Brahayan Stiven Gil Henao
Jhon Alexander Bedoya Carvajal
Sebastián Salamanca Mendez
Ana María Ospina Arredondo

# Informe Resultados Caso de Estudio - Aprendizaje No Supervisado

## Planteamiento del problema

Las instituciones financieras tienen un registro de deudas contraídas por diversos individuos, empresas o grupos empresariales, donde presentan un riesgo mayor o menor según el tiempo de mora de la deuda, la base de datos que se pretende analizar a continuación tiene la recopilación de información sobre las relaciones financieras con sus diferentes clientes, dando un registro de la deudas y pagos efectuados a las diferentes entidades y sus productos, algunas de las variables que se en encontrarán en esta base son:  tipo de entidad, descripción, renglón, vencimientos, saldo a la fecha.

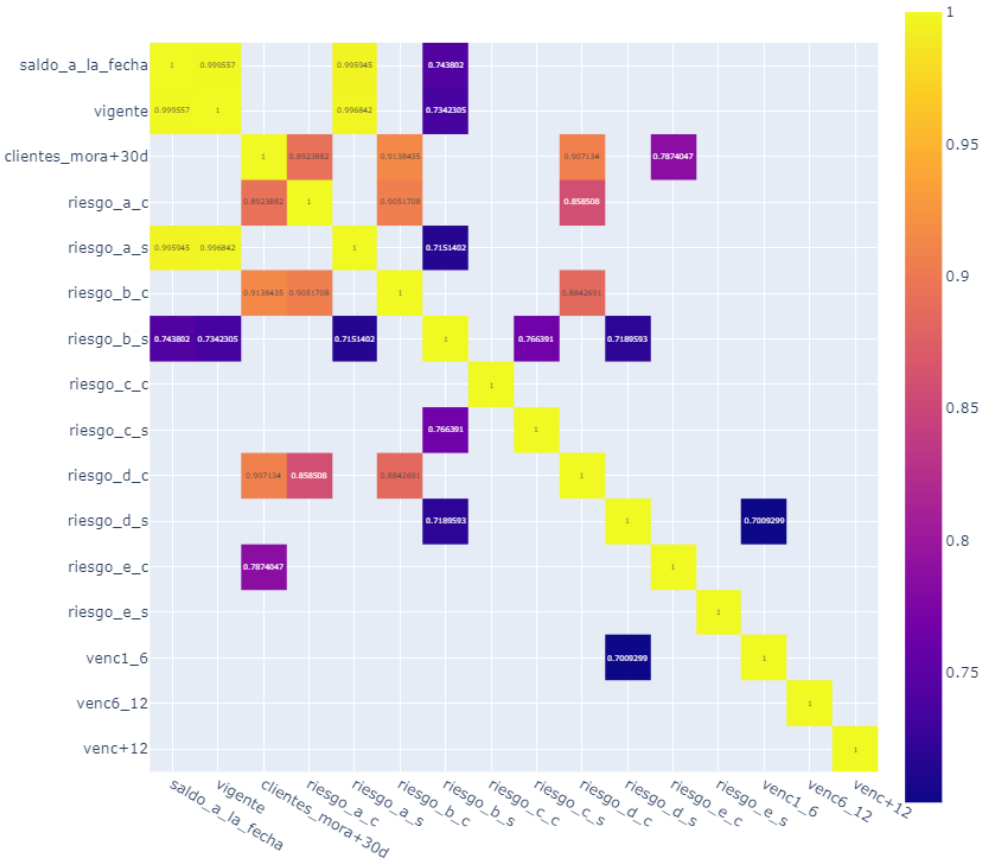
## Limpieza y transformación

Para la solución a esta base de datos se observó la información de cada una de las variables, se eliminaron las columnas que no aportan información significativa, además, se realizó la transformación de los nombres para facilitar su manejo.
Para la columna "tipo_entidad" se decide reemplazar el código por el nombre correspondiente según la superintendencia.

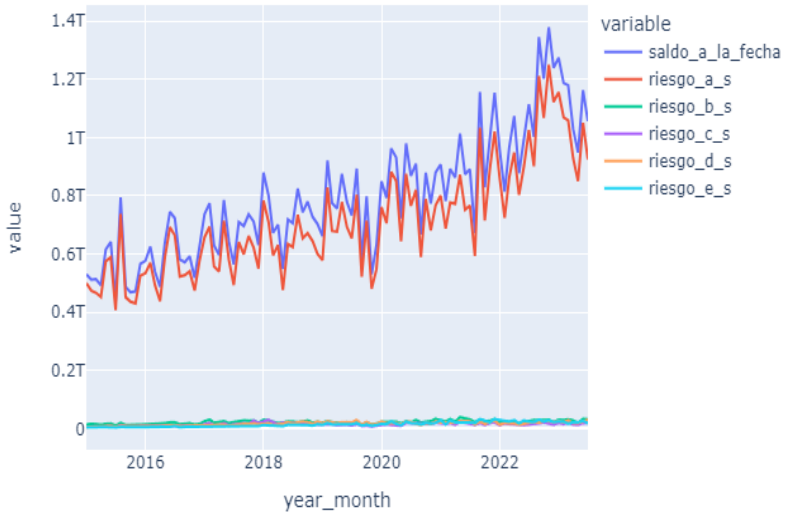
Para las variables numéricas se analiza si hay presencia de valores negativos, dado que una deuda es de valores positivos o cero en caso de que haya sido saldada, los negativos encontrados fueron eliminados bajo el supuesto que sea un error en la recopilación de datos, además se realizó el tratamiento de datos nulos llenandolos con 0 suponiendo que no hubo abono. Si no hay registros de abonos durante un mes dado, el saldo a la fecha y el saldo vigente deben ser iguales, también se eliminan estos valores donde su suma sea diferente.  La variable Renglón contiene más información que Unicap, por lo que se decide trabajar solo con la primera variable. Finalmente, se exporta la base de datos limpia y se procede a trabajar con ella.

## Análisis de correlación

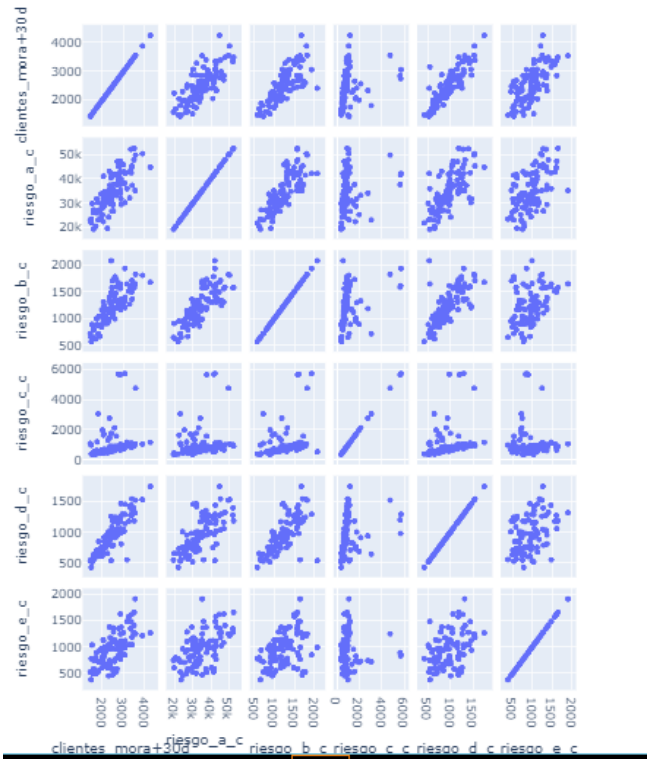
Para realizar el análisis de correlaciones se decidió tomar primero una muestra de 25000 datos aleatorios a los cuales se les conservó porcentajes similares con base en la variable "tipo_entidad". Se realizó un gráfico de correlación entre las variables numéricas, en el cual se encontró una alta correlación entre "saldo_a _fecha" y "saldo_vigente", dado que saldo vigente  representa el saldo luego de los  abonos de los clientes a cada una de las cuentas en cada mes, se decidió descartarla ya que nos proporciona la misma información que la suma de los abonos dados a cada cuenta vencida.




Respecto a las altas correlaciones con algunos riesgos en sus saldos, se debe a que estas representan la cantidad  total de lo que hay en cada una de las cuentas que se tiene de cada uno de los riesgos, un ejemplo de lo que se puede observar en el gráfico es que “riesgo_a_s” está altamente correlacionado con “saldo_a _la_fecha” dado a que la mayoría de clientes está categorizado como clientes con saldo de riesgo tipo a, además analizado otros gráficos se puede ver un comportamiento casi igual entre ambas variables en el tiempo, por lo que se decide descartar las columnas de saldos y dejar saldo total a la fecha.

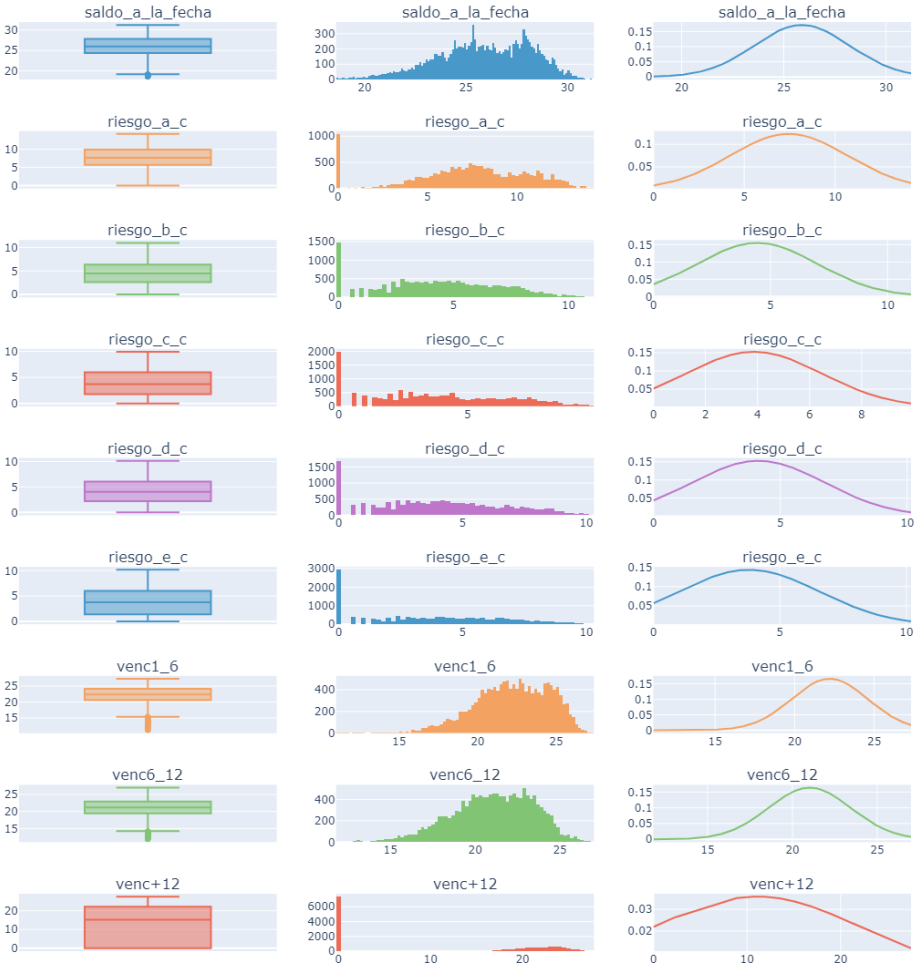


Para el manejo de la alta correlaciòn que existe entre el riesgo con el número de clientes con más de 1 mes en mora se decide hacer un gráfico de disperciòn para observar con qué variables maneja una relaciòn lineal, y se observa la alta relaciòn con 4 variables, por lo cual se decidió descartarla.




## Análisis exploratorio de datos

Para el análisis exploratorio de los datos se analizan los tipos de entidad presentes en la base de datos donde cerca del 70% son establecimientos bancarios, se mira también los diferentes tipos de préstamos de las entidades y se evidencia que todos los tipos de préstamos se realizan con una frecuencia muy parecida, se realizó el análisis para las variables de riesgo y vencimiento pero por la escala de esta no era clara la interpretación por lo tanto se decide aplicar logaritmo a estos valores para que todas queden en una escala menor.




Al tener los datos en una escala menor si era posible su interpretación y se evidencio que el saldo a la fecha tenía mayormente valores altos, las variables de riesgo tenían muchos valores en 0 y había una acumulación de clientes de riesgo tipo A mayor que las de los otros tipos, se evidencio que los abonos en las cuentas vencidas entre 1-12 meses eran altos, y no había muchas cuentas vencidas a más de 12 meses.

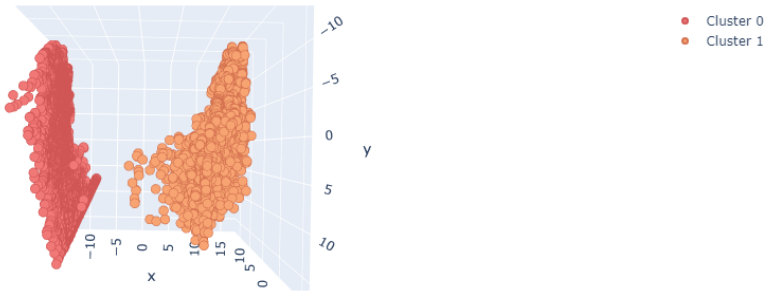
## Selección de características y Modelos

Se aplica un PCA a los datos, obteniendo 3 componentes que conservan el 95% de la varianza de los datos.

El primer modelo ajustado es un K-Means, se ajustó tanto con los datos completos como con los componentes resultados de la PCA. Se aplica la técnica del codo con el coeficiente de inercia obteniendo un número óptimo de clusters de 3. Se instancia el modelo K-Means aleatorio con 3 clusters, 100 inicializaciones y un máximo de iteraciones de 1000. Se obtuvieron las predicciones y se evaluó el modelo. Los mismos pasos anteriores se aplicaron con los 3 componentes obtenidos en la PCA, obteniendo en este un número óptimo de clusters de 2 y permitiendo ser graficados.

El segundo modelo fue un DBSCAN, de igual manera tanto con los datos completos como con los componentes obtenidos en el PCA. Con los datos completos se obtuvo un número óptimo de clusters de 3 y mucho ruido, con los componentes obtenidos en el PCA toma todos los datos como ruido.

Ya que el modelo DBSCAN no dio buenos resultados, se decide ajustar también un modelo jerárquico aglomerativo. Tanto con los datos completos como con los componentes resultantes del PCA el dendrograma nos muestra un número óptimo de clusters de 2. Ambos tuvieron un rendimiento similar respecto a las métricas, pero se decide tomar los clusters suministrados por el modelo jerárquico aglomerativo ajustado con los componentes obtenidos en la PCA ya que nos permite verlos gráficamente.




## Conclusiones y recomendaciones

Se termina con dos grandes grupos de datos que dividen las carteras de esta forma

**Menor abono y menor riesgo(Cluster 0)**

- Menor abono a las cuentas vencidas
- Mayor cantidad de clientes con riesgo tipo A que son los de menor riesgo
- Menor cantidad de clientes con riesgo tipo B,C,D,E que son lo de mayor riesgo de incumplimento
- Presenta valores más bajos de saldo a la fecha.

**Mayor abono y mayor riesgo(Cluster 1)**

- Mayor abono a las cuentas vencidas
- Menor cantidad de clientes con riesgo tipo A que son los de menor riesgo
- Mayor cantidad de clientes con riesgo tipo B,C,D,E que son lo de mayor riesgo de incumplimento
- Presenta valores más Altos de saldo a la fecha.