Maestría en Ingeniería de Información - MINE4101
Proyecto 2024-2

Ciencia de datos aplicada
Proyecto final – Segunda entrega

Segmentación de compañías para ventas de plataforma EMIS

Integrantes:

- Juan Daniel Castrellón (201729285)
- Kevin Camilo Becerra Walteros (201812779)
- Laura Andrea Roncancio Pava (201815149)
- Javier Alejandro Gómez Muñoz (201217975)

1. Definición de la problemática y entendimiento del negocio

ISI Markets es el proveedor líder de datos macroeconómicos, industriales, empresariales y de inteligencia soberana. ISI Markets tiene la fuente más completa de datos confiables y conocimientos prácticos sobre mercados emergentes y desarrollados.

El grupo está compuesto por cuatro marcas (CEIC, EPFR, EMIS y REDD) que comparten un propósito común de iluminar la oportunidad en mercados emergentes y desarrollados donde la información puede ser difícil de obtener e incluso más difícil de confiar.

Teniendo en cuenta lo anterior, el proyecto se basará en datos de EMIS correspondientes a 10.000 compañías, que tienen operaciones en la ciudad de Bogotá.

El objetivo de las compañías al utilizar EMIS, es tener una base sólida para la toma de decisiones, según prospecciones y riesgos, para generar ventas a otras compañías que puedan generar ganancias significativas.

La problemática actual consiste en que el volumen de datos que deben analizar los equipos de ventas es muy alto y les toma mucho tiempo, incluso más del tiempo que disponen. Por lo anterior, la solución que se plantea es una categorización de las compañías y un dashboard en el cual se pueda filtrar su salud financiera y sus características para encontrar las óptimas y así disminuir el tiempo de búsqueda y tener resultados más rentables, lo que se traduce en un aumento de valor para EMIS y sus clientes.

Objetivos del proyecto:

- Desarrollar un producto de datos en forma de un dashboard que permita al equipo comercial identificar rápidamente las compañías potenciales como clientes en Bogotá con buena salud financiera y abiertas a negocios, optimizando así las campañas de ventas y estrategias comerciales.
- Reducir el tiempo de análisis de la base de datos mediante técnicas de análisis automatizado.
- Facilitar la toma de decisiones del equipo comercial con información clara y accionable.

Variables:

- Ingreso total por operación.
- Ratio financiero (ROA).
- Ratio de rentabilidad (ROE).

- Efectivo y equivalentes.
- Exportaciones e importaciones.
- Tendencia de ingresos netos por ventas.
- Propiedad, planta y equipamiento.
- Deuda.
- Número de empleados.
- Tendencia del beneficio operativo.

2. Ideación

El producto será un dashboard de inteligencia empresarial que permita al equipo comercial identificar y priorizar empresas en Bogotá con buena salud financiera y potencial de negocio. El objetivo es facilitar la toma de decisiones con datos financieros claros y filtros interactivos que optimicen el proceso de segmentación.

2.1 Potenciales usuarios y procesos actuales

Potenciales principales usuarios

- **Equipo Comercial:** Ventas y desarrollo de negocios. Requieren identificar empresas abiertas a negocios para priorizar contactos.
- **Business Manager:** Responsable de decisiones estratégicas y alianzas comerciales.

Procesos actuales

- **Equipo Comercial:** Segmenta manualmente 10.000 empresas. Sin un análisis automatizado, priorizar se vuelve complicado y consume demasiado tiempo.
- **Business Manager:** Toma decisiones estratégicas con base en múltiples reportes dispersos, lo que complica la planificación eficiente de alianzas.

Dolores actuales

- Tiempo excesivo para segmentar 10.000 empresas manualmente.
- Dificultad en la claridad de la salud financiera de los prospectos de clientes.
- Acceso a la información, dificultando decisiones rápidas.

Beneficios del producto de datos

- Automatizar la segmentación de empresas con base en KPIs financieros.
- Ofrecer filtros dinámicos para encontrar empresas con alto potencial de negocio en menor tiempo.
- Visualización clara de las empresas más rentables y abiertas a colaboraciones.

2.2 Requerimientos del producto de datos

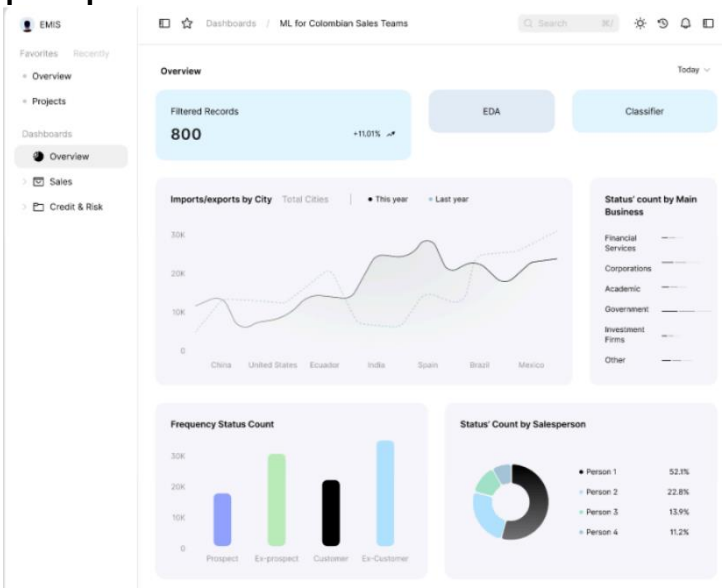
- Filtros por sector (NAICS/CIIU), ingresos anuales, deuda y ROA/ROE.
- Visualización del segmento al que pertenece una compañía, a partir del modelo de segmentación.

- **<u>Indicadores clave</u>**
  - Total Operating Revenue
  - Net Sales Revenue Trend

- Return on Assets (ROA) y Return on Equity (ROE)

- **Visualizaciones**
  - Gráfico de líneas: Tendencia de ingresos netos por ventas.
  - Gráfico de barras: Comparación del EBITDA entre empresas.
  - Gráfico de pastel: Distribución de empresas por sector (NAICS).
- **Tabla Interactiva**
  - Detalle por empresa: Nombre, ingresos, sector, deuda, ROA, y ROE.
  - Iconos con enlaces para acceder a más información financiera o reportes detallados.
- **Exportación de Reportes**
  - Opción de descargar análisis en PDF/Excel.
  - Botón de exportación rápida para preparar listas de contacto del equipo comercial.

2.3 Mockup del producto



2.4 Componentes tecnológicos del producto

- **Backend:** Python para procesamiento y transformación de datos.
- **Visualización y Frontend:** Streamlit o Dash (Python): Alternativas si se busca un desarrollo más personalizado.
- **Integración y Automatización:** ETL Pipeline en Python para EMIS.

2.5 Mockup del flujo del usuario

- **Inicio del usuario**

El equipo comercial accede al dashboard y selecciona sus filtros iniciales (sector, rango de ingresos y ubicación, etc.).

- **Visualización del análisis**

Se muestran gráficas de tendencias y tablas interactivas con las empresas más relevantes; el usuario puede ordenar los resultados por ROA, ROE, o deuda para identificar las mejores oportunidades.
En esta visualización se muestra el segmento al que la empresa pertenece.

- **Toma de Decisiones**

Con los resultados, el usuario exporta un reporte en formato PDF con las empresas seleccionadas. El equipo comercial prioriza los contactos según el análisis realizado.

3. Responsible

3.1. Implicaciones éticas

- **Sesgos en los modelos de IA:** Los algoritmos de inteligencia artificial pueden estar sujetos a sesgos si los datos de entrada no son representativos. Por ejemplo, si se usan datos incompletos o desactualizados, se podrían excluir empresas con potencial de crecimiento.
- **Impacto en la competencia y transparencia:** Si la información analizada se comparte sin control, se podría facilitar un acceso desigual a datos estratégicos, beneficiando solo a algunas empresas.

3.2. Privacidad y confidencialidad

- **Protección de datos sensibles:** En el análisis financiero, es posible manejar información confidencial sobre la situación económica de empresas. El mal uso o filtración de esta información puede perjudicar la reputación o competitividad de las empresas involucradas.
- **Confidencialidad comercial:** La salud financiera de una empresa es un activo sensible. Si se difunde sin autorización, podría afectar su posición en el mercado.

3.3. Aspectos regulatorios

Se debe asegurar de que el uso de datos cumpla con las regulaciones locales y globales, como:

- **Ley 1581 de 2012 en Colombia:** Regula la protección de datos personales.
- **ISO 27001:** Estándar para la gestión de seguridad de la información.
- **Política de Habeas Data:** Derecho que tienen las personas naturales y jurídicas de conocer, actualizar y rectificar información personal y financiera que se maneja sobre ellas.

3.4. Transparencia

- **Explicabilidad de los Modelos de IA:** El equipo de ventas debe ser capaz de entender cómo se seleccionaron las empresas abiertas a negocios. Si los modelos son cajas negras (black boxes), podría existir falta de confianza en las recomendaciones.

4. Enfoque analítico

Dado que se busca realizar un análisis general, se tienen varias preguntas de negocio:

- ¿Qué empresas en Bogotá tienen ingresos operativos anuales mayores a 1 millón de dólares y podrían ser clientes potenciales?
- ¿Qué sectores muestran mayores ingresos y pueden ser prioritarios para nuevas alianzas?
- ¿Cuáles son las empresas con un ROA y ROE superiores al promedio del mercado, indicando buena gestión de recursos y rentabilidad?
- ¿Qué porcentaje de empresas posee activos líquidos significativos, como efectivo y equivalentes, que indiquen capacidad para responder a obligaciones inmediatas?

- ¿Qué industrias o sectores presentan las mejores oportunidades para cerrar acuerdos comerciales, considerando su rentabilidad y crecimiento reciente?
- ¿Cuáles son las empresas con un número significativo de empleados, lo que puede indicar mayor capacidad operativa para proyectos conjuntos?
- ¿Cuáles empresas han reducido su deuda a corto o largo plazo, mejorando su capacidad de negociación?
- ¿Qué compañías presentan niveles de endeudamiento preocupantes que podrían limitar su apertura a nuevos negocios?

Las **métricas** que se usarán para evaluar la calidad del modelo son:

- **Tasa de oportunidad de los prospectos:** Mide el porcentaje de prospectos que muestran interés después del primer contacto comercial. Su propósito es evaluar qué tan bien los clústeres generados identifican empresas con interés real en nuevas oportunidades.
- **Tasa de intensión de los prospectos:** Mide el porcentaje de prospectos que interactúan con el equipo comercial tras la primera comunicación. Su propósito es validar la efectividad del análisis en identificar empresas dispuestas a negociar.
- **Coste por lead calificado:** Mide el costo de trabajo del equipo de ventas (salarios y recursos) por cada prospecto calificado. Su propósito es evaluar la eficiencia del proceso de priorización de empresas para optimizar recursos.

Técnicas de Machine Learning y Estadísticas

- **PCA (Análisis de Componentes Principales):**
  - **Propósito:** Reducir la dimensionalidad del dataset, identificando patrones financieros clave que faciliten la segmentación.
- **Clustering No Supervisado:**
  - **PCA + K-means:** Agrupa las empresas según características similares en términos de ingresos, deuda y rentabilidad.
  - **PCA + HDBSCAN:** Detecta clústeres basados en densidades, útil para encontrar grupos de empresas con patrones más complejos o irregulares.
  - **PCA + Hierarchical Clustering:** Detecta patrones latentes complejos mediante una combinación de deep learning y clustering no supervisado.

5. Recolección de datos

Para llevar a cabo este análisis, se empleó EMIS NEXT, una plataforma especializada que facilita el acceso a datos financieros y corporativos de diversas empresas. Esta herramienta permite una recolección de información precisa mediante filtros avanzados, como el perfil de *Company Screener*, que en este caso se configuró con los siguientes criterios:

- Ubicación: Empresas localizadas en Bogotá, Colombia.
- Disponibilidad de Cuentas Financieras: Se incluyeron únicamente las empresas con cuentas financieras completas, lo que permite obtener información clave como ingresos, activos, pasivos y otros elementos contables.

La información obtenida se estructura en un formato tabular donde cada fila representa una empresa y cada columna, un atributo o indicador relevante. Estos datos incluyen tanto los elementos de identificación de la empresa (nombre, sector, ubicación) como sus principales cuentas financieras. Entre estas cuentas

se encuentran los ingresos, costos, activos, pasivos y patrimonio, lo cual es fundamental para realizar un análisis comparativo efectivo.

Utilidad de los datos

Esta información proporciona una base sólida para la prospección de clientes al permitir:

- Análisis Comparativo: Contrastar la situación financiera de diferentes empresas en Bogotá.
- Evaluación de Indicadores: Examinar métricas clave de desempeño financiero como el margen de utilidad y el retorno sobre activos.
- Segmentación por Sector: Clasificar a las empresas por sector económico, lo cual permite identificar patrones específicos y tendencias que facilitan la segmentación de prospectos.

Diccionario de datos

Atributo	Tipo	Descripción
Total operating revenue	Float	Ingresos operativos totales de la empresa en el periodo más reciente.
Industry (NAICS)	String	Código de clasificación industrial de la empresa según el sistema NAICS.
Import	String	Importaciones.
Export	String	Exportaciones.
Property, plant, equipment	Float	Valor total de las propiedades, planta y equipos de la empresa.
Cash and cash equivalents	String	Total de efectivo y equivalentes de efectivo disponible en la empresa.
Net sales revenue trend	Float	Tendencia de crecimiento o decrecimiento en las ventas netas en el periodo evaluado.
Operating profit trend	Float	Tendencia de crecimiento o decrecimiento en las ganancias operativas de la empresa.
Number of Employees	String	Cantidad de empleados en la empresa.
Return on Equity (ROE) (%)	Float	Rendimiento sobre el capital propio de la empresa, expresado como porcentaje.
Return on Assets (ROA) (%)	Float	Rendimiento sobre los activos de la empresa, expresado como porcentaje.
Short Term Debt	Float	Total de deuda a corto plazo de la empresa.
Long Term Debt	Float	Total de deuda a largo plazo de la empresa.
Quick Ratio (x)	Float	Ratio rápido (Quick Ratio) de la empresa, que mide su liquidez inmediata.

6. Entendimiento de los datos

Se realizó un entendimiento general de los datos mediante técnicas de análisis univariado y multivariado según los intereses del negocio en cuestión.

- **Análisis univariado**
  - Property, plant and equipment:

Esta es una variable numérica la cual tiene la distribución que consiste en una concentración de datos pequeños, y ciertos outliers a la derecha, es decir, empresas que tienen valores significativamente altos comparados con el resto de las empresas. Las estadísticas indican que el valor mínimo es de $287.55 y el máximo de alrededor de $95.000.000, siendo el percentil 75% $3500, es decir, el 75% de las empresas tienen

menos de 3500 USD en activos de propiedad, planta o equipo. En total, este campo tiene 5683 datos, es decir, poco más de la mitad de las empresas obtenidas no tienen el valor en nulo.

- Cash and cash equivalents:

La distribución de los datos relacionados con el cash y cash equivalents tiene una distribución que también cuenta con colas muy largas. El valor mínimo es de 0 USD, y el máximo de alrededor de 20.000.000 USD, con un percentil 75% de 6000 USD, que recalca la existencia de outliers en la cola derecha de la distribución.

- Net sales revenue trend:

Nuevamente nos encontramos con una distribución sesgada a la derecha con valores extremos altos. Los valores mínimos de la misma son de –100%, siendo la empresa con mayor Net Revenue de alrededor de 1.400.000$. La mediana de esta se encuentra en 12.75%, lo cual resalta la alta cantidad de valores extremos a la derecha.

- Operating profit trend:

Acá nos encontramos con una variable que tiene muchos valores extremos tanto a la derecha como a la izquierda de la distribución. Sin embargo, cabe notar que existen outliers más alejados de la distribución central a la derecha que a la izquierda. En este caso, el valor mínimo es de –500.000% mientras que el valor máximo es de 2.000.000%, donde el percentil 50 tiene un valor de –1%.

- Importaciones y exportaciones:

Para el análisis de importaciones y exportaciones, se seleccionaron el top 10 de países que más aportan al total de importaciones y exportaciones en empresas en Colombia, y se seleccionó que tanto porcentaje del producto de esa empresa en cuestión se exporta al exterior, o que tanta materia prima o productos se importan en cuestión. Acá nos encontramos con grandes potencias como lo son China y Estados Unidos. Mientras China representa el socio más importante para importaciones a Colombia, Estados Unidos es el país que más representa exportaciones para el país. Asimismo, entre los países importantes encontramos todos los países que comparten frontera con Colombia, a excepción de Venezuela. Además, a esta lista se suman Italia, Alemania y España como representantes del continente europeo que tienen relaciones comerciales con empresas colombianas.

- NAICS:

Para el análisis de NAICs, se seleccionó la primera actividad comercial que estaba en la columna de NAICS y se realizaron diferentes análisis, principalmente de carácter multivariado, sobre las mismas.

En primer lugar, se obtuvieron los 15 NAICS más importantes en Bogotá, donde encontramos que el 3.3% de las empresas radicadas en Bogotá se dedican a la venta de sistemas de computadores y Software, siendo el NAIC más relevante en la ciudad. Asimismo, también nos encontramos con una alta cantidad de NAICS relacionados con constricciones, tal como la venta de maquinaria, construcción de puentes y carreteras, venta de materiales de construcción, construcción de edificios residenciales, entre otras, por lo que este sector también es de importancia dentro del contexto bogotano. Por otro lado, también encontramos otros NAICs relacionados con industria farmacéutica, venta de bienes y servicios o transporte, que también representan una actividad significativa dentro de la realidad Bogotana.

- Long y short term debt:

Estas es una variable nuevamente que tiene muchos valores outliers en la cola de la derecha. Asimismo, al entender mejor la variable, nos damos cuenta de que la mayoría de las empresas, más del 75% cuentan con una deuda a largo plazo de 0 USD, y existen pocas empresas que cuentan con deudas tan altas a largo plazo. El mismo fenómeno se cumple con la deuda a corto plazo, siendo que entre un 50 y un 75% de las empresas no cuentan con deuda a corto plazo. Esta variable puede ser de interés, pues puede ayudar

al equipo a comprender y segmentar que empresas son más propensas a contar con deuda.

- Quick ratio:

La variable representada en el boxplot del Quick Ratio (x) muestra una gran cantidad de valores atípicos en la cola derecha. Observamos que la mayoría de las empresas tienen un Quick Ratio bajo, concentrándose en un rango pequeño, mientras que unas pocas presentan valores significativamente elevados, alcanzando hasta 3500. Este patrón indica que la mayoría de las empresas mantienen un nivel de liquidez rápida controlado, mientras que algunas poseen valores excepcionalmente altos, lo que podría reflejar situaciones específicas de liquidez o estructuras financieras particulares. Esta variable es relevante para analizar, ya que permite identificar y segmentar aquellas empresas con características financieras únicas en términos de su capacidad de cubrir deudas a corto plazo.

- Total operating revenue:

La variable representada en el boxplot del Total Operating Revenue muestra, nuevamente, una alta concentración de valores en un rango bajo, con una cantidad considerable de valores atípicos hacia la derecha. La mayoría de las empresas tienen ingresos operativos totales reducidos, mientras que pocas alcanzan cifras muy altas, llegando hasta aproximadamente 1.4e8. Este fenómeno sugiere que solo una minoría de empresas genera ingresos operativos significativamente elevados en comparación con la mayoría, lo cual podría ser indicativo de un mercado concentrado o de empresas con características de operación únicas. Analizar esta variable puede proporcionar al equipo una comprensión de las diferencias en la generación de ingresos operativos y ayudar a segmentar las empresas según su capacidad de ingresos.

- ROA:

La variable representada en el boxplot del Return on Assets (ROA) (%) evidencia, una vez más, una alta concentración de valores en la parte baja, con numerosos valores atípicos extendiéndose hacia la derecha. La mayoría de las empresas presentan un ROA bajo, mientras que unas pocas alcanzan valores extremadamente altos, hasta alrededor de 250,000%. Este comportamiento sugiere que la mayoría de las empresas tienen un retorno moderado sobre sus activos, mientras que unas cuantas empresas muestran una eficiencia de uso de activos inusualmente alta, posiblemente debido a condiciones o eventos excepcionales.

- ROE:

La variable representada en el boxplot del Return on Equity (ROE) (%) muestra una amplia dispersión de valores, incluyendo outliers tanto en la cola negativa como en la positiva. La mayoría de las empresas tienen un ROE cercano a cero, mientras que algunas presentan valores extremadamente negativos, hasta alrededor de -600,000%, y otras, valores excepcionalmente positivos, cercanos a 200,000%. Esta distribución sugiere que existen empresas con rendimientos negativos importantes en relación con el patrimonio, lo cual podría reflejar pérdidas significativas o altos niveles de apalancamiento. Al mismo tiempo, algunas empresas tienen rendimientos muy positivos, indicando una alta eficiencia en la generación de beneficios con respecto al patrimonio.

- Employees:

La variable representada en el boxplot de Employees presenta una alta concentración de valores bajos, con una cola derecha extensa que indica la presencia de valores atípicos. La mayoría de las empresas cuentan con un número de empleados reducido, mientras que unas pocas alcanzan cifras mucho mayores, llegando hasta aproximadamente 70,000 empleados. Este patrón sugiere que existen algunas empresas considerablemente más grandes en términos de personal, mientras que la mayoría operan con plantillas pequeñas. Esta variable es relevante para segmentar las empresas según su tamaño y permite analizar las diferencias en la estructura de recursos humanos,

lo cual puede reflejar también variaciones en la capacidad operativa y la escala de las operaciones empresariales.

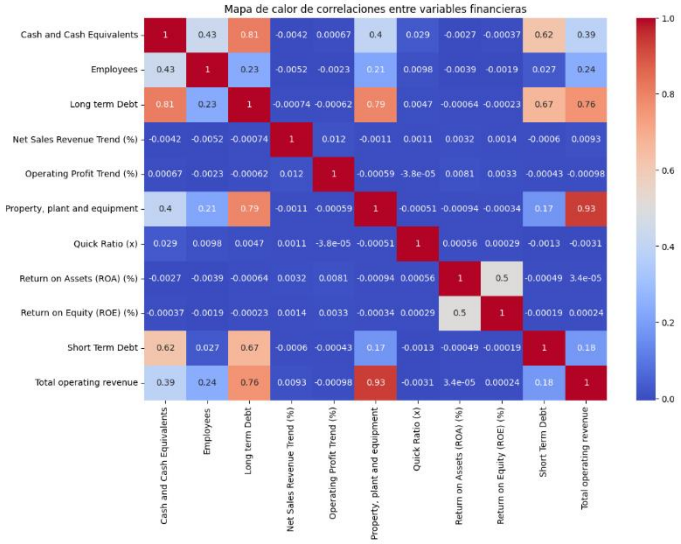
- **Análisis multivariado**

Debido a las necesidades expresadas por el negocio, es de interés entender principalmente el comportamiento de las empresas con que son outliers. Especialmente, queremos entender que sectores componen las mejores oportunidades para invertir. Por ello, el principal componente del análisis multivariado será realizar paretos para cada una de las variables numéricas respecto al NAIC. Con ello, es posible entender mejor los comportamientos de los diferentes sectores de Bogotá, además que permite comprender el estado actual de los mismos. A continuación, se presentan los resultados de los diferentes paretos realizados para este propósito.

- **Cash and Cash Equivalents**: Observamos que unos pocos sectores acumulan la mayoría del efectivo y equivalentes de efectivo. Sectores como las actividades financieras y de seguros dominan este recurso, lo que podría indicar una alta liquidez en estas industrias.
- **Employees**: La distribución de empleados muestra una concentración significativa en ciertas industrias, reflejando la naturaleza intensiva en mano de obra de sectores específicos como la manufactura y los servicios. Esto indica que estos sectores son grandes empleadores y su rendimiento puede tener impactos significativos en el empleo general.
- L**ong Term Debt**: El endeudamiento a largo plazo también está concentrado en pocos sectores, siendo el sector de generación de energía y las finanzas los que llevan la mayor carga. Esto sugiere un apalancamiento considerable en estos sectores, posiblemente debido a la naturaleza de los proyectos y el alto capital necesario.
- **Net Sales Revenue Trend**: Este gráfico de Pareto muestra que ciertos sectores dominan las tendencias de ingresos por ventas netas, lo que sugiere un crecimiento desigual en los ingresos de ventas en algunas industrias clave.
- **Operating Profit Trend**: Aunque es un gráfico con poca variación, muestra que la generación de energía eléctrica es el único sector que destaca en términos de tendencia de beneficios operativos, indicando una rentabilidad estable o en crecimiento dentro de este sector.
- **Property, Plant, and Equipment**: Finalmente, la inversión en propiedades, plantas y equipo está concentrada en sectores específicos como extracción de gas natural y transporte de petróleo. Esto es coherente con la necesidad de infraestructura intensiva en capital en estas industrias.
- **Quick Ratio (Q/R)**: La mayoría de los sectores presentan niveles bajos de liquidez rápida, con una acumulación significativa en la mitad inferior de la gráfica. Sin embargo, algunos sectores específicos, como la construcción y los servicios profesionales, dominan en términos de capacidad de pago inmediato.
- **Return on Assets (ROA)**: Este gráfico de Pareto muestra una concentración del rendimiento sobre los activos en pocos sectores, con una variación significativa. Los sectores de tiendas de alimentos especializados, transporte de mercancías generales y materiales de construcción son los principales contribuyentes al ROA, lo que sugiere un uso eficiente de activos en estos casos.
- **Return on Equity (ROE)**: El rendimiento sobre el patrimonio está concentrado principalmente en sectores como el transporte de oleoductos y tiendas de alimentos especializados. Esto indica que en estos sectores, las empresas generan altos retornos en relación con el capital de los accionistas, lo que podría reflejar estrategias efectivas de apalancamiento.

- **Short Term Debt**: La deuda a corto plazo se concentra en sectores como fondos de beneficios para empleados, extracción de gas natural y arrendadores de bienes raíces, lo que señala una dependencia de financiación a corto plazo en estos sectores, posiblemente por su necesidad de capital circulante.
- **Total Operating Revenue**: Los ingresos operativos totales están distribuidos en una gran variedad de sectores, con una curva de Pareto que muestra cómo ciertos sectores dominan en generación de ingresos, como el comercio mayorista de productos derivados del petróleo y la extracción de gas natural, reflejando la importancia de estos sectores en la economía.

Por otro lado, también se intentó encontrar relaciones entre las diferentes variables numéricas con las que contamos, sin embargo, no encontramos correlaciones significativas entre las diferentes variables, por lo que se asumirá que no existen correlaciones entre las diferentes variables según se muestra en la siguiente gráfica:



Por último, también se realizó un análisis por PCA, que logró disminuir la dimensionalidad a 14 variables, las cuales explicaban alrededor del 60% de la varianza total de los datos. Las dimensiones resultantes del análisis por componentes nos arrojaron conclusiones interesantes sobre el set de datos:

- Un primer componente refleja una combinación de variables relacionadas con la estructura financiera de las empresas. El mismo se compone de una combinación lineal de 0.50 Long term debt, 0.47 cash and cash equivalents, 0.44 Property plant and equipment, 0.44 Total operating revenue y 0.32 Short term debt.
- Un Segundo componente podría estar capturando la relación entre deuda a corto plazo y las tendencias de las ganancias operativas, lo que sugiere que aquellas empresas con mayores deudas a corto plazo podrían tener un desempeño financiero distinto en términos de crecimiento de ingresos y beneficios. Este se compone de 0.56 Short term debt, -0.44 Emplyee, -0.33 total operating revenue y 0.27 cash and cash equivalents.
- Un tercer componente que está relacionado con las importaciones y exportaciones, especialmente a China y Estados Unidos. Cabe resaltar que mientras los componentes asociados con China son positivos, los relacionados con USA son negativos, lo que refleja la realidad competitiva de estos dos países.

7. Conclusiones iniciales

- Predominio de micro y pequeñas empresas: La industria de Bogotá está formada principalmente por empresas de menor escala (de 0 a 50 empleados). Esta característica podría ser un reflejo de la estructura empresarial de la ciudad, donde las micro y pequeñas empresas impulsan una parte significativa de la economía local.
- Concentración de las exportaciones: Un reducido número de empresas realiza entre el 80% y el 100% de las exportaciones de Bogotá. Este hallazgo sugiere que las exportaciones de la ciudad dependen principalmente de industrias específicas, lo cual representa tanto una fortaleza para dichas industrias como un riesgo potencial en caso de cambios en el mercado.
- Relación entre deuda y efectivo disponible: Se observó que las empresas con mayor deuda a largo plazo también tienden a tener mayores cantidades de efectivo disponible. Este patrón podría indicar que las empresas con mayor capacidad de financiamiento también cuentan con mejores recursos de liquidez para sus operaciones.
- Inversión en activos fijos y ventas totales: Existe una relación directa entre las inversiones en activos fijos, las ventas totales y la deuda a largo plazo. Esto sugiere que las empresas con mayores ventas y deuda invierten también en sus activos fijos, lo que probablemente les permite expandir y mejorar sus operaciones.
- Impacto de la deuda a corto plazo en el desempeño financiero: Las empresas con mayores niveles de deuda a corto plazo podrían mostrar un desempeño financiero distinto, especialmente en términos de crecimiento de ingresos y beneficios. Este comportamiento sugiere que la gestión de deuda a corto plazo es un factor clave en el rendimiento financiero de las empresas.
- Correlaciones financieras para agrupación: Las correlaciones entre las variables analizadas permiten trazar un perfil general de la salud financiera de las empresas. Estos resultados serán la base para el siguiente paso en el análisis: seleccionar el mejor modelo de agrupación para clasificar a las empresas en diferentes clústeres de acuerdo con su rendimiento y segmentación NAICS, lo que permitirá una segmentación más precisa para la prospección.

8. Preparación de datos

- **Carga de datos:** Los datos (en formato JSON/CSV/XLSX) son cargados desde un archivo fuente utilizando la función load_data.

- **Selección inicial de variables:** Basado en el EDA, se identificaron y seleccionaron las variables relevantes para reducir la dimensionalidad del dataset y centrarse en características clave para el análisis, como:
  - Variables financieras: Cash and Cash Equivalents, Total operating revenue.
  - Indicadores operativos: Net Sales Revenue Trend (%), Operating Profit Trend (%).
  - Métricas de solvencia: Quick Ratio (x), Long term Debt, Short Term Debt.
  - Otros: Industry (NAICS), Number of Employees.

- **Limpieza de datos:** Se hace para mejorar la calidad de los datos, evita errores durante el modelado y asegura consistencia.
  - Imputación de valores faltantes en variables numéricas y categóricas.
  - Eliminación de filas con valores nulos persistentes.
- **Procesamiento de la Industria (NAICS):** La agrupación en categorías amplias simplifica el análisis y hace que los resultados sean más interpretables.

- Limpieza de la columna: Eliminación de ruido en los códigos NAICS.
- Extracción del sector principal: Se extrajeron los dos primeros dígitos del código NAICS.
- Mapeo a categorías amplias: Usando un diccionario predefinido, los códigos NAICS se agruparon en sectores simplificados como "Manufactura" o "Servicios Públicos".

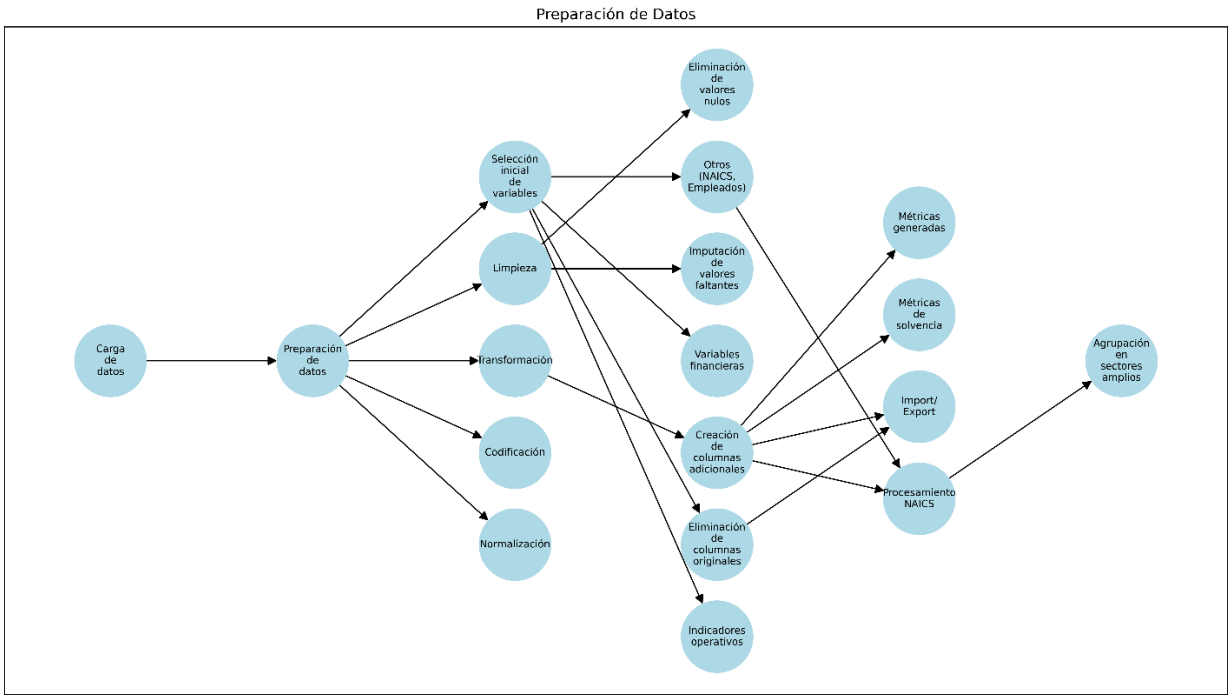
- **Preprocesamiento de Importaciones y Exportaciones:** Permite un análisis más granular de los datos de comercio internacional.
  - Se utilizaron expresiones regulares para extraer países y porcentajes de las columnas Import y Export.
  - Creación de columnas adicionales: Para los 10 países con mayor participación, se crearon columnas individuales que reflejan el porcentaje asociado.
  - Eliminación de Columnas Originales: Las columnas originales Import y Export se eliminaron para reducir dimensionalidad.

- **Limpieza y transformación:**
  - Métricas Generadas:
    - Debt to Assets: Relación entre deuda total y activos.
      Debt to Assets=Long Term Debt + Short Term Debt/Property, Plant and Equipment
    - Relative_Growth: Diferencia entre tendencias de ingresos netos y utilidades operativas.
      Relative Growth=Net Sales Revenue Trend – Operating Profit Trend

  - Los valores infinitos resultantes fueron reemplazados por NaN.
  - Los valores faltantes generados en esta etapa fueron imputados con 0.

- **Normalización:** Todas las variables numéricas fueron escaladas utilizando StandardScaler para llevarlas a la misma escala, mejorando la eficiencia del modelo de clustering.
- **Codificación:** Las variables categóricas (como los sectores NAICS) fueron codificadas utilizando OneHotEncoder.

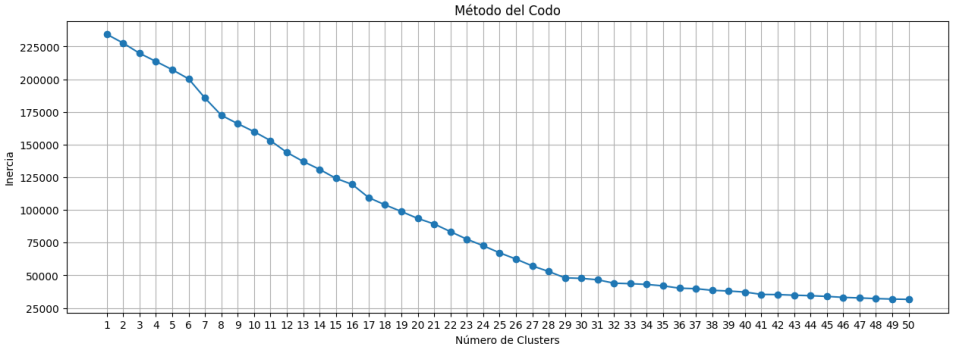
Se adjunta el diagrama de bloques funcional con los diferentes procesos implementados:



9. **Estrategia de validación y selección de modelo:** En total, se seleccionaron tres modelos con los cuales se podría realizar el agrupamiento: K-Means, DBSCAN y Hierarchical Clustering. La métrica de interés sobre la cual se realizó la validación del modelo fue el índice de silueta, el cual indica que tan bien separados están los grupos, y que tanto se parecen dos puntos que pertenecen a un mismo clúster.
   Para la selección del modelo se optó por el método del codo con el fin de optimizar los hiperparámetros. Mediante el mismo, se espera encontrar un equilibrio entre el valor de cada hiperparámetro y la reducción de los errores del modelo. De esta forma, para diferentes valores de hiperparámetros se realiza una gráfica contra el error medio y se encuentra el denominado codo de la gráfica. Una vez se definió un rango de valores, se buscó el que mejores resultados entregó en términos del índice de silueta.

10. **Construcción y evaluación del modelo**
    Para entrenar el modelo se separaron los datos en un dataset de entrenamiento y validación. Las métricas de los modelos se calcularon a partir de la validación con el fin de verificar que el algoritmo funcione igualmente bien para datos vistos como para aquellos no vistos. A continuación, el proceso de construcción de los dos modelos seleccionados

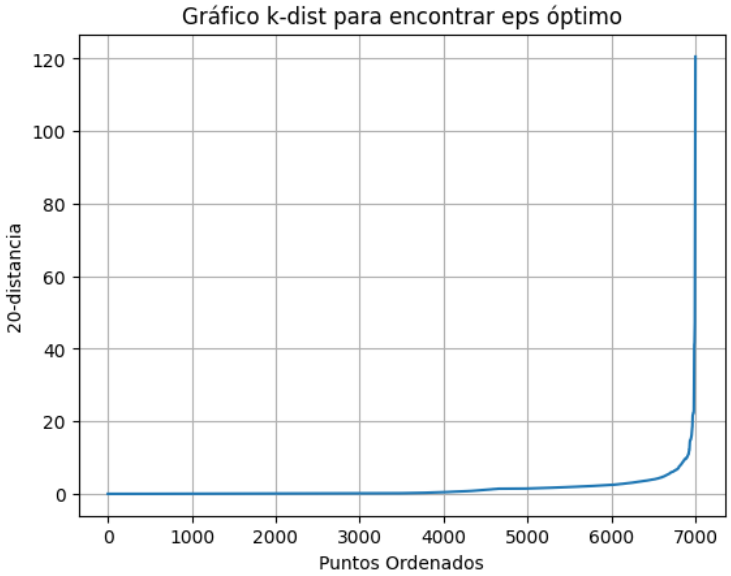
- **K-Means:** Para K-Means solamente es necesario ajustar el número que el algoritmo va a generar. Para ello, según se mencionó previamente, se ejecutó el método del codo con clústeres desde 1-50. A continuación, se muestra la gráfica resultante:



Como se ve, a partir de un valor entre 20 y 30 clústeres, el modelo deja de mejorar significativamente su error. Por tanto, sabemos que allí se debe encontrar el número óptimo de clústeres. Con una búsqueda más exhaustiva encontramos que 25 clústeres nos entregan un índice de silueta óptimo cercano a 0.45. Además de mostrar una gráfica de las siluetas con un menor número de puntos con una silueta significativamente baja. Esto indica que actualmente la calidad de los clústeres es moderada, pero podría ser mejorada. Aun así, la mayoría de los datos se encuentran correctamente segmentados, por lo que se espera que pueda mejorar el proceso actual de selección de empresas

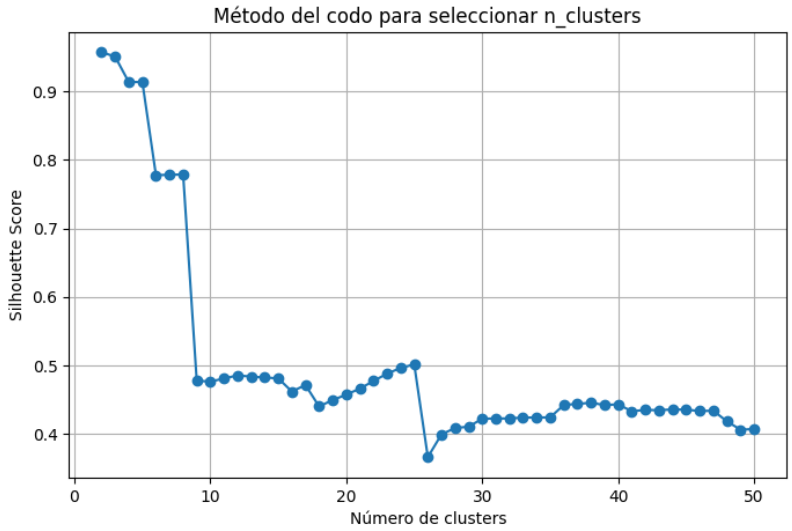
- **DBScan:** Para el caso de DBScan es necesario ajustar dos hiperparámetros eps y min_samples. EPS define el radio mínimo para considerar a dos puntos como vecinos. Ajustar este valor es crucial, debido a que valores muy pequeños puede resultar en la generación de muchos clústeres, mientras que valores muy altos puede causar que datos no relacionados queden ajustados en un mismo clúster. Por otro lado, min simples se refiere al mínimo número de puntos necesarios para considerar a una región con muchos puntos un clúster independiente.

  Para la selección de los min simples, se optó por dos puntos. Debido a la naturaleza del problema, donde se tienen datos muy dispersos y outliers muy grandes, usar más de dos puntos puede causar que se queden por fuera muchos de los puntos outliers, los cuales son importantes para el análisis. Por otro lado, la selección del eps se hizo mediante el método del codo, obteniendo el siguiente resultado:



Como se ve, en este caso el número de épsilon óptimo es un valor entre 10-20. Con una búsqueda del mayor índice de silueta encontramos que 20 es un valor óptimo. Sin embargo, al realizar el gráfico de silueta es posible visualizar que solamente se generan dos clusters. Estos corresponden entonces a datos considerados “normales”, mientras que se genera un segundo cluster con outliers. Debido a los objetivos planteados para el proyecto, se descartó entonces la posibilidad de usar este algoritmo.

- **Hierarchical Clustering:** Por último, se encontró un modelo óptimo de hierarchical clustering. La selección de hiperparámetros se hizo nuevamente haciendo uso del método del codo. En el caso de Hierarchical, el hiperparámetro principal es el número de clústeres. A continuación, se muestra una gráfica del número de clústeres vs el índice de silueta.



Según muestra la imagen, el índice de silueta decrece significativamente a medida que aumenta el número de clústeres. Siendo el número óptimo de clústeres 2 con un índice superior al 0.9. Sin embargo, al ejecutar el algoritmo con clústeres óptimo, nos encontramos con el mismo problema que tiene DBScan, y es que se crean una cantidad menor de clústeres, por lo que los datos quedan agrupados en

un gran clúster con todos los datos considerados “promedios” y otros clústeres con una cantidad de datos pequeña que no aportan valor para la empresa.

De esta forma, se seleccionó el modelo de KMeans, el cual es la que más valor está aportando para los objetivos planteados para el proyecto.

11. Construcción del producto de datos

- **Modelo de Machine Learning seleccionado:** K-Means, entrenado y optimizado para agrupar empresas en clusters basados en métricas financieras y operativas.
  - Entrenamiento: Entrenado con los datos preprocesados utilizando un pipeline que incluye imputación, escalado y creación de características adicionales como Debt_to_Assets y Relative_Growth.
  - Validación: Se seleccionó el número óptimo de clusters (k=25) de acuerdo a las métricas.
  - Almacenamiento: El modelo se guardó en un archivo kmeans_model.joblib para su integración en otros componentes.

- **API REST:** Creada con FastAPI de la siguiente forma:
  - **Ruta home (/):** Verifica que la API esté funcionando correctamente.
  - **Ruta /classify/:**
    - Recibe los datos de empresas en formato JSON.
    - Preprocesa los datos utilizando el pipeline definido (run_pipeline).
    - Clasifica las empresas en uno de los clusters generados por el modelo K-Means.
    - Devuelve una respuesta en formato JSON con las empresas y el cluster asignado.
    - **Despliegue en entorno de desarrollo:** El servidor de la API se ejecuta con Uvicorn de forma local, lo que permite su integración con otras aplicaciones.

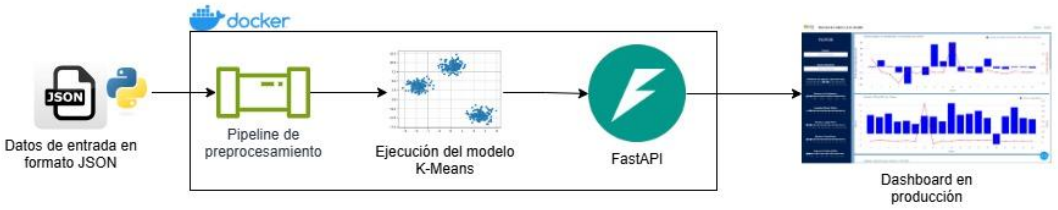
- **Dashboard:** Creado a través del framework Dash para visualizar los resultados del clustering y permitir la interacción con los datos clasificados.
  - **Características:**
    - **Filtros:**
      - Selección de clusters y sectores industriales.
      - Ajustes dinámicos para métricas clave como tendencia de ingresos operativos o número de empleados
    - **Visualizaciones:** Algunos gráficos que se incluyeron fueron:
      - Gráfico de tendencias de rentabilidad y crecimiento por cluster.
      - Gráfico de relación entre ROA (Return on Assets) y ROE (Return on Equity) por cluster.

- **Despliegue:**
  - **Modelos y Pipelines:**
    - kmeans_model.joblib: Modelo de clustering.
    - preprocessor.pkl: Pipeline de preprocesamiento.
  - **API REST:**
    - Integración del modelo con FastAPI.
    - Desplegable como un contenedor Docker para portabilidad.
  - **Dashboard:**
    - Implementado como una aplicación Dash.

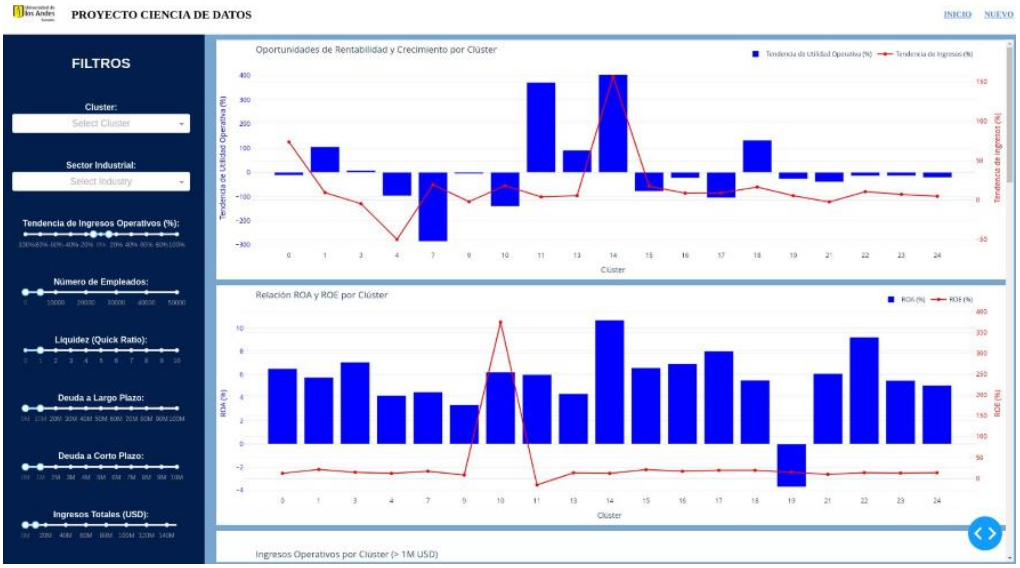
- Capaz de consumir datos desde la API o trabajar directamente con datos procesados.

- **Despliegue con Docker:**
  - Contiene un Dockerfile para empaquetar y desplegar toda la solución.
  - El contenedor incluye el modelo, la API y las dependencias necesarias.

Se anexa la arquitectura de solución del producto de datos:



Producto de datos finalizado:



Se anexa en este link un manual de uso para el producto de datos.

12. Retroalimentación por parte de la organización

Primera interacción: Definición inicial del problema y planteamiento del caso.

El día 30 de septiembre de 2024, se llevó a cabo una reunión con el Lead Developer para explorar las necesidades organizacionales y definir el problema a resolver. Durante esta sesión, se discutieron las posibilidades de implementar un modelo de machine learning enfocado en segmentar empresas. Se identificaron las principales variables que definen a los clientes potenciales y se planteó la idea de realizar un análisis PCA para priorizar las características relevantes. También se exploraron herramientas de AutoML para crear modelos de clustering. Este primer acercamiento permitió establecer los objetivos iniciales del proyecto y trazar un plan de trabajo.

***Segunda interacción*: Ajustes en los criterios y filtros del análisis**

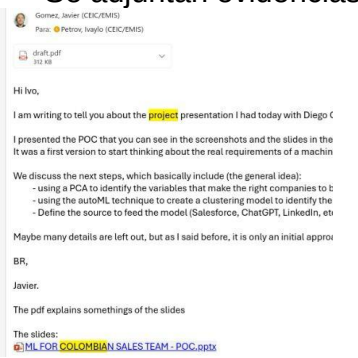
El día 21 de octubre de 2024, se llevó a cabo una reunión con el Senior New Business Developer y con copia a la Directora Latam, donde se presentó la segunda corrida del modelo. Durante la sesión, se analizaron los filtros empleados,

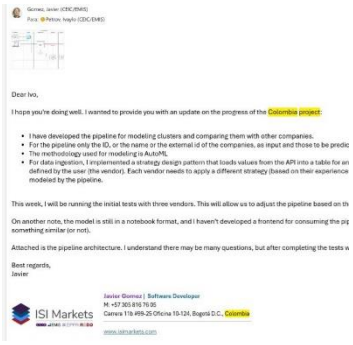
como las industrias específicas (NAICS), los rangos de empleados e ingresos. También se entregaron los resultados intermedios, incluyendo estadísticas descriptivas, datos preprocesados y los clusters encontrados. Esta interacción fue crucial para ajustar el enfoque hacia sectores estratégicos, como construcción, salud y minería, y garantizar que el análisis estuviera alineado con las expectativas organizacionales.

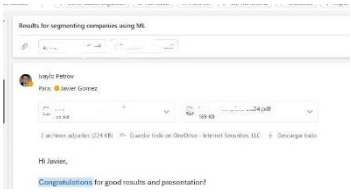
<u>Tercera interacción: Presentación de resultados y entrega de productos consumibles</u>

El día 29 de noviembre de 2024, se realizó una presentación con el Lead Developer, el Senior New Business Developer y la Directora Latam. En esta sesión, se presentaron los resultados finales del análisis utilizando un modelo de K-Means Clustering. El modelo agrupó exitosamente las empresas en clusters específicos basados en características clave. Durante la reunión, se entregaron dos productos consumibles: una API que permite acceder dinámicamente a los clusters y un dashboard interactivo que facilita la visualización de los datos y su análisis. Estos productos fueron diseñados para ser directamente utilizables en la toma de decisiones estratégicas, cerrando así la entrega del proyecto.

Se adjuntan evidencias de estas comunicaciones:


Gomez, Javier (CEIC/EMIS)
Para: Petrov, Ivaylo (CEIC/EMIS)
draft.pdf
312 KB
Hi Ivo,
I am writing to tell you about the project presentation I had today with Diego
I presented the POC that you can see in the screenshots and the slides in the
It was a first version to start thinking about the real requirements of a machin
We discuss the next steps, which basically include (the general idea):
- using a PCA to identify the variables that make the right companies to b
- using the autoML technique to create a clustering model to identify the
- Define the source to feed the model (Salesforce, ChatGPT, LinkedIn, et
Maybe many details are left out, but as I said before, it is only an initial approx
BR,
Javier.
The pdf explains somethings of the slides
The slides:
ML FOR COLOMBIAN SALES TEAM - POC.pptx


Dear Ivo,
I hope you're doing well. I wanted to provide you with an update on the progress of the Colombia project:
- I have developed the pipeline for modeling clusters and comparing them with other companies.
- For the pipeline only the ID, or the name or the external id of the companies, as input and those to be predic
- The methodology used for modeling is AutoML
- For data ingestion, I implemented a strategy design pattern that loads values from the API into a table for an defined by the user (the vendor). Each vendor needs to apply a different strategy (based on their experience modeled by the pipeline.
This week, I will be running the initial tests with three vendors. This will allow us to adjust the pipeline based on th
On another note, the model is still in a notebook format, and I haven't developed a frontend for consuming the pip something similar (or not).
Attached is the pipeline architecture. I understand there may be many questions, but after completing the tests w
Best regards,
Javier
ISI Markets


Results for segmenting companies using ML
Ivaylo Petrov
Para: Javier Gomez
Hi Javier,
Congratulations for good results and presentation!

13. Conclusiones

- Se cumplió el objetivo del proyecto ya que se realizó el dashboard para el análisis de los usuarios, donde se les facilita encontrar prospectos según los servicios que ofrecen y se logró hacer la clasificación de las 10.000 empresas de Bogotá por su cluster.
- Dada la diversidad de las empresas, una dificultad grande fue agruparlas con una buena métrica de silueta.
- El producto de datos contribuye a un aumento significativo al valor del negocio, esto ya que se estima que la tasa de oportunidades aumente en 30% a 40%, la tasa de interacción aumente en un 15% a 25% y el coste por lead calificado disminuya en 40% a 50%.
- El modelo arroja una clasificación de 25 clusters, los cuales contribuyen al análisis de los equipos de ventas, ya que si bien siguen teniendo índices independientes, les ayuda a tomar un grupo de empresas que comparten tipo de actividad y filtrar por cada variable, haciendo que el análisis deje de tomar una semana entera al analizar empresa por empresa, por ejemplo, filtrar las empresas de comercio mayorista con mayor número de empleados o mayores ingresos y ofrecerse para un servicio en particular de alto volumen.

- Dado que las empresas no están obligadas a reportar sus datos financieros, hay algunos faltantes, aquí habría una oportunidad de mejorar las salidas de los modelos.
- Algunos clústeres destacan por oportunidades de crecimiento y utilidad operativa, mientras que otros muestran altos niveles de riesgo debido a bajos retornos (ROA y ROE) o liquidez limitada. La distribución de ingresos sugiere concentración en sectores específicos, como comercio mayorista y minorista, mientras que otros sectores tienen menor participación. Esto evidencia una diversificación económica desigual y la necesidad de estrategias personalizadas para mejorar la sostenibilidad financiera de cada clúster.
- Los clústeres con altos ingresos operativos y rentabilidad (ROA) representan las mejores oportunidades para prospectar, especialmente en sectores como comercio mayorista y minorista. Sin embargo, clústeres con baja liquidez o alta deuda pueden requerir un análisis más detallado para adaptar las propuestas de valor a sus necesidades específicas.
- El mejor modelo obtenido, es suficiente para dar solución a la oportunidad del negocio.