**PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ**

**FACULTAD DE CIENCIAS SOCIALES**


# Probabilidad de default de portafolios de deuda corporativa en economías emergentes

Trabajo de investigación para obtener el grado académico de Bachiller en Ciencias Sociales con mención en Finanzas presentado por:

Estrella Torres, Maykol Alexander

Vega Nuñez, Johan Jose

Asesor:
Bendezú Medina, Luis Alfonso

Lima, 2023

# Informe de Similitud

Yo, *Bendezú Medina, Luis Alfonso,* docente de la Facultad de Ciencias Sociales de la Pontificia Universidad Católica del Perú, asesor(a) del Trabajo de Investigación de Bachillerato titulado *Probabilidad de default de portafolios de deuda corporativa en economías emergentes* del/de la autor (a)/ de los(as) autores(as) *Estrella Torres, Maykol Alexander y Vega Nuñez, Johan Jose* dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 8 %. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 14/02/2024.
- He revisado con detalle dicho reporte y el Trabajo de Investigación de Bachillerato, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 14 de febrero del 2024

<table>
<tr><td colspan="2">Apellidos y nombres del asesor / de la asesora:<br>Bendezú Medina, Luis Alfonso</td></tr>
<tr><td>DNI: 40675969</td><td rowspan="2">  <br>Firma</td></tr>
<tr><td>ORCID: 0000-0002-6125-38676</td></tr>
</table>

Resumen

En los últimos años, se ha producido una intensa producción de investigación académica con respecto a los modelos que estiman o predicen los eventos de incumplimiento de pago, debido al mayor interés de las empresas por mantener una mejor gestión de riesgo de crédito. Por ello, el presente trabajo tiene como principal objetivo comparar la capacidad predictiva de los modelos tradicionales o estadísticos (Regresión Logística) contra modelos Machine Learning (XGBoost y Random Forest). Para lo cual, se emplea una muestra de compañías latinoamericanas que emitieron bonos corporativos durante el período de 1990 a 2022, analizando así un total de 389 empresas. Asimismo, se usó Bloomberg, como fuente de información para extraer los ratios financieros con frecuencia trimestral en el periodo ya mencionado, obteniendo así 51,060 observaciones. Una de las características de este trabajo es que se usaron las variables del modelo de puntaje Z de Altman junto con otros ratios financieros complementarios, para así mejorar la precisión de los modelos. Sin embargo, para la determinación del mejor modelo predictivo, se usaron las métricas de clasificación como el AUC (Área bajo la curva ROC), Precisión, Recall y Score F1, y los resultados mostraron que los modelos de Machine Learning tuvieron un mejor rendimiento de clasificación con respecto a la Regresión logística, siendo el modelo Random Forest el que mejor performance según las métricas de evaluación.

Palabras clave: Probabilidad de default, modelo Logit, Random Forest, precisión.

**Índice**

## Índice de Tablas

## Índice de Figuras

## 1. Introducción

El presente trabajo de investigación tiene como principal objetivo identificar el modelo más eficiente al predecir la probabilidad de "default" de un portafolio de deuda corporativa, cuyos emisores sean medianas y grandes empresas pertenecientes a economías emergentes. Asimismo, el trabajo busca realizar un repaso de los diferentes modelos de riesgo de crédito, tanto tradicionales como contemporáneos, a través de una revisión de literatura teórica y empírica.

Esto como consecuencia de que, en los últimos cuarenta años, los desequilibrios macroeconómicos y la exposición a choques externos ocasionaron el debilitamiento de las condiciones financieras de muchas compañías localizadas en países emergentes (Naciones Unidas, 2014). Las cuales, se declaraban en bancarrota debido a que muchas de ellas afirmaron no contar con el nivel de liquidez suficiente para cubrir con el pago de sus deudas, por lo que estudiar los modelos de riesgo de crédito se volvieron cada vez más importantes.

Es así como, en la década de 1980, el constante incremento en los precios del crudo y las mayores presiones inflacionarias, obligaron a empresas de Argentina, Brasil, México, Uruguay y Venezuela a tener que reprogramar el pago de su deuda, debido a los menores niveles de productividad y los mayores costos de financiamiento (Bauer y Yamey, 2019). Asimismo, en 1994, la devaluación del peso mexicano y la reducción de las reservas internacionales, provocaron que aquellas empresas mexicanas que tenían deuda en moneda extranjera tuvieran que declararse en quiebra debido a que era imposible hacer frente a sus obligaciones financieras (Watkins, Dijk & Spronk, 2017).

Por su parte, la crisis asiática de 1997 provocó el deterioro de los mercados bursátiles latinoamericanos, luego de que las pérdidas se extendieron sobre los mercados de Japón, Hong Kong, Europa y Estados Unidos (Naciones Unidas, 1998). Asimismo, en los años 2000 ocurrió el "dotcom crash", que se caracterizó por ser una burbuja asociada a las empresas de internet, la que cuando estalló provocó la quiebra de una gran cantidad de empresas tecnológicas (Ofek y Richardson, 2003).

Consecuentemente, para responder la pregunta sobre qué modelo de riesgo de crédito predice mejor la probabilidad de incumplimiento en la cadena de pagos, será necesario realizar una comparación entre los modelos de riesgo tradicional (usualmente un modelo econométrico) y los modelos contemporáneos, en los que usaremos modelos factoriales para medir el riesgo de crédito de un portafolio de deuda corporativa, así como una regresión logística para calcular la probabilidad de default.

En ese sentido, la estructura del informe consistirá en que, como primer y segundo punto, se realizará una revisión de la literatura tanto teórica como empírica sobre los modelos a utilizar más adelante en la medición de riesgo de crédito del portafolio de deuda corporativa. Asimismo, se tratará de identificar todos los aportes que se hayan realizado -hasta el momento- sobre la medición de riesgo de crédito para economías emergentes, para luego, en el tercer punto, presentar algunos de los hechos estilizados. Finalmente, en el cuarto punto, se realizará el planteamiento de la hipótesis que responderá a la pregunta de investigación, que será resuelta con los resultados y conclusiones halladas en el presente informe.

## 2. Revisión básica de literatura teórica

A fin mantener un determinado orden, se empezará detallando los modelos estructurales de riesgo de crédito, los cuales tiene una relación entre los activos y los pasivos de una compañía, y que derivan de una base inicial, el cual es el modelo de Merton. En ese sentido, según (Jorion & GARP (Global Association of Risk Professionals), 2010) , se propone un caso en el que una firma tiene como un activo un valor de "V" y como pasivos un bono que se denotará como "B". Si el valor de "V" es mayor al de "B", entonces se deduce que la compañía pagará el nominal del bono, por lo que la compañía cumplirá con sus obligaciones, mientras que, si el activo "V" es menor a "B", no se podrá cumplir con las obligaciones que se tienen con los prestamistas. Asimismo, se asume que este valor V sigue un proceso estocástico, lo cual permite que se pueda desarrollar un modelo Black Scholes con la finalidad de encontrar el valor patrimonial siguiendo la siguiente fórmula:

$$E = V * N(d1) - e - rT * B * N(d2) \dots\dots\dots (i)$$

$$d1 = log(V/B) + \frac{(r + \sigma^2/2) * T}{\sigma\sqrt{T}}, d2 = d1 - \sigma * \sqrt{T} \dots\dots\dots\dots\dots\dots\dots\dots\dots (ii)$$

A partir de este modelo, se deriva el famoso modelo KMV, el cual agrega otras estimaciones como DD ("Distance to Default") o EDF ("Expected Default Frequency") con la finalidad de medir el riesgo de default de un activo listado en bolsa o de información pública. Por su parte, para McNeil et al. (2015), la metodología de EDF supone que la distancia de la probabilidad del incumplimiento debería clasificar a las compañías, siendo un mayor DD, una mayor probabilidad de incumplimiento.

$$DD = \frac{ln(V/B) + (r - \sigma 2/2)}{\sigma * \sqrt{T}} \dots\dots\dots (iii)$$

De esta forma, también se puede deducir que el EDF = N(-DD). Sin embargo, a pesar de que ambos modelos entregan una perspectiva de mercado con respecto al riesgo, se tienen ciertas críticas que son necesarias agregar. Por ejemplo, ambos modelos no se podrían utilizar en compañías no listadas o no públicas, puesto que los datos, tanto en el modelo KMV y el modelo de Merton, no distinguen los diferentes

tipos de deuda ni tampoco los diferentes tipos de bono que se puede emitir en el mercado. No obstante, para el presente trabajo se supera la presente crítica, ya que el enfoque de la investigación es el de medir compañías que hayan emitido bonos en una cierta ventana de tiempo.

Por otro lado, el modelo CreditMetrics puede evaluar un vector de activos y no solo uno a diferencia de los otros modelos ya mencionados. Así, Zhang (2018) menciona que CreditMetrics estima la volatilidad si es que existe algún cambio en la calificación crediticia, principalmente con la finalidad de evaluar el riesgo de crédito de un portafolio de deuda corporativa y para realizar este método, primero se tiene que presentar una regla de clasificación. Sin embargo, el paso más importante es establecer una probabilidad de transición con respecto a la nota crediticia que recibe una compañía de una clasificadora durante los próximos t años. Luego de realizar la matriz de transición, se pasa a definir el horizonte de riesgo al que se quiere exponer para luego aplicar un factor de descuento de una curva forward. Terminado esto, se pasa a desarrollar una simulación de Montecarlo, ya que es necesario encontrar la distribución conjunta de los activos del portafolio. En este punto, es importante mencionar que las transacciones se deben a las correlaciones que se tiene entre los activos con las obligaciones.

Por último, también se hará mención del modelo de distribución de pérdidas de un portafolio de deuda propuesto por Oldrich Vasicek (2002). En ese sentido, vamos a suponer que se tiene una cartera de préstamos, el cual está sujeto a un incumplimiento, por lo que va a generar pérdidas para el prestamista, y un portafolio financiado por deuda y capital. En consecuencia, esto haría que la calidad crediticia de este prestatario también dependiera de la probabilidad de que las pérdidas del portafolio excedan a su capital. Es así como empezaremos asumiendo que los activos de una firma “i” al tiempo “T” se representan de la siguiente manera:

$$log\ A_i\ (T)\ =\ log\ A_i\ +\ \mu_i\ *\ T\ -\ 0.5\ *\ {\sigma_i}^2\ *\ \sqrt{T}\ *\ X_i\ \ldots\ldots\ldots\ (iv)$$

Donde $X_i$ es una distribución normal, por lo que la probabilidad de default para el préstamo $i_{th}$ se comporta de la siguiente manera:

$$p_i = P[A_i(T) < B_i] = P[X_i < c_i] = N(c_i), \ldots\ldots\ldots (v)$$

donde:

$$ci = \frac{(\log B_i - \log A_i - \mu_i * T + 0.5 * \sigma_i 2 * T)}{\sigma_i * \sqrt{T}} \ldots\ldots\ldots (vi)$$

Ahora será necesario suponer que el portafolio consta de n-préstamos en igual cantidad. Por lo tanto, la "p" seguirá siendo la probabilidad de incumplimiento y también hay que suponer que las firmas están correlacionadas p y que todas tienen una misma fecha de cumplimiento T. De la misma forma, se tiene un $z_i$, el cual sigue una distribución de Wiener.

$$z_i = bx + a\varepsilon_i, \ldots\ldots\ldots (vii)$$

$$b = \sqrt{\rho}, a = \sqrt{1 - \rho} \ldots\ldots\ldots (viii)$$

El término bx se puede interpretar como la exposición que pueda tener una firma a un factor como el entorno macroeconómico y el término $\varepsilon$ representaría los riesgos idiosincráticos de la propia firma, por lo que se puede denotar de la siguiente manera:

$$P_k = P\left[L = \frac{k}{n}\right] \ldots (ix)$$

$$\binom{n}{k} P[A_{it} < B_1, \ldots, A_{kT} < B_k, A_{k+1T} \geq B_{k+1}, \ldots, A_{nT} \geq B_{nT}] \ldots (x)$$

$$\binom{n}{k} \int_{-\alpha}^{\alpha} \left[N(-\frac{c + b\mu}{a})\right]^k \left[1 - N(-\frac{c + b\mu}{a})\right]^{n-k} dN(\mu) \ldots (xi)$$

Es, por esto, que la integración final es la distribución condicional de las pérdidas del portafolio condicionado al estado macroeconómico, el cual se mide mediante las desviaciones estándar.

## 3. Revisión de estudios empíricos previos

En los últimos años, se realizaron diferentes investigaciones entre académicos y profesionales, con respecto a modelos tanto estadísticos (regresiones Logit o Probit) como computacionales (modelos Random Forest, SVM, entre otros), que ayuden a predecir eventos de incumplimiento de pago o medir el riesgo de crédito en un portafolio de deuda corporativa.

En primer lugar, en Barboza et al. (2017) utilizaron las siguientes técnicas: Bagging, Boosting, Random Forest, SVM con kernel linear, SVM con kernel de función radial, Redes Neuronales Artificiales (ANN por sus cifras en inglés), regresiones logísticas (Logit) y Análisis Discriminante Múltiple (MDA). Los autores muestran dos evaluaciones, en la cual una se basa en variables escogidas por ellos mismos y otra que usa solo las variables del modelo Z-score de Altman. En la primera evaluación, se observa que el modelo con menor predicción es el modelo MDA con un área bajo la curva (AUC %) de 63.68%. A pesar de eso, el modelo Logit llega a obtener un AUC de 90.10%, con lo cual es superior que modelos como ANN (90.08%), SVM RBF (85.17%) y SVM lineal (67.2%). Sin embargo, los modelos restantes son superiores en capacidad de predicción que el modelo Logit, siendo el modelo Bagging el que mejor destacó con un AUC de 92.97% y con un error tipo II de 13.23 %. En la segunda evaluación, se destacó que los modelos tradicionales (MDA y Logit) fueron superados en capacidad de predicción por los modelos Machine Learning con excepción del modelo SVM Lineal. Por su parte, tanto el modelo Logit y MDA obtuvieron un AUC de 86.16% y 67.4%, mientras que los demás modelos obtuvieron un AUC arriba de los 89.54%, con excepción del modelo SVM Lineal, el cual obtuvo un 66.31% de AUC y un error tipo II de 36.76%. En este estudio, se utilizaron datos de empresas norteamericanas en una ventana de tiempo de 1985 y 2013 usando Compustat, y con respecto a las firmas insolventes, los autores tomaron la base de datos Salomón del NYU. Luego de un muestreo, se usaron 133 empresas insolventes y 13300 empresas solventes.

De la misma forma, pero comparando solo modelos Machine Learning, Shetty (2022) usa el modelo XGBoost, SVM y Algoritmos de Modelos de Aprendizaje Profundo (Deep Learning) con una data de 3728 entre pequeñas y medianas

empresas belgas, las cuales 1864 se declararon en quiebra desde 2002 y 2012. El estudio tuvo como resultado que los modelos tuvieron una f1-score de 82% y 83% llegando a solo utilizar el ratio de liquidez, el rendimiento sobre los activos (ROA) y un ratio de solvencia. Una de las diferencias entre ambos casos de estudio es que en el caso del primero se llega a utilizar las mismas variables que usa el modelo Altman Z-Score junto con otros indicadores de crecimiento como el crecimiento de los activos, crecimiento de las ventas, crecimiento del número de empleados, margen operativo, cambio en el rendimiento del capital y cambio del precio - valor contable. Sin embargo, en ambos sí se puede observar un excelente desempeño de los modelos Machine Learning.

Por otra parte, Hsiao y Gao (2016), hicieron una comparación entre el modelo Altman Z-Score, el modelo KMV y el modelo Naïve con la finalidad de comparar la capacidad predictiva que tienen estos modelos. En términos sencillos, el modelo Naïve parte también de un modelo Merton, el cual fue explicado en párrafos anteriores. Sin embargo, en este modelo se asume que la volatilidad de la deuda estaría relacionada con la volatilidad del patrimonio de la siguiente manera, siendo D, el nominal de un bono, y E, el patrimonio.

$$Naive\ \sigma_D = 0.05 + 0.25 * \sigma_e \ \ldots\ldots\ldots (xii)$$

Entonces para hallar la volatilidad de la compañía se deriva de un ponderado entre el peso del patrimonio y de la deuda.

$$Naive\ \ \sigma_v = \frac{E}{E+F} * \ \sigma_e \ + \frac{F}{E+F} * \ Naive\ \sigma_D \ \ldots\ldots\ldots (xiii)$$

Para finalmente terminar en:

$$Naive\ DD \ = \frac{ln\ [((E+F))] + (\mu A \ - \ 0.5 \ * \ Naive\ \sigma V) \ * \ T}{Naive\ \ \sigma_v \sqrt{T}} \ldots\ldots\ldots (xiv)$$

$$PDNaive \ = \ N(-Naive\ DD) \ \ldots\ldots\ldots (xv)$$

Esta investigación utilizó la data de firmas estadounidenses en el periodo de la crisis financiera del 2007 al 2012 tanto empresas financieras como no financieras públicas, de las cuales en total se utilizaron 17647 firmas, de las cuales 392 firmas se declararon en bancarrota durante el periodo. La conclusión de los autores fue que el modelo Z-Score fue el que tuvo menor desempeño en general con un AUC de 80.6%, mientras que el modelo KMV sostenía un mejor performance en la capacidad de predicción de bancarrota para empresas no financieras con un área bajo la curva de 90.4%, mientras que el modelo Naïve era el mejor predictor para empresas financieras con un 86.2%. En general, la investigación también concluyó que el mejor modelo tanto, para firmas financieras y no financieras, era el modelo Naïve con un AUC de 87.5%.

En Yeh et al. (2012) se realiza la comparación entre modelos Random Forest, SVM, Rough Set Theory (RST) y entre otros modelos híbridos como Random Forest y árboles de decisión. Además, el estudio hace otra comparación entre el desempeño de modelos con solo variables contables y/o financieras contra de otro que utiliza las mismas variables, pero que agrega el factor DD (Distance to Default) del modelo KMV. Los resultados arrojan que agregando el factor DD del modelo KMV presentan mejoras significativas en el accuracy. Por ejemplo, el modelo híbrido Random Forest y RST manifiestan un accuracy de 73.2% sin el factor DD, mientras que agregando esta variable la métrica ya mencionada mejora a 93.4%. Además, es importante considerar que la investigación trabaja con datos de empresas taiwanesas que cotizan en su respectivo mercado, lo cual hace un total de 2470 empresas tecnológicas en un periodo del 2003 al 2008. Por lo tanto, en ambas investigaciones, se puede encontrar que los modelos basados en el movimiento de mercado o agregando variables relacionadas a esta, tienen una mejor predictibilidad al momento de seleccionar eficientemente las compañías que entraron en bancarrota.

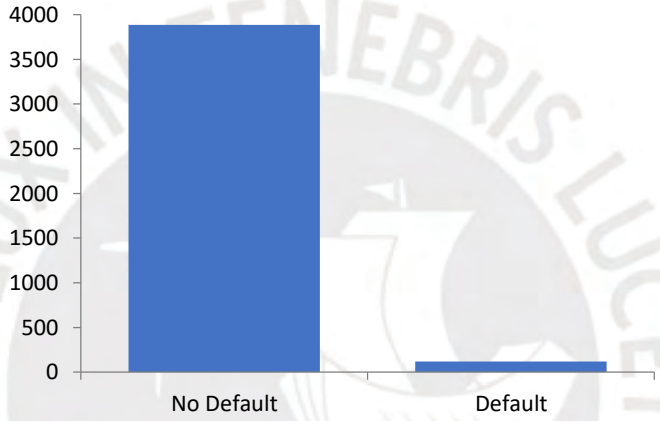
Hasta el momento se han desarrollado modelos que solo incluyen variables contables, de mercado y otros que enlazan estas dos variables con la finalidad de alcanzar una mejor eficiencia. Sin embargo, ahora se presentará un estudio que relaciona las variables propias de una firma con su entorno macroeconómico. En el trabajo realizado por Bonfim (2009), remarca que el riesgo de crédito y el contexto macroeconómico estarían enlazados, ya que, si bien los incumplimientos o las declaraciones de bancarrota se pueden explicar por temas idiosincráticos de una

compañía, el modelo de predictibilidad podría ser clasificar estas siniestralidades con mejor rendimiento si es que se utilizan variables macroeconómicas. Para esta investigación, se usó dos bancos de datos del Banco Central de Portugal, uno que provee uno que provee la lista de firmas y sus exposiciones al crédito, mientras que la otra provee una amplia información financiera y contable anual de una serie de compañías portuguesas. Al concatenar ambas tablas se tiene un total de 113,119 observaciones para un total de 33,084 de empresas para un periodo de 1996 a 2002. Además, es importante mencionar que para este caso se utiliza como variable dependiente al incumplimiento de un préstamo. La construcción de las variables financieras se usaron ratios de rentabilidad, apalancamiento de la firma, estructura de financiamiento, liquidez e inversión, mientras que las variables macroeconómicas son el crecimiento del PBI, rendimiento de los bonos soberanos de Portugal a 10 años, la variación del indicador de la bolsa de Portugal, crecimiento del crédito entre otras. A eso se debe agregar que también se usó variables “dummy” por años. La presente investigación muestra una serie de resultados, pero se hará mención de que se realizó un modelo Logit base con solo variables financieras contra esas mismas variables, pero agregando o las variables “dummy” por año o agregando las variables macroeconómicas. En resumen, se realizaron una comparación de un modelo base contra 8 modelos (1 agregando variables “dummy’s” y otro los otros 7 agregan las distintas variables macroeconómicas). Se encontró que el modelo base obtuvo un Pseudo-R Cuadrado de 0.037, mientras que el modelo con “dummy’s” alcanzó un 0.046, mientras que los modelos que agregan variables macroeconómicas pasaron el 0.042 (Modelo 2) y el máximo fue de 0.045 (Modelo 5). Por lo tanto, se puede observar que los modelos a evaluar obtuvieron un margen de mejora con respecto a su modelo base.

## 4. Hechos estilizados

Con respecto a la base de datos, se utilizarán bonos latinoamericanos que fueron emitidos desde 1990 a 2020, de los cuales 135 compañías de 3914 han tenido un evento de incumplimiento con respecto a sus obligaciones. De esos datos, se observa que el mayor número de empresas que tuvo este tipo de eventos fue Brasil con 48 casos, seguido por México con 37 casos, mientras que Argentina ha tenido 27 casos, mientras que los demás países de la región han tenido menores registros de Default financiero.

Figura 1. Emisores Default y no Default




Fuente: (Bloomberg) Elaboración: Propia

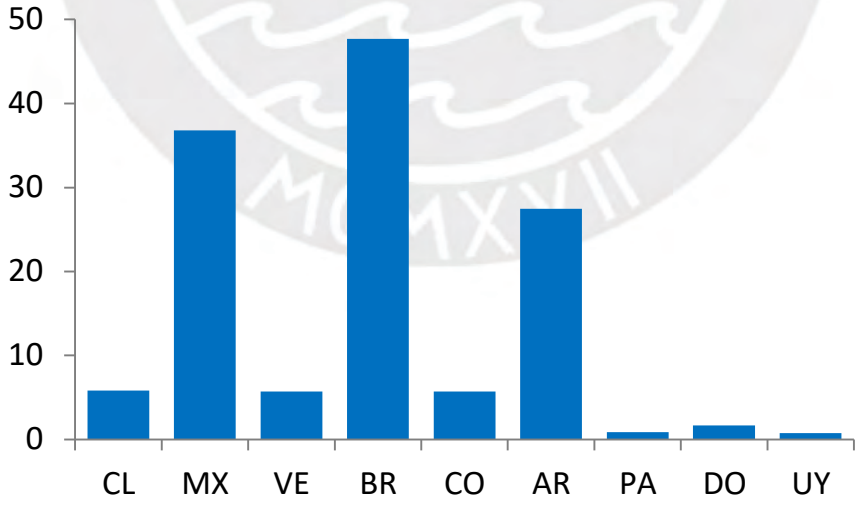
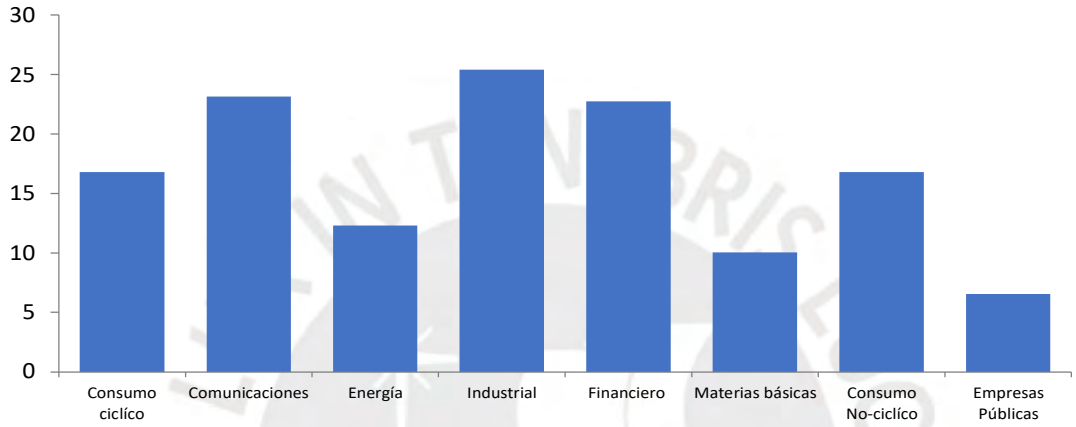



Figura 2. Emisores en Default por países

Fuente: (Bloomberg) Elaboración: Propia

Por otro lado, no se observa una determinada tendencia con respecto al sector que ha tenido mayor número de incumplimiento, pero sí se observa que el sector con menor frecuencia de esta clase de eventos es de utilidad pública, mientras que los sectores Industriales, Financieros y de Comunicaciones tienen una mayor densidad, así como se observa.

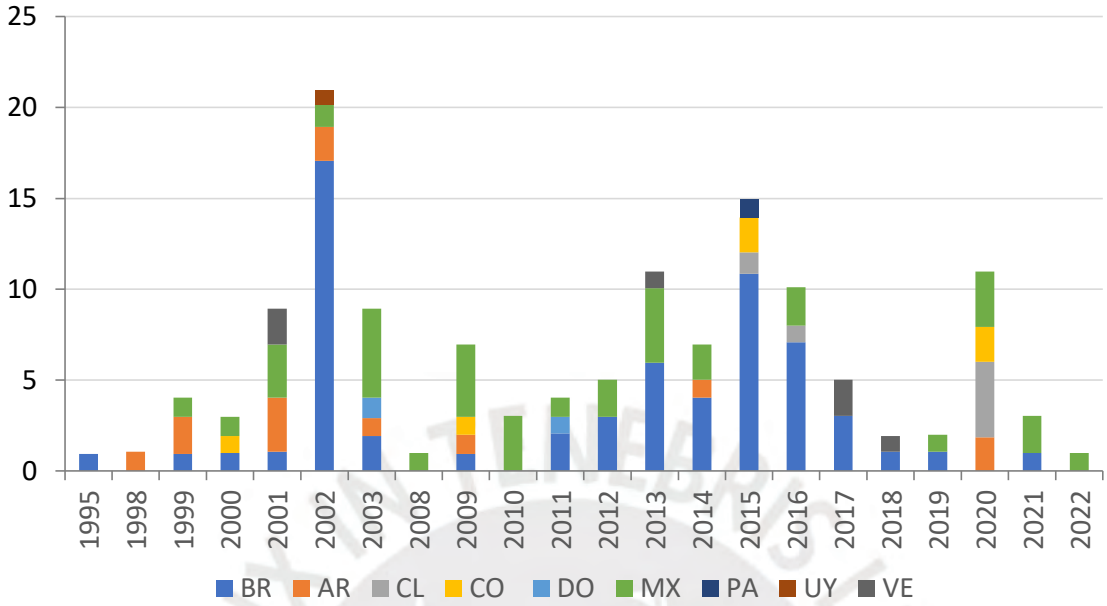
Figura 3. Emisores en Default por actividad económica




Fuente: (Bloomberg) Elaboración: Propia

Sin embargo, una tendencia muy marcada es cuando se observa los incumplimientos por años y agrupados por países, siendo dos momentos importantes. El primer pico marcado y el año en el que mayor número se dio este siniestro fue el 2002, y en gran parte, por empresas argentinas. Este hecho probablemente haya tenido como producto a la crisis del “corralito”. Asimismo, el otro pico de defaults se dio en el año 2015, y en su mayoría por empresas de Brasil. Esto probablemente por la recesión económica que tuvo este país entre el 2014 y el 2016. A partir de esto, se puede suponer que el entorno en el que se desenvuelven las compañías puede ser significativamente para explicar estos eventos de default, ya que incluso un periodo de estrés como el año 2020 no ha tenido la misma relevancia como los otros dos eventos, pero resaltando que fue el año en el que más diversificación hubo en defaults con respecto a países, siendo Chile el más concentro con 4 firmas las que entraron en default.

Figura 4. Emisores en Default por año y por países




Fuente: (Bloomberg) Elaboración: Propia

Finalmente, es necesario remarcar que, para esta entrega, todavía se sigue trabajando con la base de datos tanto en su limpieza como en la ampliación (geográfica) o filtración (tamaño de las firmas). Esto con la finalidad de responder adecuadamente a la tesis y a la problemática planteada. Sin embargo, por lo ya comentado anteriormente, el trabajo de investigación va tomando una dirección clara con respecto a la construcción de la base de datos, pero este avance ha sido productivo para ver ciertas tendencias ya expuestas.

## 5. Hipótesis tentativas o preguntas para responder

El presente trabajo de investigación tiene como finalidad hacer una evaluación entre modelos econométricos, estadísticos y una serie de modelos de Machine Learning con la finalidad de encontrar qué modelo tienen mejor rendimiento al momento de clasificar a las compañías que entraron en incumplimiento. Asimismo, luego de la revisión de literatura, la hipótesis que se maneja para el presente estudio es que los modelos Machine Learning sean mejores clasificadores que los modelos tradicionales para una base de datos de compañías que colocaron bonos en una ventana de tiempo del 1990 a 2022.

## 6. Variables y metodología

Esta sección se dividirá en cuatro partes. En la sección 6.1 se realizará una descripción de las variables explicativas que se utilizarán en los modelos de predicción. En la sección 6.2 se tratará las metodologías utilizadas para predecir la probabilidad de default tanto para los modelos tradicionales (Regresión Logística) y los modelos implementados a través de Machine Learning y en la que también se abordará la Técnica Smote aplicada a datos no balanceados, los cuales serán utilices al momento de realizar las estimaciones. Finalmente, en la sección 6.3 se definirán las métricas de evaluación para los modelos tradicionales como la "Regresión Logística" y los modelos de Machine Learning.

### 6.1. Variables y base de datos

La base de datos que se utilizará comprende a 389 empresas latinoamericanas que no han incumplido sus pagos de sus obligaciones o que no se han declarado en bancarrota (126 son financieras y 263 son no financieras), y 28 empresas que han incumplido con sus pagos o que se han declarado en bancarrota durante el período de estudio. El cual abarca toda la información disponible en Bloomberg, desde el tercer trimestre del 1990 hasta el segundo trimestre del 2022, obteniendo así un total de 51 060 observaciones.

#### 6.1.1. Variables

Por su parte, las variables seleccionadas para el estudio están relacionadas con las variables elegidas por Altman (2000), las cuales son las primeras cinco variables presentadas en la tabla 1 sumado a otras variables tanto de rentabilidad, apalancamiento y de liquidez.

**a) RETAINED EARNINGS / TOTAL ASSET**

Las ganancias retenidas es un buen indicador para poder predecir el default de una compañía, debido a que esta cuenta registra las ganancias que se reinvierten en una compañía. Asimismo, la edad de una compañía estaría implícita dentro de este ratio, debido a que usualmente las compañías con menor tiempo en el mercado acumulan menores ganancias retenidas a comparación con mayor tiempo en el

mercado. Por último, en cierta manera, el presente ratio se puede interpretar como un ratio de apalancamiento, ya que firmas con altas ganancias retenidas llegan a utilizar una menor deuda.

**b) EBIT / TOTAL ASSET**

Las ganancias antes de intereses e impuestos es una buena aproximación para observar la rentabilidad operativa que tienen una compañía sin el apalancamiento. Como se sabe, en los casos de bancarrota o de incumplimiento el valor razonable de los pasivos supera al de los activos, por lo que una variable de rentabilidad es necesaria para poder clasificar a las compañías. Incluso Altman (2000) menciona que este ratio supera en poder de predictibilidad a otros como el flujo de caja.

**c) WORKING CAPITAL / TOTAL ASSET**

El Working Capital se compone entre la diferencia entre los activos líquidos menos los pasivos líquidos, con lo cual este ratio se define como una liquidez. Es, por ello, que si una firma presenta menores activos líquidos no podrá hacer frente a sus obligaciones en el corto plazo, por lo que en el tiempo se presentarán contracciones en los activos totales.

**d) TOTAL MARKET VALUE / TOTAL LIABILITIES**

El Market Value se puede definir como el número de acciones comunes multiplicado por el precio de mercado de una acción de una compañía en el caso que esta esté listada en alguna bolsa de valores. Este ratio añade un valor de mercado al patrimonio de una compañía, con lo cual, si los pasivos son mayores a la suma de los activos y al patrimonio, estaríamos ante una pronta declaración de bancarrota.

**e) SALES / TOTAL ASSET**

Este ratio nos indica la capacidad que tiene la compañía para generar sus flujos de ventas desde el lado más general desde el punto de vista del estado de resultados. Además, el presente ratio es un buen indicador de competitividad de la compañía con

respecto al entorno en el que se desarrolla. No obstante, también es importante recalcar que este ratio puede variar considerablemente con respecto a diferentes industrias.

**f) ROE**

El presente ratio se descompone entre la utilidad neta sobre el patrimonio, por lo que es un indicador de rentabilidad post apalancamiento e impuestos, con lo cual nos puede dar otra visión de su rentabilidad.

**g) ROA**

El ROA se descompone entre la utilidad neta sobre los activos totales, por lo que es otro indicador de rentabilidad post apalancamiento e impuestos, por lo cual sigue la misma lógica que el ROE con la diferencia que este busca una relación con los activos totales, por lo cual será otra variable por evaluar en el modelo.

**h) CURRENT RATIO**

El Current Ratio es otro indicador de liquidez, el cual compone una división del activo corriente sobre el pasivo corriente, el cual fue evaluado en estudios como el Shetty (2022), y en el que ha logrado un buen “performance”. Asimismo, para Altman (2000) la variable de liquidez es una de las más importantes, por lo cual será interesante su comparación.

**i) DEUDA TOTAL / EBITDA**

El presente ratio es otra opción a la pregunta de qué tan apalancada está una compañía, pero desde el punto de vista de la operatividad de una compañía, por lo que el múltiplo Deuda Total sobre EBITDA es una métrica que nos puede aportar algún grado de significancia tanto en los modelos de regresión logística, así como para los modelos de Machine Learning.

Tabla 1. Estadísticos Descriptivos de las No Default - Financieras

| Ratios | Promedio | Mediana | Desviación Estandar | Mín | Máx |
|---|---|---|---|---|---|
| RETAINED EARNINGS TO TOT ASSET | -0.03 | 0.03 | 1.41 | -48.91 | 0.74 |
| EBIT TO TOT ASSET | 0.01 | 0.01 | 0.04 | -0.98 | 0.75 |
| REVENUE TO TOT ASSET | 0.06 | 0.03 | 0.11 | -0.23 | 3.12 |
| TOT MKT VAL TO TOT LIAB | 182.44 | 1.02 | 10617.96 | -0.04 | 655460.52 |
| WORKING CAPITAL TO TOT ASSET | 0.13 | 0.08 | 0.19 | -0.67 | 1.00 |
| CUR RATIO | 2.63 | 1.53 | 5.56 | 0.00 | 121.56 |
| RETURN COM EQY | 0.15 | 0.13 | 0.26 | -1.79 | 7.01 |
| RETURN ON ASSET | 0.02 | 0.02 | 0.13 | -1.56 | 1.07 |
| TOT DEBT TO EBITDA | 9.87 | 5.43 | 62.23 | 0.00 | 2780.17 |

Fuente: Elaboración propia con datos de Bloomberg

Tabla 2. Estadísticos Descriptivos de las No Default – No Financieras

| Ratios | Promedio | Mediana | Desviación Estandar | Min | Max |
|---|---|---|---|---|---|
| RETAINED EARNINGS TO TOT ASSET | -9.23 | 0.05 | 710.69 | -74025.00 | 1.11 |
| EBIT TO TOT ASSET | -1.49 | 0.02 | 110.47 | -10871.00 | 40.26 |
| REVENUE TO TOT ASSET | 0.17 | 0.14 | 0.14 | 0.00 | 2.36 |
| TOT MKT VAL TO TOT LIAB | 2.29 | 1.60 | 9.46 | 0.03 | 860.61 |
| WORKING CAPITAL TO TOT ASSET | -0.49 | 0.05 | 61.55 | -7189.00 | 1.00 |
| CUR RATIO | 1.85 | 1.28 | 25.22 | 0.00 | 2073.63 |
| RETURN COM EQY | 0.10 | 0.12 | 0.71 | -31.98 | 22.18 |
| RETURN ON ASSET | -0.62 | 0.04 | 46.23 | -4007.16 | 8.26 |
| TOT DEBT TO EBITDA | 5.13 | 2.89 | 21.41 | 0.00 | 882.68 |

Fuente: Elaboración propia con datos de Bloomberg

Tabla 3. Estadísticos Descriptivos de las Default

| Ratios | Promedio | Mediana | Desviación Estandar | Min | Max |
|---|---|---|---|---|---|
| RETAINED EARNINGS TO TOT ASSET | -0.58 | 0.01 | 1.97 | -19.39 | 0.79 |
| EBIT TO TOT ASSET | -0.00 | 0.01 | 0.13 | -2.12 | 2.46 |
| REVENUE TO TOT ASSET | 0.13 | 0.11 | 0.10 | -0.09 | 0.55 |
| TOT MKT VAL TO TOT LIAB | 4.06 | 1.00 | 55.35 | -0.07 | 1497.17 |
| WORKING CAPITAL TO TOT ASSET | -0.08 | 0.02 | 0.57 | -7.18 | 0.86 |
| CUR RATIO | 1.36 | 1.07 | 1.19 | 0.02 | 9.34 |
| RETURN COM EQY | -0.11 | 0.06 | 2.03 | -5.27 | 45.89 |
| RETURN ON ASSET | 0.01 | 0.01 | 0.48 | -1.38 | 10.72 |
| TOT DEBT TO EBITDA | 19.79 | 4.43 | 65.26 | 0.00 | 592.21 |

Fuente: Elaboración propia con datos de Bloomberg

Fuente: Elaboración propia con datos de Bloomberg

## 6.2. Modelos de Estimación

### 6.2.1. Modelo de regresión logística

La técnica Logit o también conocida como Regresión Logística se basa -principalmente- en una combinación lineal de variables independientes que permite estimar la probabilidad de que una empresa pertenezca a alguno de los grupos en discusión (empresas que hicieron default o empresas que no hicieron default). Según Westgaard & Wijst (2001), está es una de las técnicas más utilizadas por las instituciones reguladoras, así como por las agencias de evaluación crediticia, ya que dentro de la gran gama de técnicas que podría utilizarse para estimar la probabilidad de default, ellos prefieren la regresión logística debido a que, según las características del modelo, este solo puede tomar dos valores, ceros o unos, así como la facilidad de interpretación y de estimación.

$$y = \{\, 0\ 1\ \ldots\ldots\ldots\ldots\ldots\ldots (xvi)$$

En ese sentido, codificaremos a la variable Y como 0 (no default) y 1 (default) y, por lo tanto, lo que este modelo estaría estimando es que, si la empresa entra en default o no, en base a las variables explicativas que, en este caso, serán ratios financieras (liquidez, rentabilidad, solvencia, entre otros).

Por otro lado, que la regresión logística tiene la siguiente forma:

$$Ln\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k \ldots\ldots\ldots (xvii)$$

Donde $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \ldots, \mathrm{X}_\mathrm{k})$ ,son los ratios financieros de las compañías en frecuencia trimestral y “Y” es la variable independiente, la cual predeciremos. En ese sentido, como se busca estimar la probabilidad de default, tenemos que despejar la variable de interés (y). Es así como la ecuación se formularía de la siguiente manera:

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k)}} \ldots\ldots (xviii)$$

Asimismo, estos serán los ratios financieros que se utilizarán como variables explicativas:

Tabla 4. Variables explicativas

| Variable | Formula | Abreviación |
|---|---|---|
| DEFAULT | Default | Y |
| RETAINED EARNINGS TO TOT ASSET | $\frac{RETAINEDEARNINGS}{TOTALASSETS}$ | X1 |
| EBIT TO TOT ASSET | $\frac{EBIT}{TOTALASSET}$ | X2 |
| RENEVUE TO TOT ASSET | $\frac{REVENUE}{TOTALASSET}$ | X3 |
| TOT MRK VAL TO TOT LIABILITIES | $\frac{TOTALMARKETVALUE}{LIABILITIES}$ | X4 |
| WORKING CAPITAL TO TOT ASSET | $\frac{WORKINGCAPITAL}{TOTALASSET}$ | X5 |
| CUR RATIO | $\frac{CURRENTASSET}{CURRENTLIABILITIES}$ | X6 |
| RETURN COM EQY | $\frac{NETINCOME}{TOTALEQUITY}$ | X7 |
| RETUNR ON ASSET | $\frac{NETINCOME}{TOTALASSET}$ | X8 |
| TOT DEBT TO EBITDA | $\frac{TOTALDEBT}{EBITDA}$ | X9 |

Fuente: Elaboración propia

Por su parte, cabe mencionar que según Memic (2015), en la regresión logística el valor de los coeficientes $\beta_{1,}\beta_{2,}\beta_{3,}\ldots.\beta_{k}$, determinan la dirección de la relación entre las variables explicativas $(X_{1,}X_{2},X_{3},\ldots,X_{k})$ y la variable dependiente Y. Asimismo, el autor menciona que usualmente estas variables son determinadas a través del método de máxima verosimilitud.

Por otro lado, se debe tener cuidado con la multicolinealidad, ya que según del Valle Moreno (2012), esto podría generar inestabilidad al incrementar la varianza de los coeficientes de regresión. En la siguiente tabla se muestra como que en tres de las variables explicativas existe un alta multicolinealidad (VIF> 10).

Tabla 5. Variance inflation factor 1

| | VIF | 1/VIF |
|---|---|---|
| ebit to tot asset 1 | 123.752 | .008 |
| retained earnings ~1 | 93.194 | .011 |
| working capital to~1 | 18.081 | .055 |
| return on asset 1 | 1.017 | .984 |
| revenue to tot ass~1 | 1.014 | .986 |
| return com eqy 1 | 1.013 | .987 |
| tot debt to ebitda 1 | 1.006 | .994 |
| cur ratio 1 | 1.002 | .998 |
| tot mkt val to tot~1 | 1 | 1 |
| Mean VIF | 26.786 | . |

Fuente: Elaboración propia

Al ejecutar la prueba de multicolinealidad, se evidencia que existen tres variables que podrían afectar a nuestra regresión, en consecuencia, se opta por eliminarla y obtenemos los siguientes resultados:

Tabla 6. Variance inflation factor 2

| Ratios | VIF | 1/VIF |
|---|---|---|
| revenue to tot ass~1 | 1.014 | .986 |
| return com eqy 1 | 1.013 | .987 |
| tot debt to ebitda 1 | 1.006 | .994 |
| cur ratio 1 | 1.002 | .998 |
| return on asset 1 | 1 | 1 |
| working capital to~1 | 1 | 1 |
| tot mkt val to tot~1 | 1 | 1 |
| Mean VIF | 1.005 | . |

Fuente: Elaboración propia

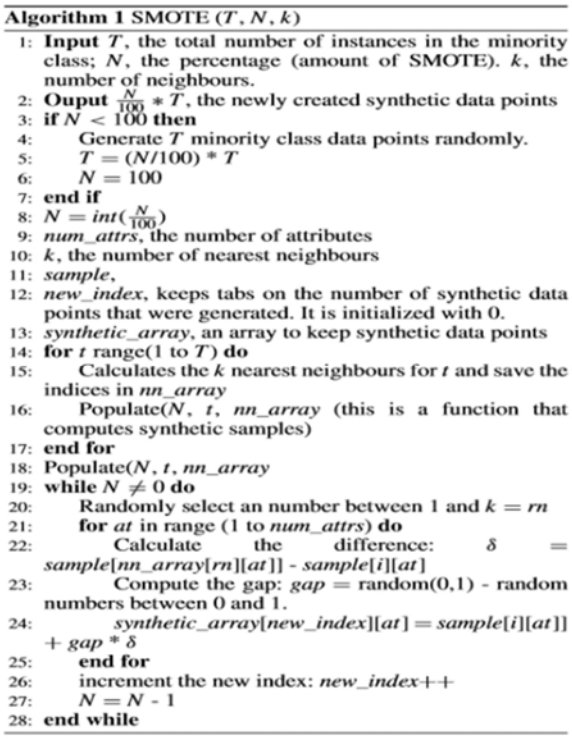
Entonces, una vez corregidas las variables, en la tabla N°6 se puede observar que ya las variables no presentan este problema, por lo que el modelo lograría tener una mejor aproximación.

### 6.2.2. Modelos de Machine Learning

#### 6.2.2.1. Técnica SMOTE aplicada en la base de datos de default

Debido a que los casos de default son pocos a comparación con los casos no default, se aplicará la técnica “The Synthetic Minority over-sampling Technique” (SMOTE), el cual es altamente utilizada para muestras no balanceadas. Por ejemplo, un caso similar es el estudio de Ileberi et. (2021), los cuales intentan probar distintos métodos de Machine Learning con la finalidad de encontrar el modelo que tenga mejor poder de predicción. El punto positivo de este método es que genera puntos sintéticos que no replican exactamente a la clase con menores valores. En consecuencia, gracias a este método, se evita el fenómeno de “over-fitting” durante el proceso de entrenamiento de la data. En la Figura N°1, se muestra el código de implementación de la técnica SMOTE, el cual se usará para la presente investigación.

Figura 5. Proceso del algoritmo SMOTE



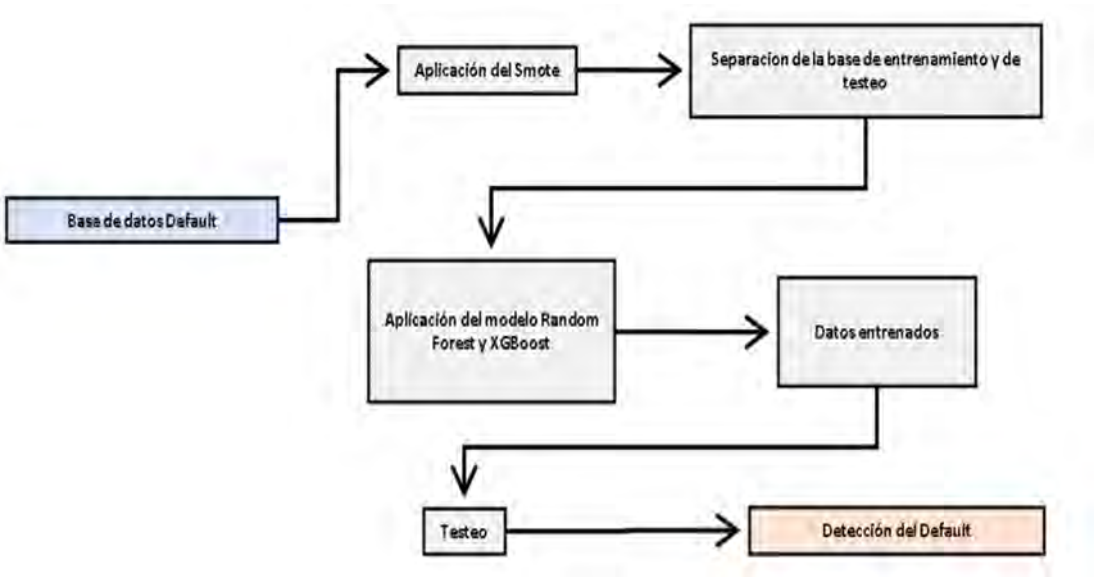
**Algorithm 1** SMOTE $(T, N, k)$

**Input** $T$, the total number of instances in the minority class; $N$, the percentage (amount of SMOTE). $k$, the number of neighbours.
**Ouput** $\frac{N}{100} * T$, the newly created synthetic data points
**if** $N < 100$ **then**
    Generate $T$ minority class data points randomly.
    $T = (N/100) * T$
    $N = 100$
**end if**
$N = int(\frac{N}{100})$
*num_attrs*, the number of attributes
$k$, the number of nearest neighbours
*sample*,
*new_index*, keeps tabs on the number of synthetic data points that were generated. It is initialized with 0.
*synthetic_array*, an array to keep synthetic data points
**for** $t$ range(1 to $T$) **do**
    Calculates the $k$ nearest neighbours for $t$ and save the indices in *nn_array*
    Populate($N$, $t$, *nn_array* (this is a function that computes synthetic samples)
**end for**
Populate($N$, $t$, *nn_array*
**while** $N \neq 0$ **do**
    Randomly select an number between 1 and $k = rn$
    **for** *at* in range (1 to *num_attrs*) **do**
        Calculate the difference: $\delta = sample[nn\_array[rn][at]] - sample[i][at]$
        Compute the gap: $gap$ = random(0,1) - random numbers between 0 and 1.
        $synthetic\_array[new\_index][at] = sample[i][at]] + gap * \delta$
    **end for**
    increment the new index: *new_index*++
    $N = N - 1$
**end while**

Fuente: Algoritmo que explica el proceso de SMOTE, Ileberi et al. (2021)

Asimismo, el proceso de implementación de esta técnica se procederá a explicar de la siguiente manera. En primer lugar, se empieza importando las paqueterías de Python tanto de Pandas como de SMOTE del módulo de imblearn. Después de eso, se invoca al Dataframe de los Default para luego separarlos en tanto variables explicativas, X, así como el objetivo, y. En esta instancia, se especifica a la función SMOTE la clase minoritaria, así como su respectiva semilla. Finalmente, se aplica esta técnica de “over-sampling” para continuar con los siguientes pasos de Machine Learning, así como se ve en la figura 6.

Figura 6. Proceso de los modelos Machine Learning




Fuente: Elaboración propia

#### 6.2.2.2. Modelo Random Forest (RF)

El modelo Random Forest (RF) fue creado por Breiman (2001) y pertenece a los distintos modelos supervisados. Este modelo se basa en los árboles de decisión (Decision Tree) y tiene un nivel de precisión similar a los modelos Boosting. Es más, el RF puede aportar mejores resultados cuando existen valores atípicos y un alto de nivel de ruido en la base de entrenamiento del modelo. Asimismo, el RF está constituido por varios subgrupos, los cuales terminan dando como resultado un mismo número de árboles de clasificación. Breiman (2001) resalta que la clase preferida se define por una mayoría de votos, por lo cual ofrece un mayor número de pronósticos precisos, y específicamente, evitando el sobreajuste de los datos.

#### 6.2.2.3. Modelo XGBOOST

El modelo XGBoost es un modelo escalable que refuerza tramo por tramo y realiza los ajustes necesarios al algoritmo del modelo “Gradient Boosting”. Por esta razón, se pasará a explicar brevemente este último modelo, al cual se define una función de pérdida, un entrenamiento del árbol más débil y una adicionalmente

agregar una técnica aditiva que une los árboles más débiles con la finalidad de optimizar la función de pérdida. Es por ello, que el modelo XGBoost aumenta un factor de regulación a la función de pérdida con lo cual se generan grupos más sencillos y generativos. Este factor ayudaría a que se pueda controlar la complejidad del modelo y así pueda evitar el "overfitting" de los datos.

### 6.3. Métricas de evaluación

Para el presente trabajo de investigación, se aplicarán las siguientes técnicas de evaluación: la curva ROC (Curva de característica operativa del receptor), el AUC (Área bajo la curva ROC), la exactitud (o más conocido como "accuracy"), Score F1 y precisión. En primer lugar, se empezará explicando la precisión y la exhaustividad (Recall), ya que son elementos claves que servirán para desarrollar nuestras métricas de evaluación. Es, por ello, que la precisión la podemos definir como los casos verdaderamente positivos (TP) sobre los casos verdaderamente positivos (TP) más los casos verdaderamente negativos (FP), mientras que la exhaustividad (o Recall) mide la cantidad verdaderos positivos que es capaz de identificar el modelo.

$$Precisión = \frac{TP}{TP + FP} \dots (xix)$$

$$Recall = \frac{TP}{TP + FN} \dots (xx)$$

En segundo lugar, el score F1 combina ambas medidas con la finalidad de un valor, mediante una media armónica en el cual asumimos que tanto la precisión y el "Recall" son de igual importancia. Por otro lado, la exactitud (Accuracy) representa el porcentaje entre los verdaderos positivos más los verdaderos negativos sobre el total de casos observados.

$$F1 = 2 * (\frac{precisión * recall}{precision + recall}) \dots (xxi)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots (xxii)$$

Finalmente, la curva ROC (curva de característica operativa del receptor) muestra la relación entre el "Recall" o tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) en diferentes umbrales de clasificación. En el mismo sentido, el

AUC, como su propio nombre lo dice, mide el área por debajo de la curva ROC, esta medida representa el grado o medida de separabilidad y comprende valores de 0 a 1. Por lo tanto, mientras más se acerca el AUC a 1, se mostrará una buena medida de separabilidad.

$$FPR = \frac{FP}{FP + TN} \ldots (xxiii)$$

## 7. Resultados

### 7.1. Logístico

Los resultados de la regresión logística realizados en Stata muestran que el AUC es 0.76, lo que significa que tiene una capacidad relativamente alta para predecir los escenarios en los que la variable dependiente (y) puede ser 0 (no default) o 1 (defaults). Esto en línea con lo expuesto por Valle (2017), quién menciona que un modelo con un AUC menor a 0.5 es ineficiente para predecir dichos escenarios. Asimismo, los resultados de la tabla también señalan que la mayoría de las variables utilizadas en el modelo de regresión son estadísticamente significativas al 1%. Entonces, podemos decir que los ratios financieros son muy importantes para predecir la probabilidad de default de un portafolio de deuda corporativa de empresas en países emergentes.

Tabla 7. Regresión logística

| Default | Coef. | St.Err. | t-value | p-value | [95 % Conf | Interval] | Sig |
|---|---|---|---|---|---|---|---|
| revenue_to_tot_ass~1 | 1.307 | .602 | 2.17 | .03 | .128 | 2.487 | ** |
| working_capital_to~1 | 0 | 0 | -0.93 | .353 | 0 | 0 | |
| cur_ratio_1 | -.194 | .179 | -1.08 | .28 | -.545 | .158 | |
| return_com_eqy_1 | -.352 | .055 | -6.39 | 0 | -.459 | -.244 | *** |
| return_on_asset_1 | .308 | .114 | 2.70 | .007 | .084 | .532 | *** |
| tot_debt_to_ebitda_1 | .003 | .001 | 3.33 | .001 | .001 | .005 | *** |
| tot_mkt_val_to_tot~1 | 0 | 0 | -1.44 | .15 | -.001 | 0 | |
| Constant | -7.3 | .182 | -40.22 | 0 | -7.656 | -6.945 | *** |
| Mean dependent var | 0.001 | SD dependent var | | | 0.027 | | |
| Pseudo r-squared | 0.035 | Number of obs | | | 51060 | | |
| Chi-square | 126.306 | Prob >chi2 | | | 0.000 | | |
| Akaike crit. (AIC) | 589.457 | Area Under The Curve. (AUC) | | | 0.7586 | | |

*** p<.01, ** p<.05, * p<.1

Fuente: Elaboración propia

### 7.2. Machine Learning

La tabla 8 muestra los resultados de la evaluación tanto del modelo logístico y de los Machine Learning (ML) con lo cual se comprueba la hipótesis inicial. Se trabajó con el lenguaje de programación Python con la versión 3.9 y los resultados de los modelos ML fueron obtenidos mediante la paquetería scikit-learn. También, el umbral

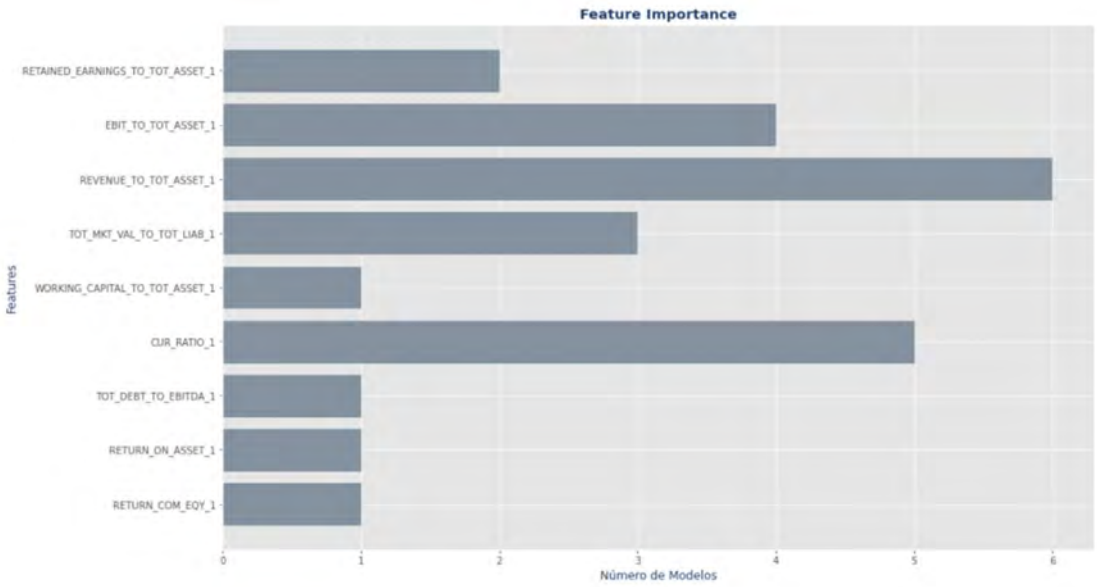
para definir las falsas positivas y las verdaderas negativas fue de 0.5. En líneas generales, se muestra que el modelo Random Forest (RF) y el XGBoost tienen similar AUC, mientras que el AUC del modelo de regresión logística es menor a los modelos ML.

Tabla 8. Evaluación de modelos

| Modelo | AUC | Accuracy | Score F1 | Precision | Recall |
|---|---|---|---|---|---|
| LogisticRegression | 0.792 | 0.784 | 0.734 | 0.956 | 0.596 |
| XGBoost | 0.924 | 0.885 | 0.879 | 0.931 | 0.832 |
| RandomForestClassifier | 0.939 | 0.903 | 0.898 | 0.945 | 0.856 |

Fuente: Elaboración propia

Figura 7. Importancia por variables

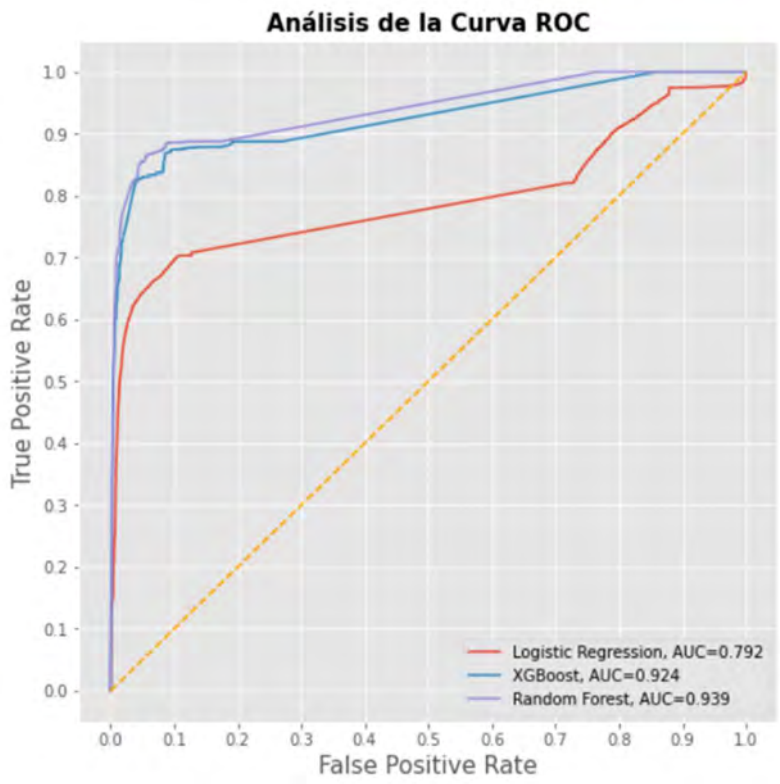



Fuente: Elaboración propia

Asimismo, otro paso importante a resaltar es que se ha realizado un proceso de clasificación de qué variables son más eficientes, los cuales se muestran en la figura 2. La importancia de este gráfico para el presente trabajo de estudio es que la mayoría de las variables del modelo de Altman (2000) han sido importantes a excepción de la variable de liquidez, es decir, el Working Capital sobre los Activos

Totales. No obstante, fue la otra variable de liquidez (Current Ratio) la que tomó mayor importancia para este caso de empresas latinoamericanas. Una diferencia por resaltar, puesto que estas variables se pasarán a entrenar a los datos como primera etapa inicial y con lo cual terminar comprobando el poder de discriminación de los modelos evaluados.

Figura 8. Curvas Roc de los modelos ML




Fuente: Elaboración propia

Finalmente, en la tabla 5, se observa que el AUC de los modelos Random Forest (RF) y XGBoost son de 93.9% y 92.4% respectivamente, mientras que, para el modelo de regresión logística, el AUC es de 79.2%. Estos resultados se ven mejor reflejados en la Figura 3, ya que los modelos ML tienen curvas ROC mayores a los de la curva ROC del modelo logístico. En el mismo sentido, el accuracy de los modelos ML siguen el mismo sentido con un 90.3% para el Random Forest y un 88.5% para el modelo XGBoost; en cambio, el “accuracy” del modelo logístico es de 78.4%. Sin embargo, para el caso de la precisión, fue el modelo logístico el que mejor performance ha tenido con un 95.6%, mientras que los modelos ML tuvieron un rendimiento de 93.1% para

el caso del XGBoost y un 94.5% para el caso del Random Forest (RF). A pesar de que el modelo logístico ha tenido mejor rendimiento en el caso de esta métrica, las diferencias no son significativas.

## 8. Agenda y limitaciones

En línea con lo propuesto por Barboza et al. (2017), y en medio de un contexto en el que medir la probabilidad de default es muy importante para todas las compañías, dado que reducirían su exposición al riesgo. Se encontró que los modelos de Machine Learning, predicen mejor la probabilidad de default para los portafolios de deuda corporativa de empresas ubicadas en economías emergentes. La capacidad de aprendizaje que tienen los modelos de Machine Learning es una de las principales características por las cuales estos modelos tienen una mejor predicción que los modelos tradicionales.

Para este primer curso, se ha logrado avanzar en gran medida la base de datos, incluso llegando a tener buenos rendimientos con respecto a los resultados mostrados anteriormente. No obstante, esta todavía no está completa totalmente. Esto se debe principalmente a que los terminales de Bloomberg tienen un límite al momento de extraer datos masivamente, por lo que, por el momento, solo se ha considerado ratios financieros o, en su defecto, variables idiosincráticas de las propias compañías. En consecuencia, uno de los objetivos para el siguiente curso será completar esta primera base junto con las variables descriptivas de cada bono emitido. Por otro lado, otro de los retos propuestos es lograr una mejor definición de la variable dependiente, ya que, por ahora, estamos tomando un dato dado por la terminal de Bloomberg. Por último, con respecto a los resultados, se hará una revisión más exhaustiva de otros modelos con la finalidad de aumentar el rigor del presente trabajo.

## 9. Conclusiones

En línea con lo propuesto por Barboza et al. (2017), y en medio de un contexto en el que medir la probabilidad de default es muy importante para todas las compañías, dado que reducirían su exposición al riesgo. Se encontró que los modelos de Machine Learning, predicen mejor la probabilidad de default para los portafolios de deuda corporativa de empresas ubicadas en economías emergentes. La capacidad de aprendizaje que tienen los modelos de Machine Learning es una de las principales características por las cuales estos modelos tienen una mejor predicción que los modelos tradicionales.

## 10. Bibliografía


Altman, E. (2000). Revisiting the Z-Score and ZETA®. *Predicting Financial Distress of Companies*(2-16). https://doi.org/https://pages.stern.nyu.edu/~ealtman/PredFnclDistr.pdf

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405-417. https://doi.org/https://doi.org/10.1016/j.eswa.2017.04.006.

Bartual, C., Fernando, G., Guijarro, F., & Romero-Civera, A. (2012). PROBABILITY OF DEFAULT USING THE LOGIT MODEL: THE . *International Scientific Conference "Whither Our Economies"*. https://doi.org/https://riunet.upv.es/bitstream/handle/10251/61415/Bartual%2C%20C.%20-%20PROBABILITY%20OF%20DEFAULT%20USING%20THE%20LOGIT%20MODEL.pdf?sequence=4

Bauer , P., & Yamey , B. (1989). La crisis de la deuda externa del Tercer Mundo. ¿Real o imaginaria? *Revista De Ciencia Política*, 79-93. https://doi.org/https://revistapolitica.uchile.cl/index.php/RP/article/view/54351

Bernanke, B. (2010). Monetary policy and the housing bubble. *Annual Meeting of the American Economic Association, Atlanta, Georgia,.* BIS. https://www.bis.org/review/r100113a.pdf

Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking & Finance, 33*(2), 281-299. https://doi.org/https://doi.org/10.1016/j.jbankfin.2008.08.006

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Carvalho, P., Curto, J., & Primor, R. (2022). Macroeconomic determinants of credit risk: Evidence from the Eurozone. *International Journal of Finance & Economics, 27*(2). https://doi.org/https://doi.org/10.1002/ijfe.2259

CEPAL, NU. (1998). *Impacto de la crisis asiática en América Latina = Impact of the Asian crisis on Latin America.* CEPAL.

Covitz, D., Liang, N., & Suarez, G. (2013). The Evolution of a Financial Crisis: Collapse of the Asset-Backed Commercial Paper Market. *The Journal of Finance, 68*(3). https://doi.org/https://doi.org/10.1111/jofi.12023

Cuadra, H. (2000). *Globalización y mercados emergentes en los noventa: Crisis financieras en México y el sudeste asiático (Doctoral dissertation, Tesis, Universidad de Colima).*

Del Valle Moreno, J., & Guerra Bustillo, W. (2012). La multicolinealidad en modelos de regresión lineal múltiple. *Revista Ciencias Técnicas Agropecuarias, 21*(4), 80-

83. https://doi.org/ http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2071-00542012000400013&lng=es&tlng=pt.

Del Valle, R. (2017). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones.*

Federal Deposit Insurance Corporation. (1997). History of the Eighties--lessons for the Future: An examination of the banking crises of the 1980s and early 1990s. En T. Curry, *The LDC Debt Crisis.* Federal Deposit Insurance Corporation.

Hestie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer.

Hsiao, I., & Gao, L. (2016). *Models of Bankruptcy Prediction Since the Recent Financial Crisis: KMV, Naïve, and Altman's Z-score.* Lund University.

Ileberi, E., Sun, Y., & Wang, Z. (2021). Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost. *IEEE Access, 9*(9). https://doi.org/10.1109/ACCESS.2021.3134330

Jorion, P., & GARP (Global Association of Risk Professionals). (2010). *Financial Risk Manager Handbook: FRM Part I / Part II.* Wiley.

McNeil, A., Rüdiger, F., & Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition.* Princeton university press.

Memić, D. (2015). Assessing credit default using logistic regression and multiple discriminant analysis: empirical evidence from Bosnia and Herzegovina. *Interdisciplinary Description of Complex Systems: INDECS, 13*(1), 128-153. https://doi.org/10.7906/indecs.13.1.13

Ocampo, J. (2014). *La crisis latinoamericana de la deuda desde la perspectiva histórica.* Cepal. https://doi.org/https://hdl.handle.net/11362/36761

Ofek, E., & Richardson, M. (2001). Dotcom Mania: The Rise and Fall of Internet Stock Prices. *NBER Working Paper, 8630*. https://doi.org/https://ssrn.com/abstract=293243

Shetty, S., Musa, M., & Brédart, X. (2022). Bankruptcy Prediction Using Machine Learning Techniques. *Journal of Risk and Financial Management, 15*(1). https://doi.org/10.3390/jrfm15010035

Sjur, W., & Van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach,. *European Journal of Operational Research, 135*(2), 338-349. https://doi.org/10.1016/S0377-2217(01)00045-5.

Trabelsi, S., He, R., He, L., & Kusy, M. (2015). A comparison of Bayesian, Hazard, and Mixed Logit model of bankruptcy prediction. *Computational Management Science, 12*(1), 81-97. https://doi.org/10.1007/s10287-013-0200-8

Vasicek, O. (15 de 11 de 2022). *Bank of Greece.* https://www.bankofgreece.gr/MediaAttachments/Vasicek.pdf

Watkins, K., Van Dijk, D., & Spronk, J. (2017). Crisis macroeconómica y desempeño de la empresa individual. La experiencia mexicana. *El Trimestre Económico*. https://doi.org/10.20430/ete.v76i304.504

Yeh, C.-C., Lin, F., & Hsu, C.-Y. (2012). A hybrid KMV model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems, 33*, 166-172. https://doi.org/10.1016/j.knosys.2012.04.004.

Zhang, Y. (s.f.). *On Credit Risk Management Models: Creditmetrics vs. KMV.*