Universidad de los Andes
Maestría en Inteligencia analítica de Datos
Aprendizaje No Supervisado
Presentado por: Felipe Cortes, Jorge Osorio, Andres Ruiz

Proyecto: Nivel de Riesgo de Intermediarios Financieros

Resumen

La entidad en estudio es una entidad clave para facilitar el acceso al crédito en Colombia, asumiendo parte del riesgo de los préstamos otorgados por intermediarios financieros a empresas y trabajadores independientes. Esta enfrenta el reto de reducir el riesgo de siniestros, que se materializan cuando un deudor incumple con el pago de su crédito, generando pérdidas para la entidad.

Actualmente, las comisiones se calculan en función del riesgo del intermediario financiero (IF), pero algunos IF, a pesar de manejar grandes volúmenes de crédito, son penalizados por la tarifa actual, lo que afecta las relaciones comerciales. El objetivo de este proyecto es analizar los datos históricos de los IF para agruparlos según sus características y patrones de siniestralidad.

Utilizando técnicas de aprendizaje no supervisado, como el clustering, se analizaron más de 78,000 registros con variables clave como el volumen de desembolso, el saldo de la deuda y el tiempo hasta el siniestro. El resultado principal permitió identificar grupos de IF con mayor siniestralidad, lo que ofrece a la entidad una oportunidad para ajustar sus políticas de tarifas y tomar decisiones más informadas sobre la gestión del riesgo.

Estos hallazgos ayudarán a optimizar las condiciones comerciales, ajustando las tarifas de manera más justa y controlando mejor el riesgo de impago, que es fundamental para la sostenibilidad del negocio.

Introducción

La entidad asume la responsabilidad de mitigar el riesgo asociado al incumplimiento crediticio de los beneficiarios hacia bancos u otros intermediarios financieros. Cuando se materializa el riesgo de impago y ocurre el siniestro, esta cubre una porción del saldo pendiente, lo que supone un desafío estratégico clave que involucra recursos públicos. La administración eficiente de estos riesgos, apoyada en análisis de datos, puede generar importantes beneficios tanto para la entidad como para el sistema financiero en general.

Actualmente, las comisiones de la entidad se calculan en función del riesgo de los intermediarios financieros (IF). Sin embargo, algunos IF, a pesar de manejar grandes volúmenes de crédito, resultan penalizados por la estructura tarifaria, lo que pone en riesgo las relaciones comerciales. Este proyecto se enfoca en identificar patrones ocultos en los datos históricos para optimizar las condiciones de negociación y mitigar el riesgo de siniestros.

A través del análisis de clustering utilizando K-Means, se logró identificar grupos de IF caracterizados por variables como el promedio de días hasta que ocurre el siniestro y el total desembolsado por la garantía. Con esta segmentación, se buscó responder la siguiente pregunta: ¿Cómo podemos identificar y agrupar a los intermediarios financieros con características similares para ajustar las condiciones comerciales y reducir el riesgo de siniestros?

La literatura revisada respalda el uso de técnicas de machine learning en la gestión del riesgo crediticio. Un estudio sistemático sobre la aplicación de machine learning en la gestión de riesgo de crédito financiero destaca similitudes clave, como el enfoque en la identificación de patrones de riesgo crediticio y la importancia de modelos interpretables. Mientras que algunos estudios utilizan algoritmos supervisados o híbridos, en nuestro caso aplicamos técnicas no supervisadas como el clustering, específicamente orientado a la segmentación de intermediarios financieros para optimizar la toma de decisiones y mitigar riesgos.

Otros estudios, como un modelo de identificación de indicadores de gestión de riesgo mediante reducción de variables, comparten nuestro interés en reducir la complejidad de los datos, utilizando técnicas como el Análisis de Componentes Principales (PCA). Aunque este enfoque también podría ser utilizado en el proyecto de la entidad para reducir la dimensionalidad de los datos, nos centramos en el clustering para identificar patrones de siniestralidad.

Finalmente, una investigación sobre la integración de modelos supervisados y no supervisados para la evaluación del riesgo crediticio plantea un esquema híbrido de análisis, mientras que en el proyecto del FNG aplicamos exclusivamente aprendizaje no supervisado. Sin embargo, ambos trabajos comparten el objetivo de mejorar la evaluación del riesgo crediticio a través de la segmentación de actores clave.

La principal limitación del enfoque es la falta de consideración de factores macroeconómicos externos que podrían influir en el comportamiento de los siniestros. La recomendación principal es implementar un sistema dinámico de segmentación de IF, basado en análisis de datos continuo, para ajustar las tarifas de manera más justa y eficaz y para mejorar la limitante macro, esta poderla tener como argumento externo para algunos ajustes en tarifas. (recomendamos cada 6 meses analizar los comportamientos del modelo).

Materiales y Métodos

Se describen los datos utilizados, incluyendo su fuente, y el proceso de limpieza utilizado. Se presentan los datos utilizando estadísticos descriptivos en tablas y/o visualizaciones y se describe detalladamente el algoritmo utilizado. [8 puntos]

Descripción de los datos

La fuente de la cual extraemos los datos fue directamente de la base de datos de la entidad. No especificamos nombres de tablas, ni de la BD por temas de confidencialidad. A continuación, mostraremos la descripción de los datos realizada para los datos extraidos.

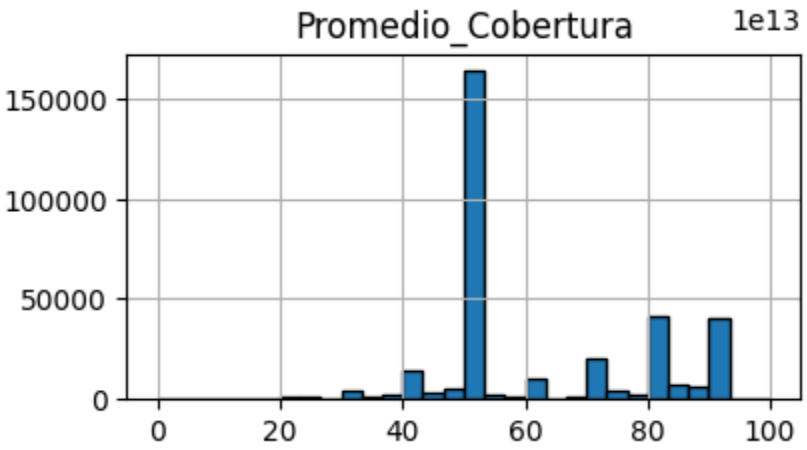
```
RangeIndex: 326768 entries, 0 to 326767
Data columns (total 19 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   Nit_Intermediario              326768 non-null  int64
 1   Nombre_Intermediario           326768 non-null  object
 2   Total_Desembolso               326768 non-null  float64
 3   Total_Saldo                    326768 non-null  float64
 4   Promedio_Cobertura             326768 non-null  float64
 5   Promedio_Dias_Hasta_Siniestro  326768 non-null  int64
 6   Tipo_Cartera                   326768 non-null  object
 7   Producto                       326768 non-null  object
 8   Nombre_producto                326768 non-null  object
 9   Programa                       326767 non-null  object
 10  Estado_Gtia                    326767 non-null  object
 11  Region_Gtia                    326767 non-null  object
 12  Municipio_Gtia                 326767 non-null  object
 13  Departameno_Gtia               326767 non-null  object
 14  Ruralidad                      326767 non-null  object
 15  Tipo_identificacion            326767 non-null  object
 16  Sector                         326767 non-null  object
 17  Tamaño                         326767 non-null  object
 18  Macrosector                    326767 non-null  object
dtypes: float64(3), int64(2), object(14)
```

Tenemos 19 variables con un total de 326767 registros

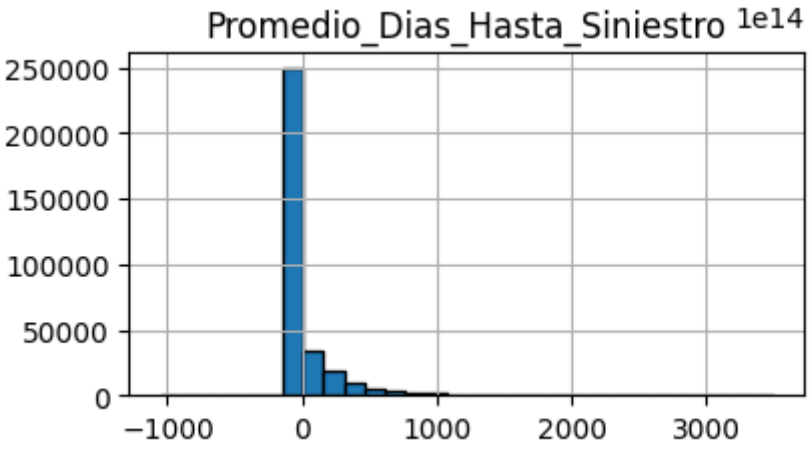
	Nit_Intermediario	Total_Desembolso	Total_Saldo	Promedio_Cobertura
count	3.267680e+05	3.267680e+05	3.267680e+05	326768.000000
mean	8.459302e+09	2.517682e+09	1.078137e+10	61.149385
std	3.423482e+08	9.516188e+10	5.008020e+11	17.064842
min	8.000116e+09	0.000000e+00	0.000000e+00	0.000000
25%	8.001479e+09	1.000000e+07	6.683547e+06	50.000000
50%	8.600030e+09	4.000000e+07	3.200000e+07	50.000000
75%	8.600343e+09	1.840000e+08	1.610072e+08	80.000000
max	8.909263e+09	3.177004e+13	1.730654e+14	100.000000

	Promedio_Dias_Hasta_Siniestro
count	326768.000000
mean	71.604306
std	202.131998
min	-1056.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	3499.000000

Total_Desembolso tiene una media de aproximadamente 2,517,682,000 con una desviación estándar de 95,161,880,000, lo que indica una gran variabilidad en los desembolsos, que van desde 0 hasta un máximo de 31,770,040,000,000. **Total_Saldo** presenta una media de 10,781,370,000 y una desviación estándar de 500,802,000,000, lo que refleja una dispersión significativa en los saldos, desde 0 hasta más de 173,065,400,000,000. El **Promedio_Cobertura** tiene una media de 61.15 y una desviación estándar de 17.06, sugiriendo variabilidad en la cobertura promedio que varía entre 0 y 100. Finalmente, **Promedio_Dias_Hasta_Siniestro** muestra una media de 71.60 y una desviación estándar de 202.13, con valores que van desde -1,056 hasta 3,499 días. Cabe destacar que la variable **Nit_Intermediario** debe considerarse como un identificador textual en lugar de numérico, ya que se trata de una etiqueta de identificación y no de una métrica cuantitativa. La alta dispersión en estas variables resalta la necesidad de un análisis más profundo para entender los patrones y variaciones en los datos, especialmente en contextos financieros o de seguros.

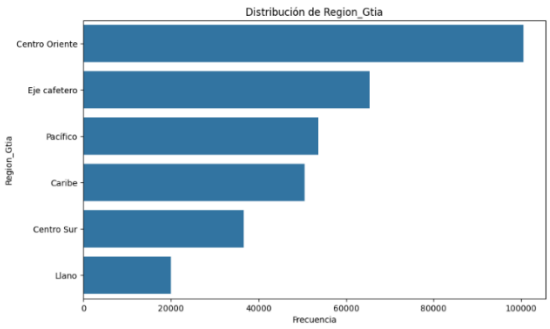


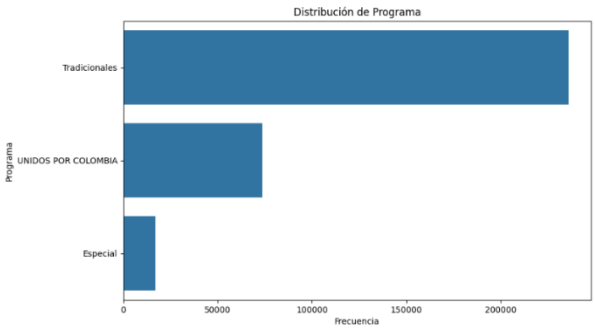
Vemos que la mayor parte de las coberturas estan en el 50%

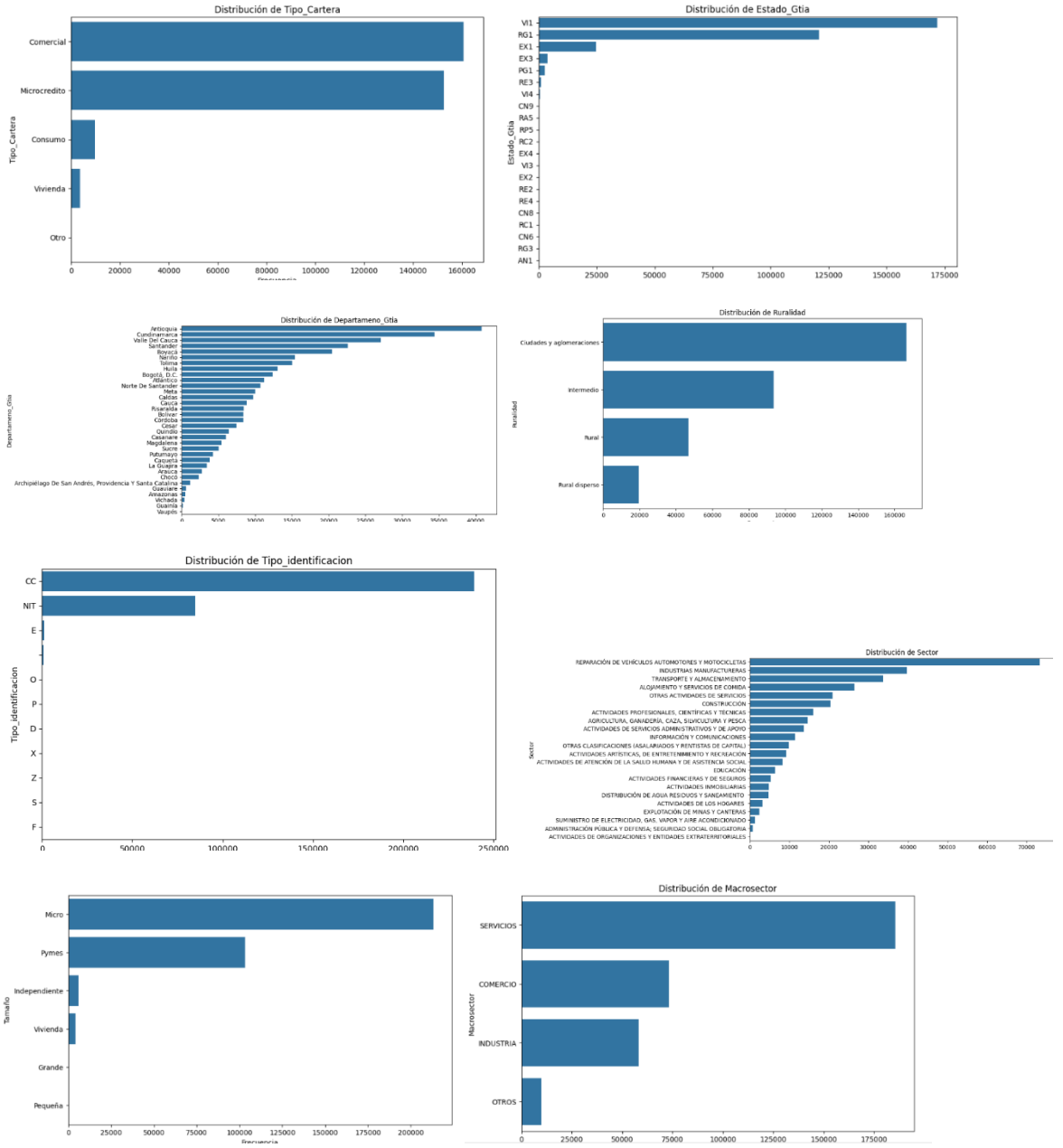


Se realizo una limpieza en los datos, la cual consistio en eliminar esos registros que reportaban valores negativos en la variable de promedio de días

hasta el siniestro, ya que no es posible que existan valores de este tipo para esta variable.

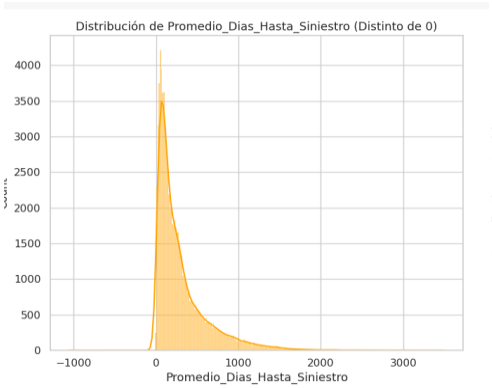


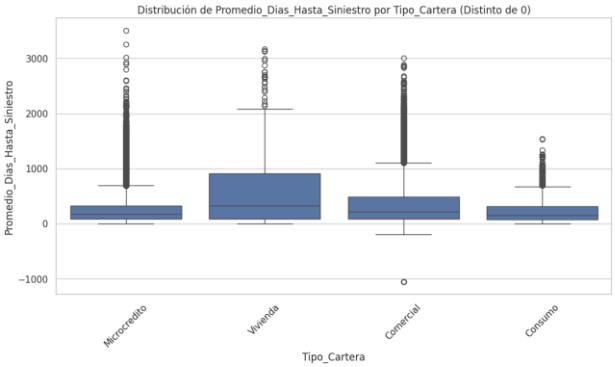


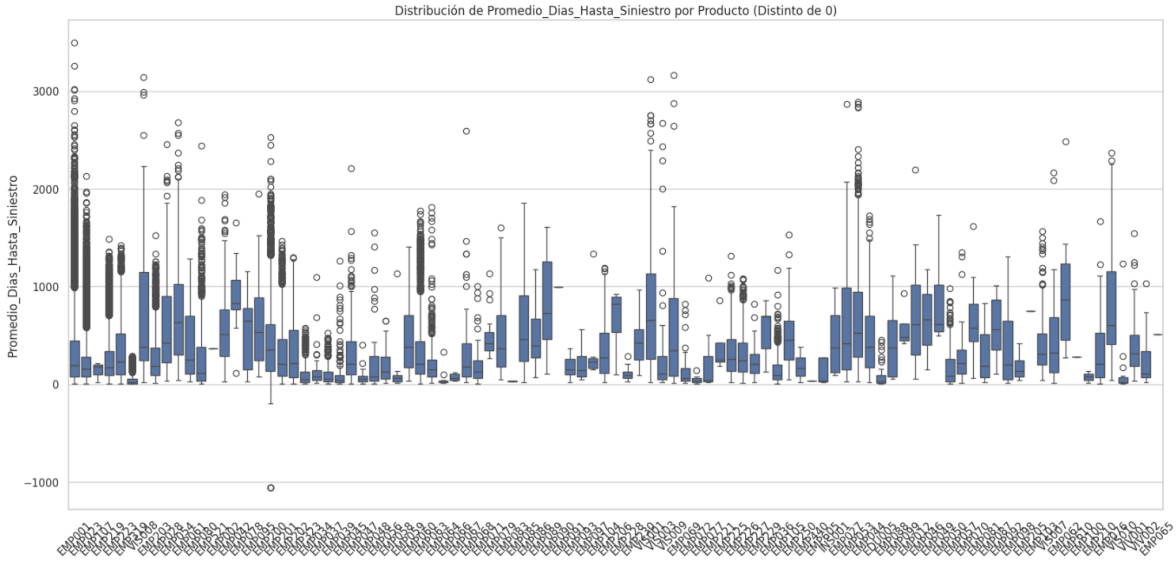


El análisis de los gráficos de barras nos permite entender la distribución de la muestra. Observamos que la mayoría de los datos corresponden a personas naturales, con una concentración en la cartera comercial y, en segundo lugar, en microcrédito. Como era de esperarse, la mayor parte de los datos se encuentran en estados vigentes y en casos de muerte natural, que no han generado reclamaciones, lo cual es un indicador positivo para el negocio. Esta variable es crucial para el análisis, ya que determina la siniestralidad.

Además, notamos que la mayor parte de la actividad se concentra en el macrosector de servicios, y a nivel geográfico, los datos están concentrados en las principales ciudades, con un aumento significativo en Nariño. Este último punto es relevante debido a la nueva estrategia de atender zonas con complejidades en orden público.







Observamos que el período desde la emisión de la garantía hasta que se siniestra se concentra, en promedio, alrededor de los 500 días. También identificamos que el microcrédito es el tipo de producto en el que las reclamaciones se realizan más rápidamente. Además, el producto EMP319 destaca como clave, dado que, a pesar de ser un producto reciente, muestra un comportamiento relevante en este contexto.

Limpieza de los datos.

Tal como comentamos anteriormente realizamos una limpieza de la variable promedio de días hasta siniestro, ya que presentaba valores negativos que no cumplen con la regla definida para este campo en la que solo se aceptan valores positivos. Aparte de eliminar registros con valores negativos, también eliminamos los registros con valores iguales a 0. Luego de realizar dicho filtrado, nos quedamos un total de 78419 filas, lo cual es una cantidad considerablemente grande de registros.

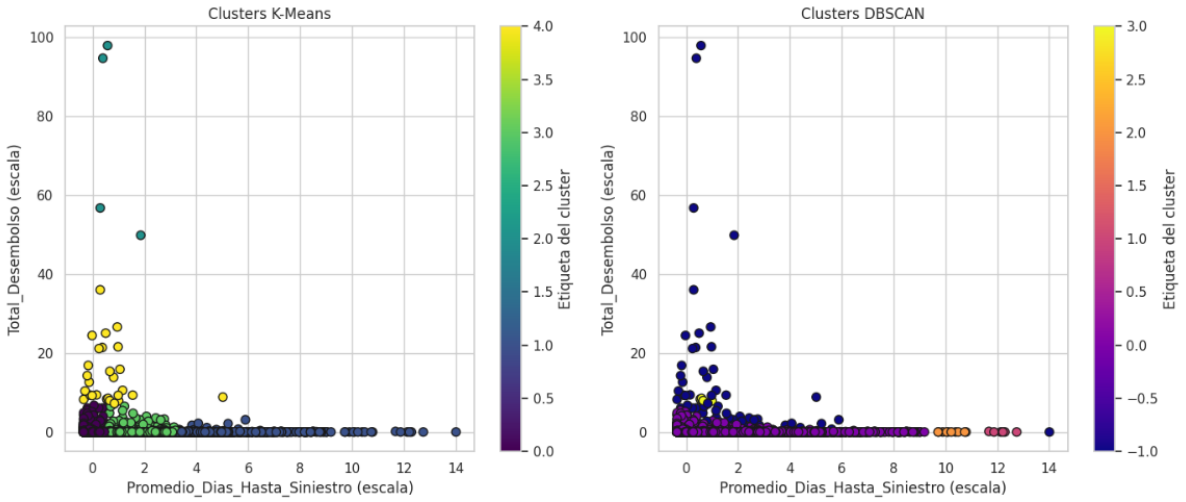
Explicación del Algoritmo Empleado

Para el desarrollo de este proyecto decidimos aplicar un algoritmo de “clustering” el cual nos permitiera generar clusters teniendo en cuenta principalmente variables como el promedio de días hasta siniestro, y el total desembolsado. Utilizamos el algoritmo de KMeans y el algoritmo de DBSCAN para generar los clusters, teniendo como objetivo inicial comparar los clusters generados por estos dos algoritmos y comparar las similitudes o diferencias, a partir de ello,

nos quedaríamos finalmente con los resultados de alguno de estos dos algoritmos. Como criterios para seleccionar alguno de esos dos algoritmos, nos fijamos básicamente en la forma o calidad de los clusters generados, así como la capacidad de poder interpretar o caracterizar cada cluster según tendencias en los datos de las observaciones que pertenecen a cada uno, y teniendo en cuenta el criterio o la opinión de un experto que trabaja directamente con la entidad el cual hace parte de este proyecto.

Para interpretar los resultados, también tuvimos en cuenta la definición de cada uno de los algoritmos, donde el KMeans es un algoritmo basado en centroides el cual debemos predefinir el número de clusters que debemos generar, y el DBSCAN es un algoritmo basado en densidades en el que se debe predefinir el radio de vecindario y el número de puntos considerados suficientes para definir un vecindario.

A continuación, se muestra una comparación de los clusters generados por ambos algoritmos usando como parametro para K-Means 5 Clusters y para DBSCAN un eps de 0.5 y un min_samples de 5:



De estos gráficos podemos sacar las siguientes conclusiones:

- Analizando la asignación de clusters en el gráfico de dispersión del total_desembolsado vs el promedio de días hasta el siniestro por cada algoritmo, nos podemos dar cuenta que en el algoritmo de KMeans se forman clusters que parecen tener tendencias que visualmente se pueden observar para los diferentes rangos de las dos variables analizadas, a diferencia del DBSCAN en el que podemos ver que una gran parte de los datos sobre todo aquellos con valores relativamente grandes del total desembolsado fueron clasificados como ruido.
- Tal como lo mencionamos anteriormente con K-Means pusimos por parámetro generar 5 clusters, de los cuales los 5 parecen tener tendencias claras y ser coherentes. Por ejemplo: El cluster 0, son observaciones en las que se tienen un total desembolsado bajo y un promedio de días bajo. Si comparamos esto con los resultados de DBSCAN obtenemos que este algoritmo genero 4 clusters donde uno de ellos (Cluster 0), comprende valores bajos de total desembolsado, pero en cuanto a la variable de promedio días hasta siniestro las observaciones de este cluster tienen valores

prácticamente en la mayor parte del rango de esta variable, lo cual no hace una segmentación adecuada que nos permita caracterizar las observaciones pertenecientes a ese cluster.

- En conclusión, DBSCAN resulta menos eficaz para este conjunto de datos. Aunque es robusto frente a outliers y detecta puntos como ruido (en este caso, muchos de los puntos con alto desembolso), los clusters generados no presentan una segmentación clara. Por ello, para este conjunto de datos K-Means parece ser más adecuado debido a la interpretabilidad de cada uno de los clusters que creó.

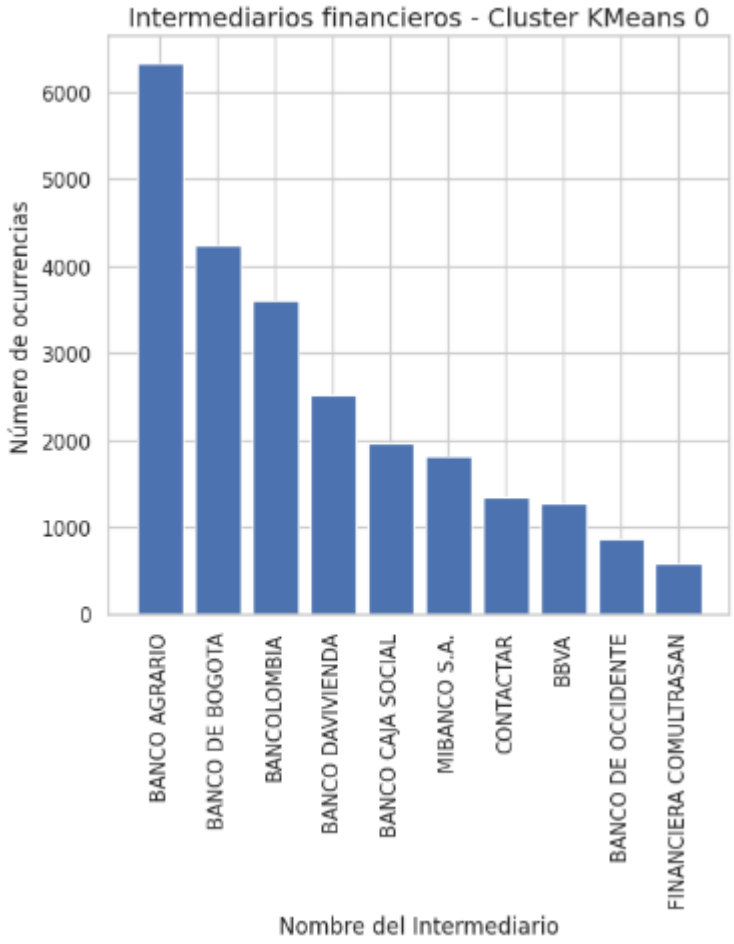
Resultados y Discusión

A continuación, podemos ver el resumen de los 5 clusters para cada método de clustering. En el k-means vemos que el cluster 0 es el más riesgoso ya que solo en 13 días se está siniestrando el crédito, es decir, el beneficiario del crédito no paga y se declara incapaz de pagar solo 12 días después de otorgada la garantía por parte de la entidad; el cluster menos riesgoso es el 1 con más de 1000 días, sin embargo, lo que llama la atención es que el total promedio de desembolso para estos dos clusters es muy similar, 1,091,342,000 para el cluster 0 y 1,439,671,000 para el cluster 1. Respecto a la calidad de la agrupación, pareciera que hay poca diferencia entre algunos clústeres en los días promedio, sin embargo, las entidades financieras tienen políticas estrictas en cuanto a ciertos umbrales apoyados en data histórica, y la diferencia más corta entre los clusters que es entre el 2 y el 4 que es de solo 29 días cae justo en un dato estadístico de entidades financieras que dice que cuando un cliente esta tan cerca de los 30 días de mora (por ejemplo) el riesgo de impago incrementa significativamente.

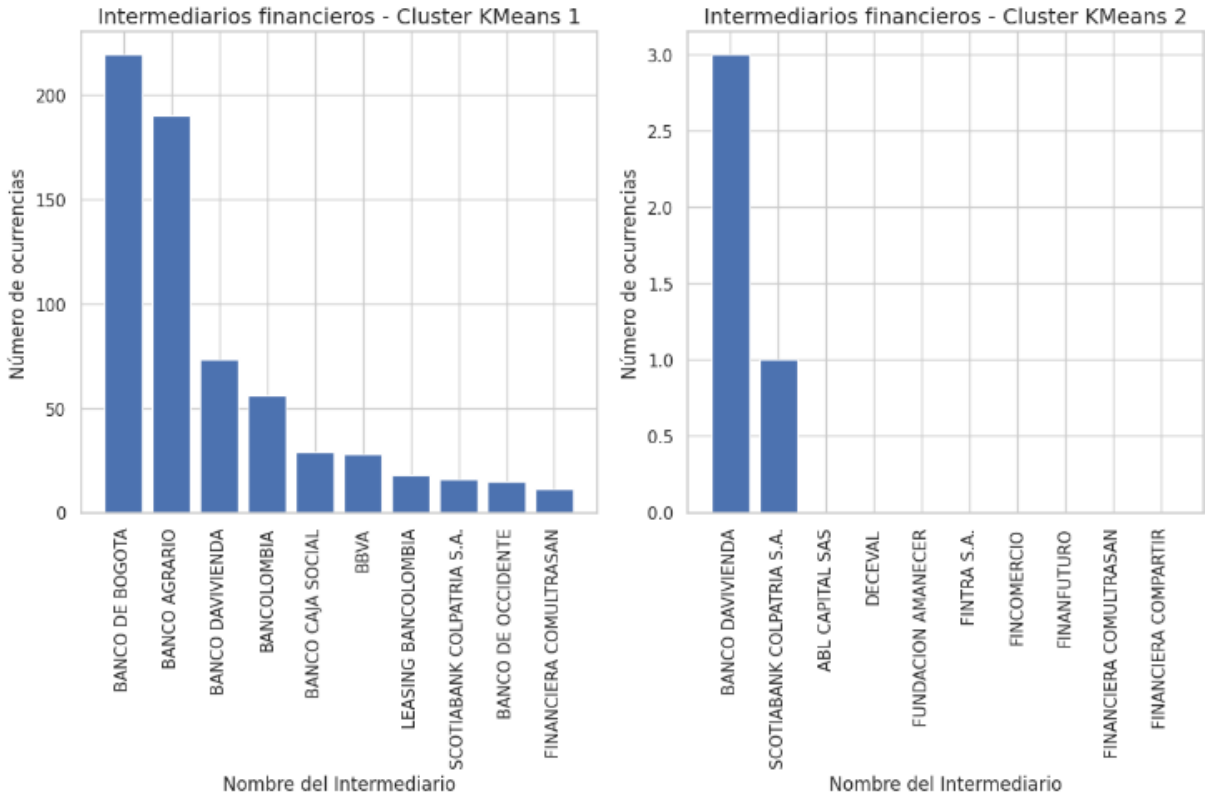
En el caso del DBSCAN, se observa que el cluster “ruido” (-1) tiene un promedio de días de 405 y el más alto promedio de desembolso, lo que indica que estos puntos no pertenecen a ningún cluster. El resto de clusteres tiene diferencias significativas, y particularmente se observa que dos clusteres tienen días promedio mayor a 2,400 días, lo que ya es muy alto y puedo no representar una real diferencia en cuanto análisis. Esto ratifica que parece que k-means hizo un mejor clustering que dbscan, por lo que los gráficos siguientes serán tomados desde este método de k-means.

```
Número de clusters para K-Means: 5
Descripción de clusters K-Means:
                Promedio_Dias_Hasta_Siniestro  Total_Desembolso
Cluster_KMeans
0                                   12.989815      1.091342e+09
1                                 1088.346561      1.439671e+09
2                                  229.250000      3.805480e+12
3                                  382.941498      2.447499e+09
4                                  200.310345      7.103448e+11
Número de clusters para DBSCAN: 5
Descripción de clusters DBSCAN:
                Promedio_Dias_Hasta_Siniestro  Total_Desembolso
Cluster_DBSCAN
-1                                 405.660714      6.934774e+11
 0                                  69.464969      1.066186e+09
 1                                2546.333333      1.256534e+09
 2                                2151.000000      4.327280e+07
 3                                 208.200000      4.149375e+11
```

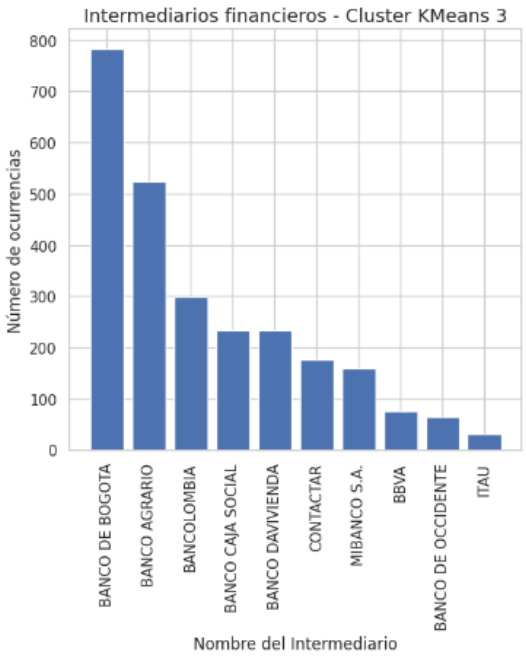
Ahora veremos unos gráficos agrupando los intermediarios financieros por la cantidad de ocurrencias en cada cluster:

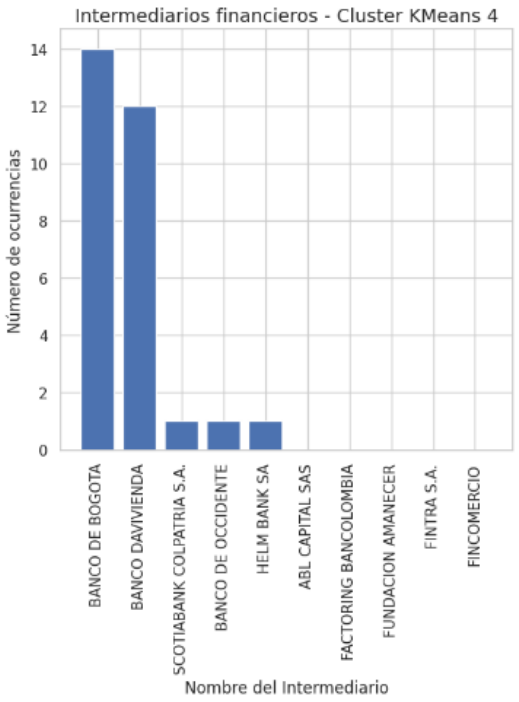


El cluster 0 es el que tiene mayor cantidad de observaciones y más riesgo. Vemos que el Banco Agrario es el que tiene una mayor cantidad de ocurrencias, seguido del Banco de Bogotá y Bancolombia. Esto lleva a abrir preguntas como: ¿Por qué estos 3 bancos son los que más ocurrencias tienen en el cluster más riesgoso? Respuesta a estas preguntas pueden ir desde el hecho que son los banco que más solicitan la garantía de la entidad, hasta le hecho de que su análisis financiero no es el adecuado para los créditos que están colocando.



Ahora, si vemos el cluster 1, que es el menos riesgoso, vemos que los 4 primeros bancos del cluster 0 se repiten, lo que podría llevar a una conclusión de que en realidad estos bancos aparecen en ambos clusters debido a que son los que más solicitan garantías a la entidad. El cluster 2 tiene muy pocas ocurrencias, y es justamente una de las desventajas del k-means, al sentirse obligado a poner todas las observaciones en un cluster.





Conclusiones y recomendaciones

Las conclusiones más precisas se pueden tomar referenciando los días promedios hasta siniestro y el promedio de desembolsos. Los montos de desembolsos de menor riesgos son en promedio de COP 1,091,342,000, lo que quiere decir que desembolsos de intermediarios en promedio menores a este monto podrían requerir más atención o monitoreo ya que tienen solo 12 días de promedio hasta el siniestro. Los desembolsos promedio por encima de COP 2,000,000,000 requieren una atención menor ya que los días promedio hasta siniestro empiezan a subir, aunque son aún menores 1 año, lo que igual puede ser considerado poco para una entidad financiera. El cluster 1 que es el menos riesgoso ya que está compuesto por operaciones con más de 3 años hasta que se siniestre, tienen un promedio de desembolsos de COP 1,400,000,000, por lo que podría indagarse más en las operaciones de estos montos para tratar de identificar características similares que hagan de estas observaciones las menos riesgosas para la entidad.

Los resultados lamentablemente no son concluyentes, por lo que se recomienda realizar clustering con otras variables para tratar de encontrar clusters que nos brinden insights más significativos, por ejemplo, promedios de días de atraso e intermediario financiero, o promedios de días de atraso y departamento. Con los resultados podrían establecerse estrategias más precisas y diferenciadas en cuanto a límites de desembolso por departamento, sector o intermediario.

Bibliografía:

Hermitaño Castro, J. A. (2022). Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática. *Interfases*, 15, 160-178. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=9039554

Guillén, R., & Torrealba, A. (s.f.). Modelo de identificación de indicadores de gestión de riesgo financiero mediante la reducción de variables o razones financieras. Recuperado de http://iies.faces.ula.ve/investiga/RGuillen/acpl_rn_rg_amt_gacl.pdf

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128, 301-315. Recuperado de: https://dl.acm.org/doi/10.1016/j.eswa.2019.02.033