

Uso de aprendizaje no supervisado para identificar perfiles de riesgo en entidades de crédito

Grupo 1, Aprendizaje No supervisado, Maestría en Inteligencia Analítica de Datos, Universidad de los Andes

Resumen: En estos tiempos de incertidumbre económica, y ante el cierre de entidades financieras de alto renombre mundial, se hace necesario, tanto para entidades públicas como para grandes inversionistas privados, encontrar metodologías de alta eficiencia y fácil ejecución en poco tiempo para poder determinar el perfil de riesgo de las entidades bancarias de Colombia. Mediante un análisis de clusterización apoyado en componentes principales para la reducción de dimensionalidad se logró identificar claramente un perfil de riesgo para distintas entidades financieras, aunque se recomienda realizar una profundización en las variables que describen el perfil, así como la inclusión de información histórica para mayor precisión

Introducción

La supervisión bancaria, como señalan Hotori, et. al. (2022), corresponde a una necesidad del sistema financiero actual, como una respuesta al control que requieren las entidades públicas ante fenómenos como la información asimétrica que se posee sobre el mercado entre entidades financieras públicas y privadas, la constante fluctuación de externalidades y los riesgos de la economía moderna. Incluso, entidades de alto reconocimiento a nivel mundial como Credit Suisse (Portafolio, 2022) o Silicon Valley Bank (SVB) (Ferguson, 2023), que recientemente tuvieron que declararse en bancarrota debido a la falta de supervisión en tiempo real, hacen prioritario realizar la pregunta: ¿Cuál es una posible metodología que permita realizar una categorización y análisis del riesgo de las distintas entidades financieras de un país, en este caso particular de Colombia?

Este tipo de análisis resulta de interés para dos grupos:

- Entidades de supervisión bancaria en Colombia: Si bien, como señalan Estrada y Gutierrez (2009), la labor de supervisión bancaria en Colombia por parte del Banco de la República, así como de la Superintendencia Financiera de Colombia (SFC), ha resultado satisfactoria en el control de crisis y monitoreo periódico de las entidades bancarias, resulta necesario establecer criterios de medición que permitan tomar decisiones ante cambios súbitos y que sean de construcción e implementación más acelerada
- Inversionistas extranjeros: Autores como Kandrak y Schlusche (2017) evidencian que, en escenarios donde la supervisión bancaria es mucho menor, era mayor la incidencia de pérdidas significativas en la inversión. Si bien la volatilidad del mercado extranjero y el score de riesgo de Colombia ha ido en aumento, un sistema eficaz de supervisión bancaria brinda confianza a los inversionistas.

Como es tradicional en los perfiles de riesgo, no se cuenta con categorías o definiciones fijas de los perfiles de riesgo, por lo que modelos de aprendizaje supervisado no se consideran como una respuesta al problema propuesto. Para este caso particular se considerará como un problema de clasificación no supervisada, por lo que se revisarán a

detalle algoritmos de clustering. También, de ser necesario, se revisará la necesidad de reducción de dimensionalidad.

Antecedentes bibliográficos

A nivel internacional, la revisión realizada por Guerra y Castelli (2021) brinda el mayor acercamiento al uso de machine learning para la supervisión bancaria. Llama la atención observar el volumen de trabajos que utilizan un enfoque supervisado para la predicción de los perfiles de riesgo (De una revisión de 41 papers publicados entre el 2000 y el 2021, únicamente 1 hace uso de técnicas de aprendizaje no supervisado).

Entre las técnicas más utilizadas se encuentra el uso de bagging y boosting, así como redes neuronales, para la determinación de perfiles de riesgo. El estudio más referenciado por bibliografía académica (y también el único que emplea técnicas de aprendizaje no supervisado), dentro de los revisados por Guerra y Castelli (2021), es el realizado por Boyacioglu et. al (2009), en el cual revisa diferentes metodologías de machine learning que permitieran identificar bancos que estuvieran en riesgo de una posible bancarrota. Dentro de las metodologías analizadas se utilizó el algoritmo de k-medias, sin embargo no obtuvo un desempeño superior frente a metodologías más robustas. En este caso, los métodos propuestos no permiten solucionar el problema ya que no se ha llegado a la bancarrota y no se tiene una única variable de respuesta, por lo que el aprendizaje no supervisado sigue siendo la metodología recomendada. Dentro de los estudios realizados con aprendizaje no supervisado más relevantes para resolver la pregunta de interés, se encuentra el estudio de Lee, Booth y Alam (2005), en el cual compara redes neuronales artificiales tanto supervisadas como no supervisadas (denominadas redes Kohonen) para la predicción de bancarrota de entidades financieras en Corea. De dicho estudio es posible concluir que, si bien las redes no supervisadas no poseen el mismo poder predictivo que aquellas que utilizan aprendizaje supervisado, permiten realizar predicciones en un tiempo más cercano, además que permite evaluar casi en tiempo real tanto la evolución de la entidad como de sus alrededores para tomar decisiones. Finalmente, en la revisión de literatura colombiana se encontró el realizado por Giraldo y Marín (2021) acerca de machine learning para riesgo de crédito, pero no para clasificar los bancos en niveles de riesgo, por lo que esta investigación supondrá un pivote académico para incentivar la profundización en el uso de machine learning para perfilar el riesgo de bancos colombianos.

Descripción de los datos

De la superintendencia financiera de Colombia, se obtuvieron datos de indicadores de salud financiera desde enero hasta mayo de 2023 (última información) para 28 diferentes entidades bancarias, dentro de los indicadores que se obtuvieron están la solvencia, IRL, Cartera/Depósitos, Cartera/Activos, Gastos Operativos/Activos, ROA, ROE, Calidad y Utilidad/ingresos.

Para hacer el análisis preliminar nos apoyamos en el uso de estadísticas descriptivas y cálculos de correlación lineal entre las variables como se aprecia en la figura 1:

Con esto se logró llegar a las siguientes conclusiones:

- Utilidad/ingresos vs. ROA y ROE: Existe una correlación lineal positiva muy fuerte entre la utilidad como proporción de los ingresos y las variables de rentabilidad. Esto sugiere que a medida que el ROA aumenta, la empresa tiende a generar una mayor utilidad en relación con sus ingresos.
- Existe una relación positiva entre solvencia e IRL. Entre más altos sean estos indicadores y están por encima del límite regulatorio, la entidad será más saludable y tendrá mayores posibilidades de enfrentar una crisis.
- Gastos operacionales/Activos: La variabilidad en esta relación es moderada, y la media indica que, en promedio, las instituciones gastan aproximadamente el 4.3% de sus activos en gastos operativos.
- ROA (Return on Assets) y ROE (Return on Equity): Ambas variables tienen medias positivas, lo que indica que, en promedio, las instituciones son rentables en términos de activos y patrimonio. Sin embargo, la variabilidad en estas tasas es alta, y hay observaciones con rendimientos negativos, especialmente en ROE como lo es por ejemplo para el caso del banco Pichincha S.A.

Estas variables nos pueden ser útiles para calcular el riesgo de las instituciones financieras, ya que ofrecen una visión de diferentes aspectos de su salud financiera, como, por ejemplo:

- Solvencia: Mide la capacidad de una institución financiera para cumplir con sus obligaciones financieras a medida que vencen. Un bajo nivel de solvencia podría indicar un mayor riesgo de incumplimiento. Esta variable es fundamental para evaluar el riesgo crediticio. En Colombia, las entidades deben tener este indicador por encima del 9%.
- IRL (Ingresos / Recursos Líquidos): Un IRL alto sugiere una mejor capacidad para generar ingresos con los recursos disponibles, lo que puede reducir el riesgo financiero. En Colombia, las entidades deben cumplir con este indicador por encima del 100%.
- Cartera/Depósitos y Cartera/Activos: Estas relaciones son relevantes para evaluar el riesgo de concentración de activos y la calidad de la cartera de préstamos. Un alto ratio de Cartera/Depósitos puede indicar una dependencia significativa de los préstamos en relación con los depósitos, lo que podría aumentar el riesgo si los préstamos no son de alta calidad. Por otro lado, un bajo ratio de Cartera/Activos podría señalar una mayor diversificación de activos, lo que podría reducir el riesgo.
- Gast_Op/Activos: Esta variable proporciona información sobre la eficiencia operativa de la institución financiera. Un alto nivel de gastos operativos en relación con los activos podría indicar ineficiencias y aumentar el riesgo de rentabilidad.

En resumen, estas variables proporcionan una visión integral de la salud financiera de las instituciones financieras y son fundamentales para evaluar el riesgo en diferentes aspectos, como el riesgo crediticio, el riesgo operativo y el riesgo de rentabilidad.

Propuesta metodológica

En este documento proponemos calcular una medida de riesgo para los bancos colombianos que recoja el comportamiento de los indicadores mencionados en la sección anterior y que contribuya a la labor de supervisión que realiza la Red. Para ello se utilizará

la metodología de componentes principales que permite la reducción de la dimensionalidad de los datos y el indicador sería el primer componente principal el cual explica la mayor varianza de los datos. Así mismo, esta metodología permite conocer el peso que tiene cada variable sobre el componente principal a partir de los *loadings*. Por otro lado, se realizará un análisis de clústeres para identificar aquellos bancos que tienen perfiles de riesgo similares para que la Red supervise en mayor medida aquellos clústeres con mayor riesgo evitando posibles futuras quiebras de bancos o contagio sobre el resto del sistema financiero. En este documento se probarán distintos métodos y a partir de los resultados se seleccionará el que genere los mejores resultados.

Resultados

1. Cálculo Indicador de riesgo mediante componentes principales

Como se mencionó anteriormente se propone calcular un indicador de riesgo para las entidades bancarias. En este caso, se hace uso de componentes principales para reducir la dimensionalidad de las variables sobre salud financiera y se selecciona el primer componente principal que es capaz de explicar del 41.5% de la variabilidad de los datos. Al usar componentes principales encontramos algunas ventajas, por un lado podemos ver la variabilidad del indicador a lo largo del tiempo y a partir de los *loadings* podemos conocer cuál variable tiene mayor relevancia dentro del indicador.

En la figura 2 se presenta el indicador de riesgo desagregado por el peso que tiene cada variable para el mes de mayo. Entre más alto sea este indicador más riesgosa es la entidad. En general vemos que a pesar de la coyuntura actual el sistema financiero colombiano se mantiene sólido. El G4 conformado por los cuatro principales bancos del país (Bancolombia, Banco de Bogotá, Davivienda y BBVA) mantienen valores negativos y son muy similares entre ellos. Por otro lado, vemos que las tres entidades bancarias con mayor riesgo son Banco Pichincha, Banco Falabella y Banco Unión. En cuanto al primer banco, en el mes de marzo la calificadora de riesgo Fitch redujo su calificación dada la profundización de las pérdidas operativas, niveles menores de capitalización y deterioros en la calidad de activos. Lo anterior debido a la decisión del banco de reformular su modelo de negocio y acelerar su transformación digital. Por el lado de Banco Falabella se observa que es una entidad que tiene una fuerte concentración en crédito de consumo, específicamente en tarjetas de crédito y en un contexto de altas tasas de interés más desaceleración económica se ha visto afectada por un aumento en los niveles de morosidad. Por último, tenemos Banco Unión que anteriormente era una compañía de financiamiento y también se encuentra en un proceso de transición al convertirse en banco. La característica común entre estas entidades es que tienen un componente de cartera sobre activos elevado, es decir son entidades con una alta exposición a riesgo crediticio que se han visto afectadas por el aumento en la morosidad.

2. Identificación de clusters

En un primer intento se utiliza k-medias y k-medoides sobre las 9 variables seleccionadas y con el método del codo se seleccionan 3 clusters, sin embargo vemos que la asignación de las entidades a estos clusters presenta problemas. Por ende, se decide usar los métodos de clusterización sobre los tres primeros componentes principales que explican cerca del

81.5% de la variación de los datos. Con componentes principales, se observa una mejora frente a cuando se realiza con las 9 variables iniciales como se observa en los gráficos a continuación y entre k-medias y k-medoides se observa una mejor clasificación en el primero. La caracterización de clusters por cada uno de los métodos puede verse en la figura 3. Cuando se utiliza k-medias sobre los tres componentes, las entidades menos riesgosas quedan agrupadas en el primer cluster mientras que la mayoría de entidades (inclusive las más riesgosas) quedan clasificadas en el cluster 2 y la más riesgosa que es pichincha queda en el 3. Sin embargo, utilizando k-medias y k-medoides vemos que la clasificación no es del todo perfecta y por esta razón realizamos otro intento con cluster jerárquico. Los resultados de este cluster se pueden ver en la figura 4. Al realizar el ejercicio de clusterización con los datos estandarizados, por el método de clúster jerárquico se evaluaron los diferentes métodos de enlace entre los puntos, Se evidencio que en algunos casos se genera una superposición de los datos de los clusters lo cual no permite un análisis claro de las agrupaciones generadas, en este ejercicio el modelo con el mejor score de Silhouette corresponde a al modelo con enlace tipo single con un score de 0.56 generando 4 grupos diferentes. posteriormente se realizó el ejercicio de clusterización con una reducción de la dimensionalidad por medio de PCA de igual manera evaluando los diferentes métodos de enlace entre los puntos, Comparando estos resultados con los obtenidos antes de realizar PCA se evidencia una mejora significativa en la generación de las agrupaciones, siendo inclusive evidente al ojo humano las diferentes agrupaciones generadas, lo cual permite un análisis más claro de las agrupaciones, en este ejercicio el modelo con el mejor score de Silhouette corresponde a al modelo con enlace tipo “average” y “centroid” para los cuales se obtuvo un score de 0.601, en la gráfica de este método se observa cómo al reducir la dimensionalidad, las observaciones que para este caso son cada una de las entidades financieras se segmentan de tal manera que forman 5 grupos y esto con respecto al perfil de riesgo financiero que se pretende analizar puede ser muy diciente ya que según los valores de los 2 primeros componentes mostrados en la gráfica y teniendo en cuenta que entre más alto es el indicador más riesgosa es la entidad se pueden ver los grupos de entidades muy riesgosas , poco riesgosas, sólidas, etc en cuanto a riesgo financiero.

Conclusiones y trabajo futuro

El uso de reducción de dimensionalidad permite realizar una calificación preliminar de las distintas entidades financieras, permitiendo realizar un perfilamiento y además permitiendo una mejor segmentación al momento de realizar grupos (clusters)

De los algoritmos utilizados, el algoritmo de clustering jerárquico permitió identificar con mayor claridad los distintos grupos por nivel de riesgo. Esto sumado a la poca capacidad computacional que requiere, lo convierte en el método ideal para plantear una solución analítica de fácil implementación

En posteriores etapas se sugiere incrementar el número de variables que contribuya a construir componentes loadings más robustos, así como la incorporación de información histórica, para ver una evolución del riesgo a lo largo del tiempo y también poder determinar comportamientos o perfiles atípicos con mayor precisión

Anexos

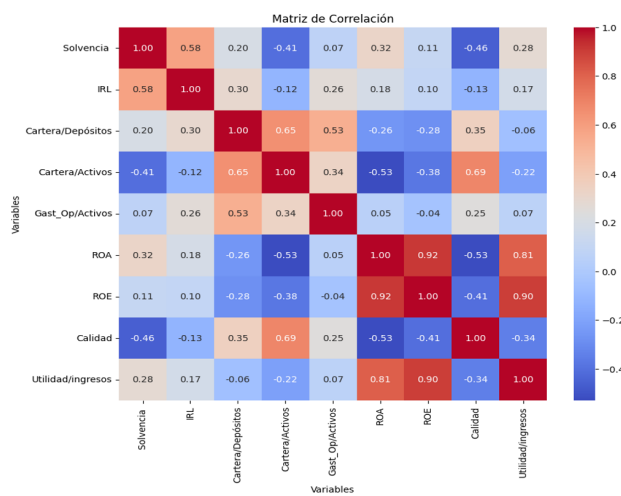


Figura 1: Matriz de correlación lineal entre variables en mapa de calor

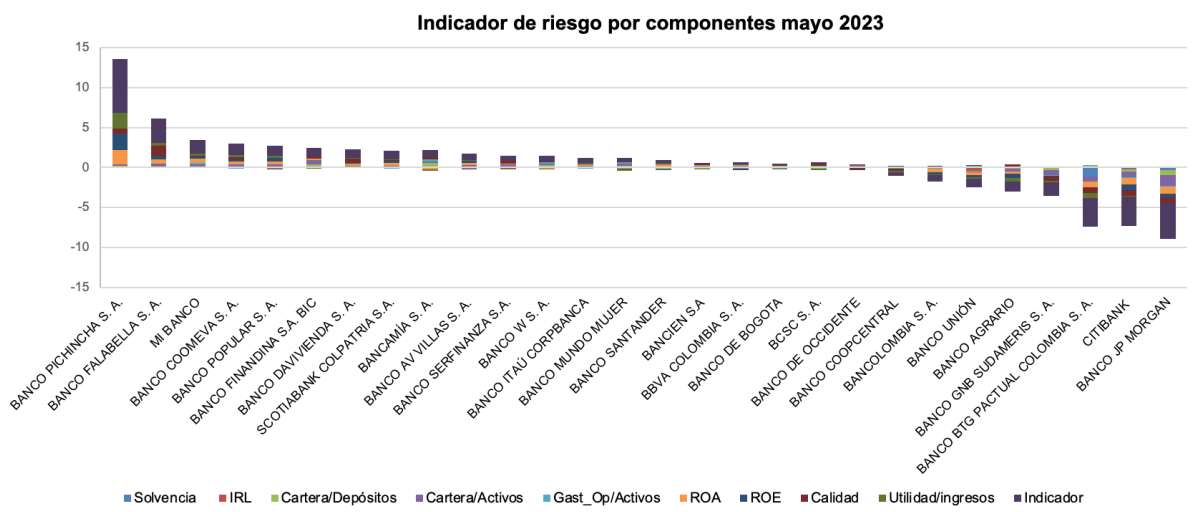
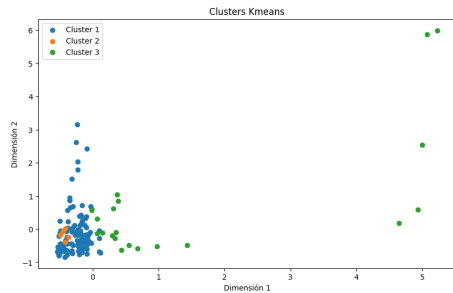
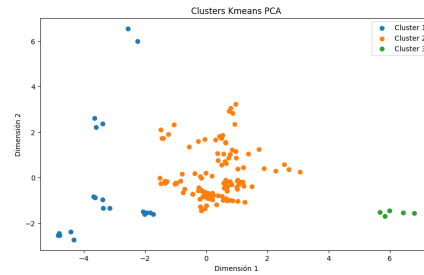


Figura 2: Indicador de riesgo desagregado para cada entidad financiera

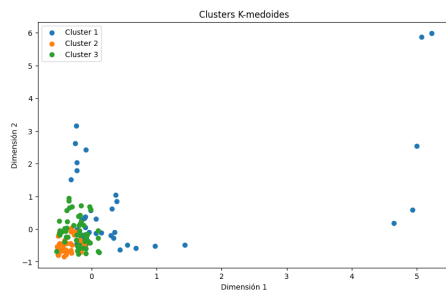
Clusters k-medias con las 9 variables



Clusters k-medias con los tres primeros componentes



Clusters k-medoides con las 9 variables



Clusters k-medoides con los tres primeros componentes

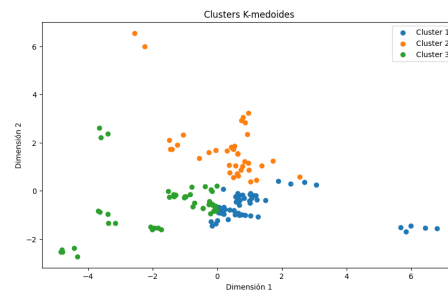
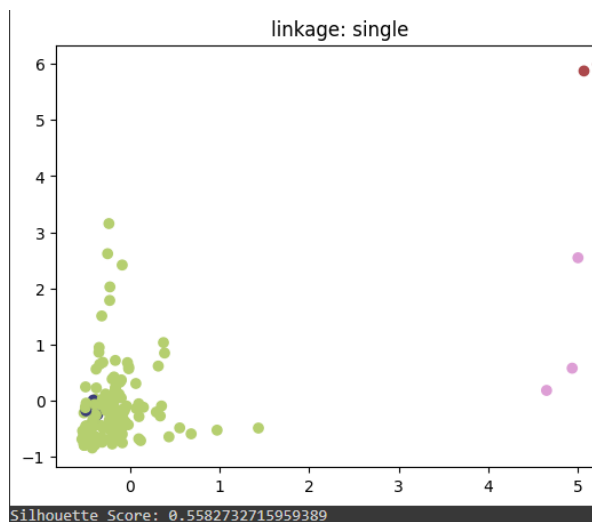


Figura 3: Resultados de la creación de clusters por las metodologías de k-medias

Clusters Jerarquico con las 9 variables, enlace “single”



Clusters Jerarquico con los tres primeros componentes enlace “Average”

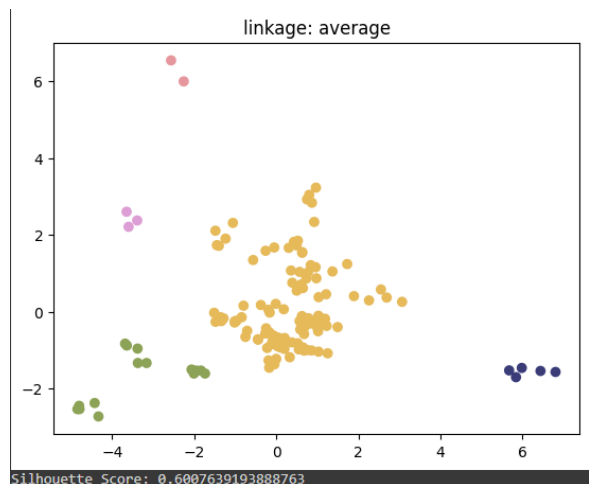


Figura 3: Resultados de la creación de clusters por las metodologías de cluster jerárquico

Bibliografía

Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*, 36(2), 3355-3366.

Estrada, D. A., & Gutiérrez, J. (2009). Supervisión y regulación del sistema financiero: modelos, implicaciones y alcances. *Perfil de coyuntura económica*, (13), 53-75.

Ferguson, A. (Marzo 28 de 2023). Por qué colapsó Silicon Valley Bank y qué puede significar. CNN.
<https://cnnespanol.cnn.com/2023/03/13/explainer-silicon-valley-bank-que-puede-significar-trax/>

Guerra, P., & Castelli, M. (2021). Machine learning applied to banking supervision a literature review. *Risks*, 9(7), 136.

Hotori, E., Wendschlag, M., & Giddey, T. (2022). *Formalization of banking supervision: 19th–20th centuries*. Springer Nature.

Kandrac, J., & Schlusche, B. (2017). The effect of bank supervision on risk taking: Evidence from a natural experiment.

Lee, K., Booth, D., & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications*, 29(1), 1-16.

Ossa Giraldo, W., & Jaramillo Marin, V. (2021). *Machine Learning para la estimación del riesgo de crédito en una cartera de consumo* (Doctoral dissertation, Universidad EAFIT).

Portafolio.co. (Marzo 15 de 2023) P. La caída de acción de Credit Suisse y su relación con quiebra de SVB. Consultado en
<https://www.portafolio.co/economia/finanzas/credit-suisse-registra-caida-historica-tras-caida-de-silicon-valley-bank-579943> el 4 de septiembre de 2023