

# Predicción de Activos financieros en función de los pasivos más relevantes de la institución financiera Banregio Grupo Financiero I.B.M, mediante los modelos de Regresión Lineal y ARIMA

Cristóbal Salinas  
Facultad de Ciencias Físico-Matemáticas,  
Universidad Autónoma de Nuevo León,  
San Nicolás de los Garza, Nuevo León

26 de marzo de 2023

## Resumen

En la presente investigación se busca predecir de la institución financiera Banregio Grupo Financiero I.B.M., los activos que se generan cada mes en función de los depósitos de los clientes y préstamos interbancarios, complementando también con el número de sucursales, número de tarjetas de crédito y número de cajeros.

Los modelos seleccionados para el análisis fue de regresión para ajustar una curva a los datos minimizando el error y hacer predicción para el mes de enero 2022, en el caso del modelo ARIMA, las predicciones que se hacen son explicadas por los datos del pasado y no por variables independientes..

*Palabras clave:* Activo, vista, plazo, préstamos interbancarios, regresión lineal, ARIMA.

## 1. Introduction

En el tema de Finanzas, se tiene la fórmula donde los activos son iguales a los pasivos más capital contable, nos enfocaremos los pasivos en particular los depósitos a la vista y plazo de los clientes y prestamos interbancarios por ser los rubros de mayor relevancia para Banregio y en particular, para las bancas comerciales como lo son los bancos tradicionales que conocemos como Afirme, Banorte, BBVA, entre otros, ya que con estos depósitos se pueden hacer préstamos a clientes o inversiones a plazo que generen intereses a favor del banco y puedan generar más activos.

Los activos también varían en función de la cantidad de sucursales, tarjetas de crédito y cajeros automáticos, ya que con estos se puede tener una mayor atracción de clientes que depositen sus ahorros en el banco de su preferencia.

Por definición, los activos siempre deben ser mayores que los pasivos para evitar problemas de liquidez del banco y evitar afectar tanto a los clientes como al sistema financiero y económico del país, como menciona Muñoz [2021], la banca pública y privada, como medida precautoria, tendrá que ocupar los montos de sus activos líquidos (fondos disponibles e inversiones) para hacer frente a sus pasivos.

Según Sánchez Mayorga [2012] Las investigaciones en la predicción de valores futuros de activos financieros, han venido desarrollándose en base a comportamientos anteriores de los mismos casos de estudio, usándose herramientas estadísticas, obteniéndose resultados muy útiles en la mayoría de los casos.

### **1.1. Objetivo**

Predecir los activos de Banregio con la mayor precisión posible de acuerdo a la información observada de la parte pasiva relevante que son los depósitos a la vista, depósitos a plazo, préstamos interbancarios en complemento con las tarjetas de crédito, sucursales y cajeros.

## **2. Literatura relacionada**

Dentro de la literatura, se encuentran pocos trabajos relacionados referentes a la predicción de activos financieros, en particular, en la predicción de la cartera de crédito.

Un trabajo relacionado a este tema es el de Romero Correa [2015], en el cual propone un modelo econométrico (ARIMA) que permite proyectar o predecir el comportamiento de variables, como lo es el activo más significativo de dicho tipo de entidades, que es la cartera de créditos referente a las cooperativas de ahorro y crédito.

También Duque Chávez [2012] menciona que utilizar variaciones y regresiones de datos estadísticos es con el propósito de realizar proyecciones hacia el futuro de activos, pasivos, cartera, patrimonio y captaciones, en este trabajo, selecciona los balances consolidados de las Cooperativas de Ahorro y Crédito reguladas por la Superintendencia de Bancos y Seguros publicados en su página web.

### 3. Marco teórico

### 4. Metodología

Para este estudio, se considera el periodo de diciembre 2016 a diciembre 2021 y la información financiera del total de la parte activa y de los depósitos en vista y plazo, préstamos interbancarios, así como la cantidad de sucursales, cajeros y tarjetas de crédito de la institución de banca múltiple Banregio. La información tomada es de los estados financieros y se publica mensualmente en la página de la Comisión Nacional Bancaria y de Valores.

Se utilizó el modelo DBSCAN a los datos ya que este el algoritmo trabaja con datos con formas extrañas, además, por ser no supervisado, da un poco más de flexibilidad de elegir los parámetros iniciales, los cuales podrían definirse por ejemplo,  $\text{min sample} = 3$  por ser datos financieros, Banxico y la CNBV piden información trimestral y eps dependerá de la estrategia de negocio. En este caso, Banregio se concentra en personas físicas, personas físicas con actividad empresarial y MyPYMES, por lo que se podría decir que un eps de 0.275 MDP es una buena aproximación de acuerdo a los estados financieros.

DBSCAN es un algoritmo no supervisado muy conocido en materia de Clustering. Fue presentado en 1996 por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiawei Xu.

A partir de unos puntos y un número entero  $k$  definido, el algoritmo pretende dividir los puntos en  $k$  grupos, llamados clústeres, homogéneos y compactos

Para cada observación miramos el número de puntos a una distancia máxima  $\epsilon$  de ella. Esta zona se denomina  $\epsilon$ -vecindad de la observación.

Si una observación tiene al menos un cierto número de vecinos, incluida ella misma, se considera una observación central. En este caso, se ha detectado una observación de alta densidad.

Todas las observaciones en la vecindad de una observación central pertenecen al mismo clúster. Puede haber observaciones centrales cercanas entre sí. Por lo tanto, de un paso a otro, se obtiene una larga secuencia de observaciones centrales que constituyen un único clúster.

En cada observación, para contar el número de vecinos a como máximo una distancia  $\epsilon$ , calculamos la distancia euclidiana entre el vecino y la observación y comprobamos si es inferior a  $\epsilon$ .

$$d(P, Q) = \sqrt{\sum_i^n (p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + \dots + (p_n - q_n)^2}$$

Cualquier observación que no sea una observación central y que no tenga

ninguna observación central en su vecindad se considera una anomalía.

Para confirmar el mínimo de observaciones para formar un cluster es de  $k=3$ , se aplicó el método del codo, calculando la distancia entre cada punto y la recta del punto inicial al punto final, tomando como criterio la primera distancia mayor.

Se aplicó también el método de regresión lineal múltiple para predecir los valores de la variable independiente (Activo) del modelo.

En una regresión se busca ajustar una curva a los datos minimizando el error. La regresión más sencilla es la regresión lineal donde se pretende predecir valores  $Y$  a partir de determinados  $n$  variables mediante la ecuación lineal  $Y = b_0 + X_1b_1 + \dots + X_nb_n$ , donde  $b_0$  coincide con una constante o intercepción, mientras que  $b_i, i \in 1, 2, \dots, n$  son la pendiente para cada  $X$ .

En series temporales, un modelo autorregresivo integrado de promedio móvil o ARIMA por sus siglas en inglés, se deriva de sus tres componentes AR (Autoregresivo), I(Integrado) y MA (Medias Móviles). Es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Se trata de un modelo dinámico de series temporales, es decir, las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes y fue desarrollado a finales de los sesenta del siglo XX. Box y Jenkins (1976).

El modelo ARIMA consiste en la combinación de un término autorregresivo (AR) y un término de promedio móvil (MA) con un elemento diferenciador (I). En general estos modelos se referencian con la palabra ARIMA (p,d,q). Donde (p) se refiere al orden del modelo autorregresivo; (d), al término de diferenciación, y (q), al término de media móvil con q términos de error. La estructura general de estos modelos  $Y_t : t$  tiene la forma de un modelo ARMA como se muestra en la siguiente ecuación:

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \dots - \theta_q \epsilon_{t-q} \quad (1)$$

Donde  $\phi$  corresponde al coeficiente autorregresivo a determinar,  $\theta$  coeficiente de media móvil a determinar,  $\epsilon$  término de error y  $Y_{t-p}$  es el registro normalizado de la serie a modelar. Para el término del diferencial se debe considerar una evaluación del orden. Los diferenciales pueden ser de primer o segundo orden, siguiendo la forma mostrada en las ecuaciones (2) y (3) respectivamente.

$$W_t = Y_t - Y_{t-1} \quad (2)$$

$$W_t^{(2)} = W_t^{(1)} - W_{t-1}^{(1)} \\ W_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \quad (3)$$

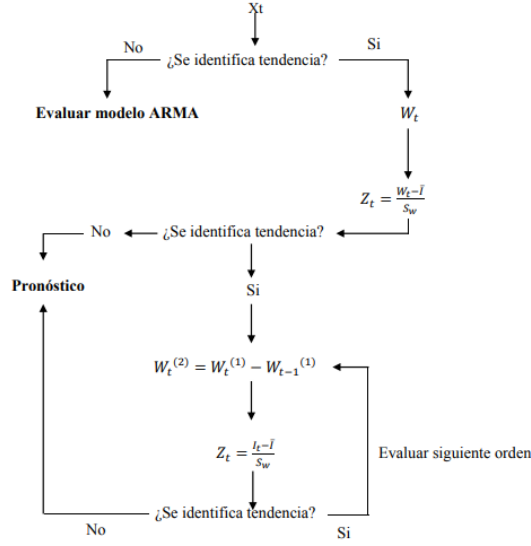


Figura 1: Diagrama de flujo Arima

$$W_t = Y_t - 2Y_{t-1} + Y_{t-2}$$

$$Z_t = \frac{W_t - \bar{W}}{S_w} \quad (4)$$

Donde  $W_t$  es el término diferenciador,  $Y_t$  se refiere al registro de la serie normalizado en el tiempo  $t$ , el término  $Z_t$  se refiere al dato del registro estandarizado en el tiempo  $t$  y se obtiene de la ecuación (4), promedio de los registros diferenciados,  $S_w$  desviación estándar del registro diferenciado. La metodología de determinación de un ARIMA se muestra en la figura 1.

Cabe resaltar que en la presente investigación se busca un enfoque de aproximación a una realidad la cual es reconocida como compleja y por tanto no se espera un ajuste exacto.

## 5. Resultados

La siguiente figura muestra el uso del método de codo para selección de número de grupos.

Se observa que la primer mayor distancia es de 165.69, por lo tanto, se confirma que el número de grupos adecuados es 3.

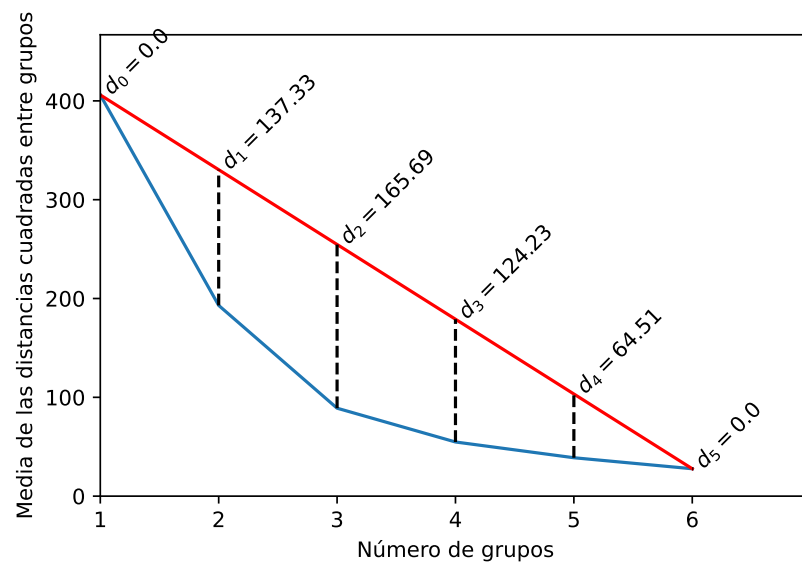


Figura 2: Selección de número de grupos

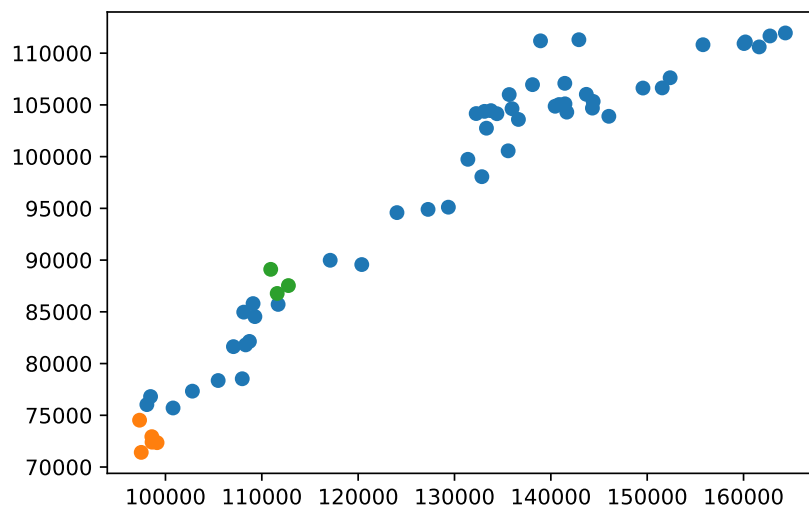


Figura 3: Agrupación de los k grupos

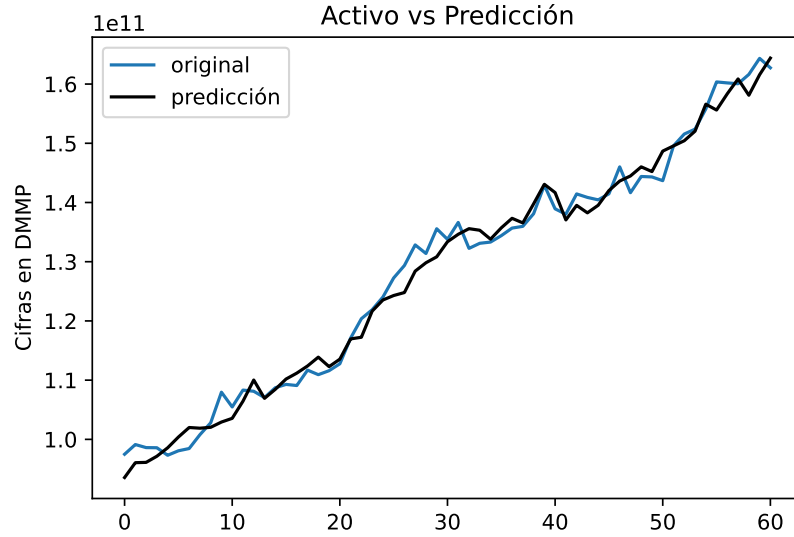


Figura 4: Comparación del activo original y sus predicciones

En la figura 3 se puede ver el agrupamiento de los 3 cluster, considerando el eje y como activos en función del eje x como la suma del resto de las variables.

Al aplicar un segundo método, Regresión Lineal Múltiple, se calculo el Error Porcentual Absoluto Medio (MAPE por sus siglas en ingles) obteniendo de resultado 0.0167 y al graficar los activos originales con sus respectivas predicciones se obtuvo la gráfica de la figura 4.

## 6. Discusión

Los resultados obtenidos aplicando el método DBSCAN coincide con lo esperado, es decir, el grupo adecuado que sugiere este método es de 3, como se sospechaba de un inicio por la medición trimestral que señala la regulación bancaria al revelar información en esa periodicidad.

Al revisar los clusters, se muestran los 3 grupos, se esperaba que estos tuvieran aproximadamente la misma cantidad de puntos, ya que, se tiene cierta estacionalidad cada trimestre por el tipo de clientes que tienen las mismas características entre sí para poder tener una mayor certeza en su comportamiento de los depósitos, pero no fue así.

De acuerdo con el método de Regresión Lineal Múltiple, la gráfica de las predicciones va en contra de la teoría, ya que esta no es una línea recta.

## 7. Conclusiones

DBSCAN no es un buen método para agrupar los datos de este estudio ya que se tienen factores que no necesariamente se tienen que tratar con este modelo, en el caso del método de codo, la cantidad de grupos arrojada es adecuada ya que es un dato que se tiene que cumplir por ser regulatorio.

Al hacer uso de la Regresión Lineal Múltiple, se tiene un buen ajuste del modelo de acuerdo con el cálculo del MAPE, pero no hay que basarnos solo en esta medición, también hay que complementar con la gráfica del modelo.

En la gráfica de los valores originales y las predicciones de los activos, se observa la misma tendencia a la alta y en varios puntos coincide pero por teoría, la línea negra debería ser una línea recta, por lo tanto, y aunque coincida en varios puntos, no es un modelo que ajuste de forma adecuada a estos datos.

## 8. Trabajo a futuro

Aplicar otros métodos para encontrar mejores predicciones y un adecuado tratamiento de los datos.

Una sugerencia de modelo podría ser el método XGBoost, como menciona Espinoza Zúñiga [2020] conviene utilizarlo porque puede manejar grandes bases de datos con múltiples variables, valores perdidos, sus resultados son muy precisos y es muy bueno en velocidad de ejecución y también trabaja con valores numéricos y Las ventajas de este algoritmo hace que se aplique en campos como: identificación de huellas digitales, seguridad vial y análisis de mercados financieros, entre otros.

## Referencias

- Dimitri Muñoz. Efectos de los riesgos de liquidez de la economía ecuatoriana en los activos y pasivos financieros de la banca pública y privada, periodo 2009-2018. Master's thesis, Pontificia Universidad Católica del Ecuador, Quito, 2021.
- Julio César Sánchez Mayorga, Ximena Millán Solarte. Medición del riesgo de liquidez. una aplicación en el sector cooperativo. *Entramado*, 2012. ISSN 1900-3803. URL <https://www.redalyc.org/articulo.oa?id=265424601007>.



Víctor Hugo Romero Correa. Implementación de modelo para pronosticar variables en cooperativas: caso de cobelén con cartera de créditos. Master's thesis, Universidad EAFIT, Escuela de Economía y Finanzas, Medellín, 2015.

Verónica Duque Chávez. Análisis, evolución y proyección del sector cooperativo controlado por la superintendencia de bancos y seguros en el ecuador en el período comprendido 2002 - 2009. Master's thesis, Universidad Andina Simón Bolívar, Sede Ecuador, 2012.

Javier Jesús Espinoza Zúñiga. Aplicación de algoritmos random forest y xgboost en una base de solicitudes de tarjetas de crédito. 2020.