del modelo final K-Prototypes utilizando **k = 4**. Esta configuración permite capturar la heterogeneidad del universo empresarial manteniendo un equilibrio entre granularidad analítica e interpretabilidad de los resultados.

El modelo fue inicializado mediante el método **Cao**, el cual es recomendado en la literatura para conjuntos de datos mixtos, ya que genera centroides iniciales más representativos y reduce la variabilidad entre ejecuciones. Adicionalmente, se utilizaron varias inicializaciones del modelo (n_init = 3) con el objetivo de reducir la sensibilidad a los valores iniciales, y se fijó una semilla aleatoria (random_state = 42) para que los resultados puedan reproducirse de manera consistente.

El algoritmo se aplicó sobre la base de datos mixta previamente preparada, la cual combina variables numéricas estandarizadas y variables categóricas en formato nominal. Como resultado, el modelo asignó a cada empresa un grupo o clúster, de acuerdo con su similitud en términos financieros y estructurales.

Estas etiquetas de clúster fueron incorporadas a la base de datos original, lo que permitió relacionar los resultados del modelo con la información de cada empresa y facilitar su análisis posterior. Finalmente, se obtuvo una base de datos enriquecida, en la que cada empresa quedó clasificada dentro de uno de los cuatro segmentos identificados.

4.4.4 Caracterización y perfilamiento de los clústeres

Una vez asignadas las etiquetas de clúster, se procedió a describir cada grupo mediante la construcción de una matriz de perfil por clúster, cuyo objetivo es resumir de forma ordenada las principales diferencias entre los segmentos identificados.

Para las variables numéricas se utilizaron medidas más representativas, como la mediana y los cuartiles primero (Q1) y tercero (Q3), en lugar del promedio. Esta decisión se tomó debido a que los datos financieros suelen presentar valores extremos y distribuciones desiguales, por lo que estas medidas permiten reflejar mejor el comportamiento típico de cada grupo.

En el caso de las variables categóricas, se identificaron las categorías más frecuentes dentro de cada clúster (TOP 1, TOP 2 y TOP 3), junto con su respectivo número de empresas. De esta manera, no solo se identifica la categoría principal, sino también la variedad interna existente en cada segmento.

Adicionalmente, se calculó el tamaño de cada clúster en términos del número de empresas y su participación porcentual respecto al total analizado, lo que permite visualizar con claridad el peso que tiene cada grupo dentro del conjunto empresarial estudiado.

La matriz de perfil resultante constituye una herramienta central del análisis, ya que permite interpretar los clústeres desde una perspectiva integral, combinando dimensiones financieras, estructurales y geográficas. Esta caracterización sirvió como base para el análisis detallado de cada segmento presentado en la sección de resultados.

4.5 Evaluación e interpretación

4.5.1 Segmentación Estratégica mediante K-Prototypes

Dada la naturaleza híbrida de los datos (variables categóricas y numéricas) y la alta dispersión identificada en el análisis descriptivo, se implementó el algoritmo de aprendizaje no supervisado K-Prototypes. Esta técnica permitió agrupar las empresas no por una clasificación preestablecida, sino por la similitud intrínseca en su estructura financiera, laboral y geográfica. Como resultado, se identificaron cuatro conglomerados (clusters) que representan la realidad multidimensional del sector empresarial en el país.

4.5.2 Visualización y Validación de los Conglomerados

Para validar la separación de los grupos, se aplicó una reducción de dimensionalidad mediante el Análisis de Componentes Principales (PCA).

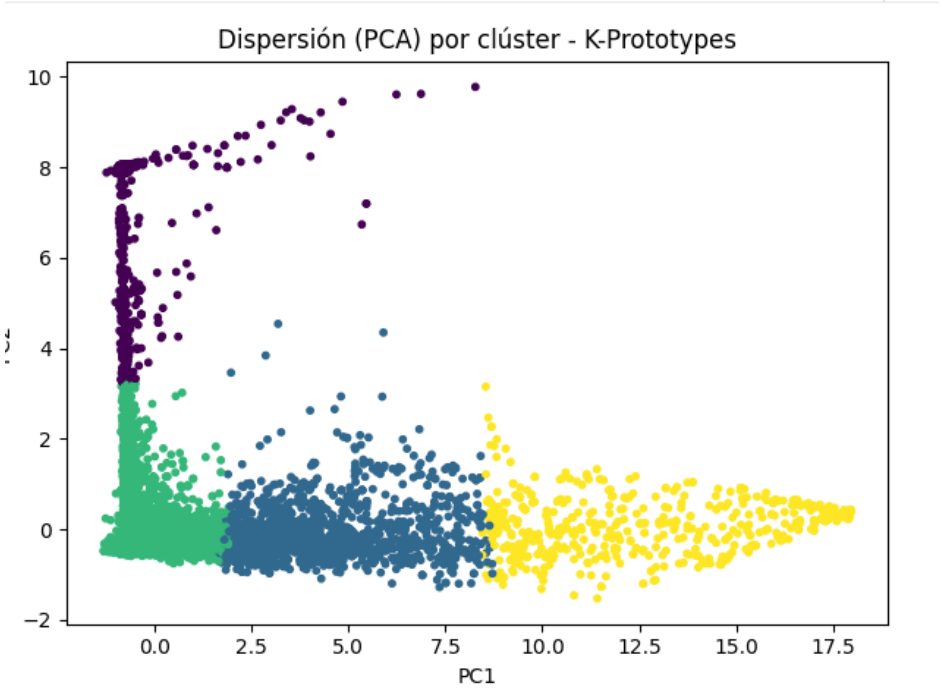


Ilustración 9 Grafico de Disperción (PCA) por clusters – K-Prototypes

Fuente: Elaboración propia.

En la Ilustración 9 se aprecia una segmentación clara de los datos en el espacio. La presencia de grupos bien definidos confirma que el algoritmo logró captar diferencias estructurales importantes entre las empresas. Mientras la mayor parte de la muestra se concentra en una zona de menor variabilidad, asociada principalmente a microempresas, los grupos correspondientes a empresas medianas y grandes se desplazan hacia los extremos de los componentes principales, lo que evidencia una separación clara en términos de escala operativa y capacidad financiera.

4.5.3 Caracterización Detallada de los Perfiles Empresariales

A continuación, se explican los cuatros modelos de empresas identificados, los cuales van más allá de una clasificación sectorial por CIIUU.

4.5.3.1 Cluster 2: El Tejido Microempresarial de Base (91.89%)

Este cluster constituye el núcleo masivo de la economía nacional, albergando a 33,234 empresas. Se define como el "tejido básico" del país, caracterizado por una escala operativa

mínima con una mediana de **4 empleados** y activos corrientes que rondan los **$41,329**. Geográficamente, tiene su mayor anclaje en la **región Sierra** y está compuesto predominantemente por Sociedades Anónimas y S.A.S. de reciente creación. Representa un segmento de subsistencia o de servicios locales con capacidad de inversión limitada.

4.5.3.2 Cluster 1: El Segmento Mediano Consolidado (4.65%)

Conformado por 1,685 empresas, este grupo marca el primer salto hacia la madurez corporativa. Con una mediana de **25 empleados** y activos que superan los **$2.0 millones**, este segmento posee la infraestructura necesaria para generar valor agregado intermedio. A diferencia del cluster anterior, su presencia es mayoritaria en la **región Costa**, sugiriendo una orientación hacia actividades productivas y comerciales de mayor escala.

4.5.3.3 Cluster 3: Las Empresas Tractoras y de Capital Intensivo (1.60%)

Representa la élite del parque empresarial (578 empresas). Estas compañías actúan como motores de la economía, con una mediana de **114 empleados** y una robustez financiera que supera los **$8.3 millones** en activos corrientes. Son empresas con alta capacidad de arrastre en las cadenas productivas, concentradas principalmente en la **región Costa**, donde operan bajo estructuras de capital sólidas y una gestión de pasivos compleja para sostener operaciones masivas.

4.5.3.4 Cluster 0: Microempresas de Servicios en la Costa (1.85%)

Este grupo minoritario (671 empresas) presenta un perfil particular: posee una masa laboral pequeña (mediana de 4 empleados), pero con los activos más bajos de todo el estudio (mediana de **$18,072**). Su concentración es casi exclusiva en la **región Costa**. Este perfil sugiere modelos de negocio "livianos", enfocados en la prestación de servicios especializados o comercio minorista de baja inversión de capital, pero formalizados legalmente.

Tabla 3 Matriz de caracterización por cluster emprearial

INDICADOR	CLUSTER 2 (MASIVO)	CLUSTER 1 (MEDIANO)	CLUSTER 0 (NICHO COSTA)	CLUSTER 3 (TRACTORAS)
NÚMERO DE EMPRESAS	33,234	1,685	671	578
PARTICIPACIÓN (%)	91.89%	4.65%	1.85%	1.59%
REGIÓN DOMINANTE	Sierra	Costa	Costa	Costa
TIPO DE CÍA. (PRINCIPAL)	Anónima	Anónima	Anónima	Anónima
MEDIANA DE EMPLEADOS	4	25	4	114
ACTIVO CORRIENTE (MEDIANA)	$41,329	$2,001,174	$18,072$	$8,360,391
VENTAS OPERATIVAS (MEDIANA)	$91,155	$3,279,920	$4,067$	$16,109,762
PATRIMONIO (MEDIANA)	$17,366	$1,091,276	$15,099$	$6,224,166

Fuente: Elaboración propia basada en resultados del modelo K-Prototypes.

4.5.4 Evaluación de la calidad del agrupamiento mediante Silhouette Score

Para evaluar la calidad del agrupamiento obtenido con el algoritmo K-Prototypes, se calculó el Silhouette Score, métrica que permite medir simultáneamente la cohesión interna de los clústeres y la separación entre ellos. Este indicador toma valores entre −1 y 1, donde los valores cercanos a 1 indican que las empresas fueron asignadas correctamente a sus respectivos grupos.

Debido a que el Silhouette Score solo puede calcularse con variables numéricas, se construyó una matriz adicional únicamente para fines de validación. En esta matriz, las variables categóricas fueron transformadas usando One-Hot Encoding, mientras que las variables numéricas fueron ajustadas con RobustScaler. Es importante aclarar que este procedimiento se utilizó solo para evaluar la calidad de los clústeres y no formó parte del entrenamiento del modelo K-Prototypes.

El valor obtenido del Silhouette Score global fue de 0.7657, lo cual indica una alta cohesión interna y una separación clara entre los clústeres, confirmando la consistencia del número de grupos seleccionado y la solidez del modelo de segmentación propuesto.

4.5.5 Discusión de las Hallazgos

El análisis mediante K-Prototypes revela una **fragmentación estructural** en el ecosistema empresarial. La transición del Cluster 2 al Cluster 1 no es lineal; representa una barrera de escalabilidad donde las empresas deben multiplicar drásticamente su dotación de personal y su base de activos.

Asimismo, se identifica un **sesgo geográfico de la escala**: mientras la Sierra domina numéricamente con una base microempresarial atomizada, la Costa concentra los perfiles de mayor capitalización y empleo (Clusters 1 y 3). Esto permite concluir que la heterogeneidad empresarial en el Ecuador no se explica principalmente por el sector económico al que pertenecen las empresas, sino por el nivel de capital que manejan y su ubicación geográfica. Los valores que inicialmente se identificaban como extremos no representan errores en los datos, sino que corresponden a un grupo específico de empresas con una dinámica contable y operativa muy distinta al resto del tejido empresarial del país.

4.6 Implicaciones Estratégicas y Recomendaciones por Cluster Empresarial

La segmentación obtenida a través de técnicas de aprendizaje no supervisado muestra que existe una alta heterogeneidad en el conjunto de empresas analizadas, tanto en su tamaño, capacidad financiera y ubicación geográfica. En este contexto, plantear recomendaciones diferenciadas por clúster resulta clave para lograr un mayor impacto en la formulación de políticas públicas, estrategias financieras y decisiones institucionales.

4.6.1 Cluster 2: El Tejido Microempresarial de Base (91.89%)

Este clúster concentra a la gran mayoría de las empresas analizadas y representa la base del sistema productivo nacional. Su baja escala operativa, el reducido nivel de activos y el pequeño número de empleados reflejan un tipo de empresa orientada principalmente a la subsistencia, a la prestación de servicios locales y a actividades de bajo valor agregado. Su mayor presencia en la región Sierra y su carácter relativamente reciente evidencian una alta dinámica de creación empresarial, aunque también muestran una marcada fragilidad estructural.

Recomendaciones:

Desde un enfoque estratégico, este segmento requiere políticas enfocadas principalmente en su sostenibilidad básica y en procesos de formalización progresiva, más que en estrategias inmediatas de expansión. Se recomienda fortalecer programas de capacitación en gestión administrativa, contable y financiera, orientados a mejorar el manejo del flujo de caja, la planificación básica y la toma de decisiones informadas.

En cuanto al acceso al financiamiento, este debe estructurarse de manera cuidadosa, priorizando productos de bajo riesgo enfocados en capital de trabajo y acompañados de asistencia técnica. Desde una perspectiva analítica, estas empresas no deberían ser evaluadas bajo los mismos criterios que los segmentos de mayor tamaño, ya que esto podría generar exclusión financiera innecesaria. En su lugar, este clúster se beneficia de modelos de evaluación

simplificados y de políticas diferenciadas que reconozcan su papel como base del empleo y de la actividad económica local.

4.6.2 Cluster 1: El Segmento Mediano Consolidado (4.65%)

El Segmento Mediano Consolidado representa una etapa de transición clave dentro del ciclo de madurez empresarial. Estas empresas han superado la fase de subsistencia y cuentan con una infraestructura organizacional y financiera que les permite generar valor agregado intermedio. Su mayor presencia en la región Costa sugiere una orientación hacia actividades productivas y comerciales de mayor escala y alcance.

Recomendaciones:

Para este cluster, las estrategias deben centrarse en consolidación, eficiencia operativa y escalamiento controlado. Se recomienda promover el acceso a financiamiento estructurado para inversión productiva, modernización tecnológica y mejora de procesos, con el objetivo de fortalecer su competitividad y reducir vulnerabilidades financieras.

Adicionalmente, este segmento constituye un candidato natural para programas de encadenamiento productivo, actuando como nexo entre microempresas y grandes compañías. Desde el punto de vista institucional y financiero, estas empresas justifican esquemas de evaluación crediticia diferenciada y sistemas de monitoreo preventivo, dado que aún se encuentran en una etapa donde la intervención oportuna puede evitar procesos de deterioro financiero y facilitar su evolución hacia estructuras empresariales más robustas.

4.6.3 Cluster 3: Las Empresas Tractoras y de Capital Intensivo (1.60%)

Este cluster agrupa a las empresas de mayor tamaño, complejidad financiera y capacidad operativa del estudio. Su elevada dotación laboral y su sólida estructura de activos las posicionan como actores estratégicos dentro de las cadenas productivas, con un alto impacto en

términos de empleo, inversión y generación de valor agregado. Su concentración en la región Costa refuerza su rol en actividades industriales, comerciales y logísticas de gran escala.

Recomendaciones:

Las principales recomendaciones para este segmento se orientan hacia la gestión avanzada del riesgo, la sostenibilidad financiera y el impacto sistémico. Debido a su mayor complejidad operativa, se recomienda implementar mecanismos de seguimiento continuo sobre indicadores clave como liquidez, nivel de endeudamiento y estructura de pasivos, así como realizar análisis que permitan evaluar su sensibilidad frente a posibles choques macroeconómicos.

Este tipo de empresas resulta especialmente adecuado para esquemas de financiamiento de largo plazo, emisiones de deuda, alianzas estratégicas y procesos de internacionalización. Desde una perspectiva institucional, este clúster debe ser considerado como un segmento de alto impacto sistémico, ya que cualquier deterioro en su situación financiera podría generar efectos relevantes sobre proveedores, clientes y el empleo. Por ello, se justifica la aplicación de modelos analíticos y esquemas de supervisión diferenciados.

4.6.4 Cluster 0: Microempresas de Servicios en la Costa (1.85%)

Este cluster minoritario, que representa el 1.85% del total de empresas (671 casos), presenta un perfil caracterizado por estructuras laborales reducidas (mediana de 4 empleados) y los niveles de activos más bajos del estudio (mediana de USD 18,072). Las empresas se concentran casi exclusivamente en la región Costa y desarrollan principalmente actividades de servicios y comercio minorista de baja inversión de capital. Este patrón sugiere modelos de negocio livianos, con mínima infraestructura productiva, pero formalizados legalmente, estrechamente vinculados a dinámicas comerciales locales. en actividades de servicios y comercio de baja inversión de capital. Su localización geográfica sugiere un fuerte vínculo con economías urbanas y dinámicas comerciales locales.

Recomendaciones:

Para este segmento, las estrategias deben enfocarse en la estabilidad operativa y el fortalecimiento de capacidades básicas, más que en procesos de expansión. Se recomienda impulsar educación financiera y gestión empresarial elemental, orientadas al control de costos, planificación de ingresos y orden administrativo. En términos de financiamiento, estas empresas se benefician de instrumentos flexibles y de corto plazo, destinados principalmente a capital de trabajo. Desde una perspectiva analítica, es fundamental evaluar este cluster con indicadores acordes a su escala, evitando comparaciones con empresas de mayor tamaño, y aprovechar su concentración regional para políticas focalizadas de apoyo al desarrollo local.

CAPITULO V: CONCLUSIONES

5.1 Conclusiones

Se concluye que la aplicación del algoritmo K-Prototypes sobre la base de datos de la SuperCias 2024 permitió identificar de manera clara patrones estructurales y financieros que no pueden ser detectados mediante análisis estadísticos tradicionales. El modelo evidenció que el universo empresarial ecuatoriano se organiza en cuatro perfiles bien diferenciados, definidos principalmente por la profundidad de su capital y su estructura laboral, más que por el sector económico al que pertenecen.

- Se determinó que las variables contables relacionadas con el activo corriente, las ventas operativas y el número de empleados son las que tienen mayor peso en los procesos de segmentación. El análisis mostró que el sector de actividad económica (CIIU), por sí solo, no es suficiente para explicar el comportamiento financiero de las empresas, siendo la capacidad de acumulación de activos y la estructura de pasivos los factores que realmente diferencian a los grupos empresariales en el contexto actual.

- El proceso de depuración, integración y normalización de la base de datos resultó fundamental para transformar una base masiva y heterogénea en un conjunto de datos adecuado para la aplicación del algoritmo. Se concluye que este tratamiento previo permitió reducir el sesgo generado por la asimetría propia de las variables financieras ecuatorianas, logrando que el modelo de clustering sea más robusto y que los grupos obtenidos representen realidades económicas consistentes y no distorsiones provocadas por valores extremos.

- La implementación del algoritmo K-Prototypes validó la necesidad de utilizar técnicas híbridas en el análisis empresarial. Al integrar simultáneamente variables numéricas (como utilidad y patrimonio) con categóricas (como región y tipo de compañía),

el modelo logró captar la "naturaleza mixta" de las empresas a diferencia de la estadística presentada en los canales oficiales de la SUPERCIAS y una estadística básica que se puede aplicar en la base (Media, Mediana y Moda).

- Finalmente, se concluye que existe una brecha estructural marcada entre los clústeres identificados, la cual fue validada mediante técnicas de reducción de dimensionalidad como el PCA. El modelo permitió confirmar que el 91.89% de las empresas se concentra en una microescala operativa (Clúster 2), mientras que un reducido grupo del 1.60% (Clúster 3) concentra la mayor capacidad de tracción económica. Este resultado evidencia que la conformación de los grupos empresariales está fuertemente influenciada por la escala de capital, reflejando una desconexión significativa entre la gran masa empresarial y los líderes del mercado.

Referencias

Alpaydin, Ethem. (2020). *Introduction to machine learning*. The MIT Press.

Ayaquica-Martínez, I. O., Martínez-Trinidad, J. Fco., & Carrasco-Ochoa, J. A. (2006). *Conceptual K-Means Algorithm Based on Complex Features* (pp. 491–501). https://doi.org/10.1007/11892755_51

Everitt, Brian. (2011). *Cluster analysis*. Wiley.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 9. https://doi.org/10.1186/s41044-016-0014-0

Hair, J. F. . (2019). *Multivariate data analysis*. Cengage.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. En *Data Mining and Knowledge Discovery* (Vol. 12).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Knox, S. W. (2018). *Machine Learning*. Wiley. https://doi.org/10.1002/9781119439868

Li, Z., & Li, P. (2015). Clustering Algorithm of Quantum Self-Organization Network. *Open Journal of Applied Sciences*, *05*(06), 270–278. https://doi.org/10.4236/ojapps.2015.56028

Meza Fabián Íñiguez Mauricio Becerra Oswaldo Rivera Johanna Vera Juan Carlos Chiluiza Jorge Ortega Rocío Malla Franklin Simbaña Adriana Valverde Sheldon López Marcelo Sotaminga, P., & Chang Calvache David Hurtado Vicente Palacios, F. (2022). *Mensaje de la Ministra Vianna Maino*.

Ping, Y., Li, H., Guo, C., & Hao, B. (2025). kProtoClust: Towards Adaptive k-Prototype Clustering without Known k. *Computers, Materials and Continua*, *82*(3), 4949–4976. https://doi.org/10.32604/cmc.2025.057693

Porter, M. E. (2007). *The Five Competitive Forces That Shape Strategy*.

Provost, Foster., & Fawcett, Tom. (2013). *Data science for business*. O'Reilly.

¿Qué es la agrupación en clústeres k-means? | IBM. (s/f). Recuperado el 3 de enero de 2026, de https://www.ibm.com/mx-es/think/topics/k-means-clustering

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus External Cluster Validation Indexes. *International Journal of Computers and Communications*, *5*(1), 27–34.

What is TDSP? (s/f). Recuperado el 3 de enero de 2026, de https://www.datascience-pm.com/tdsp/

www.supercias.gob.ec. (s/f). Recuperado el 3 de enero de 2026, de https://www.supercias.gob.ec/portalscvs/index.htm

ANEXOS

ANEXO A Aplicación del Algoritmo K-Prototypes para la segmentación de compañías ecuatorianas 2022

A.1 Propósito del Anexo

El presente anexo tiene como finalidad documentar y analizar la aplicación del algoritmo K-Prototypes al conjunto de datos del año 2022, con el objetivo de evaluar la estabilidad temporal de los patrones de segmentación empresarial identificados en el análisis principal correspondiente al año 2024.

Para garantizar la comparabilidad de resultados, el procedimiento aplicado en este anexo replica de forma estricta la metodología utilizada en el cuerpo principal del trabajo, manteniendo las mismas variables, criterios de depuración, técnicas de preprocesamiento y configuración del modelo.

De este modo, el análisis del año 2022 permite verificar si los clústeres identificados responden a estructuras empresariales persistentes y no a comportamientos coyunturales de un solo período.

A.2 Metodología Aplicada

La segmentación de las compañías correspondientes al año 2022 se desarrolló siguiendo el enfoque del Team Data Science Process (TDSP), aplicando el algoritmo K-Prototypes, adecuado para conjuntos de datos que combinan variables numéricas y categóricas.

El análisis utilizó:

- Variables numéricas relacionadas con la escala económica y financiera (empleados, activos, pasivos, patrimonio, ventas, utilidad neta y liquidez corriente).
- Variables categóricas asociadas a la forma jurídica y localización regional.

Las variables numéricas fueron estandarizadas y los valores extremos tratados mediante winsorización, mientras que los registros con información incompleta en variables críticas fueron excluidos, con el fin de preservar la consistencia del modelo.

A.3 Determinación de Numero de Clusters

Para definir el número óptimo de clústeres, se aplicó el **método del codo**, evaluando el comportamiento del costo intra-cluster del algoritmo K-Prototypes para valores de k entre 2 y 6.

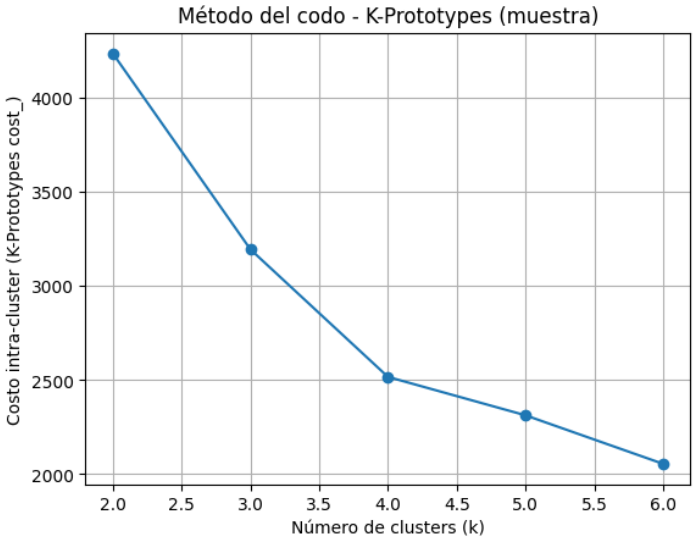


Ilustración 10 Método del Codo Base 2022

Los resultados evidenciaron una disminución significativa del costo hasta $k = 4$, seguida de reducciones marginales progresivamente menores para valores superiores. Este patrón indica un punto de equilibrio entre cohesión interna y complejidad del modelo en **cuatro clústeres**, por lo que se seleccionó este valor como óptimo.

La elección de $k = 4$ resulta consistente con el análisis realizado para el año 2024, reforzando la comparabilidad interanual de los resultados.

A.4 Validacion del Modelo de Segmentación

La calidad del modelo fue evaluada mediante dos enfoques complementarios:

- **Visualización PCA**: la proyección de las observaciones en dos componentes principales mostró una separación clara entre los clústeres, particularmente asociada a diferencias en escala económica y estructura financiera.

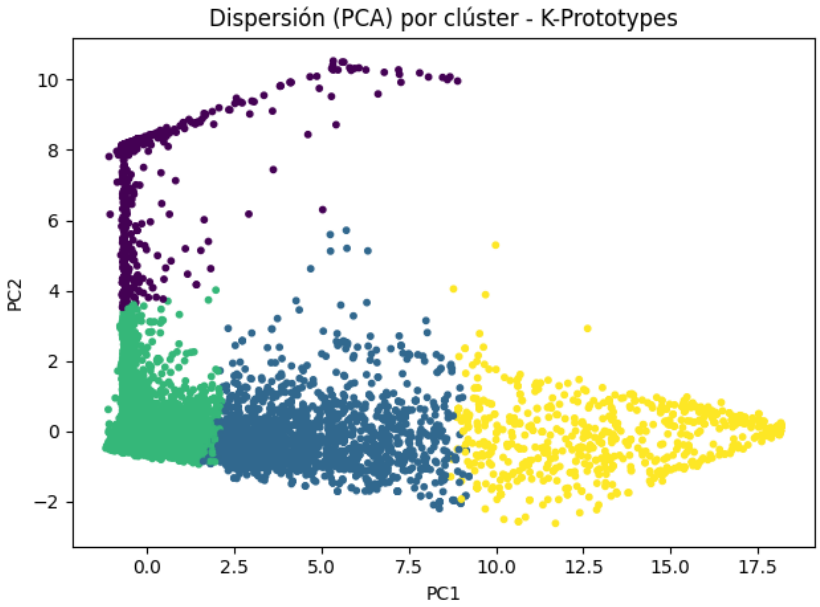


Ilustración 11 Grafico PCA 2022

- **Silhouette Score global**: el valor obtenido fue de **0.7964**, lo que indica una alta cohesión interna de los grupos y una adecuada separación entre ellos.

En el contexto de datos empresariales heterogéneos y de gran volumen, este resultado evidencia un **desempeño robusto del algoritmo**, confirmando la validez de la segmentación obtenida para el año 2022.

A.5 Caracterizacion de los Clusteres – Año 2022

A.5.1 Cluster 2: Tejido Microempresarial de Base (92.64 %)

Este clúster agrupa 44.557 empresas, representando el 92.64 % del total, lo que lo convierte en el segmento dominante del tejido empresarial ecuatoriano en 2022. Las empresas que lo conforman presentan una estructura laboral reducida, con una mediana de 4 empleados, y niveles limitados de capitalización. El activo corriente mediano alcanza los USD 53.140, mientras que las ventas operativas medianas se sitúan en USD 103.227, reflejando modelos de negocio de pequeña escala y alta dependencia de la operación diaria.

Desde el punto de vista financiero, el patrimonio mediano es de USD 20.584, acompañado de una liquidez corriente baja (mediana de 0.24), lo que sugiere escasa holgura financiera y una mayor exposición a restricciones de corto plazo. Geográficamente, este clúster se concentra principalmente en la región Sierra, y jurídicamente predomina la compañía anónima, seguida de la responsabilidad limitada y las sociedades por acciones simplificadas. Este grupo representa la base productiva formal del país, caracterizada por su masividad, baja escala de capital y elevada vulnerabilidad ante cambios en el entorno económico.

A.5.2 Cluster 1: Empresas Medianas Consolidadas (4.23 %)

El clúster 1 está compuesto por 2.034 empresas, equivalentes al 4.23 % del total, y agrupa organizaciones con un nivel intermedio de madurez económica y operativa. Estas empresas presentan una mediana de 49 empleados, junto con un activo corriente mediano de USD 3.871.266 y ventas operativas medianas de USD 6.566.422, lo que evidencia una capacidad productiva significativamente superior a la del segmento microempresarial. El patrimonio mediano asciende a USD 2.125.501, mientras que la liquidez corriente se sitúa en torno a 0.80, indicando una gestión financiera más equilibrada, aunque aún expuesta a tensiones propias de operaciones de mayor escala.

Estas empresas se concentran principalmente en la región Costa, con presencia relevante en la Sierra, y están constituidas mayoritariamente como compañías anónimas. Este clúster constituye un segmento estratégico para el crecimiento económico, con potencial para la expansión, la generación de empleo y la articulación de encadenamientos productivos.

A.5.3 Cluster 3: Empresas Tractoras y de Capital Intensivo (1.54 %)

El clúster 3 agrupa 743 empresas, lo que representa el 1.54 % del total, y corresponde a las organizaciones de mayor escala económica identificadas en el análisis. Estas empresas cuentan con una mediana de 203 empleados, así como niveles elevados de capitalización. El activo corriente mediano supera los USD 16.408.565, mientras que las ventas operativas medianas alcanzan los USD 29.345.314.

El patrimonio mediano es de USD 12.827.824, acompañado de una liquidez corriente cercana a 0.85, lo que refleja estructuras financieras robustas, aunque con un nivel de apalancamiento acorde a su escala operativa. Este clúster se localiza mayoritariamente en la región Costa y está conformado principalmente por compañías anónimas, con presencia relevante de sucursales extranjeras.

Estas empresas cumplen un rol de empresas tractoras, con alto impacto en la generación de empleo, la inversión y la dinamización de sectores productivos.

A.5.4 Cluster 0: Microempresas de Servicios en la Costa (1.58 %)

El clúster 0 está compuesto por 763 empresas, equivalentes al 1.58 % del total, y presenta un perfil microempresarial particular. Las empresas de este grupo cuentan con una mediana de 3 empleados y los niveles más bajos de activos y ventas del conjunto analizado. El activo corriente mediano es de USD 21.130, mientras que las ventas operativas medianas alcanzan apenas USD 17.690.

Un rasgo distintivo de este clúster es su liquidez corriente extremadamente elevada, lo que sugiere estructuras operativas muy livianas, con mínimos compromisos financieros y bajo nivel de endeudamiento. Estas empresas se concentran casi exclusivamente en la región Costa y corresponden principalmente a actividades de servicios y comercio de baja inversión de capital, formalizadas, pero de limitada complejidad económica.

A.6 Síntesis del Anexo

El análisis del año 2022 confirma la existencia de cuatro perfiles empresariales claramente diferenciados, definidos principalmente por la escala de capital, la estructura laboral y la localización geográfica.

La alta consistencia interna de los clústeres y la robustez de las métricas de validación evidencian que la segmentación obtenida refleja patrones estructurales persistentes del tejido empresarial ecuatoriano, lo que respalda la comparación con los resultados del año 2024 desarrollados en el cuerpo principal del trabajo.

ANEXO B Aplicación del Algoritmo K-Prototypes para la segmentación de compañías ecuatorianas 2023

B.1 Propósito del Anexo

El presente anexo tiene como objetivo documentar y analizar la aplicación del algoritmo K-Prototypes al conjunto de datos correspondiente al ejercicio fiscal 2023, con el fin de evaluar la consistencia interanual de los patrones de segmentación empresarial identificados en el análisis principal del año 2024 y en el anexo correspondiente al año 2022.

Para garantizar la comparabilidad de los resultados, el procedimiento aplicado en este anexo replica de manera estricta la metodología utilizada en los otros períodos de estudio, manteniendo constantes las variables analizadas, los criterios de preprocesamiento y la configuración del modelo.

B.2 Metodología Aplicada

La segmentación empresarial del año 2023 se desarrolló siguiendo el enfoque del Team Data Science Process (TDSP), empleando el algoritmo K-Prototypes, adecuado para conjuntos de datos que combinan variables numéricas y categóricas.

Se utilizaron:

- Variables numéricas asociadas a la escala económica y financiera de las empresas (número de empleados, activos corrientes, pasivos corrientes, patrimonio, ventas operativas, utilidad neta y liquidez corriente).
- Variables categóricas relacionadas con el tipo de compañía y la región geográfica.

Las variables numéricas fueron escaladas mediante RobustScaler, mientras que las variables categóricas se transformaron mediante One-Hot Encoding exclusivamente para fines

de validación. Los valores atípicos fueron tratados mediante winsorización y los registros con información incompleta en variables críticas fueron excluidos, con el objetivo de preservar la estabilidad del modelo.

B.3 Determinación de Numero de Clusters

Para determinar el número óptimo de clústeres, se aplicó el método del codo, evaluando el costo intra-cluster del algoritmo K-Prototypes para valores de $k$entre 2 y 6.

El análisis del gráfico evidencia una reducción significativa del costo hasta $k = 4$, a partir del cual las mejoras marginales disminuyen de forma progresiva. Este comportamiento confirma la existencia de un punto de equilibrio adecuado en cuatro clústeres, por lo que se seleccionó este valor como óptimo para el año 2023, en coherencia con los resultados obtenidos para los años 2022 y 2024.

B.4 Validacion del Modelo de Segmentación

La calidad del modelo fue evaluada mediante dos enfoques complementarios:

- Análisis de Componentes Principales (PCA): la proyección de las observaciones en dos componentes principales mostró una separación clara entre los clústeres, particularmente asociada a diferencias en escala económica y estructura financiera.

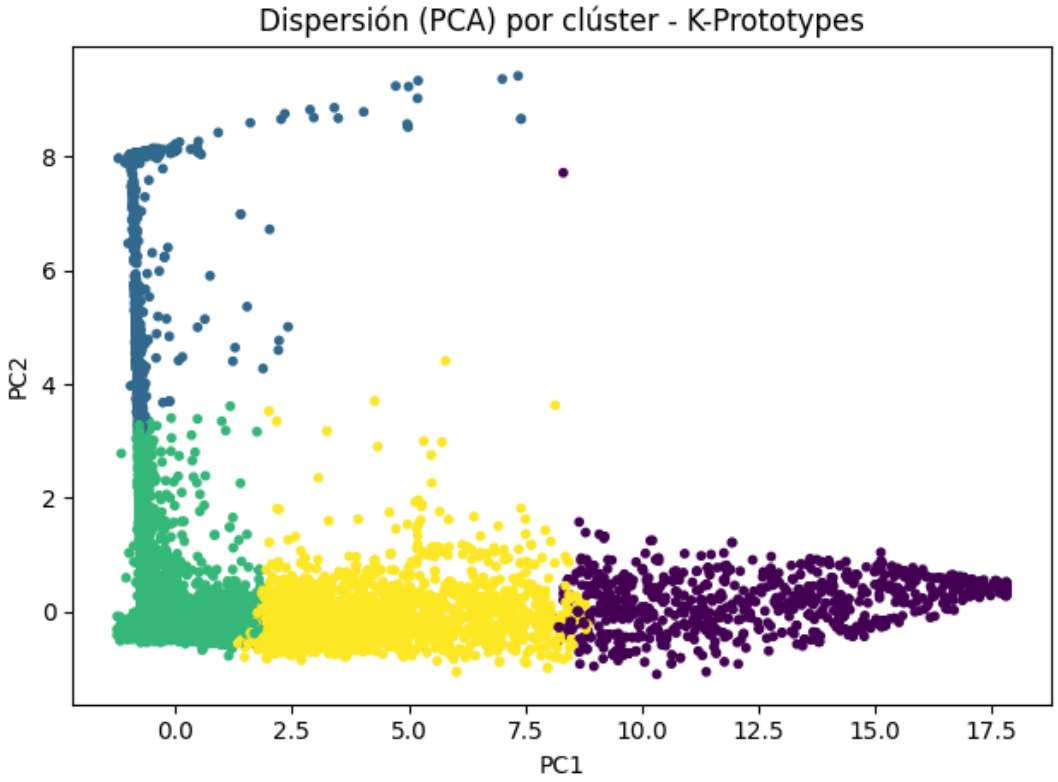


Ilustración 12 Grafico PCA 2023

- Silhouette Score global: el valor obtenido fue de 0.7767, lo que indica una alta cohesión interna dentro de los clústeres y una adecuada separación entre grupos.

Este resultado confirma la robustez del modelo de segmentación aplicado al año 2023 y refuerza la estabilidad del esquema de clústeres a lo largo del período de análisis.

B.5 Caracterizacion de los Clusteres – Año 2022

B.5.1 Cluster 2: Tejido Microempresarial de Base (91.58 %)

Este clúster agrupa 48.848 empresas, representando el 91.58 % del total, consolidándose nuevamente como el segmento dominante del tejido empresarial ecuatoriano.

Las empresas de este grupo presentan una mediana de 4 empleados, junto con niveles reducidos de capitalización. El activo corriente mediano asciende a USD 98.794, mientras que

las ventas operativas medianas se sitúan en USD 103.227, reflejando modelos de negocio de pequeña escala y alta dependencia del flujo operativo. Desde el punto de vista financiero, el patrimonio mediano alcanza los USD 47.180, con una liquidez corriente baja (mediana de 0.24), lo que evidencia una limitada capacidad de maniobra financiera en el corto plazo.

Geográficamente, este clúster se concentra principalmente en la región Costa, seguido por la Sierra, y está compuesto mayoritariamente por compañías anónimas, acompañadas por sociedades de responsabilidad limitada y sociedades por acciones simplificadas. Este segmento representa la base productiva formal del país, caracterizada por su alta masividad y baja profundidad de capital.

B.5.2 Cluster 1: Microempresas Formalizadas de Baja Escala (1.79 %)

El clúster 1 está conformado por 956 empresas, equivalentes al 1.79 % del total, y agrupa empresas de muy pequeña escala, pero con estructuras formales consolidadas. Estas organizaciones presentan una mediana de 4 empleados, junto con un activo corriente mediano de apenas USD 32.928, y ventas operativas medianas de USD 6.566.422. A pesar de su tamaño reducido, algunas empresas muestran niveles puntuales de facturación elevados, lo que explica la dispersión observada en el PCA.

El patrimonio mediano se sitúa en USD 27.278, y la liquidez corriente alcanza valores cercanos a 0.80, indicando una gestión financiera básica pero funcional. Estas empresas se localizan principalmente en la región Costa y corresponden a actividades de servicios y comercio de baja inversión inicial.

B.5.3 Cluster 3: Empresas de Escala Media-Alta (4.92 %)

El clúster 3 agrupa 2.623 empresas, lo que representa el 4.92 % del total, y corresponde a empresas con una estructura laboral y financiera significativamente superior al promedio.

Estas organizaciones cuentan con una mediana de 68 empleados, junto con un activo corriente mediano de USD 5.019.336 y ventas operativas medianas de USD 29.345.314.

El patrimonio mediano asciende a USD 3.070.202, acompañado de una liquidez corriente cercana a 0.85, lo que refleja una estructura financiera más sólida y capacidad de absorción de riesgos operativos. Este clúster se concentra mayoritariamente en la región Costa y está compuesto principalmente por compañías anónimas, cumpliendo un rol relevante en términos de generación de empleo y dinamización económica.

B.5.4 Cluster 0: Empresas Tractoras y de Alta Intensidad de Capital (1.71 %)

El clúster 0 está conformado por 910 empresas, equivalentes al 1.71 % del total, y agrupa a las organizaciones de mayor escala económica del año 2023. Estas empresas presentan una mediana de 238 empleados, junto con niveles elevados de capitalización. El activo corriente mediano supera los USD 23.222.584, mientras que el patrimonio mediano asciende a USD 17.196.041.

Desde el punto de vista financiero, estas empresas presentan estructuras complejas, con niveles de pasivo acordes a su escala, y desempeñan un rol clave como empresas tractoras, con alto impacto en inversión, empleo y encadenamientos productivos.

B.6 Síntesis del Anexo

El análisis correspondiente al año 2023 confirma nuevamente la existencia de cuatro perfiles empresariales claramente diferenciados, definidos principalmente por la escala económica, la estructura laboral y la localización geográfica, más que por el sector de actividad.

La consistencia de los clústeres, junto con un Silhouette Score elevado y patrones de separación estables en el PCA, evidencia que la segmentación obtenida mediante K-Prototypes refleja estructuras empresariales persistentes en el tiempo. Estos resultados refuerzan la validez

del enfoque longitudinal desarrollado en la investigación y constituyen un insumo clave para la interpretación integrada de los años 2022, 2023 y 2024 presentada en las conclusiones generales del trabajo.

ANEXO C Código de la aplicación del algoritmo K-Prototype para las Bases 2022 – 2023 – 2024