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

El FNG (Fondo Nacional de Garantías) representa la entidad con la que el Gobierno Nacional de Colombia canaliza recursos para catalizar y democratizar el acceso al crédito en el país. Su función es ser garante en las operaciones crediticias requeridas por micro, pequeñas, medianas y grandes empresas, así como por trabajadores independientes, para impulsar sus actividades productivas.

El resultado principal del presente trabajo es utilizar el Aprendizaje No Supervisado para descubrir patrones y obtener información producto del análisis de datos, que permita disminuir el riesgo que asume el FNG al momento de otorgar la cobertura de garantías solicitada por el cliente e intermediario financiero, así como tomar mejores decisiones en pro de la mitigación de riesgos financieros.

Introducción

El FNG asume la responsabilidad de mitigar el riesgo asociado a la posibilidad de incumplimiento crediticio de los beneficiarios hacia entidades generadoras de crédito, bancos u otros intermediarios financieros.

Cuando se materializa el riesgo de impago por parte de un beneficiario y se produce el siniestro, el FNG se encarga de asumir una porción del saldo pendiente de la obligación en ese momento. Esto plantea un desafío estratégico fundamental que implica recursos públicos y cuya administración, con ayuda de análisis de datos, representa un punto de inflexión con diversos beneficios para la entidad y distintos actores en el sistema financiero global.

A través del análisis de datos contenidos en las bases del FNG pretendemos identificar patrones o relaciones ocultas en el conjunto de datos, de tal manera que podamos tomar decisiones orientadas a disminuir el riesgo de siniestro, por ejemplo, identificar grupos de intermediarios financieros (clustering) que tengan características similares y con base en esto definir diferentes condiciones de negociación diferentes. También podría aplicarse un Sistema de Recomendación, de tal manera que un algoritmo nos sirva para realizar ajustes a las métricas de evaluación de riesgos. También nos interesaría identificar datos que no sigan el patrón general del conjunto, lo cual sería útil para detectar mayores riesgos en cierta población de datos y fraudes.

Ya planteado el problema, la pregunta que nos gustaría resolver con este proyecto es la planteada a continuación: ¿Cómo podemos identificar y segmentar a los intermediarios financieros con características similares para optimizar las condiciones de negociación y mitigar el riesgo de siniestros en el FNG?

Revisión de la Literatura

A continuación, mostraremos 3 estudios que encontramos relacionados con la aplicación del machine learning en campos relacionados con la gestión de riesgo de créditos financieros. Para cada uno de ellos, mostraremos las similitudes y las diferencias que tienen con el proyecto que realizaremos.

A). Aplicación de Machine Learning en la gestión de riesgo de crédito financiero: Una revisión sistemática

Similitudes:

- Ambos proyectos se centran en la gestión del riesgo crediticio. Esto es una similitud clave, ya que en ambos casos se busca mitigar el riesgo asociado con el incumplimiento de obligaciones crediticias. Esto coincide con la motivación del FNG de identificar patrones y segmentar intermediarios financieros para reducir este tipo de riesgos.
- Ambos enfoques consideran la aplicación de algoritmos de machine learning para mejorar la gestión del riesgo de crédito. En el caso del FNG, la intención es utilizar clustering para segmentar intermediarios financieros, mientras que para el otro proyecto analizan las ventajas de distintos algoritmos de machine learning hibridos en la gestión de riesgos.
- Tanto en el proyecto investigado como el que se desea hacer para el FNG la interpretatibilidad que nos deben ofrecer los modelos toma gran importancia, ya que lo que se busca es encontrar patrones que nos permiten mejorar la toma de decisiones con el objetivo de disminuir el riesgo de siniesto.

Diferencias:

- El trabajo que deseamos hacer tiene un contexto especifico, en el cual el análisis se basa en los intermediarios financieros que trabajan con el FNG. Para el caso del proyecto investigado, ellos realizan una revisión más general de la gestión de riesgos de crédito.
- El enfoque que nosotros planteamos para el proyecto del FNG esta alineado con el uso de técnicas machine learning no supervisado como lo es el clustering o el de sistemas de recomendación. En el proyecto investigado tienen más un enfoque de aplicación de machine learning supervisado u de algoritmos híbridos.
- Tenemos un enfoque centrado en encontrar patrones en los datos acerca de los intermediarios financieros, por otro lado, ellos tratan de implementar modelos predictivos más complejos que muchas veces carecen de interpretabilidad.

B.) Modelo de identificación de indicadores de gestión de riesgo financiero mediante la reducción de variables o razones financieras

Similitudes:

- Ambos estudios se centran en la gestión del riesgo financiero, lo que constituye una similitud clave. En el caso del proyecto del FNG, el objetivo es mitigar el riesgo crediticio mediante la identificación y segmentación de intermediarios financieros. En el estudio revisado, el objetivo es desarrollar un modelo de pronóstico de riesgo financiero utilizando técnicas multivariantes.
- El proyecto que se desea plantear para el FNG como el modelo presentado en la fuente analizada utilizan técnicas para reducir la complejidad de la información con la que se trabaja. En el FNG, esto lo realizaremos a través de técnicas de clustering para identificar patrones y segmentar intermediarios financieros. En el estudio revisado, se utiliza el Análisis de Componentes Principales (PCA) para reducir la redundancia en los datos y facilitar el desarrollo de modelos predictivos.
- Es posible que también lleguemos a usar PCA como proceso previo al análisis de clustering, todo esto para reducir la dimensionalidad de los datos que vamos a usar y por así decirlo agrupar características de variables de intermediarios financieros que podrian estar correlacionadas.

Diferencias

- El estudio investigado se centra en la construcción de un modelo predictivo especifico para la gestión del riesgo financiero en entidades bancarias utilizando técnicas avanzadas de machine learning. Para el caso del FNG nos orientamos en la segmentación de intermediarios financieros para optimizar la toma de decisiones que permita la mitigación de riesgos.

- El estudio aplica algoritmos tanto de análisis supervisado como de no supervisado tal como lo es aplicar redes neuronales y el uso de PCA, en nuestro caso tendremos un enfoque de aplicar técnicas de machine learning no supervisado como lo es el clustering.
-

C.) Integration of unsupervised and supervised machine learning algorithms for credit risk assessment (biblioteca U)

Similitudes

- Ambos trabajos comparten el objetivo de mejorar la evaluación del riesgo crediticio. En el proyecto del FNG, se busca mitigar el riesgo de siniestros mediante la identificación y segmentación de intermediarios financieros. De manera similar, el estudio se enfoca en mejorar los modelos de evaluación de crédito para discriminar entre solicitantes buenos y malos.
- En ambos casos, se propone el uso de técnicas de aprendizaje no supervisado para mejorar la precisión en la evaluación del riesgo. El proyecto del FNG menciona la posibilidad de usar PCA y clustering, mientras que el estudio aplica técnicas no supervisadas entre las cuales mencionan tambien el clustering en diferentes etapas del proceso. Además, en el estudio proponen un esquema hibrido en el que usan modelos no supervisados junto con modelos supervisados.

Diferencias

- El estudio propone una integración de aprendizaje supervisado con no supervisado en múltiples etapas del proceso, mientras que el proyecto del FNG solo nos centraremos en la aplicación de modelos de aprendizaje no supervisado enfocados en la identificación y segmentación de los intermediarios financieros.
- El enfoque del estudio está orientado hacia la mejora del credit scoring en general, aplicable a un conjunto amplio de datos crediticios. El proyecto del FNG, por otro lado, tiene un enfoque más específico en la gestión de riesgos relacionados con intermediarios financieros y la mitigación de siniestros.

Descripción de los datos

1.

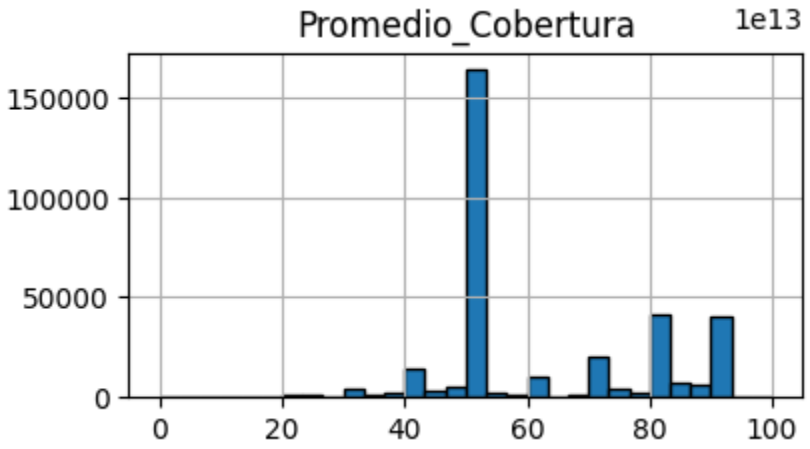
```
RangeIndex: 326768 entries, 0 to 326767
Data columns (total 19 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   Nit_Intermediario               326768 non-null  int64
 1   Nombre_Intermediario            326768 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

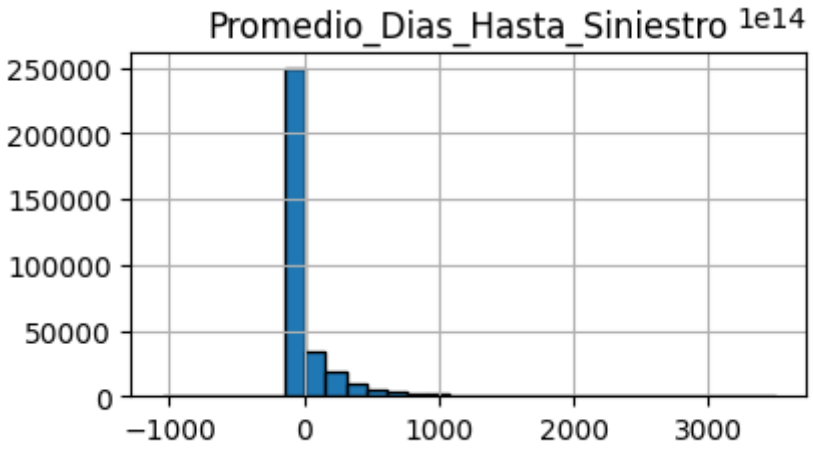
	Nit_Intermediario	Total_Desembolso	Total_Saldo	Promedio_Cobertura
count	3.267680e+05	3.267680e+05	3.267680e+05	326768.000000
mean	8.459302e+09	2.517682e+09	1.078137e+10	61.149385
std	3.423482e+08	9.516188e+10	5.008020e+11	17.064842
min	8.000116e+09	0.000000e+00	0.000000e+00	0.000000
25%	8.001479e+09	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

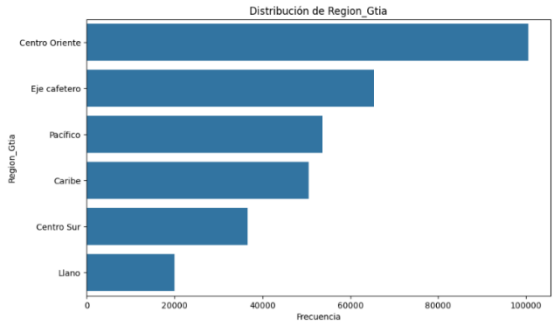
Total_Desembolso tiene una media de aproximadamente 2,517,682,000 con una desviación estándar de 95,161,880,000, lo que indica una gran variabilidad en los desembolsos, que van desde 0 hasta un máximo de 31,770,040,000,000. **Total_Saldo** presenta una media de 10,781,370,000 y una desviación estándar 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.

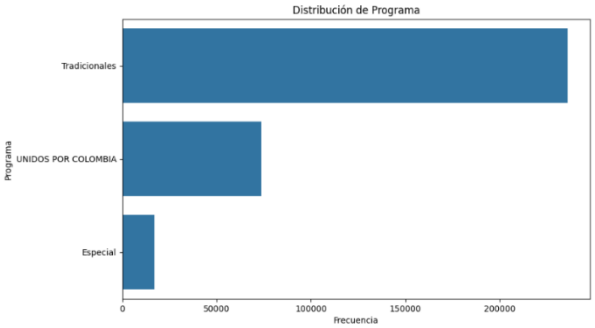


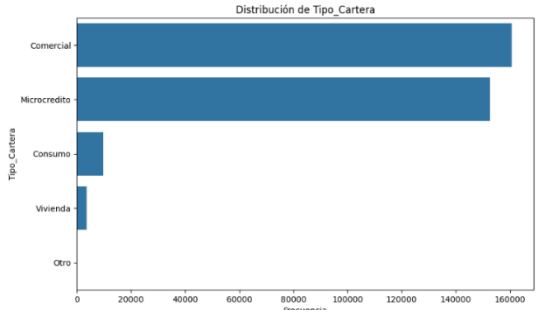
Vemos que la mayor parte de las coberturas estan en el 50%

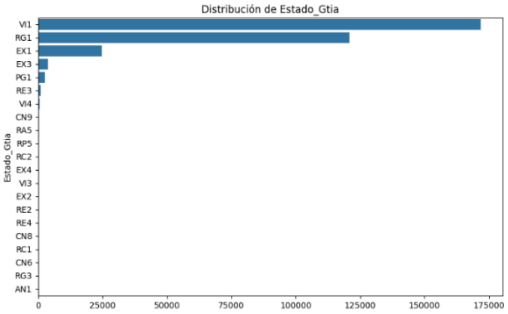


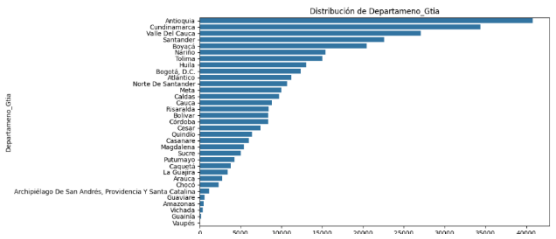
Hay que realizar una limpieza ya que hay valores negativos en los dias promedio del desembolso hasta el siniestro lo cual no es posible

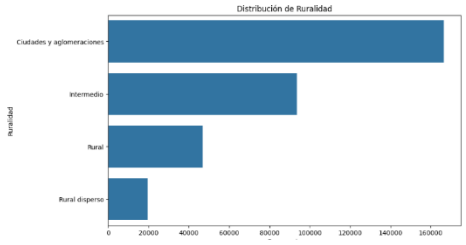


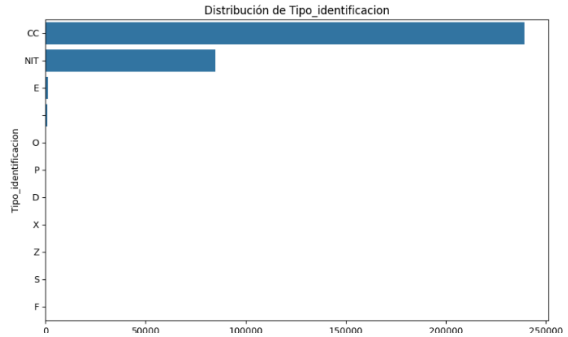


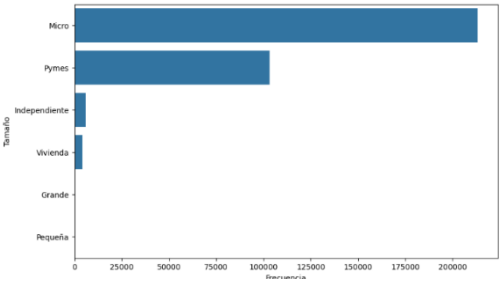


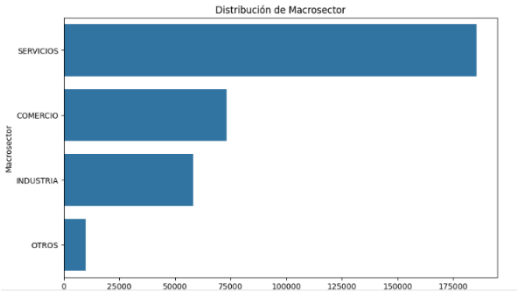






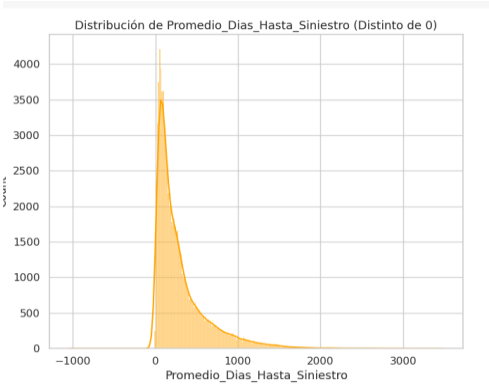


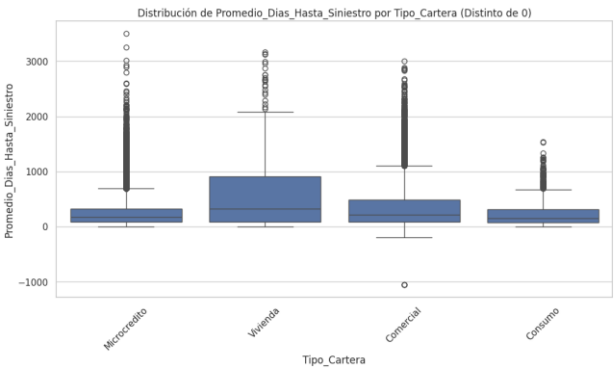


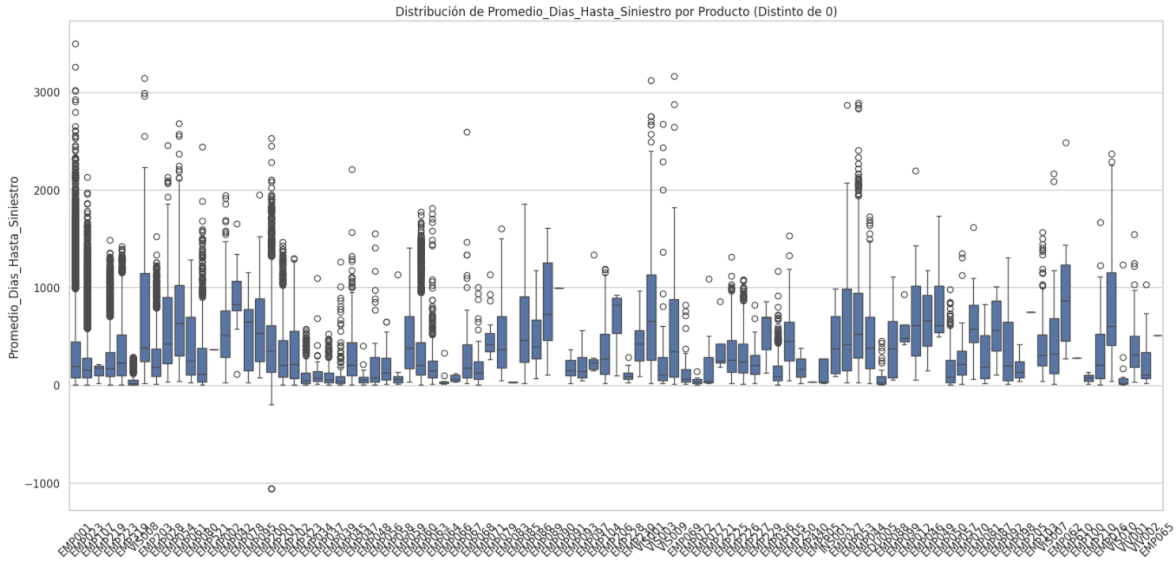


El análisis de los gráficos de barras nos permite entender la distribución de la muestra. Observamos que la mayoría de los datos corresponden a personas naturales, con una concentración en la cartera comercial y, en segundo lugar, en microcrédito. Como era de esperarse, la mayor parte de los datos se 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.





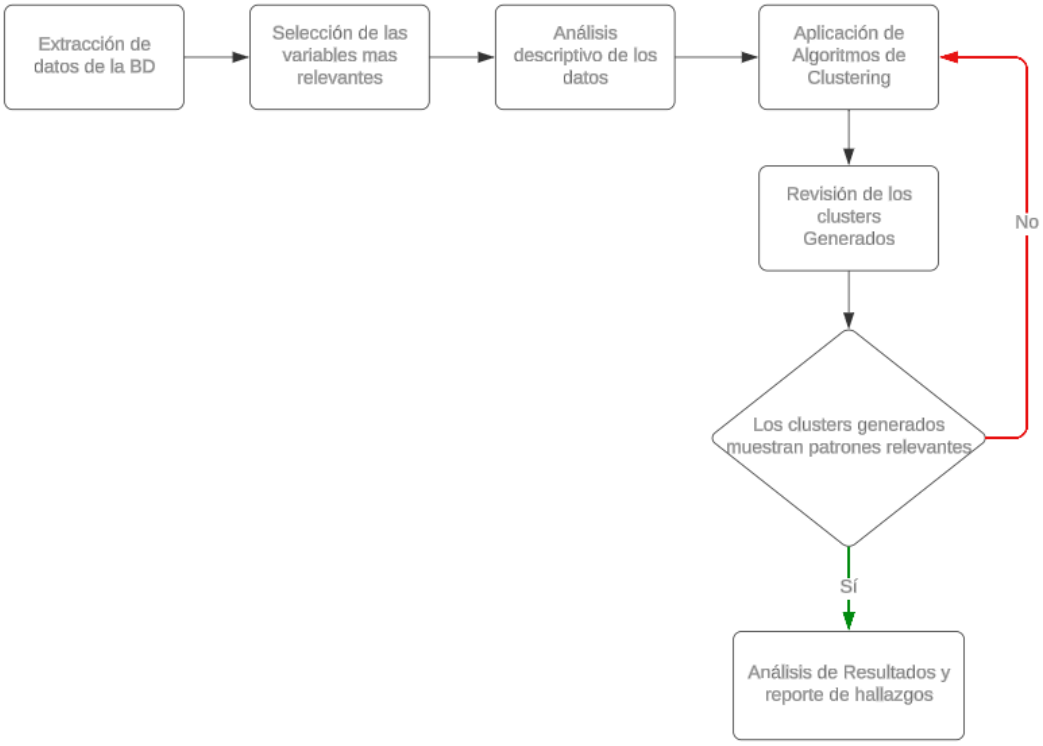


Observamos que el período desde la emisión de la garantía hasta que se siniestra se concentra, en promedio, alrededor de los 500 días. También identificamos que el microcrédito es el tipo de producto en el que las reclamaciones se realizan más rápidamente. Además, el producto EMP319 destaca como clave,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

Metodología del proyecto

A continuación, hablaremos acerca de la metodología que utilizaremos a lo largo del desarrollo del proyecto, la cual podría llegar a variar dependiendo de la calidad de los resultados obtenidos, o de la necesidad que pueda surgir de incluir algún paso adicional para mejorar dichos resultados

El siguiente diagrama de flujo muestra el paso a paso de la metodología inicial en la que trabajaremos a lo largo del proyecto:



1. **Extracción de datos de la BD:** En este proceso realizamos la extracción de los datos que utilizaremos para trabajar en el proyecto. Esta información fue extraida directamente de la base de datos del FNG y básicamente tiene información relevante de los intermediarios financieros.
2. **Selección de las variables más relevantes:** En este paso realizamos la selección de las variables más relevantes que debería tener el dataset a analizar. Para este paso fue muy importante la opinión de un experto el cual hace parte de nuestro equipo de trabajo. El bajo sus conocimientos del negocio selecciono las variables más relevantes para el análisis que se deseaba hacer con los intermediarios financieros.
3. **Análisis descriptivo de los datos:** En este proceso con ayuda de python se realizó el análisis descriptivo de los datos con el fin de poder familiarizarnos con las variables que vamos a trabajar a lo largo del proyecto. Se realizo tanto un análisis de las variables cuantitativas como cualitativas, se conoció la dimensión de nuestro dataset e identificamos patrones preliminares encontrados a partir del análisis descriptivo realizado.
4. **Aplicación de Algoritmos de clustering:** Posteriormente al análisis descriptivo de los datos, procederemos a la implementación de algoritmos de clustering vistos en clase. El objetivo de esta sección es aplicar diferentes métodos de clustering a los datos con los que vamos a trabajar.
5. **Revisión de los clusters generados:** Luego de aplicar los algoritmos de clustering, procederemos a analizar los clusters generados por cada algoritmo. De la mano del experto y de los conocimientos adquiridos de los datos en los análisis descriptivos evaluaremos la calidad de los clusters generados a partir de la información que podamos extraer de ellos. Es decir, evaluaremos si los clusters generados nos permiten identificar grupos de intermediarios financieros que tengan algún patrón en común, patrón que pueda servir de análisis para el FNG como soporte en la toma de decisiones.

En caso de no encontrar información relevante acerca de los intermediarios financieros optaremos por ejecutar un nuevo algoritmo ya sea cambiando los parámetros del algoritmo de clustering utilizado, o cambiando las variables a utilizar en el análisis.

6. **Análisis de resultados y reporte de hallazgos:** Para esta última parte, mostraremos los resultados obtenidos con la explicación de como estos clusters generados nos sirven de insumo para soportar la toma de decisiones en el FNG, y que conclusiones relevantes nos permite extraer de los intermediarios financieros analizados.

Debemos considerar que esta metodología puede cambiar a partir de los resultados que vayamos obteniendo en el desarrollo del proyecto. Probablemente debemos incluir una fase de preprocesamiento en la cual se deba hacer una limpieza de los datos con los que vamos a trabajar y su respectiva estandarización. Además, podríamos optar por incluir otros algoritmos de análisis no supervisado, tal como podría ser aplicar PCA a las variables cuantitativas que contamos para reducir la dimensionalidad del dataset con que trabajaremos.

Esta metodología es la correcta para el proyecto planteado porque aborda de manera detallada cada etapa necesaria para obtener resultados significativos y aplicables a la toma de decisiones del FNG. El proceso comienza con la extracción de datos directamente desde la base de datos, garantizando que la información sea relevante. La selección de variables basada en la experiencia de un experto nos asegura que el análisis se enfoque en los aspectos o características más relevantes para los intermediarios financieros. El análisis descriptivo de los datos nos permite familiarizarnos con las variables y detectar patrones preliminares, facilitando una mejor comprensión del comportamiento de los datos.

El uso de algoritmos de clustering es adecuado para este proyecto, ya que permite identificar grupos de intermediarios financieros con características comunes lo cual nos permite resolver la pregunta que planteamos en el objetivo de este proyecto mediante la obtención de insights valiosos. La revisión continua de los clusters generados, en conjunto con la experiencia adquirida del estudio de los datos y con el apoyo de los conocimientos de negocio del experto, nos garantiza que los grupos identificados sean significativos y útiles para un análisis posterior. En resumen, esta metodología es la ideal para generar resultados que sean útiles para la toma de decisiones en el FNG.

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