Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Modelos Analíticos de Priorización Empresarial y Potenciación del Empleo Formal en Santiago de Cali

Trabajo de Grado

Presentado por:

Laura Daniela Rey Gaitán

Bogotá, D.C.

2026

Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

Modelos Analíticos de Priorización Empresarial y Potenciación del Empleo Formal en Santiago de Cali

Trabajo de Grado

Presentado por: Laura Daniela Rey Gaitán

Bajo la dirección de: Yudy Constanza Castaño

Maestría en Business Analytics

Escuela de Administración

Escuela de Ingeniería Ciencia y Tecnología

13 de enero de 2026

Bogotá, D.C., Colombia

2026

Mi gratitud a la directora de este proyecto, por su acompañamiento comprometido y por brindarme la estructura necesaria para avanzar con seguridad. A mi familia, por ser la fuerza que sostiene mi perseverancia y mi motivación. Y a Camilo, porque sin su apoyo incondicional, cariño y presencia constante, este logro no habría sido posible.

Contenido

Preliminares ***8***

Declaración de exoneración de responsabilidad 10

Lista de figuras 11

Lista de tablas 12

Resumen Ejecutivo 15

Abstract 17

1. Introducción ***19***

2. Contexto institucional y socioeconómico del proyecto ***22***

3. Planteamiento del problema ***28***

4. Hipótesis del proyecto ***29***

5. Preguntas analíticas ***31***

6. Objetivos ***32***

6.1 Objetivo general 32

6.7 Objetivos específicos 32

7. Alcance y limitaciones ***33***

7.1 Alcance del Proyecto 33

7.2 Limitaciones del Proyecto 35

8. Marco Teórico ***36***

8.1 Técnicas utilizadas 37

9. Metodología 42

9.1 Fuentes de información utilizadas 44

Fuente: Elaboración propia. 48

9.2 Decisiones estratégicas de procesamiento 48

9.3 Procesos generales de ingeniería de variables 51

9.4 Arquitectura analítica del proyecto 55

10. Pregunta analítica 1: ¿Qué características empresariales comparten mayoritariamente las empresas que han logrado generar empleo formal en los años recientes en Santiago de Cali? 56

10.1 Construcción de Dataset A: 56

10.2 Análisis del tejido empresarial que ha generado empleo en los últimos años 59

10.3 Conclusiones 64

11. Pregunta de análisis 2: ¿Qué grupos o perfiles empresariales pueden identificarse dentro del conjunto de empresas que generan empleo formal? 65

11.1 Variables a incluir en el modelo, creación de nuevas variables y su tratamiento 65

11.2 Selección de número óptimo de Clústeres 68

11.3 Aplicación de modelos de clusterización 70

11.4 Evaluación de los modelos 70

11.5 Caracterización de Clústeres 72

11.6 Conclusiones 78

12. Pregunta de análisis 3: ¿Cuál es la probabilidad de que una empresa incremente su número de empleados en el año siguiente a partir de las características empresariales que presenta en el periodo actual? 79

12.1 Construcción de la base predictiva 80

12.2 Modelos aplicados 82

12.3 Evaluación de modelos y elección 84

12.4 Comportamiento e interpretabilidad del modelo 89

12.5 Implementación del modelo 92

13. Base Maestra de Priorización 93

14. Conclusiones 94

15. Recomendaciones estratégicas para la Secretaría de Desarrollo Económico 96

15.1 Adopción del perfil de "empresa estrella" como como eje de focalización estratégica 96

15.2 Desarrollo de programas diferenciados para empresas generadoras de empleo en riesgo 97

15.3 Estrategia de asignación diferenciada de cupos de fortalecimiento empresarial según el perfil de clúster 98

15.4 Estrategia de robustecimiento operativo de la Secretaría 99

16. Pasos a seguir 101

16.1 Construcción e implementación de un dashboard de consulta de la Base de Priorización 101

16.2 Robustecimiento anual del modelo predictivo y construcción de un pipeline automatizado 102

16.3 Desarrollo de un proyecto de evaluación de impacto de programas e incentivos de la Secretaría 103

17. Referencias bibliográficas 105

Preliminares

Declaración de originalidad y autonomía

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi propia cuenta y que, por lo tanto, su contenido es original.

Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

Laura Daniela Rey Gaitán

Firmado en Bogotá, D.C. 15 de diciembre de 2025

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de su autor. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

Laura Daniela Rey Gaitán

Firmado en Bogotá, D.C. 15 de diciembre de 2025

Lista de figuras

Figura 1 _____ 24

Figura 2 _____ 43

Figura 3 _____ 44

Figura 4 _____ 49

Figura 5 _____ 50

Figura 6 _____ 53

Figura 7 _____ 55

Figura 8 _____ 58

Figura 9 _____ 62

Figura 10 _____ 63

Figura 11 _____ 67

Figura 12 _____ 68

Figura 13 _____ 69

Figura 14 _____ 73

Figura 15 _____ 90

Figura 16 _____ 94

Lista de tablas

Tabla 1 _____ 45

Tabla 2 _____ 45

Tabla 3 _____ 52

Tabla 4 _____ 72

Tabla 5 _____ 73

Tabla 6 _____ 86

Tabla 7 _____ 99

Abreviaturas:

AUC: Área Under the Curve (Área bajo la curva)

AUC–ROC: Area Under the Receiver Operating Characteristic Curve

AUC–PR: Area Under the Precision–Recall Curve

CIIU: Clasificación Industrial Internacional Uniforme

CRISP–DM: Cross Industry Standard Process for Data Mining

CV: Cross–Validation (Validación Cruzada)

DANE: Departamento Administrativo Nacional de Estadística

DIAN: Dirección de Impuestos y Aduanas Nacionales

EDA: Exploratory Data Analysis (Análisis Exploratorio de Datos)

ETL: Extract, Transform, Load

FN: False Negative

FP: False Positive

F1–Score: Medida armónica entre precisión y recall

GEIH: Gran Encuesta Integrada de Hogares

LR: Logistic Regression

ML: Machine Learning

MIPYME: Micro, pequeña y mediana empresa

OHE: One–Hot Encoding

PIB: Producto Interno Bruto

PoC: Proof of Concept

RF: Random Forest

ROC: Receiver Operating Characteristic

SHAP: SHapley Additive exPlanations

SIC: Superintendencia de Industria y Comercio

SMOTE: Synthetic Minority Over–sampling Technique

TPR: True Positive Rate

TN: True Negative

TP: True Positive

Resumen Ejecutivo

La Secretaría de Desarrollo Económico de Santiago de Cali enfrenta el reto estratégico de focalizar sus esfuerzos, programas, incentivos e instrumentos de fortalecimiento empresarial hacia las empresas con mayor potencial para aportar a la meta distrital de creación de 100.000 nuevos empleos entre 2024 y 2027. Ante un panorama de recursos limitados y la necesidad de identificar patrones de generación de empleo, este proyecto desarrolla una solución integral que combina análisis descriptivo, segmentación y modelación predictiva bajo la metodología CRISP–DM, utilizando información histórica empresarial entre 2020 y 2024.

En primera instancia, el proyecto identificó las empresas registradas en el 2024 que generaron empleo en por lo menos 2 años de los 5 observados, sobre estas empresas se desarrolló un análisis descriptivo detallado, lo cual permitió caracterizar factores estructurales relevantes y construir un perfil mayoritario entre estas empresas.

Posteriormente, con el fin de identificar grupos homogéneos dentro de esas empresas y comprender perfiles empresariales con dinámicas similares, se aplicaron técnicas de segmentación. Estos clústeres permitieron distinguir empresas ancla con alto desempeño, empresas con comportamientos intermedios, que se sitúan como perfil de enfoque para el desarrollo de políticas e incentivos generales, y empresas con señales tempranas de vulnerabilidad, lo cual constituye un insumo clave para diseñar intervenciones diferenciadas para la Secretaría.

Finalmente, se construyó un modelo predictivo supervisado para estimar la probabilidad de que cada empresa aumente su planta laboral en el año siguiente. Este modelo permitió identificar las características con mayor capacidad explicativa y, una vez aplicado a la base de datos correspondiente, se integró a una Base Maestra de Priorización de las empresas de Santiago de

Cali con su pertenencia a cada Clúster identificado y su potencial de generación de empleo en 2025, consolidando así una herramienta analítica directamente utilizable por la Secretaría de Desarrollo Económico. En conjunto, el proyecto demuestra la viabilidad técnica y estratégica de incorporar analítica avanzada en la gestión pública local y aporta insumos para el diseño de programas más focalizados que acerquen a la administración al cumplimiento de la meta trazada.

Palabras clave: empleabilidad, modelos estadísticos, decisiones basadas en datos, CRISP-DM, programas empresariales.

Abstract

The Secretariat of Economic Development of Santiago de Cali faces the strategic challenge of focusing its efforts, programs, incentives, and business-strengthening instruments on the companies with the greatest potential to contribute to the district goal of creating 100,000 new formal jobs between 2024 and 2027. In a context of limited resources and the need to identify patterns of employment generation, this project develops a comprehensive solution that integrates descriptive analysis, segmentation, and predictive modeling under the CRISP–DM methodology, using historical business data from 2020 to 2024.

First, the project identified companies registered in 2024 that generated employment in at least two of the five years analyzed. A detailed descriptive analysis was conducted on these firms, enabling the characterization of relevant structural factors and the construction of a predominant profile among them.

Subsequently, segmentation techniques were applied to identify homogeneous groups within these companies and to understand business profiles with similar dynamics. The resulting clusters made it possible to distinguish anchor firms with high performance, firms with intermediate behaviors—considered priority profiles for the design of general policies and incentives—and firms with early signs of vulnerability, providing key insights for differentiated interventions by the Secretariat.

Finally, a supervised predictive model was built to estimate the probability that each company will increase its workforce in the following year. Once applied to the corresponding database, the model was integrated into a Master Prioritization Database for companies in Santiago de Cali, incorporating cluster membership and projected employment-generation potential for 2025. The result is an analytical tool directly usable by the Secretariat of Economic

Development. Overall, the project demonstrates the technical and strategic feasibility of incorporating advanced analytics into local public management and provides inputs for designing more targeted programs that support the administration in achieving its employment goal.

Keywords: employability, statistical models, data-driven decisions, CRISP-DM, business programs.

1. Introducción

La administración distrital de Santiago de Cali, liderada por el alcalde Alejandro Eder, ha definido una meta ambiciosa: lograr la creación de 100.000 nuevos empleos entre 2024 y 2027. Esta meta no solo responde a la urgencia de dinamizar el mercado laboral, sino también a una visión que entiende el empleo formal como el punto de partida para superar desigualdades y construir movilidad social real, para lo cual resulta relevante poner la lupa en el desarrollo y fortalecimiento del tejido empresarial como principal determinador de la generación de empleo formal en la ciudad.

Sin embargo, Santiago de Cali cuenta con más de 100.000 empresas registradas y la Secretaría de Desarrollo Económico cuenta con recursos limitados para desarrollar programas e incentivos que potencien el ecosistema sin perder de vista la meta de generación de empleos formales. En este contexto, la Secretaría de Desarrollo Económico enfrenta un reto estructural: identificar cuáles empresas, sectores y comportamientos empresariales tienen mayor capacidad de aportar a la generación de empleo.

Tradicionalmente, la asignación de apoyos y beneficios empresariales ha respondido a convocatorias abiertas y, muchas veces, generales, que atienden principalmente las necesidades que son presentadas por los diferentes sectores y grupos económicos con criterios cualitativos que, aunque valiosos, no le apuntan necesariamente a la potencialidad que tenga una empresa de aportar a la meta de generación de empleo planteada por la administración.

La ciudad cuenta con un universo empresarial diverso, caracterizado por una amplia presencia de micro y pequeñas empresas y un número reducido de empresas medianas y grandes con mayor capacidad de absorción laboral. En este escenario, la analítica de datos puede convertirse en un instrumento transformador: construir soluciones analíticas como técnicas

descriptivas, segmentación empresarial y modelos predictivos para comprender y anticipar el comportamiento laboral de las empresas de Cali.

El estudio parte de una premisa fundamental: las características empresariales contienen información valiosa sobre el comportamiento futuro de las empresas, incluyendo su capacidad para generar empleo. Variables como el tamaño, la antigüedad, el nivel de ingresos, la estabilidad financiera, el sector económico y la dinámica territorial han mostrado aportar señales sobre la evolución de las empresas. No obstante, la realidad empresarial es compleja y multifactorial. Por ello, este proyecto reconoce que, aunque los datos estructurados capturan una parte esencial del fenómeno, existen también elementos no observados —estrategias internas, gobernanza corporativa, relaciones comerciales, shocks de mercado, dinámicas regulatorias— que influyen en la decisión de contratar o expandirse.

El proyecto se desarrolla bajo los lineamientos de la metodología CRISP–DM, que provee un marco estructurado y replicable para la implementación de soluciones analíticas. En la fase de comprensión del negocio, se definió el propósito central: orientar la focalización de programas de fortalecimiento empresarial hacia las empresas con mayor potencial de aportar a la generación de empleo en el corto plazo. En la fase de comprensión y preparación de datos, se integraron los Registros Mercantiles reportados entre 2020 y 2024 y se aplicaron procesos rigurosos de limpieza, transformación y enriquecimiento de variables. Posteriormente, la fase de modelado permitió construir y comparar distintas técnicas de segmentación y predictivas.

A partir de estos modelos, el proyecto no solo construye segmentos empresariales homogéneos mediante técnicas de clustering, permitiendo comprender perfiles estructurales, y sectoriales, sino que también le asigna una probabilidad de generación de empleo a estas empresas mediante modelos predictivos.

Con el fin de construir un insumo directo para la planeación institucional, se consolidó una Base Maestra de Priorización que contiene el listado de las empresas que han generado empleo durante por lo menos 2 de los últimos 5 años en la ciudad con su asignación a uno de los Clústers identificados y la probabilidad de que generen empleo en el año.

Este documento se estructura de la siguiente manera: primero, se presenta el contexto institucional y socioeconómico, con el planteamiento del problema y las preguntas analíticas a resolver, después el marco conceptual que fundamenta el análisis descriptivo, la segmentación y la modelación predictiva; posteriormente se describe la metodología aplicada; luego se exponen los resultados, incluyendo el perfil mayoritario de las empresas generadoras de empleo, los clústeres diferenciados y el desempeño de los modelos predictivos aplicados; y finalmente se presentan las conclusiones y las recomendaciones estratégicas que orientan la implementación institucional de esta solución en la Secretaría.

2. Contexto institucional y socioeconómico del proyecto

Santiago de Cali es la tercera ciudad más poblada de Colombia, una de las economías regionales más importantes y un eje estratégico de desarrollo nacional, no sólo por su ubicación geográfica privilegiada —en el valle del río Cauca, con conexiones clave hacia el Pacífico a través del puerto de Buenaventura— sino también por su robusta estructura empresarial y productiva que alimenta el crecimiento regional.

En 2024, el Producto Interno Bruto ("PIB") de Colombia alcanzó los 1.706.447 miles de millones de pesos colombianos, donde el Valle del Cauca contribuyó el 9,7%, destacando a Cali por sí sola con un aporte significativo del 8,7% del PIB nacional (DANE, 2024). Esta magnitud sitúa a Cali como un motor clave del crecimiento regional y nacional, lo que a su vez refuerza el valor estratégico de articular políticas públicas de fortalecimiento empresarial e innovación para amplificar su impacto.

A su vez, Cali ha enfrentado históricamente importantes desafíos en materia social. En los últimos años, cerca del 23,6 % —aproximadamente 535.000 habitantes— se encuentra en condición de pobreza monetaria y alrededor del 7,8 % de su población —unas 178.000 personas— vive en pobreza extrema (El País, 2024; Álban Comunes, 2024). Este panorama refleja una realidad compleja marcada, entre otros factores, por los flujos continuos de desplazamiento forzado provenientes del Pacífico y del Cauca, regiones profundamente afectadas por la violencia y la desigualdad.

Miles de familias llegan a Cali en búsqueda de protección, oportunidades y un nuevo comienzo, pero lo hacen enfrentando barreras estructurales de acceso al empleo formal, educación, vivienda y seguridad alimentaria. A ello se suma una tasa de desempleo urbano que

en 2024 alcanzó el 12,6 %, y que históricamente golpea con mayor fuerza a mujeres, jóvenes y población migrante (DANE, 2024).

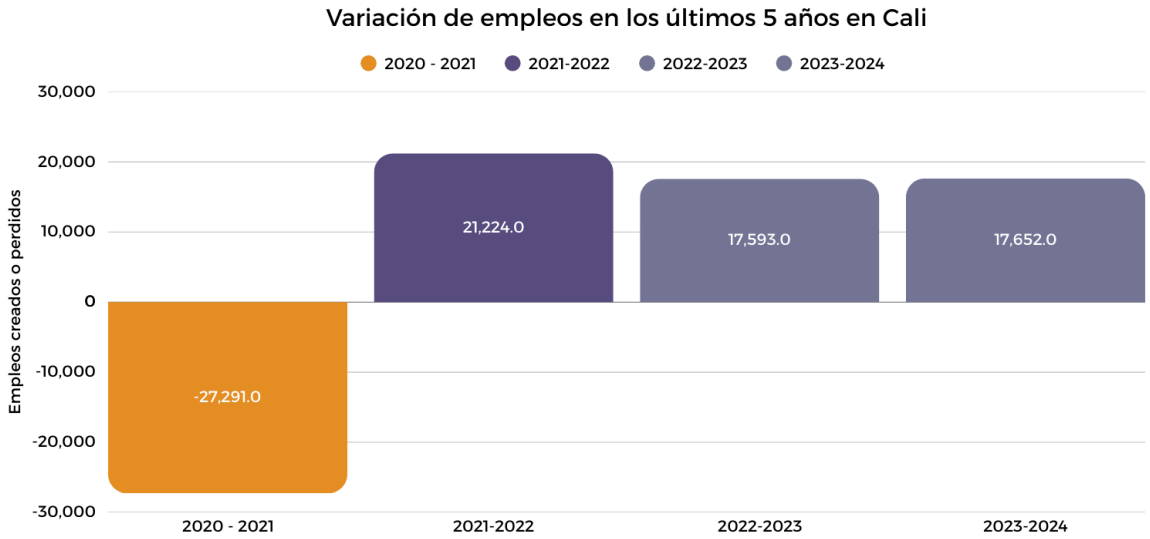
Teniendo en cuenta este contexto, la administración distrital, liderada por el alcalde Alejandro Eder, ha orientado su estrategia de desarrollo hacia un enfoque que reconoce el papel central del sector privado en la dinamización económica. Este enfoque, alineado con corrientes contemporáneas que resaltan la importancia de entornos institucionales favorables —caracterizados por reglas claras, protección de derechos de propiedad y estabilidad macroeconómica— sostiene que la inversión privada, la creación de empresas y la expansión del empleo formal constituyen motores fundamentales del progreso social (Clausen, 1983).

En consecuencia, la generación de empleo formal y digno se ha convertido en uno de los indicadores prioritarios para la administración, al considerarse el mecanismo más efectivo para reducir brechas socioeconómicas y promover movilidad social. En esta línea, el Plan Distrital de Desarrollo *Cali Capital Pacífica de Colombia,* establece como meta la creación de 100.000 nuevos empleos formales durante el cuatrienio 2024 a 2027 , propósito que se erige como eje orientador de administración pública.

La variación del empleo en Cali durante los últimos cinco años permite comprender la magnitud del desafío que enfrenta la administración distrital al trazarse dicha meta. Tal como lo evidencia la Figura 1, la ciudad experimentó entre 2020 y 2021 una pérdida neta de 27.291 empleos, reflejo directo de los efectos económicos de la pandemia y de la fragilidad estructural del mercado laboral local.

Figura 1

Serie histórica: variación en la generación de empleos en los últimos 5 años en Cali.



Fuente: Elaboración propia.

Posteriormente, entre 2021 y 2022, se observó un rebote significativo con la creación de 21.224 empleos, impulsado por la reapertura económica. Sin embargo, en los ciclos siguientes 2022–2023 y 2023–2024 el crecimiento se estabilizó en torno a 17.500 empleos anuales, evidenciando que la capacidad histórica de la economía caleña para generar nuevos puestos de trabajo se sitúa por debajo de los 20.000 empleos por año.

En este contexto, la meta establecida por la administración actual de crear 25.000 empleos formales al año, para un total de 100.000 empleos durante el cuatrienio, constituye un objetivo tan ambicioso como estratégico. Pues, alcanzar dicha meta supone superar de manera sostenida el comportamiento histórico reciente del mercado laboral: la ciudad no ha generado 25.000 empleos nuevos en ningún año de los últimos cinco, ni siquiera en el periodo de mayor recuperación económica (2021–2022). Lograr este volumen de empleo implica un incremento del 40 % frente al ritmo promedio de generación observado entre 2022 y 2024.

En consecuencia, este desafío se convierte en un mandato transversal para todas las dependencias del gobierno distrital; sin embargo, tal como lo establece el Decreto 0516 de 2016, la responsabilidad principal recae en la Secretaría de Desarrollo Económico, entidad encargada de liderar, articular y asegurar el cumplimiento de esta misión.

Su rol es decisivo, pues entre sus funciones misionales se encuentran: “determinar y articular las estrategias de mercadeo y promoción para el posicionamiento del Municipio de Santiago de Cali”; “liderar el fortalecimiento del emprendimiento, las mipymes y las grandes empresas”; y “determinar estrategias para incentivar el empleo, el emprendimiento y la generación de ingresos en el Municipio de Santiago de Cali”. Estas competencias convierten a la Secretaría en el actor institucional clave para traducir la visión estratégica del gobierno en políticas, programas e intervenciones concretas que impacten el tejido productivo y promuevan la generación de empleo formal en la ciudad.

La Secretaría de Desarrollo Económico, consciente de su responsabilidad estratégica en la transformación productiva de la ciudad, ha puesto en marcha diversas acciones que se estructuran en dos grandes líneas: el posicionamiento de Cali a nivel nacional e internacional y el fortalecimiento del empresariado local. Ambas dimensiones se conciben como complementarias y necesarias para ampliar las oportunidades económicas, atraer inversión y avanzar en la meta distrital de generar empleo formal y sostenible.

En cuanto al posicionamiento de ciudad, la administración parte de la premisa de que los grandes eventos internacionales, el turismo interno y el turismo extranjero funcionan como dinamizadores directos e indirectos de la economía local, al estimular la demanda de bienes y servicios, activar sectores conexos y consolidar confianza en el territorio como destino atractivo

para invertir. Bajo esta lógica, Cali ha sido sede de eventos de alto impacto como la COP16, la Maratón de Cali, la FIFA Women's World Cup Fan Festival, y la Macrorrueda de Negocios de ProColombia, entre otros. Todos estos escenarios se integran bajo la narrativa —"Cali es donde debes estar"—, una marca territorial diseñada para reforzar la percepción de Cali como un nodo urbano vibrante, competitivo y con capacidad institucional para acoger eventos globales y oportunidades de negocio.

No obstante, es evidente que la generación de empleo asociada a eventos esporádicos conlleva riesgos en términos de sostenibilidad, dado que se trata, en muchos casos, de ocupaciones temporales que no necesariamente logran transformar de manera estructural las dinámicas sociales y económicas de la ciudad. Precisamente por ello, el fortalecimiento del sector empresarial adquiere una relevancia central. Es esta dimensión la que permite pasar de impactos coyunturales a transformaciones duraderas, capaces de sostener el empleo formal, incrementar la productividad y consolidar un tejido económico resiliente en el largo plazo.

Con este propósito, la Secretaría de Desarrollo Económico ha impulsado un conjunto de iniciativas orientadas a fortalecer el tejido empresarial local y promover su crecimiento competitivo. Entre los programas e incentivos emblemáticos se destacan "Mi Primera Exportación", los alivios tributarios en el Impuesto Predial Unificado y el Impuesto de Industria y Comercio, así como las estrategias "Conexiones de Valor Empresarial", "BI ON" y los "Nodos de Innovación". Todas estas acciones buscan aumentar la productividad, facilitar el acceso a mercados nacionales e internacionales y consolidar un ecosistema empresarial más sofisticado y resiliente.

Uno de los avances más significativos ha sido la consolidación de la estrategia Conexiones de Valor Empresarial, con la cual se han gestionado más de 900 conexiones de valor para mipymes de sectores como servicios, industrias creativas, manufactura, comercio y turismo, articulando a 56 organizaciones públicas y privadas en un ecosistema de apoyo empresarial. Asimismo, la Secretaría ha acompañado a 1.153 mipymes en la construcción de vínculos comerciales con empresas ancla, impulsando encadenamientos productivos que han fortalecido ingresos, competitividad y acceso a nuevos mercados. A esto se suma la creación de 32 espacios de innovación social que dieron origen a los Nodos de Innovación, donde 68 organizaciones han cocreado soluciones para sectores estratégicos como BPO–KPO–ITO[1], gastronomía, hotelería, logística y economía verde.

Con el desarrollo de las iniciativas y programas enunciados anteriormente, la Secretaría de Desarrollo Económico ha demostrado un trabajo juicioso, consistente y orientado a fortalecer el tejido empresarial de la ciudad. Sus iniciativas reflejan un entendimiento claro de las necesidades del sector productivo y una articulación efectiva con los gremios de la ciudad. Este conjunto de esfuerzos constituye una base institucional robusta que, sin duda, ha contribuido al fortalecimiento del empresariado caleño.

No obstante, es importante señalar que la formulación de estos programas y la definición de sus poblaciones objetivo no siempre se estructuran ni se evalúan desde la perspectiva de su potencial para generar empleo formal tanto en el diseño de los programas como en la selección y priorización de las empresas beneficiarias. Integrar este enfoque permitiría alinear aún mejor la

[1] Los sectores BPO (Business Process Outsourcing), KPO (Knowledge Process Outsourcing) e ITO (Information Technology Outsourcing) corresponden a industrias de servicios basadas en conocimiento que externalizan procesos empresariales operativos, analíticos y tecnológicos, respectivamente. Estos sectores se caracterizan por su capacidad de generar empleo formal —particularmente para población joven y talento especializado— y por impulsar la competitividad e innovación mediante servicios como atención al cliente, análisis de datos, consultoría, desarrollo de software y soluciones digitales.

oferta institucional y sus esfuerzos con la apuesta del gobierno distrital, optimizando la eficiencia de los recursos, potenciando el impacto de cada intervención y favoreciendo una toma de decisiones basada en evidencia que acelere el cumplimiento de la meta de generación de empleo formal para la ciudad.

3. Planteamiento del problema

La ciudad de Santiago de Cali enfrenta uno de los retos más ambiciosos de la administración distrital: la creación de 100.000 empleos formales durante el cuatrienio. Alcanzar esta meta exige una comprensión profunda del comportamiento del tejido empresarial local y de los factores que determinan su capacidad real de generar empleo. Este desafío se hace aún más evidente cuando se observa que, para 2024, Cali registró aproximadamente 100.000 empresas, distribuidas en sectores heterogéneos en tamaño, productividad, sector económico, entre otros. En contraste, la Secretaría de Desarrollo Económico, entidad encargada de liderar esta misión, opera con un presupuesto anual cercano a los 68 mil millones de pesos, de los cuales solo alrededor del 26 % se destina a iniciativas de fortalecimiento empresarial.

Este contexto plantea una restricción evidente: los recursos disponibles no son suficientes para intervenir de manera homogénea en todas las empresas de la ciudad, lo cual obliga a priorizar con precisión quirúrgica. Bajo estas condiciones, se vuelve imprescindible que la Secretaría oriente sus esfuerzos hacia aquellas empresas y sectores que realmente “mueven la aguja” en la generación de empleo formal.

Además, en la práctica institucional, las iniciativas de fortalecimiento empresarial se diseñan y ejecutan con criterios sectoriales, de innovación, de competitividad o de tamaño empresarial —especialmente mipymes—, pero no necesariamente incluyen un análisis

estadístico de grupos o predictivo de generación de empleo como eje de focalización. Esto significa que programas, incentivos o cupos de participación pueden asignarse de forma equitativa entre empresas diversamente clasificadas, sin considerar si su aporte potencial al indicador de empleabilidad será marginal o sustancial. En un escenario de recursos limitados, esta ausencia de priorización basada en evidencia reduce la eficiencia del gasto público de cara a la meta propuesta e impide capturar oportunidades estratégicas para cumplirla.

Por ello, se evidencia una necesidad crítica: desarrollar un análisis riguroso, sustentado en técnicas de ciencia de datos, que permita comprender cómo se comportan las empresas que generan empleo, identificar grupos o perfiles empresariales que han sido constantes en la generación de empleos, y estimar la probabilidad de que cada empresa genere empleo en el año siguiente. Con esta información, la Secretaría de Desarrollo Económico podría orientar esfuerzos, focalizar programas, priorizar cupos y diseñar intervenciones diferenciales basadas en evidencia, aumentando la eficacia institucional y contribuyendo de manera decisiva al logro de la meta de empleo formal en Cali. Este proyecto surge, entonces, de la necesidad de construir herramientas analítica que permita convertir grandes volúmenes de información empresarial en decisiones de política pública más precisas, eficientes y estratégicas.

4. Hipótesis del proyecto

La formulación de este proyecto parte de la necesidad de que la Secretaría de Desarrollo Económico oriente sus recursos hacia las empresas con mayor potencial de generar empleo formal en Santiago de Cali. Dado que el tejido empresarial es amplio y diverso, y que los recursos y el tiempo son limitados, resulta esencial identificar empresas, patrones y segmentaciones que permitan focalizar la intervención pública. Este enfoque es consistente con

la literatura, que demuestra que la generación de empleo no es homogénea y depende de características empresariales específicas (Haltiwanger, Jarmin & Miranda, 2013).

Con base en lo anterior, la finalidad de este proyecto parte de una Hipótesis General que se comparte a continuación:

> **Hipótesis general (HG):** "Si la Secretaría de Desarrollo Económico focaliza sus esfuerzos, programas y recursos teniendo en cuenta las características empresariales que comparten las empresas que históricamente generan empleo, los grupos o clústeres existentes entre dichas empresas y las unidades productivas que presentan una mayor probabilidad de generar empleo en el periodo siguiente, entonces se facilitará la expansión del empleo formal en la ciudad y se incrementará la eficacia institucional para avanzar hacia la meta de crear 100.000 nuevos empleos durante el cuatrienio."

Esta Hipótesis General orienta la necesidad y pertinencia del desarrollo del presente proyecto y esboza el espíritu del mismo. Esta hipótesis se presenta además como una posible hipótesis a evaluar en un ejercicio de análisis de impacto que pueda desarrollar de manera posterior la Secretaría. Lo cual da paso a presentar la hipótesis que será abordada en el desarrollo del presente proyecto:

> **Hipótesis analítica (HA):** "Existen variables empresariales —operativas, financieras, sectoriales y demográficas— que explican en mayor medida la generación de empleo formal, y estas permiten identificar grupos empresariales mediante técnicas de clusterización y estimar la probabilidad de contratación futura mediante modelos predictivos."

La validación de esta hipótesis orientará la comprensión de los determinantes del empleo empresarial y servirá como base para diseñar estrategias de focalización institucional más precisas. A partir de estas hipótesis se formulan las preguntas analíticas que guiarán el desarrollo metodológico del proyecto y permitirán evaluar empíricamente la afirmación aquí planteada.

5. Preguntas analíticas

A partir del problema planteado, se formulan las siguientes preguntas analíticas que orientarán el desarrollo de este proyecto:

¿Qué características empresariales comparten mayoritariamente las empresas que han logrado generar empleo formal en los años recientes en Santiago de Cali?

¿Qué grupos o perfiles empresariales pueden identificarse dentro del conjunto de empresas que generan empleo formal?

¿Cuál es la probabilidad de que una empresa incremente su número de empleados en el año siguiente a partir de las características empresariales que presenta en el periodo actual?

Las preguntas anteriores orientan el análisis cuantitativo y la metodología empleada en este proyecto, al delimitar los fenómenos empresariales que requieren ser comprendidos y modelados. A partir de ellas, se formulan los objetivos que guían el desarrollo de este proyecto, el diseño de los métodos estadísticos, las técnicas de segmentación y los modelos predictivos necesarios para potenciar la generación de empleo formal en la ciudad.

6. Objetivos

6.1 Objetivo general

Orientar la toma de decisiones y la focalización de los programas de fortalecimiento empresarial de la Secretaría de Desarrollo Económico hacia las empresas con mayor potencial de contribuir a la meta distrital de creación de 100.000 nuevos empleos entre el 2024 y el 2027, mediante el desarrollo de soluciones analíticas de carácter descriptivo, de segmentación y predictivo, aplicadas bajo la metodología CRISP–DM.

6.7 Objetivos específicos

1. Identificar agrupaciones de empresas con comportamientos similares en materia de generación de empleo, con el fin de reconocer segmentos prioritarios que orienten estrategias específicas de fortalecimiento empresarial.

2. Implementar modelos predictivos supervisados para estimar la probabilidad de que una empresa genere empleo en el año siguiente, fortaleciendo la capacidad institucional de priorización y toma de decisiones basada en evidencia.

3. Consolidar una Base Maestra de Priorización que contenga las empresas activas registradas en el 2024 que reportaron generación de empleo en por lo menos 2 de los 5 años observados en el proyecto, así como su pertenencia a los Clústers identificados y su probabilidad de generar empleo para el año siguiente.

7. Alcance y limitaciones

7.1 Alcance del Proyecto

El presente proyecto contempla el siguiente alcance metodológico y analítico:

Propósito analítico integral

Desarrollar un conjunto de soluciones analíticas de carácter descriptivo, de segmentación y predictivo que permitan:

- identificar las características empresariales que comparten en mayor medida las empresas asociadas a la generación de empleo formal en Santiago de Cali en los últimos 5 años,
- reconocer grupos de empresas con comportamientos similares mediante técnicas de clusterización,
- y estimar la probabilidad de que una empresa incremente su planta de personal en el año siguiente.

El propósito final es orientar la toma de decisiones y la focalización de programas de fortalecimiento empresarial hacia las empresas con mayor potencial de contribuir a la meta distrital de generar 100.000 nuevos empleos formales durante el cuatrienio.

Población objetivo

El estudio se concentra en empresas registradas en Santiago de Cali en las bases oficiales suministradas por la Secretaría de Desarrollo Económico que tienen como fuente principal el Registro Mercantil que se encuentra bajo dominio de la Cámara de Comercio de Cali. Se consideran principalmente empresas formales con información disponible entre los años 2020 y 2024, independientemente de su tamaño o sector económico.

Fuentes y estructura de información

El proyecto se sustenta en dos conjuntos de datos que son construidos durante el desarrollo del proyecto:

- **Dataset A (descriptivo–segmentativo):** Empresas activas en 2024 integradas con su comportamiento histórico de generación de empleo (2020–2023) y con variables operativas, financieras, demográficas y sectoriales reportadas en el 2024.
- **Dataset B (predictivo):** Base de datos longitudinal (panel) que contiene información anual por empresa concatenada verticalmente, utilizada para modelar la contratación futura.

Metodología aplicada

Se emplea el enfoque CRISP-DM, con técnicas de:

- análisis estadístico descriptivo,
- clusterización no supervisada para identificar perfiles empresariales,
- y modelos predictivos supervisados para estimar la probabilidad de generación de empleo.

Entregables principales

El proyecto contempla como entregables:

1. La identificación de grupos o clústeres empresariales que generan empleos en Cali con comportamientos diferenciados, que puedan orientar la intervención estratégica de estos grupos.
2. Un modelo predictivo validado para estimar la probabilidad de generación de empleo en el año siguiente, en formato pkl.

3. Una Base Maestra de Priorización que contenga las empresas activas registradas en el 2024 que reportaron generación de empleo en por lo menos 2 de los 5 años observados en el proyecto, así como su pertenencia a los Clústers identificados y su probabilidad de generar empleo para el año siguiente.

7.2 Limitaciones del Proyecto

El estudio presenta una serie de limitaciones que deben considerarse en la interpretación de los resultados:

Limitaciones asociadas a los datos

- Disponibilidad y completitud: Algunas variables empresariales no se encuentran disponibles para todos los años o para todas las empresas, lo cual puede restringir el número de observaciones útiles en los modelos.
- Calidad y consistencia: Registros incompletos, diferencias en criterios de captura entre años o ausencia de registros financieros pueden afectar el desempeño del modelo y la precisión de los clústeres.
- Granularidad temporal: El uso de datos de periodicidad anual impide capturar dinámicas intra-anuales de contratación, expansión o contracción empresarial, que podrían contener señales relevantes.
- Representatividad sectorial: Aunque se incluye el universo de empresas activas en Cali con información disponible, algunos sectores pueden estar subrepresentados debido a brechas naturales de información como la informalidad empresarial.

Limitaciones asociadas a los modelos

- Sensibilidad a decisiones técnicas: La clusterización depende de decisiones como número de grupos, métricas de distancia y variables seleccionadas. Esto implica supuestos metodológicos que pueden modificar la segmentación final.
- Riesgo de sobreajuste (overfitting): Los modelos predictivos requieren validación rigurosa para asegurar que la probabilidad estimada generalice adecuadamente a años futuros.

Restricciones institucionales para la aplicación del modelo

- La aplicación del modelo predictivo a nuevas bases que reciba la Secretaría, requiere de un proceso de transformación, el cual depende exclusivamente de la voluntad de los equipos.
- Las recomendaciones derivadas del análisis deben complementar, no reemplazar, el criterio técnico de los equipos de desarrollo económico.

Aún con estas limitaciones, este proyecto representa una oportunidad única de aplicar el poder de los datos al servicio del desarrollo local. Al incluir criterios que priorización que se enfoquen en la generación de empleo la administración podrá tomar decisiones más informadas, eficientes y equitativas, con un impacto directo sobre el empleo formal, la competitividad y la calidad de vida de los ciudadanos.

8. Marco Teórico

Esta sección presenta los fundamentos teóricos, conceptuales y metodológicos que orientan el desarrollo de este proyecto. La articulación conceptual presentada a continuación no solo permite comprender los pilares que soportan la estrategia analítica, sino que establece la

coherencia metodológica necesaria para interpretar los resultados del análisis y su utilidad para la política pública.

8.1 Técnicas utilizadas

Analítica descriptiva

La analítica descriptiva constituye la primera fase del proceso CRISP–DM y su objetivo es comprender el comportamiento histórico de las variables relevantes mediante estadísticas univariadas, bivariadas y multivariadas. Este tipo de análisis permite identificar patrones, distribuciones, tendencias y posibles relaciones entre variables (Han, Kamber & Pei, 2012).

En este proyecto, la analítica descriptiva permitió caracterizar las empresas que han sido constantes en la generación de empleo formal entre 2020 y 2024, analizando atributos financieros (activos, pasivos, rentabilidad), demográficos (antigüedad), sectoriales (CIIU) y territoriales (comunas).

Segmentación empresarial

La segmentación empresarial busca agrupar empresas en conjuntos homogéneos según características similares. Este proceso es útil para la política pública, ya que permite distinguir grupos de empresas con necesidades diferenciadas (Dolničar, Grün & Leisch, 2018).

La heterogeneidad del tejido empresarial caleño hace pertinente identificar segmentos que reflejen comportamientos financieros y laborales. Estos segmentos se obtienen mediante técnicas de clusterización, que se describen a continuación.

Clustering y técnicas de aprendizaje no supervisado

La clusterización es un método de aprendizaje no supervisado que busca particionar un conjunto de datos en grupos basados en su similitud (Jain, 2010). El algoritmo elegido en este

proyecto fue K-Means, debido a su interpretabilidad y eficiencia computacional en comparación con el rendimiento de Agglomerative.

K-Means agrupa observaciones minimizando la distancia dentro de cada clúster:

$$\min_{C} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

donde Ck es el conjunto de observaciones del clúster k y μk su centroide.

Métricas de calidad del Clustering

Para garantizar la calidad de la segmentación, se utilizaron métricas internas:

- Índice Silhouette:

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

donde a(i) es la distancia promedio al propio clúster y b(i) la mínima distancia a otro clúster.

- Calinski–Harabasz:

$$CH = \frac{\mathrm{tr}(B_k)/(K-1)}{\mathrm{tr}(W_k)/(n-K)}$$

- Davies–Bouldin:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right)$$

Estas métricas guiaron la selección de K=4, lo que permitió identificar perfiles estratégicos para la Secretaría.

Empleo formal

El empleo formal se define como el número de trabajadores reportados por las empresas al registro mercantil, y es un indicador clave de desarrollo económico (OIT, 2018). En este proyecto se analizó mediante la variable de planta de personal y su variación anual. El objetivo final fue modelar la probabilidad de que una empresa genere empleo en el año siguiente.

Modelos predictivos supervisados

Los modelos predictivos supervisados estiman la ocurrencia de un evento futuro a partir de datos históricos (Hastie, Tibshirani & Friedman, 2009). En este caso, el evento es la generación de empleo en el año siguiente (variable binaria).

Los modelos utilizados fueron:

- *Regresión Logística*: Modelo lineal que estima la probabilidad de pertenecer a la clase positiva:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}}$$

- *Random Forest*: Algoritmo de ensamblaje basado en múltiples árboles de decisión construidos mediante bootstrap. Reduce varianza y captura relaciones no lineales.
- *XGBoost*: Método basado en gradient boosting que minimiza una función de pérdida aditiva:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i)$$

con cada *ft* optimizado secuencialmente.

Estos modelos permiten estimar probabilidades, generar rankings y evaluar patrones predictivos a nivel de empresa.

Métodos de Evaluación de Modelos de Machine Learning

La evaluación rigurosa del desempeño de los modelos de ML es crucial para seleccionar el mejor candidato y entender su capacidad de generalización a datos no vistos.

Validación cruzada

La validación cruzada estratificada evalúa la capacidad de generalizar del modelo. Consiste en dividir los datos en k particiones:

$$CV = \frac{1}{k}\sum_{i=1}^{k} f_i$$

donde *fi* es la métrica del pliegue *i*.

Permite evitar sobreajuste y seleccionar el modelo con mejor desempeño.

Matriz de Confusión

Tabla que visualiza el rendimiento detallando Verdaderos Positivos (TP), Verdaderos Negativos (TN), Falsos Positivos (FP) y Falsos Negativos (FN) (Powers, 2011). Es fundamental para entender dónde falla el modelo. Para modelos de clasificación binaria se utilizan métricas derivadas de la matriz de confusión.

- *Accuracy:* Representa la proporción total de predicciones correctas sobre el conjunto completo de casos. Es útil para obtener una visión general del desempeño del modelo.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precisión:* Indica qué porcentaje de los casos clasificados como positivos realmente lo eran. Es clave para que la firma contacte únicamente a las empresas adecuadas y evite esfuerzos en organizaciones que no corresponden.

$$Precision = \frac{TP}{TP + FP}$$

- *Sensibilidad:* Expresa qué porcentaje de las empresas que realmente pertenecen a la clase positiva fueron detectadas por el modelo. Es esencial para no dejar pasar oportunidades valiosas.

$$Recall = \frac{TP}{TP + FN}$$

- *F1:* Útil como métrica única cuando se busca un balance entre Precision y Recall (Powers, 2011).

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- *ROC – AUC*: La curva ROC grafica el Recall (TPR) vs. la Tasa de Falsos Positivos (FPR = FP / (FP + TN)) a diferentes umbrales de clasificación. El AUC mide el área bajo esta curva, proporcionando una medida agregada del rendimiento del clasificador independientemente del umbral (Powers, 2011). Un AUC cercano a 1 indica un excelente poder discriminatorio.

$$AUC = \int_0^1 TPR(FPR)\, dFPR$$

SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE es una técnica para tratar desbalance de clases creando observaciones sintéticas de la clase minoritaria (Chawla et al., 2002). Su función es mejorar la capacidad del modelo de detectar la clase positiva:

$$x_{\text{new}} = x_i + \lambda(x_{zi} - x_i), \quad \lambda \in [0, 1]$$

Se aplicó antes del entrenamiento para balancear la probabilidad de predicción.

Feature Importance

La importancia de variables determina el grado en que cada variable contribuye al modelo. En árboles se calcula a partir de la reducción de impureza (Gini o entropía).

$$FI_j = \sum_{t \in splits(j)} \Delta impurity_t$$

Permite identificar las variables más influyentes: antigüedad, tamaño, activos, variación del empleo, rentabilidad, entre otras.

SHAP (SHapley Additive exPlanations)

SHAP explica la contribución marginal de cada variable a la predicción individual siguiendo los valores de Shapley:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(x) - f_S(x) \right]$$

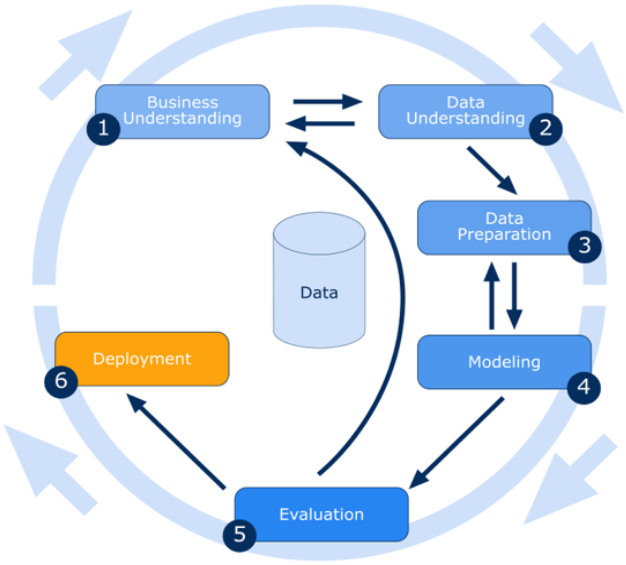
9. Metodología

Para el desarrollo del presente proyecto se ha elegido la metodología CRISP-DM (“Cross-Industry Standard Process for Data Mining”), ampliamente reconocida en la industria y la academia por el esquema de eficiencia que le aporta a proyectos que requieren de procesos de análisis de datos y modelado estadístico.

Esta metodología permite estructurar el proceso de análisis en fases claramente definidas, facilitando la comprensión, replicabilidad y trazabilidad del trabajo analítico. Uno de los aportes más relevantes de esta metodología consiste en entenderlo no como proceso lineal sino circular que consulta constantemente el entendimiento del negocio para realmente lograr soluciones que apunten a los KPIs (“Key Performance Indicators”).

Figura 2

Ilustración de modelo CRISP-DM



Fuente: Tomado de Roy (2018).

De esta manera CRISP-DM es una metodología adecuada para este tipo de análisis porque pone en el centro del proceso la comprensión del problema de negocio y la utilidad del modelo, permitiendo conectar los resultados del análisis con decisiones concretas. Además, permite trabajar de forma flexible y estructurada desde la exploración inicial hasta la implementación práctica de hallazgos, lo que garantiza una alineación constante con los objetivos institucionales de impacto.

Como parte del proceso metodológico, construimos un pipeline visual tipo LEGO para aterrizar de manera concreta las etapas del modelo CRISP-DM al contexto específico de este proyecto. Esta herramienta me permitió no solo organizar y estructurar las fases del análisis —desde el entendimiento del negocio hasta el desarrollo de los modelos y su posible implementación—, sino también visualizar con mayor claridad cómo cada paso debía responder directamente a los objetivos estratégicos definidos desde la Secretaría de Desarrollo Económico.

Figura 3

Actividad LEGO en el marco del desarrollo del proyecto empresarial

Fuente: Elaboración propia.

9.1 Fuentes de información utilizadas

Las fuentes de información empleadas en este proyecto corresponden exclusivamente los reportes de Registro Mercantil que deben realizar las empresas de forma anual ante la Cámara de Comercio de su jurisdicción, es este caso, la Cámara de Comercio de Cali. Estos registros son compartidos con la Secretaría de Desarrollo Económico mediante archivos xlsx año a año una vez el periodo de registro culmina —31 de diciembre—. En este caso contamos con las bases empresariales anuales de los años 2020, 2021, 2022, 2023 y 2024 compuestas inicialmente por registros no solo realizados por empresas ubicadas en la ciudad de Cali sino también en municipios aledaños como Jamundí, Yumbo o Dagua.

Las dimensiones de cada una se las bases se comparten a continuación:

Tabla 1

Dimensiones de las bases de Registro Mercantil anualizadas.

Año	Registros	Columnas
2020	91.635	16
2021	97.023	21
2022	109.845	14
2023	102.377	18
2024	199.053	25

Fuente: Elaboración propia.

Estas bases constituyen registros administrativos oficiales que contienen información detallada por empresa, incluyendo atributos operativos, demográficos, financieros entre otros fundamentales para el análisis. En este punto se resalta un primer reto, las bases no cuentan con la misma estructura, e incluso muchas de las variables a pesar de que representan lo mismo no cuentan con nombres unificados en todas las bases, lo cual exigió realizar un análisis del contenido de cada variable para asignar nombres unificados. A continuación se encuentra un diccionario de las variables que fueron identificadas en las bases:

Tabla 2

Variables identificadas en las fuentes de información utilizadas.

Variable	Entendimiento
MATRICULA	Se refiere al número de matrícula con el que se registra la empresa.
ENTE JURIDICO	Indica el tipo de sociedad o persona natural.

TIPO REGISTRO	Indica si es mercantil o si es una entidad sin ánimo de lucro (ESAL).
FECHA_MATRICULA	Mes, Día, Año de Indica la fecha en la que se realizó la matricula por primera vez.
FECHA_RENOVACION	Mes, Día, Año de Indica la fecha en la que se realizó la renovación de la matricula en el periodo de la base.
FECHA_CANCELACION	Mes, Día, Año de Indica la fecha en la que operó la cancelación de la matricula en caso de que haya sucedido.
NIT	Es el identificador de una empresa en Colombia
DIGITO CHEQUEO	Es el número de seguridad que acompaña al NIT de la empresa
RAZON SOCIAL	Es el nombre con el que se constituye una empresa y aparece en los documentos públicos y privados de constitución de la empresa.
REP_LEGAL	Es la persona natural o jurídica facultada para actuar en nombre y representación de otra persona natural o jurídica, ante terceros y las autoridades
DIRECCION_COMERCIAL	Indica la dirección geográfica anclada a la empresa
EMAIL	Indica el correo electrónico que se encuentra registrado para la empresa ante la Cámara de Comercio

COMUNA	Indica a la comuna en la cual se ubica la dirección comercial de la empresa en la ciudad de Cali
CODIGO CIIU	Es la Clasificación Industrial Internacional Uniforme (CIIU) que facilita la clasificación de la actividad económica principal de una empresa.
DESC CIIU	Es la descripción en texto del código en cuestión.
SECTOR	Indica el sector industrial al que pertenece la empresa de acuerdo con su actividad económica CIIU.
TAMANO957	Clasifica la empresa por su tamaño de acuerdo con la Ley 590 de 2000
DESC_TAMANO	Clasifica la empresa por su tamaño de acuerdo con el Decreto 957 de 2019.
COD_ESTADO	Indica si el estado de la matricula es activo o cancelado.
ACTIVOS_LIQUIDADOS	Reporta los activos que la empresa ha vendido o convertido en efectivo.
TOT_ACTIVOS	Es la suma de todos los bienes y derechos que tiene la empresa.
TOT_PASIVO	Es la suma de todas las deudas y obligaciones que tiene la empresa.
INGRESOS_OPERACIONALES	Son los ingresos generados por la actividad principal de la empresa.
UTILILIDAD_BRUTA	Representa la ganancia después de restar los costos directos de producción a los ingresos operacionales

UTILILIDAD_PERDIDA	Muestra si la empresa ganó o perdió dinero después de restar todos sus costos, gastos e impuestos a los ingresos.
PATRIMONIO	Es el valor residual de la resta de los pasivos a los activos de la empresa.
NUMERO DE TRABAJADORES	Indica el número de trabajadores vinculados a la empresa al momento de realizar la matricula o renovación.
CANT. ESTABLECIMIENTOS	Indica la cantidad de establecimientos físicos que tiene la empresa en general.

Fuente: Elaboración propia.

9.2 Decisiones estratégicas de procesamiento

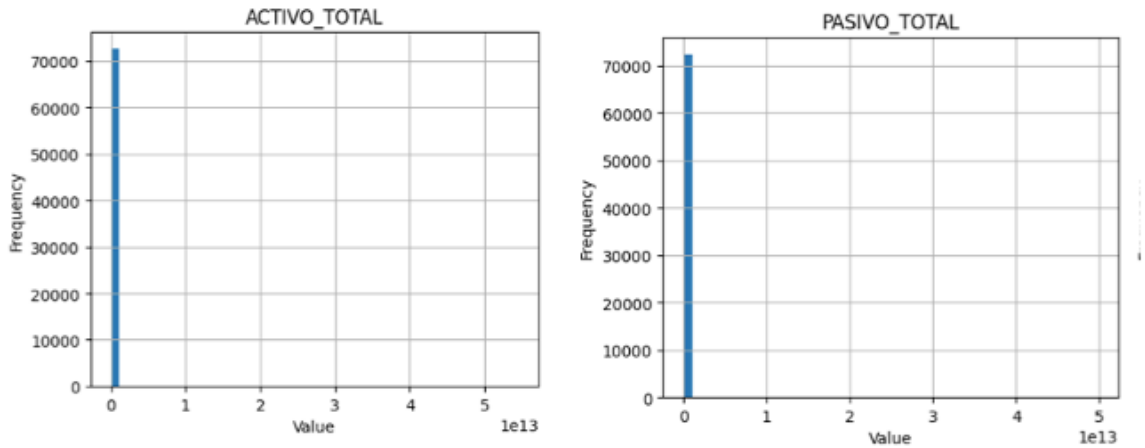
El preprocesamiento de datos constituyó una fase esencial para asegurar la calidad del análisis. Como primera medida de filtrado se decidió eliminar los registros de las empresas que reportaban una ubicación diferente a Cali, lo cual redujo el número de registros de las bases en alrededor en un 4.8%. En el manejo de valores faltantes o nulos se adoptó un enfoque diferenciado según el tipo de variable. Para las variables categóricas, se asignaron categorías como “Otro” cuando la ausencia era estructural; para las numéricas de tipo financiera, se imputaron valores en 0 buscando tener una radiografía realista de los reportes empresariales.

En este punto se identificó una necesidad muy relevante para el desarrollo de los análisis que constituyó una verdadera decisión estratégica de procesamiento. Durante la fase de preprocesamiento, el análisis exploratorio de las variables ACTIVO_TOTAL y PASIVO_TOTAL reveló, como se puede observar en la Figura 4, una distribución altamente sesgada hacia valores extremadamente bajos. La mayoría de los registros se concentraban en un

punto cercano a cero, mientras que un número muy reducido de empresas presentaba valores significativamente superiores. Este fenómeno produjo histogramas prácticamente “colapsados” en el extremo izquierdo, lo cual evidenció un universo empresarial con una proporción elevada de microempresas o unidades productivas con activos mínimos, y una distribución marcada por gran dispersión y elevada asimetría.

Figura 4

Histogramas de variables financieras en las bases crudas.



Fuente: Elaboración propia.

A partir de esta evidencia visual y estadística, se identificó la necesidad de establecer un punto de corte mínimo de activos que permitiera garantizar que las empresas incluidas en el análisis tuvieran un nivel operativo real que participara en el proceso de generación de empleo formal y no respondiera a una dinámica de creación de empresas o personas jurídicas “de papel”. Para ello utilizamos la base del 2024 para filtrar solo las empresas que hubieran reportado INGRESOS_OPERATIVOS diferentes de 0.

Al analizar el nuevo comportamiento de los activos en la Figura 5, identificamos que establecer un punto de corte de $10 millones de pesos en activos se justifica tanto estadística como económicamente. Desde el punto de vista estadístico, dicho umbral se sitúa ligeramente por encima del primer cuartil (Q1 = $8 millones), lo que permite excluir el 25 % de empresas con menor capitalización, caracterizadas por activos muy cercanos a cero, pero conserva alrededor del 75 % de las empresas con capacidad patrimonial relativamente mayor.

Figura 5

Comportamiento de la variable TOT_ACTIVOS para empresas que registran utilidades.

	TOT_ACTIVOS
count	20,446
mean	8,221,177,980
std	388,777,163,750
min	0
25%	8,000,000
50%	127,001,028
75%	1,036,604,941
max	54,521,486,000,000

Fuente: Elaboración propia.

Desde la perspectiva económica, estas unidades con activos inferiores a $10 millones se asemejan a los microestablecimientos de subsistencia descritos en la literatura latinoamericana, los cuales operan con capital muy reducido y tienen un aporte marginal al empleo formal (Pagés & Madrigal, 2019). Una vez consultado con los equipo operativos y estratégicos de la Secretaría se concluyó que mantenerlas en el análisis no solo introduciría ruido y podría distorsionar tanto la identificación de patrones como los resultados de la clusterización y de los modelos

predictivos, sino que también permitiría que los modelos aprendan de empresas de papel que realmente no cuentan con operaciones reales que impacten los indicadores de empleabilidad.

Por ello, el umbral de $10 millones se adopta como un criterio conservador pero razonable para garantizar que la muestra analizada esté compuesta por empresas con un mínimo de escala económica y una capacidad real de generación de empleo. De esta forma buscamos reflejar la dinámica real del universo empresarial caleño a la vez que buscamos enfocarnos solo en aquellas empresas que podemos denominar efectivamente “activas”.

En cuanto a registros que imposibilitaban la identificación de la empresa, como por ejemplo, valores en la columna NIT duplicados o en 0 fueron excluidos de forma estricta. Adicionalmente, se aplicaron transformaciones logarítmicas log1p y winsorización.

Las variables numéricas también fueron estandarizadas mediante StandardScaler cuando formaban parte del pipeline predictivo de modelos que así lo requerían, mientras que las variables categóricas fueron codificadas a través de One Hot Encoding, garantizando que los modelos pudieran interpretarlas correctamente.

9.3 Procesos generales de ingeniería de variables

La ingeniería de características siguió criterios de relevancia, coherencia temporal, parsimonia e interpretabilidad. En particular se crearon las siguientes variables en cada una de las bases anuales:

- ANTIGÜEDAD: Se creó con el propósito de analizar la madurez de las empresas. Se calculó con la variable “Mes, Día, Año de FECHA_MATRICULA” obteniendo únicamente el año y calculándolo con el año actual “2025”.

- GEN_REP: Indica el género del Representante Legal de la empresa, este proceso fue realizado mediante un análisis de los nombres reportados en las bases, contrastados con nombres comunes por cada género. Los valores contenidos en esta variable son F para Femenino, M para Masculino o No indica cuando el nombre no facilitó la identificación del género.
- COMUNA_CLUS: Cali cuenta con 22 comunas oficiales y 15 corregimientos, sin embargo al ser esta una variable categórica, resulta relevante agrupar las comunas en la medida en la que compartan comportamientos sociodemográficos cercanos. Para lo cual identificamos por cada comuna la Tasa de Ocupación Laboral, la Tasa de Desempleo, población y el estrato de moda en la comuna.

Se realizó una normalización de las variables cuantitativas utilizadas para garantizar que todas tuvieran la misma escala y peso dentro del análisis, dado que algunas se encontraban en porcentajes y otras en valores absolutos. Con estos datos estandarizados, se aplicó un modelo de clusterización K-means con cuatro grupos, lo que permitió clasificar las comunas según su similitud en condiciones laborales y socioeconómicas, identificando patrones que agrupan comunas con características favorables, intermedias o menos favorables en materia de empleo y estructura social. De esta forma los clúster construidos se comportan de la siguiente forma:

Tabla 3

Clústeres de comunas y corregimientos.

Clúster	Comunas	Descripción del Clúster
0	2, 3 17 y 19	Alta ocupación, bajo desempleo, alta población y estrato medio-alto

1	1, 4, 6, 13, 14, 15, 20 y 21	Baja ocupación, alto desempleo, baja población y estrato bajo
2	22	Muy alta ocupación, casi sin desempleo, estrato alto
3	5, 7, 8, 9, 10, 11, 12, 16 y 18	Media ocupación, medio desempleo, población moderada, estrato medio
4	18 Corregimientos de Cali	Corregimientos

Fuente: Elaboración propia.

- ESTABLECIMIENTOS_CAT: Dada la alta concentración de valores cercanos a 1 en la variable de ESTABLECIMEINTOS se decidió convertir esta variable a categórica creando tres categorías “1”, “2 – 5” y “6+”.
- CATEGORIA_EMPLEADOS: Esta variable se crea de igual forma identificando la gran cantidad de empresas que tienen en su mayoría pocos empleados lo cual podemos identificar mediante los cuartiles que se identifican en la Figura 6 . Teniendo en cuanta esto se creó esta variable de tipo categórica con las siguientes categorías: “<=3”, “<=10”, “<=50” y “+50”.

Figura 6

Métricas variable de empleados reportados por la empresa.

	EMPLEADOS_T
count	72,528.00
mean	19.98
std	195.39
min	1.00
25%	1.00
50%	3.00
75%	8.00
max	12,259.00

Fuente: Elaboración propia.

- DEUDA: La creación de la variable DEUDA, que toma el valor de 0 cuando la empresa no registra pasivos y 1 cuando sí los presenta, resulta fundamental para el análisis porque permite identificar de manera clara y directa la presencia de obligaciones financieras dentro de cada unidad empresarial.
- EMPLEADOS_T1: Refleja el número de empleados que tuvo una empresa en el año siguiente al registro es decir en el t + 1.
- VAR_E: Indica la variación en número de empleados que tuvo una empresa de un año determinado t al año posterior t +1. Esta variable es determinante para la construcción de la Variable Objetivo o Target.
- OBJ_: Esta es la variable objetivo de análisis, piedra angular en el desarrollo de los problemas analíticos. Dado que nuestro interés es centrar la atención en aquellas empresas que generaron empleos de un año a otro esta variable marca 1 cuando la variable VAR_E es mayor a 0 y 0 cuando es menor a 0.

9.4 Arquitectura analítica del proyecto

La arquitectura analítica se divide en dos pipelines complementarios. El Pipeline 1, basado en el Dataset A, integra la información de empresas activas en 2024 junto con su comportamiento de generación de empleo entre 2020 y 2023, información que se obtiene de la construcción de las variables EMPLEADOS_T1, VAR_E y OBJ_ año a año. Este pipeline se utiliza para realizar análisis descriptivos que permiten identificar patrones de comportamiento y posteriormente aplicar clusterización para reconocer grupos homogéneos de empresas que han demostrado en los últimos años participar activamente en la generación de empleo.

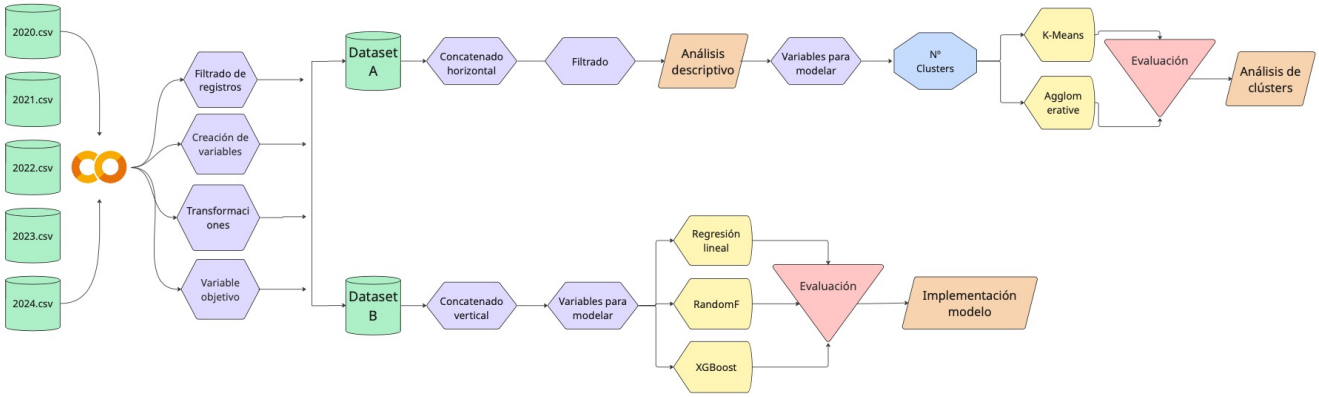
El Pipeline 2, basado en el Dataset B, está orientado a la predicción de generación de empleo de una empresa para el año siguiente basado en sus características empresariales actuales. Para ello, se construyó una base longitudinal con registros por empresa y por año, lo que permite entrenar modelos supervisados capaces de estimar la probabilidad de que una empresa genere empleo en el periodo siguiente.

Diagramas integrales de Pipelines

Los diagramas que representan visualmente cada pipeline —desde los datos crudos hasta la generación de resultados analíticos— se observan en la Figura 7. Estos diagramas describen los pasos de limpieza, transformación, ingeniería de variables y procesamiento necesarios para derivar tanto los Datasets como las fases subsiguientes de modelamiento, evaluación, y análisis facilitan lo cual facilita la comprensión del flujo metodológico de extremo a extremo.

Figura 7

Pipeline de metodología utilizada desde la obtención de fuentes crudas hasta la implementación de modelos estadísticos.



Fuente: Elaboración propia.

10. Pregunta analítica 1: ¿Qué características empresariales comparten mayoritariamente las empresas que han logrado generar empleo formal en los años recientes en Santiago de Cali?

10.1 Construcción de Dataset A:

Como se expuso en las secciones anteriores, una vez cada una de las bases anuales fueron sometida a las transformaciones generales, se seleccionó la base correspondiente al año 2024 con el propósito de obtener una "fotografía" actualizada del tejido empresarial caleño. Dado que uno de los objetivos centrales del proyecto consiste en analizar cómo se han comportado, en términos de generación de empleo, las empresas activas en el año más reciente, se procedió a integrar información histórica proveniente de las bases de años previos. Esta integración se realizó mediante un proceso de JOIN utilizando como llave única el NIT, con el fin de incorporar las variables EMPLEADOS_, VAR_E y OBJ_ para cada uno de los años disponibles.

A partir de estas variables se construyó una nueva medida denominada CREA_EMP, calculada como la suma de los valores de OBJ_, con el fin de identificar en cuántos años, dentro del periodo analizado, una empresa registró creación neta de empleo. Así mismo se construyó la medida VAR_TOT que buscar representar la variación en el número de empleos creados o perdidos durante los años observados.

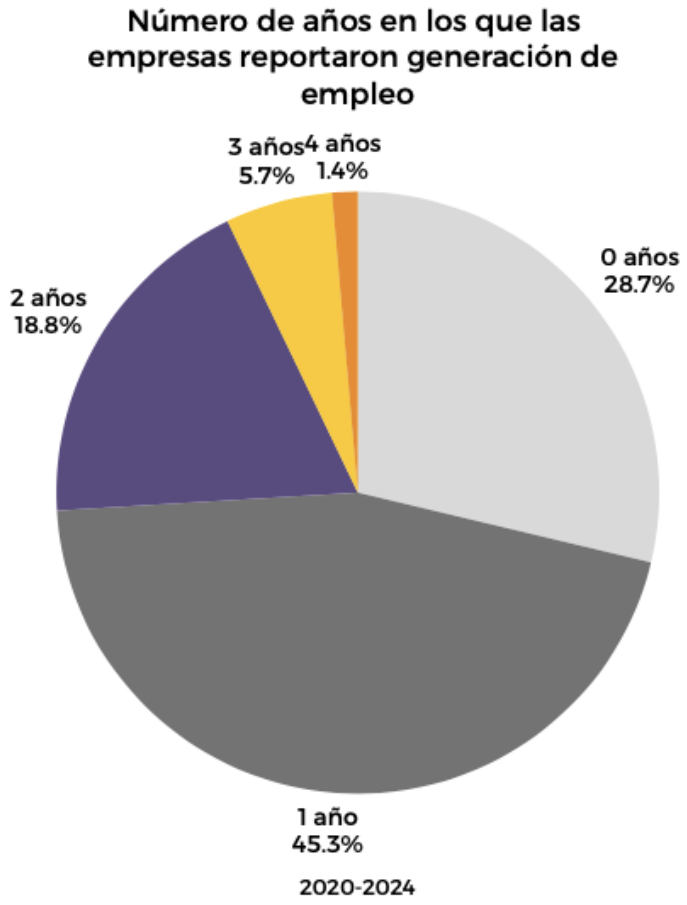
La base del año 2024 fue la única que incluyó la variable INGRESOS_OPERACIONALES, lo que permitió aplicar un filtro adicional orientado a garantizar que el Dataset A estuviera compuesto exclusivamente por empresas con actividad económica verificable. Para ello, se excluyeron las empresas que reportaban ingresos operacionales iguales a cero, lo que contribuyó a depurar unidades productivas sin evidencia de actividad económica reciente.

Como resultado, el Dataset A quedó conformado por las empresas activas que realizaron la creación o renovación de su registro mercantil en 2024, enriquecidas con información histórica sobre su comportamiento laboral. La estructura final de este Dataset comprendió 30.632 registros y 43 variables.

Dado que el propósito analítico de esta fase consiste en identificar las características empresariales que mayoritariamente comparten las empresas generadoras de empleo, se procedió a examinar la distribución y comportamiento de la variable CREA_EMP.

Figura 8

Años en los que las empresas activas en el 2024 reportaron generación de empleo durante el periodo 2020 a 2024.



Fuente: Elaboración propia.

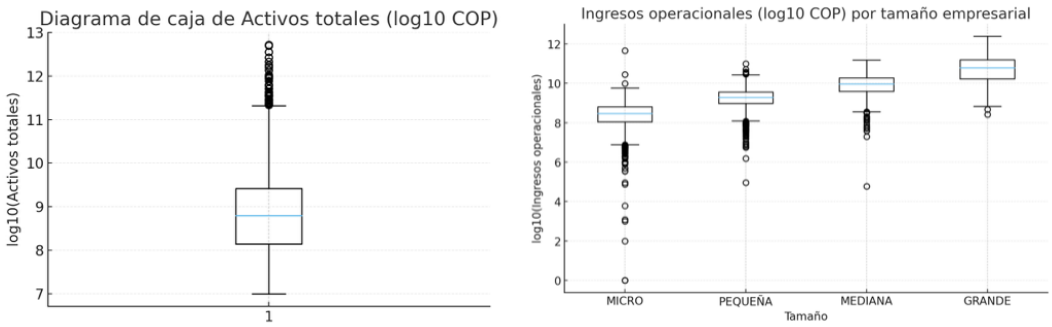
Cómo se puede observar en la Figura 8, solo el 25,9% de las empresas reportaron más de un año de generación de empleo, de modo que el análisis se centrará en este grupo de empresas que han, a pesar de los retos económicos, generado empleo en por lo menos 2 de los 5 años observados. Ello nos generó un nuevo Dataset con 6.142 empresas registradas.

10.2 Análisis del tejido empresarial que ha generado empleo en los últimos años

Variables financieras

Figura 9

Comportamiento de las variables financieras de empresas generadoras de empleo en Cali.



Fuente: Elaboración propia.

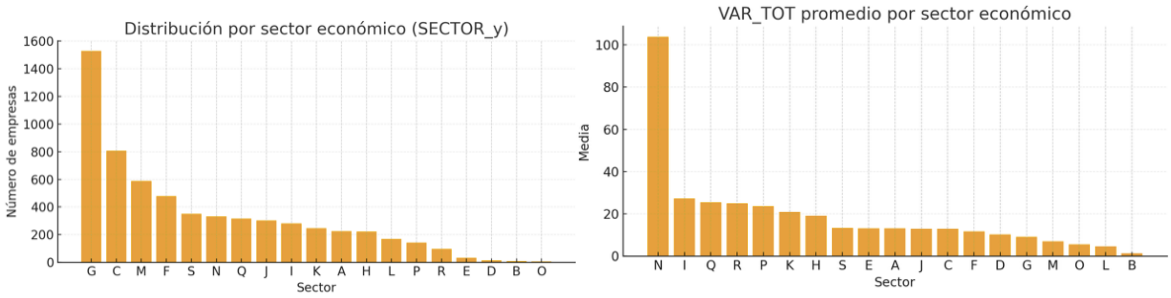
El análisis financiero muestra un tejido empresarial heterogéneo, pero con patrones claros entre las empresas que generan empleo de forma sostenida en Cali. En activos totales, el 25 % de las firmas posee montos iguales o inferiores a 139 millones de pesos, la mediana se ubica en 624 millones, y el 75 % no supera los 2,6 mil millones, lo que indica que la mayoría opera con niveles patrimoniales bajos o medios, pese a la presencia de algunas empresas de gran escala.

En ingresos operacionales, el 25 % reporta montos ≤ 282 millones, la mediana asciende a 965 millones, y el 75 % se mantiene por debajo de 3,3 mil millones. Un hallazgo relevante es que, aunque las empresas grandes tienen ingresos más altos, los boxplots muestran que micro, pequeñas y medianas presentan rangos de ingresos sorprendentemente cercanos, especialmente en los valores centrales. Esto sugiere que buena parte del tejido empresarial activo en generación de empleo opera en bandas similares de ingresos, probablemente por dinámicas de mercado, tamaño del sector atendido o estructuras de operación comparables.

Sector económico

Figura 10

Top de sectores económicos que impactan la generación de empleo y su magnitud.



Fuente: Elaboración propia.

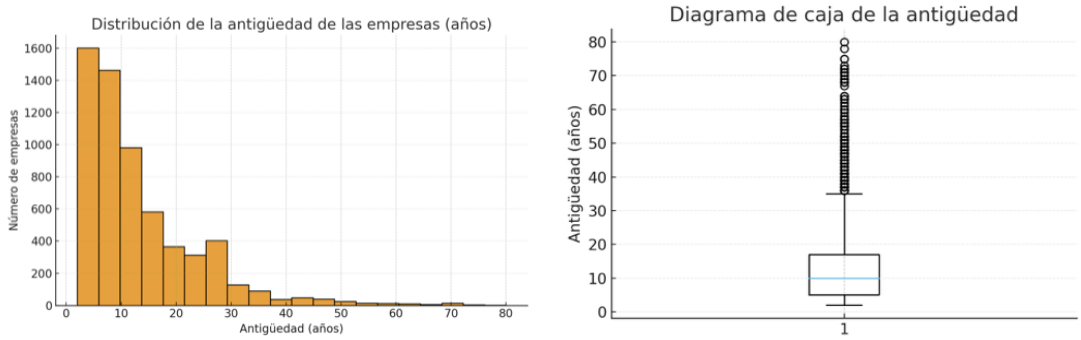
La primera gráfica muestra que los sectores con mayor número de empresas generadoras de empleo en Cali son G – Comercio (24,9 %), C – Manufacturas (13,1 %) y M – Actividades profesionales y técnicas (9,6 %). Esto evidencia que el tejido empresarial que sostiene la creación recurrente de empleo se concentra en actividades comerciales, productivas y de servicios especializados, coherentes con la estructura económica de la ciudad.

En contraste, la segunda gráfica revela que los mayores incrementos promedio de empleo (VAR_TOT) pertenecen a sectores menos numerosos: N – Servicios administrativos y de apoyo (103,8 empleos), I – Transporte y almacenamiento (27,2) y Q – Servicios sociales y de salud (25,4). Estos sectores destacan por su alta intensidad laboral y capacidad de expansión. En conjunto, las dos visualizaciones muestran una dualidad: mientras comercio, manufactura y servicios profesionales forman la base del tejido empresarial, sectores como servicios administrativos, transporte y salud concentran las mayores expansiones netas de empleo.

Antigüedad

Figura 11

Antigüedad de las empresas que han generado empleo en los últimos años.



Fuente: Elaboración propia.

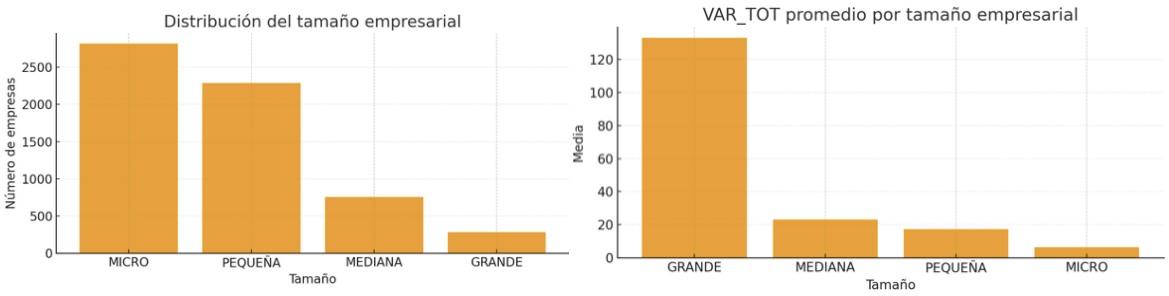
La Figura 11 relacionada con la variable ANTIGÜEDAD muestra que las empresas generadoras de empleo se ubican mayoritariamente en un rango entre 5 y 15 años de antigüedad, con una mediana de 10 años y un rango entre 2 y 80 años. La distribución es asimétrica a la derecha: con un grupo más reducido de empresas de muy larga trayectoria.

Esto sugiere que el tejido empresarial que sostiene la generación sostenida de empleo en Cali está compuesto mayoritariamente por empresas de mediana edad, más que por firmas centenarias. Cabe resaltar que este punto analiza la generación en sí y no la magnitud de la generación.

Tamaño empresarial

Figura 9

Comportamiento del tamaño empresarial frente a la generación de empleo y su magnitud.



Fuente: Elaboración propia.

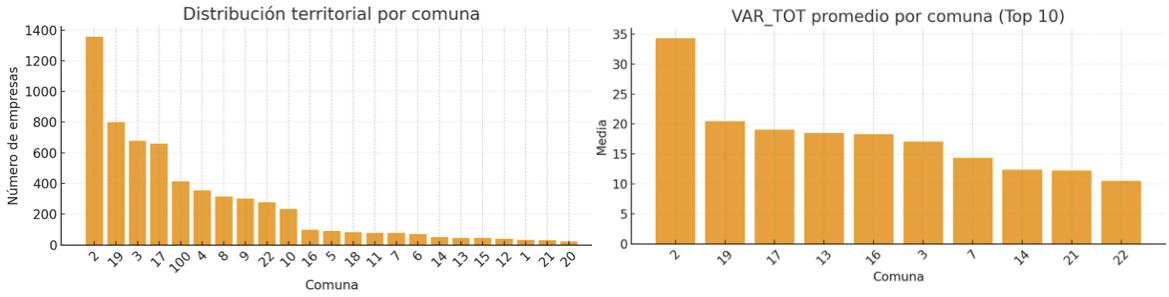
La distribución del tamaño empresarial muestra que el tejido que genera empleo recurrente en Cali está compuesto principalmente por microempresas (47 %) y pequeñas empresas (37 %), mientras que las medianas representan el 11 % y las grandes apenas el 4 %. Esto confirma que la base del ecosistema empresarial que sostiene la creación de empleo es predominantemente de pequeña escala, con estructuras operativas reducidas y uno o pocos establecimientos.

Sin embargo, al analizar la variación neta de empleo (VAR_TOT), el comportamiento cambia de forma sustancial. Aunque menos numerosas, las empresas grandes son las que más empleo generan en promedio, con incrementos cercanos a 133 trabajadores, seguidas por las medianas con alrededor de 23 y las pequeñas con cerca de 17. Las microempresas, pese a su peso numérico, presentan aumentos modestos (≈ 6 empleos). Esto evidencia una doble dinámica: las micro y pequeñas sostienen la base del empleo, pero la expansión neta más significativa proviene de un reducido grupo de empresas medianas y grandes, que concentran la mayor capacidad para crear nuevos puestos de trabajo en el periodo analizado.

Ubicación geográfica

Figura 10

Distribución territorial de las empresas generadoras de empleo.



Fuente: Elaboración propia.

La distribución territorial evidencia una fuerte concentración de empresas generadoras de empleo en unas pocas comunas de Cali. Las comunas 2, 19, 3 y 17 reúnen la mayor cantidad de firmas, con participaciones del 22,1 %, 13,0 %, 11,0 % y 10,7 % respectivamente. Este patrón refleja la presencia de corredores económicos consolidados, donde se concentran actividades comerciales, de servicios y empresariales que favorecen la operación y permanencia de unidades productivas formales.

Al analizar la variación neta del empleo (VAR_TOT) por comuna, surge un comportamiento que le apunta a los mismos corredores empresariales: la Comuna 2 lidera ampliamente con un incremento promedio cercano a 35 empleos, seguida por las comunas 19, 17, y 13, que registran variaciones que oscilan entre 15 y 22 empleos. Esto indica que, además de ser zonas con mayor densidad empresarial, algunas comunas funcionan como polos de expansión laboral, donde se localizan empresas con mayor capacidad de crecimiento. En conjunto, estos

resultados sugieren que la geografía empresarial de Cali no solo concentra empresas, sino que también distribuye las oportunidades de generación de empleo.

10.3 Conclusiones

El análisis integral del tejido empresarial que ha generado empleo en al menos dos de los últimos cinco años en Cali muestra diferentes combinación de factores estructurales, sectoriales, territoriales y financieros. Aunque la mayoría de empresas pertenece a los segmentos micro y pequeño, los mayores aportes netos de empleo provienen de un grupo reducido de empresas medianas y grandes, ubicadas en sectores intensivos en mano de obra y en zonas de alta actividad económica. Estos patrones evidencian que el ecosistema de generación de empleo en Cali es dual: amplio y diverso en su base, pero altamente concentrado en su impacto.

Asimismo, territorios como las comunas 2, 19 y 17 emergen como nodos donde se concentran unidades productivas con mayor capacidad de crecimiento, mientras que sectores como comercio, manufactura y servicios profesionales sustentan la estructura empresarial, en contraste con sectores como servicios administrativos, transporte y salud, que lideran la expansión del empleo

En atención a la primera pregunta analítica planteada se presentan las características que comparten mayoritariamente las empresas activas que han generado empleo en por lo menos 2 de los últimos 5 años en la ciudad de Cali:

- **Antigüedad:** Edad empresarial cercana a los 10 años.
- **Tamaño:** Micro o Pequeña.
- **Sectores estratégicos:** Comercio (G), Manufactura (C), Actividades profesionales y técnicas (M).

- **Territorio:** Localizada en zonas de concentración empresarial y dinamismo laboral, especialmente la Comuna 2, seguida de 19, y 3.
- **Reportes financieros:** Deuda activa, activos rondan los 600 millones.

11. Pregunta de análisis 2: ¿Qué grupos o perfiles empresariales pueden identificarse dentro del conjunto de empresas que generan empleo formal?

11.1 Variables a incluir en el modelo, creación de nuevas variables y su tratamiento

A partir del Dataset A, que contiene las empresas caleñas activas registradas en 2024 que generaron empleo en por lo menos 2 de los 5 últimos años, se definió un conjunto de variables para la construcción del modelo de clusterización. Tras varias pruebas preliminares, se optó por trabajar exclusivamente con variables numéricas en el proceso de formación de clústeres, mientras que las variables categóricas (sector económico, tamaño, comuna, género del representante, entre otras) se reservaron para la caracterización descriptiva posterior de cada clúster. Esta decisión se fundamenta en que los algoritmos de segmentación se basan principalmente en distancias euclidianas y pueden verse distorsionado cuando se incluyen numerosas variables categóricas codificadas (one-hot), que tienden a dominar la métrica de distancia y dificultan la interpretación de los grupos.

El conjunto final de variables numéricas incluidas en el modelo estuvo compuesto por: ANTIGÜEDAD (años de operación de la empresa), ACTIVOS_LOG (logaritmo de los activos totales), APALANCAMIENTO (relación pasivo total / activo total), MARGEN_RENTABLE (utilidad bruta / ingresos operacionales) y VAR_TOT (variación

neta del número de empleados en el periodo analizado). Las dos últimas variables derivadas, APALANCAMIENTO y MARGEN_RENTABLE, se construyeron para capturar la estructura de financiación y la eficiencia económica de la empresa, más allá de los valores absolutos de activos o ingresos.

- APALANCAMIENTO: Este indicador de razón (ratio) se calculó dividiendo el Pasivo Total (TOT_PASIVO) entre el Activo Total (TOT_ACTIVOS) de cada entidad. Su relevancia analítica es fundamental, puesto que mide el porcentaje de los activos de una empresa que está financiado mediante deuda (pasivos). Por lo tanto, la inclusión de esta variable en el estudio es crucial para determinar si existe una correlación entre el nivel de endeudamiento, y el riesgo que este implica, y la capacidad de las empresas para generar y sostener el empleo, medido por la variable VAR_TOT.
- MARGEN_RENTABLE: Este indicador de rentabilidad se construyó dividiendo la Utilidad Bruta (UTILIDAD_BRUTA) entre los Ingresos Operacionales (INGRESOS_OPERACIONALES) de cada empresa. El resultado permite aproximar la eficiencia con la que la empresa convierte sus ingresos operativos en ganancias antes de gastos administrativos y financieros.

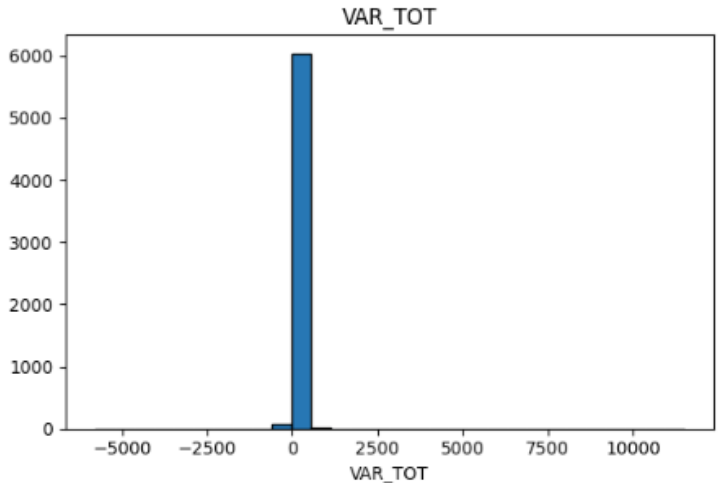
Para reducir la asimetría de las distribuciones y mitigar el impacto de valores extremos, los activos se transformaron mediante logaritmo (ACTIVOS_LOG) y todas las variables numéricas se estandarizaron posteriormente mediante StandardScaler, de modo que tuvieran media cero y desviación estándar uno antes de aplicar los algoritmos de clusterización.

Un punto particular de atención fue la variable VAR_TOT, que presenta una distribución altamente sesgada, con la mayoría de las empresas registrando variaciones modestas en el empleo y un conjunto reducido de casos con crecimientos o reducciones muy elevados. Este

comportamiento sugiere la presencia de outliers que podrían influir en la formación de clústeres. Por esta razón se exploró una transformación sign–log (VAR_TOT_tr), que aplica logaritmo a los valores absolutos conservando el signo, con el propósito de comprimir las colas de la distribución sin perder la información sobre la dirección del cambio en el empleo.

Figura 11

Comportamiento de la variable VAR_TOT en las empresas generadoras de empleo.



Fuente: Elaboración propia.

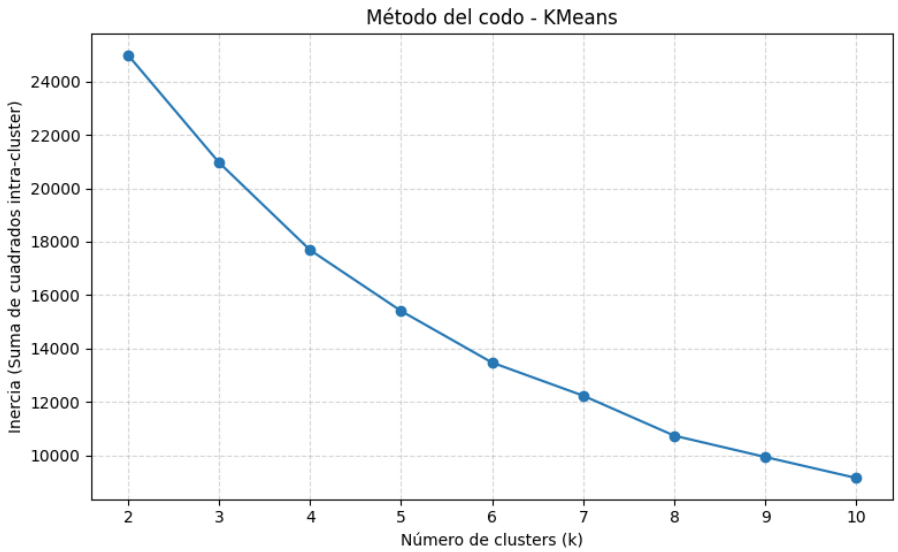
No obstante, la comparación sistemática de las métricas de validación interna mostró que los modelos estimados con VAR_TOT_tr presentaban un menor índice Silhouette y un mayor índice Davies–Bouldin, tanto para k = 4 como para k = 5 y k = 6, en comparación con los modelos que utilizaban VAR_TOT original estandarizada. En consecuencia, se decidió mantener VAR_TOT sin transformar dentro del modelo final, dado que aportaba una mejor capacidad de diferenciación entre empresas en términos de su dinámica de empleo.

En síntesis, el modelo de clusterización se construyó a partir de un conjunto compacto de cinco variables numéricas que resumen la antigüedad, la escala financiera, la estructura de endeudamiento, la rentabilidad y la variación neta del empleo.

11.2 Selección de número óptimo de Clústeres

Figura 12

Evaluación de inercia por número de Clústers.



Fuente: Elaboración propia.

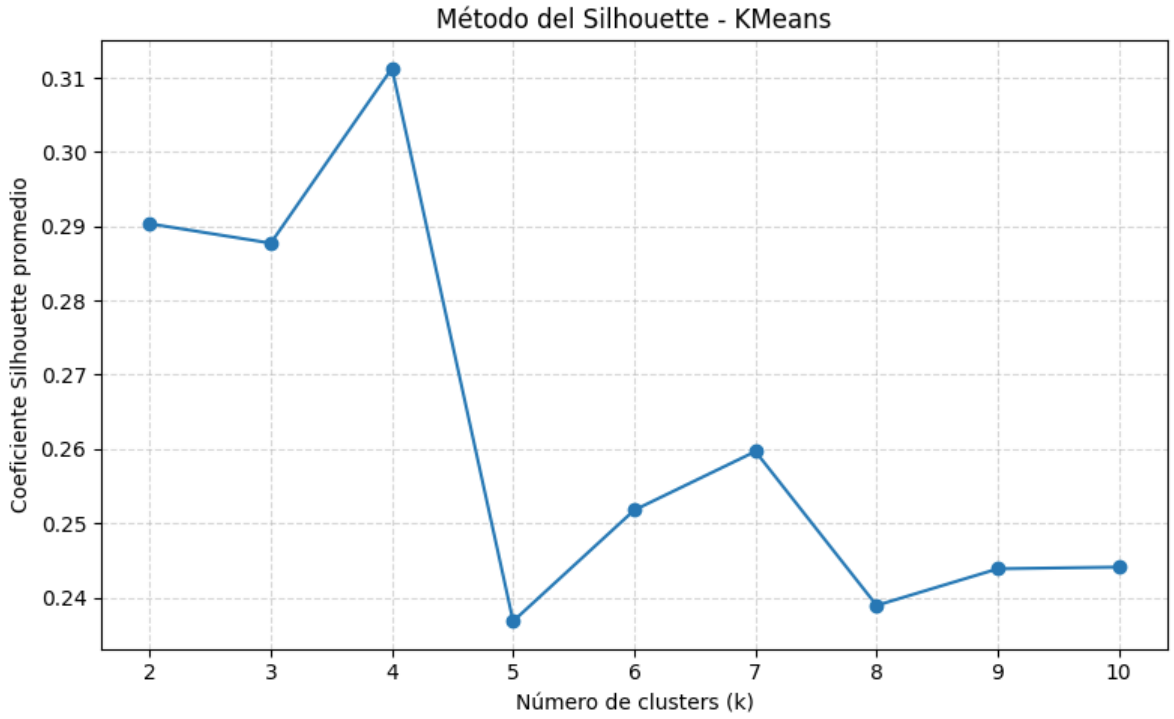
Para identificar el número adecuado de clústeres que permita segmentar a las empresas según sus características estructurales, financieras y laborales, se aplicó el método del codo (Elbow Method) utilizando el algoritmo K-Means sobre la matriz de variables previamente transformadas y estandarizadas. Este método evalúa la inercia del modelo (suma de cuadrados intra-clúster), que representa qué tan compactos son los grupos formados: a menor inercia, mayor cohesión interna del clúster.

En la gráfica resultante se observa que la inercia disminuye rápidamente entre k = 2 y k = 4, reflejando mejoras significativas en la separación de los grupos. Sin embargo, a partir de k = 4, la pendiente empieza a suavizarse y las reducciones adicionales en la inercia se vuelven progresivamente menores. El comportamiento de la curva muestra un punto de inflexión entre k

= 4 y k = 5, donde la ganancia marginal en cohesión deja de ser sustancial. Este patrón indica que, aunque incrementar el número de clústeres continúa reduciendo la inercia, las mejoras posteriores ya no justifican la complejidad adicional del modelo.

Figura 13

Evaluación de Silhouette por número de Clústers.



Fuente: Elaboración propia.

Con el fin de complementar el método del codo se aplicó el método del silhouette para evaluar la cohesión y separación de los clústeres generados por K-Means.

El coeficiente de silhouette se calculó para valores de *k* entre 2 y 10. Tal como se observa en la gráfica, el valor promedio del silhouette alcanza su máximo en k = 4, lo que indica que este número de clústeres ofrece el mejor balance entre cohesión interna y separación entre grupos.

11.3 Aplicación de modelos de clusterización

Los modelos de clusterización constituyen una herramienta clave para la segmentación empresarial, ya que permiten agrupar empresas con comportamientos y características similares. Su aplicación permite pasar de un análisis descriptivo general a una comprensión más fina de los tipos de empresas que conforman el tejido productivo.

Con el número óptimo de clústeres establecido en k = 4, se procedió a aplicar dos técnicas de segmentación: K-Means y Agglomerative Clustering. Ambas metodologías, fueron implementadas con el propósito de identificar grupos homogéneos de empresas a partir de sus características estructurales, financieras y laborales.

La aplicación de dos técnicas de clusterización tiene el propósito de comparar su desempeño y determinar cuál modelo ofrecía una segmentación más coherente, estable y útil para el análisis. Al contrastar los resultados de K-Means y Agglomerative, es posible evaluar la estabilidad de los clústeres y determinar cuál técnica ofrece una clasificación empresarial más interpretable para efectos de análisis.

11.4 Evaluación de los modelos

Con el objetivo de seleccionar el modelo de clusterización que ofreciera la segmentación más coherente del tejido empresarial de Cali, se evaluó el desempeño de K-Means y Agglomerative Clustering utilizando métricas de validación interna. Estas métricas permiten analizar, desde distintas perspectivas, la calidad de los clústeres formados, en términos de cohesión interna y separación entre grupos. A partir de estos indicadores se definió cuál de los dos algoritmos resulta más adecuado para representar la estructura de los datos y servir como base para la interpretación de perfiles empresariales.

Métricas de evaluación

- Índice Silhouette: cuantifica simultáneamente la similitud de cada observación con su propio clúster y la diferencia con los demás, tomando valores entre –1 y 1. Valores más altos indican clústeres más compactos y mejor separados.
- Índice Calinski–Harabasz (CH): relaciona la dispersión entre los centros de los clústeres con la dispersión interna de cada grupo. Valores mayores reflejan clústeres bien separados y con baja variabilidad interna.
- Índice Davies–Bouldin (DB): mide el solapamiento entre clústeres a partir de la distancia entre sus centros y su dispersión interna. A diferencia de los anteriores, en este caso valores menores son preferibles, pues indican menor similitud entre los grupos.

Análisis de métricas

En la Tabla 4 se presentan los resultados de estas métricas para ambos modelos estimados con k = 4 clústeres. K-Means alcanzó un índice Silhouette de aproximadamente 0,311, un índice Calinski–Harabasz cercano a 1505,9 y un índice Davies–Bouldin en torno a 1,062. Estos valores indican una segmentación con clústeres relativamente compactos y bien separados entre sí, especialmente si se considera la heterogeneidad inherente a los datos empresariales analizados.

En paralelo, se estimó un modelo Agglomerative Clustering con la misma configuración de variables y número de clústeres, cuyos resultados fueron inferiores en términos de calidad de clusterización: el índice Silhouette se situó alrededor de 0,295, el Calinski–Harabasz en 1345,4 y el Davies–Bouldin en 1,066.

Tabla 4

Métricas de evaluación de modelos de segmentación.

Modelo	Silhouette	Calinski–Harabasz	Davies–Bouldin
K-Means	0,31	1.505	1,06
Agglomerative	0,29	1.345	1,06

Fuente: Elaboración propia.

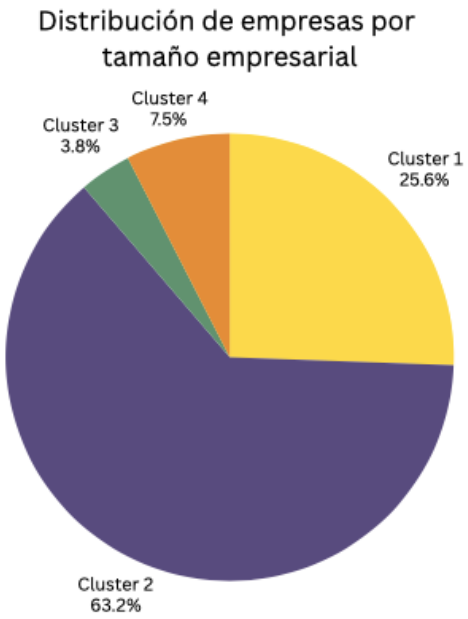
Aunque las diferencias no son abismales, K-Means presenta sistemáticamente mejores valores en las tres métricas, lo que indica clústeres algo más compactos, con mayor separación entre sus centros y menor solapamiento interno. Adicionalmente, K-Means ofrece centroides claramente definidos, lo que facilita la interpretación económica de cada grupo en términos de antigüedad, tamaño financiero, endeudamiento, rentabilidad y variación de empleo.

11.5 Caracterización de Clústeres

La aplicación del modelo K-Means, permitió identificar cuatro perfiles empresariales. La Figura 16 muestra que el clúster 2 concentra la mayor parte de las empresas de la muestra, seguido del clúster 1, mientras que los clústeres 4 y, especialmente, 3 agrupan un número mucho menor de empresas, aunque con características extremas en términos de estructura financiera y generación de empleo.

Figura 14

Número de empresas por Clúster.



Fuente: Elaboración propia.

A partir de los centroides del modelo y de los cruces con variables categóricas (sector económico, tamaño empresarial, ubicación por comuna), se construyeron los perfiles que se describen a continuación.

Tabla 5

Centroides de los Clústers creados.

Clúster	Antiguedad	Activos Log	Apalancamiento	Margen rentable	Variación empleos
1	24.7	22.4	0,48	0,10	19.4
2	8.1	19.4	0.48	0.10	9.9
3	24.8	23.5	0.62	0.00	1758
4	7.8	19	1.4	-0.65	9.1

Fuente: Elaboración propia.

Clúster 1 - Empresas estrella

El clúster 1 agrupa el segundo volumen más importante de empresas y se caracteriza por una antigüedad significativamente mayor (alrededor de 25 años en promedio) y niveles de activos superiores a los del clúster 1. Aquí se observa una estructura de tamaño medianamente equilibrada: aproximadamente 42,5 % de las empresas son pequeñas, 36 % medianas y 15 % grandes, lo que configura un segmento de empresariado consolidado.

En términos sectoriales, este clúster también se apoya en el comercio (G) y la manufactura (C), pero con mayor presencia relativa de servicios profesionales (M), salud y servicios sociales (Q) y actividades de servicios administrativos y de apoyo (N). Territorialmente, destacan las comunas 2, 3, 4 y 18. Una proporción relevante de empresas cuenta con entre 2 y 5 establecimientos e incluso con más de 6, lo que refleja cierta expansión geográfica.

Desde la perspectiva del empleo, este clúster concentra una fracción considerable de empresas con más de 50 trabajadores, pero resalta especialmente la categoría de empleados entre 10 y 50. La variación neta de empleo (VAR_TOT) promedio es de aproximadamente 19 empleados, lo cual la ubica como uno de los Clústers con mayor aporte en la generación de empleos.

La mayoría de empresas ha generado empleo en dos o tres de los últimos cinco años, y se observa una presencia algo mayor de casos con CREA_EMPLEO = 4. En términos financieros, mantiene niveles de apalancamiento moderados y márgenes positivos, con muy baja incidencia de empresas endeudadas de forma crítica. En conjunto, el clúster 1 puede interpretarse como

el núcleo de empresas consolidadas de pequeña y mediana escala que, sin ser el segmento mayoritario, tiene un peso importante en la generación estable de empleo formal.

Clúster 2 – Empresas pequeña escala

El clúster 2 es el grupo de mayor tamaño, reuniendo cerca de dos tercios de las empresas analizadas. Se compone principalmente de microempresas (61,9 %) y pequeñas empresas (35,9 %), con niveles de activos y antigüedad relativamente bajos (alrededor de 8 años de operación en promedio). En términos de empleo, predominan las empresas con menos de 10 trabajadores, y una proporción significativa se encuentra en los rangos de hasta 3 empleados, lo que evidencia estructuras organizacionales de escala reducida.

Sectorialmente, este clúster está dominado por el comercio al por mayor y al por menor (G) y las industrias manufactureras (C), complementadas por actividades de servicios profesionales, científicos y técnicos (M) y algunos servicios administrativos y de apoyo (N). Territorialmente, las empresas se concentran en las comunas 2, 3, 17 y 18, que corresponden a zonas urbanas consolidadas con alta densidad de actividad económica. La mayoría opera con un solo establecimiento y presenta niveles de apalancamiento moderados, con márgenes operacionales positivos pero acotados.

En cuanto a su dinámica laboral, la distribución de CREA_EMPLEO muestra que la gran mayoría de empresas de este clúster ha generado empleo en dos de los últimos cinco años, con una fracción menor que alcanza tres años. En conjunto, el clúster 2 representa el tejido empresarial joven y de menor escala de la ciudad: empresas formales de comercio y servicios que mantienen operaciones y contribuyen al empleo, pero con capacidades de expansión aún

limitadas. Este Clúster es el que representa mayoritariamente el conjunto macro de empresas caleñas.

Clúster 3 – Empresas ancla

El clúster 3 es el grupo más pequeño en número de empresas, pero el más extremo en términos de impacto laboral. Sus centroides muestran niveles muy altos de activos (ACTIVOS_LOG) y antigüedad similar a la del clúster 0, pero destaca sobre todo por una VAR_TOT promedio extraordinariamente elevada, combinada con el hecho de que el 100 % de las empresas tiene más de 50 empleados. Además, una parte relevante cuenta con múltiples establecimientos (2–5 y 6+), lo que sugiere operaciones de gran escala.

La estructura sectorial de este clúster es muy concentrada: aproximadamente 80 % de las empresas pertenece al sector N (actividades de servicios administrativos y de apoyo) y el 20 % al sector H (transporte y almacenamiento). Se trata, por tanto, de compañías intensivas en mano de obra, dedicadas a servicios tercerizados, logística, transporte y operaciones de soporte. Territorialmente, se localizan casi exclusivamente en las comunas 2 y 3, dentro del clúster territorial 0, lo que las sitúa en corredores estratégicos de la ciudad.

En términos de trayectoria laboral, este es el único clúster en el que la distribución de CREA_EMPLEO muestra una presencia significativa del valor 4 años de generación de empleo, además de dos y tres años. Ello, sumado a la altísima variación neta de empleo, permite identificar al clúster 3 como un núcleo hiperintensivo en creación de empleo, conformado por pocas pero muy grandes empresas que expanden de manera acelerada su planta laboral. Aunque su peso numérico en la muestra es reducido, su contribución potencial a la generación de empleo

formal en la ciudad es desproporcionadamente alta, lo que lo convierte en un segmento estratégico para la política pública.

Clúster 4 – Empresas en riesgo

El clúster 3 agrupa un número reducido de empresas, pero su perfil es claramente diferenciado en términos de riesgo. Se trata de microempresas jóvenes (alrededor de 8 años en promedio), con niveles de activos relativamente bajos y un apalancamiento muy elevado, significativamente superior al del resto de clústeres. A diferencia de los grupos 1 y 2, el margen de rentabilidad operacional es marcadamente negativo, lo que sugiere dificultades para cubrir costos y sostener su operación en el tiempo.

La mayoría de estas empresas tiene menos de 10 empleados, muchas de ellas con plantillas de hasta 3 trabajadores, y opera con un solo establecimiento. Sectorialmente, el clúster presenta una mezcla en la que destacan las actividades de transporte y almacenamiento (H) y los servicios administrativos y de apoyo (N), junto con participación de comercio (G) y manufactura (C). Territorialmente, se distribuyen en diversas comunas, con mayor presencia en las comunas 3, 18 y 23, y una participación relevante en los clústeres territoriales 0 y 3.

En términos de comportamiento laboral, la variación neta de empleo promedio es moderada, y la distribución de CREA_EMPLEO indica que la mayoría de empresas solo ha generado empleo en dos de los últimos cinco años, con una presencia reducida de casos con tres años. La vari

Considerando su alto apalancamiento, su baja escala y su rentabilidad negativa, el clúster 3 puede interpretarse como un segmento de microempresas vulnerables, cuya sostenibilidad y

capacidad de mantener empleo formal en el tiempo es limitada y podría estar en riesgo frente a shocks económicos o restricciones de liquidez.

Síntesis comparativa

La comparación de los cuatro clústeres revela una estructura empresarial estratificada en la ciudad. Por un lado, el clúster 2 agrupa el grueso del tejido empresarial formal joven, de baja escala y concentración en comercio y servicios, que sostiene una parte importante del empleo pero con crecimientos moderados. El clúster 1 representa un segmento más consolidado de pequeñas, medianas y algunas grandes empresas, con mayor diversificación sectorial y mejor desempeño en generación de empleo.

En el extremo superior, el clúster 3, aunque numéricamente pequeño, concentra a las empresas de gran escala con mayor intensidad en creación de empleo, principalmente en servicios administrativos y logística, lo que lo constituye en un objetivo prioritario para políticas de acompañamiento y alianza público–privada. En el extremo opuesto, el clúster 4 identifica a las microempresas jóvenes más frágiles, con problemas de rentabilidad y altos niveles de apalancamiento, para las cuales podrían requerirse estrategias diferenciadas de fortalecimiento financiero y técnico si se busca preservar el empleo que hoy sostienen.

11.6 Conclusiones

La pregunta de análisis abordada buscaba identificar qué grupos o perfiles empresariales pueden distinguirse dentro del conjunto de empresas que generan empleo formal en Santiago de Cali, y cuáles son los patrones que diferencian su estructura, sector económico y comportamiento operativo, con el fin de orientar programas, incentivos y estrategias focalizadas.

Con ello en mente, podemos identificar que la segmentación obtenida reveló dos grandes grupos que representan en mayor medida la fotografía actual del tejido empresarial formal generador de empleo en la ciudad: un conjunto amplio de micro y pequeñas empresas jóvenes —que representan la gran mayoría de empresas del tejido empresarial caleño—, y un grupo consolidado de pequeñas y medianas empresas con trayectorias más largas, niveles de activos superiores y desempeño más estable pero con un aporte a la generación del número de empleos realmente significativa.

De igual forma, el ejercicio también permitió identificar dos clústeres de especial interés estratégico: por un lado, un grupo reducido pero altamente determinante de grandes empresas intensivas en empleo, cuyas operaciones de servicios administrativos, apoyo logístico y actividades tercerizadas explican una proporción significativa de la variación neta del empleo en el período analizado; y por otro lado, un grupo de microempresas vulnerables, caracterizado por altos niveles de apalancamiento, márgenes negativos y estructuras operativas frágiles que comprometen su sostenibilidad y, por ende, la estabilidad del empleo que generan.

Dado el valor estratégico de este ejercicio, tanto para la lectura general de los tipos de empresas que componen cada Clúster, como para el análisis e intervención uno a uno de las empresas que se encuentran en riesgo o que se consolidan como empresas ancla, en la Base Maestra de Priorización se asignó cada empresa al Clúster al cual pertenecen.

12. Pregunta de análisis 3: ¿Cuál es la probabilidad de que una empresa incremente su número de empleados en el año siguiente a partir de las características empresariales que presenta en el periodo actual?

12.1 Construcción de la base predictiva

Como fue indicado anteriormente, la construcción del Dataset B consistió en la concatenación vertical de los registros empresariales alojados en las bases 2020, 2022 y 2023, cabe resaltar que en este caso la base 2021 no contaba con la cantidad de variables suficientes para incluirla dentro del análisis en conjunto con las demás bases. Téngase en cuenta que las bases ya habían surtido preliminarmente un proceso de transformación en materia de depuración, transformación de variables creación de variables y creación de la variable objetivo.

Habiendo aplicado la concatenación vertical de los registros anuales de los periodos indicados, obtuvimos una base conformada por 72.528 registros empresa–año, correspondientes a 48.905 empresas únicas.

Recordemos que la variable objetivo se construyó a partir de la dinámica del empleo entre dos periodos consecutivos. Para cada empresa y año t, se dispone del número de empleados en el periodo actual (EMPLEADOS_T) y del número de empleados en el año siguiente (EMPLEADOS_T1). Con esta información se definió la variable binaria VAR_OBJ2, que toma el valor de 1 cuando la empresa incrementa su planta de personal (EMPLEADOS_T1>EMPLEADOS_T) (EMPLEADOS_T1>EMPLEADOS_T) y 0 en caso contrario (EMPLEADOS_T1≤EMPLEADOS_T) (EMPLEADOS_T1≤EMPLEADOS_T).

En síntesis, el Dataset B quedó conformado por la variable objetivo VAR_OBJ y un conjunto robusto de variables explicativas que integran las dimensiones estructurales, y financieras de las empresas. Entre ellas se incluyen: el sector económico, el tamaño empresarial, la comuna y su agrupación (COMUNA_CLUS), la antigüedad, el número de establecimientos y sus categorías, el nivel de empleo actual (EMPLEADOS_T), así como los indicadores contables

de activos, pasivos, ingresos y patrimonio, junto con sus transformaciones logarítmicas y winsorizadas.

Balanceo de clases

Durante el análisis inicial se identificó que la variable objetivo presenta un desbalance medio. En el Dataset B, la distribución de la variable objetivo es: 12.324 observaciones (17,0 %) corresponden a empresas que aumentaron su número de empleados, mientras que 60.204 observaciones (83,0 %) corresponden a empresas que mantuvieron o redujeron su empleo. Si bien este no es un desbalanceo extremo, esta asimetría puede afectar negativamente el aprendizaje del modelo, favoreciendo la predicción de la clase mayoritaria. Por ello, se aplicó una técnica de balanceo que pudiera contrarrestar estos riesgos.

La técnica seleccionada fue SMOTE (Synthetic Minority Oversampling Technique), ampliamente reconocida para el tratamiento de desbalances moderados y severos. A diferencia del oversampling tradicional —que simplemente replica observaciones de la clase minoritaria—, SMOTE genera nuevas observaciones sintéticas mediante interpolaciones lineales entre instancias reales cercanas. Esta característica resulta especialmente adecuada en este proyecto por tres razones:

1. **Evita el sobreajuste** asociado a la duplicación de registros, al introducir variabilidad dentro de la clase minoritaria.
2. **Preserva la estructura multidimensional de los datos**, lo cual es relevante dado el amplio número de predictores estructurales, territoriales y financieros presentes en el modelo.

3. **Mejora la capacidad del modelo para aprender fronteras de decisión más representativas**, permitiendo distinguir mejor a las empresas que efectivamente incrementan su empleo.

12.2 Modelos aplicados

Con el fin de responder la pregunta de análisis se seleccionaron tres modelos supervisados ampliamente utilizados en problemas de clasificación binaria con múltiples predictores estructurales, financieros y territoriales: Random Forest, Regresión Logística y XGBoost. La elección de estos modelos responde tanto a la naturaleza de la pregunta como a la complejidad de la base de datos.

En primer lugar, todos permiten estimar probabilidades, requisito indispensable para el cálculo de umbrales y la interpretación de la probabilidad de aumento de empleo. En segundo lugar, estos algoritmos presentan distintos enfoques de aprendizaje: uno lineal (regresión logística) y dos no lineales basados en árboles (Random Forest y XGBoost), permitiendo capturar tanto relaciones simples como interacciones complejas entre variables.

Composición de los modelos y configuración aplicada

Cada modelo fue ajustado con parámetros específicos orientados a mejorar la capacidad predictiva, controlar el sobreajuste y adaptarse a la estructura de la base:

Regresión Logística.

La Regresión Logística se incorporó como modelo lineal de referencia y como mecanismo para producir probabilidades interpretables. Previo al entrenamiento, las variables

predictoras se estandarizaron mediante StandardScaler, garantizando estabilidad numérica y comparabilidad entre predictores que se encuentran en escalas distintas, como activos, ingresos, tamaño empresarial y variables dummy derivadas de la codificación categórica.

El modelo se entrenó con un máximo de 500 iteraciones para asegurar convergencia, especialmente en un escenario con alta dimensionalidad producto del one-hot encoding.

Random Forest Classifier.

Su configuración incluyó 500 árboles de decisión, con una profundidad máxima de 15 niveles y un mínimo de 20 observaciones por nodo hoja. Estas restricciones permiten que el modelo identifique patrones complejos sin incurrir en sobreajuste, garantizando un aprendizaje robusto sobre la base empresarial de Cali. Adicionalmente, la selección de variables bajo la regla sqrt en cada división introduce diversidad entre árboles y reduce la correlación interna del bosque, favoreciendo mayor estabilidad en las predicciones. Esta configuración busca capturar patrones complejos de las variables financieras, territoriales y de empleo, mientras reduce la varianza entre árboles.

XGBoost Classifier.

El modelo XGBoost se seleccionó por su alto rendimiento en escenarios con múltiples predictores y estructuras no lineales. Su configuración incluyó 500 árboles con una tasa de aprendizaje de 0,05, lo cual favorece un entrenamiento progresivo y controlado. La profundidad máxima de cada árbol se fijó en 6 niveles, permitiendo capturar interacciones importantes sin sacrificar generalización. Los parámetros subsample = 0.8 y colsample_bytree = 0.8 introdujeron

regularización mediante muestreo parcial de observaciones y predictores, fortaleciendo la estabilidad del modelo frente a una base empresarial amplia y variada.

La selección de modelos y umbral óptimo abre paso a la siguiente sección, donde se evalúa rigurosamente su desempeño mediante métricas adecuadas a problemas de clasificación con clases desbalanceadas.

12.3 Evaluación de modelos y elección

Con el objetivo de determinar el modelo más adecuado para predecir la probabilidad de generación de empleo por parte de las empresas de Santiago de Cali, se implementó un proceso de evaluación basado en validación cruzada estratificada. La validación cruzada consiste en dividir los datos en varios subconjuntos o folds y entrenar el modelo repetidamente, usando en cada iteración una parte distinta de los datos para validación. Este procedimiento es especialmente útil porque permite obtener métricas de desempeño más estables, robustas y libres de sobreajuste, dado que cada modelo es evaluado en múltiples particiones independientes de los datos. Así, la selección del modelo final no depende de una sola división arbitraria de entrenamiento y prueba, sino de un desempeño promedio confiable.

Dado el desbalance anteriormente identificado, se integró la técnica SMOTE dentro del pipeline utilizado en la validación cruzada actuando sobre el conjunto de entrenamiento de cada fold durante el proceso de validación. Esto significa que en cada iteración el algoritmo genera muestras sintéticas solo a partir de los datos disponibles para entrenamiento, mientras que el conjunto de validación permanece completamente intacto. En consecuencia, cada fold entrena un modelo con un conjunto balanceado y evalúa su desempeño en un subconjunto sin intervenir,

lo que produce una estimación más estable y representativa del comportamiento del modelo en escenarios reales.

Para comparar los modelos, se utilizaron varias métricas que reflejan distintos aspectos del comportamiento predictivo.

- **La exactitud (Accuracy)** indica la proporción total de predicciones correctas del modelo; sin embargo, en problemas con clases desbalanceadas no es una medida suficiente, pues puede favorecer modelos que predicen mayoritariamente la clase dominante, como es el caso.
- **La precisión (Precision)** mide qué tan correctamente clasifica el modelo a las empresas que predice como generadoras de empleo; es decir, un valor alto indica pocos falsos positivos y, por lo tanto, predicciones más confiables.
- **La sensibilidad (Recall)** indica la proporción de empresas generadoras de empleo que el modelo logra identificar; en este contexto, un valor alto implica que el modelo deja escapar muy pocos casos verdaderamente positivos.
- **F1** combina precisión y sensibilidad en un solo indicador armónico, útil cuando se requiere un equilibrio entre ambos criterios.
- **El área bajo la curva ROC (ROC-AUC)** evalúa la capacidad general del modelo para discriminar entre empresas que generan empleo y aquellas que no, independientemente del umbral de clasificación.
- **El área bajo la curva de precisión-recall (PR-AUC)** se utiliza como métrica más informativa en escenarios con desbalance de clases, ya que refleja simultáneamente la calidad de las predicciones positivas y la capacidad del modelo para capturar casos relevantes.

A continuación se presenta una tabla comparativa de los modelos evaluados (Regresión Logística, Random Forest y XGBoost), utilizando como referencia los promedios obtenidos en la validación cruzada para cada una de las métricas.

Tabla 6

Métricas de evaluación de modelos predictivos

Modelo	Accuracy	Precisión	Recall	F1	ROC - AUC
Regresión Logística	0,62	0,24	0,60	0,35	0,66
Random Forest	0,68	0,28	0,57	0,38	0,70
XGBoost	0,79	0,35	0,24	0,24	0,69

Fuente: Elaboración propia.

Regresión Logística

La regresión logística mostró un desempeño sólido como modelo base, destacándose por presentar el recall más alto entre los modelos evaluados, lo que implica que logró identificar un mayor número de empresas que efectivamente generaron empleo. No obstante, este comportamiento vino acompañado de un nivel reducido de precisión, reflejando un volumen importante de falsos positivos.

En términos de negocio, esto implicaría que la Secretaría de Desarrollo Económico recibiría un conjunto más amplio de empresas candidatas, pero con menor confiabilidad respecto a cuál de ellas verdaderamente contribuye al empleo. El modelo privilegia "no dejar escapar" empresas potenciales, incluso a costa de considerar algunas que en realidad no generan empleo.

Random Forest

El modelo Random Forest presentó los mejores resultados en términos de F1 y ROC-AUC, lo que refleja un equilibrio adecuado entre precisión y recall, además de una capacidad superior para discriminar entre empresas generadoras y no generadoras de empleo. Aunque su recall es ligeramente menor al de la regresión logística, la precisión es considerablemente mayor, lo que indica predicciones más certeras y una menor proporción de falsos positivos.

Esto resulta fundamental en un escenario donde los recursos institucionales para acompañamiento empresarial son limitados y deben dirigirse hacia actores con mayor probabilidad real de impacto laboral. En el análisis general, el Random Forest se comportó como el modelo más robusto y estable.

XGBoost

El modelo XGBoost obtuvo el mayor valor de accuracy; sin embargo, esta métrica resultó engañosa, ya que el modelo mostró un recall significativamente bajo, lo que implica que dejó sin identificar a la mayoría de empresas generadoras de empleo. A pesar de contar con una precisión relativamente competitiva, su baja sensibilidad lo convierte en un modelo poco adecuado para un problema donde la identificación correcta de empresas con impacto laboral es prioritaria. En términos de negocio, su adopción generaría un riesgo elevado de subidentificación de oportunidades de generación de empleo, lo cual contradice los objetivos estratégicos del proyecto.

En este punto, es donde la constante interacción con el negocio y el entendimiento profundo de sus necesidades —tal como lo plantea la metodología CRISP-DM— cobra una relevancia decisiva. El objetivo central de la Secretaría de Desarrollo Económico es asignar sus

recursos y esfuerzos de manera eficiente para avanzar hacia la meta distrital de los 100.000 empleos, lo que exige no solo identificar empresas con potencial, sino hacerlo con la mayor certeza posible.

En ese sentido, la métrica de precisión adquiere un papel fundamental, pues un valor más alto implica destinar esfuerzos a empresas que realmente pueden aportar a la generación de empleo y no desperdiciar recursos en casos erróneos. Sin embargo, esta métrica no puede analizarse de forma aislada: aunque XGBoost presenta la precisión más alta, lo logra a costa de dejar por fuera a la gran mayoría de empresas que sí generan empleo, limitando severamente el alcance de la Secretaría.

Tras discutir estos hallazgos con el equipo técnico de la Secretaría, se confirmó que una estrategia que ignore gran parte de las oportunidades reales no es viable. Bajo esta lógica, el modelo Random Forest ofrece un desempeño mucho más sólido y equilibrado, combinando una precisión suficientemente alta con un nivel adecuado de sensibilidad y discriminación. Esto permite focalizar los esfuerzos institucionales con mayor confiabilidad, sin renunciar a identificar un volumen significativo de empresas que realmente pueden contribuir al cumplimiento de la meta de empleo.

Una vez consolidado el Dataset B y definida la variable objetivo VAR_OBJ2, se procedió a la construcción de los conjuntos de entrenamiento y prueba (train y test), paso fundamental para garantizar la validez de los modelos predictivos. La separación entre estos conjuntos permite evaluar la capacidad del modelo para generalizar adecuadamente frente a datos no observados durante el entrenamiento, evitando fenómenos de sobreajuste (overfitting) y proporcionando estimaciones fiables de desempeño en escenarios reales.

El procedimiento consistió en: (i) realizar una partición estratificada del Dataset B en train del 80% de la base y test del 20%, conservando la proporción original de clases; y (ii) aplicar SMOTE únicamente sobre el conjunto de entrenamiento, equilibrando la cantidad de observaciones de ambas clases. El conjunto test permaneció sin alteraciones, garantizando una evaluación objetiva del desempeño predictivo del modelo frente a una distribución realista.

Como resultado, la base de train balanceada incorporó un número equivalente de casos pertenecientes a cada clase de la variable objetivo teniendo como resultado una base de 96.326 registros, distribuida en un 50% por cada clase, mientras que la base **test** conservó la distribución natural del fenómeno. Esta configuración asegura un entrenamiento robusto, reduce sesgos asociados al desbalanceo original y permite medir con mayor precisión la efectividad de los modelos predictivos construidos en las fases posteriores.

12.4 Comportamiento e interpretabilidad del modelo

Para conocer con mayor detalle el comportamiento del modelo frente a cada clase, se realizó una evaluación mediante matriz de confusión, para lo cual se dividió la base en conjuntos de entrenamiento y prueba, se aplicó SMOTE únicamente sobre el conjunto de entrenamiento y, posteriormente, se seleccionó un umbral de clasificación de 0,53. Con este umbral se generaron las predicciones sobre la base de prueba y se obtuvieron las métricas ROC-AUC (0,70), PR-AUC (0,32) y la matriz de confusión mostrada en la figura.

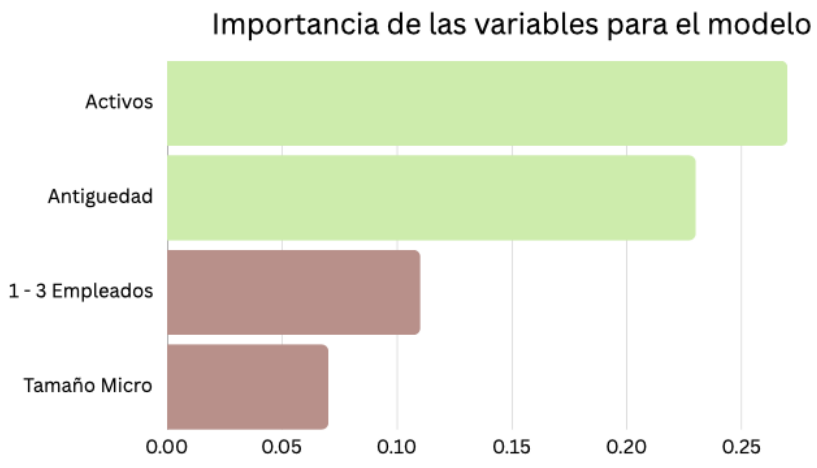
La matriz de confusión indica que el modelo clasifica correctamente 9.386 empresas como no generadoras de empleo y 1.224 como generadoras de empleo, mientras que se equivoca en 2.655 falsos positivos (empresas que no generan empleo pero el modelo las marca como

generadoras) y 1.241 falsos negativos (empresas que sí generan empleo pero el modelo no las detecta).

En términos proporcionales, el modelo logra identificar cerca de la mitad de las empresas que realmente generan empleo (recall de la clase 1 ≈ 0,50) y, entre las empresas que marca como generadoras, aproximadamente un 32 % efectivamente lo son (precisión de la clase 1 ≈ 0,32). Esto confirma varios hallazgos: (i) el modelo mantiene un desempeño coherente con la validación cruzada, (ii) está dispuesto a "arriesgarse" a incluir algunas empresas que finalmente no generan empleo, a cambio de no dejar por fuera a una proporción importante de casos positivos, y (iii) persiste un compromiso entre focalización fina y cobertura, que deberá ser tenido en cuenta por la Secretaría al momento de diseñar las estrategias de acompañamiento empresarial.

Figura 15

Importancia de variables para el modelo.



Fuente: Elaboración propia.

Teniendo en cuenta lo mencionado anteriormente, aplicamos la propiedad feature importance del modelo Random Forest para entender los pesos que tienen cada variable utilizada en el modelo en la predicción que supone. La Figura 15 de variables según el feature importance del modelo Random Forest muestra que la antigüedad de la empresa y el nivel de activos totales son, por amplio margen, los factores más determinantes para clasificar si una empresa genera empleo.

Esto confirma que las características estructurales asociadas a madurez y capacidad financiera son señales fuertes y consistentes para el modelo: las empresas consolidadas y con mayor soporte económico tienden a presentar patrones de generación de empleo más estables y predecibles. En otras palabras, el comportamiento histórico de Cali sugiere que la permanencia en el mercado y la solvencia son condiciones que permiten sostener o crear puestos de trabajo.

Para complementar el análisis de feature importance e ir más allá del simple peso de cada variable, se utilizó la técnica SHAP (SHapley Additive exPlanations). SHAP descompone la predicción del modelo en "aportantes" individuales por variable, indicando no solo qué tan importante es cada una, sino también en qué dirección empuja la probabilidad: valores positivos aumentan la probabilidad de que la empresa sea clasificada en la clase 1 (genera empleo) y valores negativos la reducen. En este caso, aplicamos SHAP sobre el modelo Random Forest final para analizar con detalle el comportamiento de tres variables categóricas tipo "dummy": CAT_EMPL_1-3 y TAMANO_MICRO.

Los resultados muestran patrones interesantes. Para CAT_EMPL_1-3, el SHAP medio es negativo cuando la variable vale 1 y positivo cuando vale 0, lo que indica que pertenecer a la categoría de empresas con 1 a 3 empleados reduce la probabilidad estimada de generar empleo

frente a las empresas que se encuentran en otros rangos de tamaño (por ejemplo, 4–10 o 11–50 empleados).

Algo similar ocurre con TAMANO_MICRO: cuando la empresa es micro (valor 1), el efecto promedio es negativo, mientras que no ser micro (valor 0) empuja la predicción hacia la clase generadora de empleo; esto es consistente con la idea de que las empresas de mayor tamaño tienen más capacidad para sostener o ampliar su planta laboral.

Este es un activo valioso para la Secretaría de Desarrollo Económico en la medida en que ratifica que las estrategias que deben plantear para lograr una mayor y más efectiva generación de empleo deben enfocarse en preservar las empresas para que generen más activos y perduren en el tiempo.

12.5 Implementación del modelo

Una vez seleccionado el modelo Random Forest como la alternativa más sólida y equilibrada para la predicción de generación de empleo empresarial —en coherencia con las necesidades del negocio y con el análisis comparativo de desempeño— se procedió a su implementación operacional. Para ello, el modelo final que fue entrenado utilizando el conjunto de entrenamiento balanceado mediante SMOTE, se consolidó en un archivo de persistencia en formato .pkl (pickle). Este formato es el estándar en Python para almacenar objetos complejos como modelos de machine learning, ya que permite conservar tanto la estructura del modelo como los parámetros aprendidos durante el entrenamiento, garantizando su reutilización en procesos futuros sin necesidad de volver a entrenarlo.

Con el modelo ya almacenado en formato .pkl, se implementó el modelo en la Base Maestra de Priorización que será entregada a la Secretaría, de modo que cada empresa que

cumpliera con el requisito de haber generado empleo en por lo menos 2 de los últimos 5 años objeto de observación se le asignó un Clúster al cual pertenece y se le asignó una probabilidad de generar empleo para el año posterior.

13. Base Maestra de Priorización

Esta herramienta contiene las empresas activas en el 2024 en Cali que generaron empleo en por lo menos 2 de los últimos 5 años de observación (2020 – 2024). La información que se relaciona con cada empresa es la reportada por las mismas en el marco del proceso de Registro Mercantil realizado durante el año 2024.

Dentro de las variables que componen esta base se encuentran el NIT, Razón Social, Fecha de Matrícula, Antigüedad, Representante Legal, Genero del Representante Legal, Dirección Comercial, Comuna, Clúster Comuna Teléfono, Email, Tamaño, Sector, Código CIIU, Descripción del Código CIIU, Activos, Pasivos, Ingresos Operacionales, Utilidad Bruta, Utilidad Perdida, Patrimonio, Apalancamiento, Margen Rentable Establecimientos, Categoría de Establecimientos, Empleados 2020, Empleados 2021, Empleados, 2022, Empleados 2023, Empleados 2024, Categoría empleados 2024, Variación de Empleados 2020-2024, Años Crea Empleo, Clúster Empresarial y Probabilidad de Generar Empleo.

Esta herramienta será compartida con la Secretaría como insumo para la toma de decisiones de focalización empresarial, contribuyendo a orientar los esfuerzos institucionales hacia unidades productivas con mayor probabilidad de aportar al cumplimiento de la meta distrital de generación de 100.000 empleos.

Figura 16

Visualización de Base Maestra de Priorización que será entregada a la Secretaría como herramienta para la toma de decisiones.

NIT	RAZON SOCIAL	CLÚSTER	% GENERAR EMPLEO	FECHA MATRICULA	ANTIGUEDAD
890300625	COOPERATIVA MEDICA DEL VALLE Y DE PROFESIONALES D	Empresa Estrella	74%	1997	28
901491488	BRANDS CLOTHING S.A.S.	Empresa Pequeña Escala	57%	2021	4
900406150	BANCO COOMEVA S.A.	Empresa Estrella	74%	2011	14
890320910	GARCES EDER S.A.S.	Empresa Estrella	78%	1981	44
890300466	TECNOQUIMICAS S.A.	Empresa Estrella	71%	1957	68
900638487	BANCA Y FINANZAS AGENTES FINANCIEROS S.A.S	Empresa Estrella	52%	2013	12
901033307	MARLIT GROUP S.A.S.	Empresa Pequeña Escala	63%	2016	9
900783257	INAGROMIN S.A.S.	Empresa Estrella	65%	2014	11
891300237	INGENIO DEL CAUCA S.A.S.	Empresa Estrella	82%	1993	32
900378212	BANCO W S.A.	Empresa Estrella	74%	2010	15
805004034	COOPERATIVA DE SERVIDORES PUBLICOS & JUBILADOS DE	Empresa Estrella	71%	1996	29
800197463	POLLOS EL BUCANERO S.A.	Empresa Estrella	78%	1993	32
805027970	COMERCIALIZADORA GIRALDO Y GOMEZ Y CIA S.A	Empresa Estrella	50%	2003	22
900368512	SOLUCIONES DE TELECOMUNICACIONES Y COMPUTO S.A.S	Empresa Estrella	57%	2010	15
890302594	MAYAGUEZ S.A.	Empresa Estrella	74%	1950	75

Fuente: Elaboración propia.

14. Conclusiones

El presente proyecto empresarial tuvo como propósito central estudiar, mediante herramientas avanzadas de analítica, las dinámicas de generación de empleo de las empresas de Santiago de Cali entre 2020 y 2024, con el fin de dotar a la Secretaría de Desarrollo Económico de una solución robusta que permita orientar la toma de decisiones, focalizar sus programas y contribuir al cumplimiento de la meta distrital de creación de 100.000 nuevos empleos.

El análisis desarrollado permitió verificar la hipótesis analítica propuesta: las características empresariales contienen señales valiosas sobre el comportamiento futuro de las firmas, pero no son las únicas determinantes de su capacidad para generar empleo. Los modelos

predictivos muestran que las características empresariales disponibles en los datos explican una parte importante del comportamiento laboral, aunque el margen de mejora que aún tienen estos modelos se traduce a que existen otros elementos —como decisiones internas, estrategias corporativas o shocks externos— que también influyen en la generación de empleo pero no están reflejados en los datos objeto de este proyecto.

En primera instancia, el proyecto confirmó que existen patrones descriptivos consistentes entre las empresas que históricamente han sido constantes en la generación de empleo. Variables como antigüedad, tamaño, estabilidad financiera y comportamiento territorial mostraron regularidades que permiten diferenciar 4 grupos de generación de empleo. Esta evidencia constituye un aporte relevante para la Secretaría, pues establece un primer mapa estructural del tejido empresarial asociado a la producción de empleo en la ciudad.

Finalmente, el modelo predictivo aportó una herramienta adicional a la Secretaría, identificar, mediante las características empresariales registradas para ese periodo por la empresa, la capacidad de la misma en generar empleo en el periodo posterior. Esta herramienta tiene una aplicación directa a la hora de asignar las plazas o cupos que pueda ofrecer un programa de incentivos empresariales. En consecuencia, el modelo predictivo no pretende ser una herramienta determinista, sino un instrumento probabilístico que orienta decisiones pero no reemplaza el juicio técnico de la Secretaría.

En suma, este proyecto no solo cumple su objetivo central, sino que aporta una herramienta conceptual, metodológica y operativa para que Cali avance hacia una política de desarrollo económico basada en evidencia, con capacidad de anticipación y con foco estratégico en el empleo como eje de prosperidad social.

15. Recomendaciones estratégicas para la Secretaría de Desarrollo Económico

La información generada a partir del análisis descriptivo, la segmentación empresarial y el modelo predictivo constituye una base sólida para que la Secretaría de Desarrollo Económico fortalezca su capacidad de toma de decisiones, focalización y planificación estratégica. A continuación se presentan las recomendaciones más relevantes derivadas de los hallazgos:

15.1 Adopción del perfil de "empresa estrella" como como eje de focalización estratégica

El análisis de clúster permitió identificar un grupo de empresas (Clúster 1) que se destacan por su desempeño superior y aportes consistentes a la generación de empleo. Estas organizaciones presentan elementos comunes: madurez empresarial, solidez financiera, niveles adecuados de activos, estabilidad sectorial y trayectorias laborales crecientes. Este conjunto se erige como un segmento prioritario debido a su potencial de contribuir de manera acelerada a la meta distrital de creación de 100.000 empleos.

Acciones sugeridas

- Incorporar el perfil de empresa estrella como criterio de elegibilidad en los nuevos programas de fortalecimiento empresarial, de forma que la focalización responda a características empíricamente asociadas con la generación de empleo.
- Desarrollar mesas de diálogo y co-creación con las empresas que integran este clúster, con el fin de identificar necesidades comunes y oportunidades de escalamiento que puedan ser atendidas con incentivos focalizados.
- Diseñar instrumentos diferenciales de acompañamiento (líneas de crédito, asistencia técnica, programas de productividad) que permitan impulsar la consolidación y expansión de este grupo estratégico.

15.2 Desarrollo de programas diferenciados para empresas generadoras de empleo en riesgo

El Clúster 4 agrupa empresas que, pese a generar empleo de forma significativa, presentan señales de riesgo financiero y operativo: altos niveles de apalancamiento y márgenes rentables bajos. Este perfil las expone a vulnerabilidad en su sostenibilidad y, por ende, al riesgo de destrucción de empleo en el corto plazo.

Acciones sugeridas

- Establecer un programa de intervención temprana dirigido a empresas del Clúster 4, priorizando acciones de saneamiento financiero, optimización operativa y asesoría estratégica para mitigar riesgos de cierre o reducción de planta laboral.
- Articular estos hallazgos con programas existentes de la Secretaría, como líneas de microcrédito o alivios empresariales, para priorizar cupos y socializar fechas de postulación y beneficios a las empresas clasificadas como “empresas en riesgo”.
- Implementar un sistema de monitoreo trimestral que permita seguir la evolución de sus indicadores financieros y laborales, anticipar deterioros y activar alertas preventivas.

15.3 Estrategia de asignación diferenciada de cupos de fortalecimiento empresarial según el perfil de clúster

Con el fin de optimizar el impacto de los programas de fortalecimiento empresarial, se propone que la Secretaría de Desarrollo Económico asigne ante nuevos programas generales de fortalecimiento empresarial los cupos disponibles basada en la composición de los clústeres identificados y en la distribución observada de la probabilidad de generar empleo al interior de cada grupo.

Para efectos del análisis y la posterior construcción de estrategias diferenciadas, se establecieron cuatro niveles de generación de empleo (Muy Alto, Alto, Medio y Bajo), definidos en función del porcentaje arrojado por el modelo predictivo desarrollado en la pregunta de análisis 3. Las categorías quedaron definidas de la siguiente manera:

- Muy alto: Superior a 70%
- Alto: Entre 50% y 70%
- Medio: Entre 30% y 50%
- Bajo: Menor a 30%

Se excluyen de esta priorización las Empresas Ancla (3,8%), dado que presentan alto desempeño en su capacidad de generación de empleo. En consecuencia, los cupos se sugieren distribuir entre tres segmentos estratégicos: Empresas Estrella (25,6%), Empresas Pequeña Escala (63,2%), y Empresas en Riesgo (7,5%).

Tabla 7

Distribución del nivel de generación de empleo por Clúster empresarial.

Grupo % Generación de Empleo	Empresas Estrella	Empresas Pequeña Escala	Empresas en Riesgo
Muy Alto	20,9%	1,2%	2,9%
Alto	56,4%	50,2%	28,2%
Medio	22%	42,9%	59,8%
Bajo	0,5%	5,5%	8,9%

Fuente: Elaboración propia.

Acciones sugeridas de acuerdo con el análisis

- Destinar el 40% de los cupos de fortalecimiento al segmento de Empresas Estrella, teniendo en cuenta el aporte mayoritario que aportan al número de empleos generados.
- Asignar el 50% de los cupos, al segmento de Empresas Pequeña Escala, teniendo en cuenta que representan mayoritariamente el tejido empresarial y que tienen una probabilidad de generación de empleo alta muy representativa.
- Asignar el 10% restante a las Empresas en Riesgo, esto toda vez que representan una cantidad menor de empresas, su probabilidad de generación de empleo se sitúa mayoritariamente en nivel Medio pero teniendo siempre presente la recomendación número 2 consistente en la intervención uno a uno de estas empresas.

15.4 Estrategia de robustecimiento operativo de la Secretaría

Además de las estrategias diferenciadas por clúster, el análisis evidencia la necesidad de fortalecer las capacidades operativas internas de la Secretaría de Desarrollo Económico para asegurar la sostenibilidad, trazabilidad y eficacia de los programas de fortalecimiento

empresarial. Las siguientes acciones sugeridas consolidan un marco operativo mínimo para avanzar hacia una gestión institucional basada en datos:

Acciones sugeridas

- Fortalecer los registros empresariales propios de la Secretaría y documentar sistemáticamente su interacción con las empresas.

Se recomienda desarrollar un repositorio unificado que capture toda la relación institucional con cada empresa: programas en los que participa, cupos asignados, resultados obtenidos, alertas generadas y requerimientos adicionales. Este registro permitirá mejorar la trazabilidad de la intervención pública y evitar duplicidades o vacíos de información entre programas, proyectos y equipos internos.

- Registrar el estado inicial de las empresas intervenidas y realizar seguimiento a sus indicadores de empleo cada seis meses.

El levantamiento de una línea base operativa es indispensable para evaluar el impacto real de las intervenciones. Este registro debe incluir variables clave como número de empleos formales, estructura financiera, niveles de riesgo y principales brechas productivas. El seguimiento semestral permitirá identificar avances, retrocesos y necesidades emergentes, facilitando la activación de intervenciones oportunas.

- Construir procedimientos y políticas de calidad de datos que integren las cifras y reportes capturados por la Secretaría.

Dada la diversidad de fuentes y formatos, se requiere establecer estándares operativos de calidad, actualización, consistencia y asignación de responsabilidades sobre los datos. Esto incluye lineamientos para validación, depuración, imputación, interoperabilidad y gobernanza,

de forma que los registros sean comparables en el tiempo y útiles para análisis avanzados, evaluación de impacto y toma de decisiones.

- Desarrollar una Base Maestra de Empresas en Cali que permita integrar fuentes externas como DIAN, Seguridad Social y registros sectoriales.

La construcción de una Base Maestra robusta permitirá consolidar un "historial empresarial" confiable que articule información tributaria, laboral, financiera y de programas públicos. Esta integración facilitará la identificación de perfiles empresariales, la detección temprana de riesgos y la focalización precisa de los programas, garantizando que las decisiones de política sean oportunas, basadas en evidencia y alineadas con las metas distritales de generación de empleo.

16. Pasos a seguir

Los resultados obtenidos en este proyecto permiten a la Secretaría de Desarrollo Económico dar un salto cualitativo hacia una gestión basada en evidencia, incorporando herramientas descriptivas, segmentación empresarial y analítica predictiva para orientar de manera más efectiva la política pública de empleo. No obstante, para garantizar la sostenibilidad, actualización y aprovechamiento pleno de estos insumos, se recomienda avanzar en una serie de pasos estratégicos que consolidan y escalan el trabajo desarrollado. A continuación, se presentan las acciones prioritarias sugeridas.

16.1 Construcción e implementación de un dashboard de consulta de la Base de Priorización

La creación de un dashboard interactivo, alojado en una plataforma corporativa (p. ej., Looker Studio dado que la Alcaldía utiliza mayoritariamente entornos de Google), permitirá que los funcionarios de la Secretaría accedan de manera ágil y visual a la Base de Priorización empresarial. Este tablero deberá incluir:

- Consulta por NIT, sector, comuna, clúster y nivel de riesgo.
- Visualizaciones sobre la probabilidad de generación de empleo y variables críticas del modelo.
- Indicadores anuales que faciliten el seguimiento de empresas priorizadas y su comportamiento laboral.

Con este instrumento, la Secretaría podrá incorporar la analítica en la operación cotidiana, fortaleciendo la transparencia, la trazabilidad de decisiones y la capacidad de respuesta ante necesidades del tejido empresarial.

16.2 Robustecimiento anual del modelo predictivo y construcción de un pipeline automatizado

Para asegurar la vigencia del modelo y su capacidad de capturar cambios estructurales en la dinámica económica local, se recomienda:

- Incorporar nuevas variables provenientes de fuentes administrativas adicionales, como por ejemplo reportes de cumplimiento tributario.
- Construir un pipeline de transformación, modelación y scoring que pueda ser aplicado a las bases de Registro Mercantil que llegan a la Secretaría anualmente.
- Monitorear métricas de desempeño (ROC-AUC, F1, precisión y estabilidad de características) para garantizar que el modelo se mantenga confiable y útil para las decisiones de política pública.

La continuidad del pipeline garantizará que la Secretaría cuente cada año con un “mapa de potencial de generación de empleo” actualizado y técnicamente sustentado.

16.3 Desarrollo de un proyecto de evaluación de impacto de programas e incentivos de la Secretaría

Con una base empresarial robusta y modelos predictivos implementados, el siguiente paso estratégico consiste en evaluar la efectividad real de los programas e incentivos otorgados por la Secretaría. Se recomienda iniciar un proyecto de evaluación de impacto que:

- Analice si las empresas beneficiarias incrementan su generación de empleo más que empresas similares no beneficiarias.
- Identifique cuáles instrumentos (créditos, asistencia técnica, formación, incentivos sectoriales) producen mayor retorno en términos de empleo.
- Permita reorientar recursos hacia las intervenciones con mayor eficacia demostrada, aumentando el impacto social y económico.

Este ejercicio consolidará un ciclo de gestión basado en evidencia, conectando diagnóstico, focalización, intervención y evaluación.

Los pasos a seguir aquí planteados constituyen la ruta para institucionalizar la analítica dentro de la Secretaría, asegurando que los hallazgos del proyecto no solo informen decisiones actuales, sino que se conviertan en herramientas permanentes de gestión pública. Con la construcción del dashboard, la actualización del pipeline predictivo y el diseño de un proyecto riguroso de evaluación de impacto, la Secretaría avanza hacia un modelo de gobernanza económica más técnico, eficiente y orientado a resultados, contribuyendo de manera tangible al cumplimiento de la meta distrital de creación de 100.000 empleos entre 2024 y 2027.

17. Referencias bibliográficas

Álban Comunes. (2024). *Pobreza extrema en Cali: balance y cifras actualizadas 2024*. https://albancomunes.com.co/?p=269

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321–357.

Clausen, A. W. (1983). *Promoting the private sector in developing countries: A multilateral approach* (Address by A. W. Clausen, President, The World Bank). World Bank. https://documents1.worldbank.org/curated/en/289631538080767870/pdf/Promoting-the-private-sector-in-developing-countries-a-multilateral-approach-address-by-A-W-Clausen.pdf

Decreto 0516 de 2016: Por el cual se definen las instancias de coordinación de la Administración Central Municipal. https://www.cali.gov.co/gobierno/publicaciones/132349/decreto-no-0516/

Departamento Administrativo Nacional de Estadística. (2024). *Cuentas nacionales departamentales*. DANE. https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales

Departamento Administrativo Nacional de Estadística. (2025). Boletín técnico: Mercado laboral – Primer trimestre de 2025. https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboralAlcaldía de Santiago de Cali. (2016).

Dolničar, S., Grün, B., & Leisch, F. (2018). *Market Segmentation Analysis*. Springer.

El País. (2024, junio 3). *Cali registró 535.000 personas en condición de pobreza monetaria durante 2024: este es el panorama*. https://www.elpais.com.co/economia/cali-registro-535000-personas-en-condicion-de-pobreza-monetaria-durante-2024-este-es-el-panorama-2548.html

Haltiwanger, J., Jarmin, R. S., & Miranda, J. (2013). *Who creates jobs? Small versus large versus young*. The Review of Economics and Statistics, 95(2), 347–361.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

International Labour Organization (OIT). (2018). *World Employment and Social Outlook*. ILO.

Jain, A. K. (2010). *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8), 651–666.

Pagés, C., & Madrigal, L. (2019). *Productividad y brechas estructurales en América Latina*. Banco Interamericano de Desarrollo.

Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies, 2(1), 37–63.

Presidencia de la República de Colombia. (2024, noviembre 7). *Presidente Petro presentó balance de la COP16 y sostuvo que la conferencia de Cambio Climático de Azerbaiyán debe avanzar hacia la soberanía energética justa*. https://www.presidencia.gov.co/prensa/Paginas/Presidente-Petro-presento-balance-de-la-COP16-y-sostuvo-que-la-conferencia-de-Cambio-Climatico-de-Azerbaiyan-241107.aspx