Un Enfoque Híbrido de Clustering y Análisis de Supervivencia No Paramétrico para la Detección de Insolvencia en PYMEs Españolas

Angel Alberto Vazquez-Sánchez[1], Marcelino Cabrera-Cuevas[4], María Elena Gómez-Miranda[3], Dionisio Buendía-Carrillo[3], and Carlos Cruz Corona[2]

[1] Departamento Inteligencia Computacional, Universidad de las Ciencias Informáticas, La Habana, Cuba
aavazquez@uci.cu
[2] Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España
carloscruzcorona@ugr.es
[3] Departamento de Economía Financiera y Contabilidad
{dbuendia, melena}@ugr.es
[4] Departamento de Lenguajes y Sistemas
mcabrera@ugr.es

Resumen El análisis del riesgo de insolvencia en pequeñas y medianas empresas representa un desafío crítico debido a la complejidad multidimensional de los indicadores financieros y a la naturaleza temporal del riesgo de quiebra. Los enfoques tradicionales, basados habitualmente en la evaluación aislada de variables financieras o en cortes univariantes, pueden pasar por alto interacciones complejas entre distintos ratios. Este estudio propone un enfoque híbrido aplicado a un conjunto histórico de datos de más de 100.000 pymes españolas durante el periodo 1999-2020. En primer lugar, se aplicó el algoritmo K-Means sobre ratios financieros estandarizados, lo que permitió segmentar la muestra en dos arquetipos financieros: «Resiliente» (67,7 %) y «Vulnerable» (32,3 %). Posteriormente, el análisis de supervivencia mediante el estimador de Kaplan-Meier y la prueba de Log-Rank confirmó la existencia de diferencias estadísticamente significativas entre las trayectorias de insolvencia de ambos grupos. Los resultados sugieren que la estratificación basada en perfiles financieros multidimensionales ofrece una estratificación con mayor poder discriminante del riesgo de insolvencia respecto a los enfoques financieros univariantes tradicionales.

Keywords: Análisis de Supervivencia · Kaplan-Meier · K-Means · Insolvencia · PYME · Log-Rank.

1. Introducción

Las Pequeñas y Medianas Empresas (PYMEs) constituyen la columna vertebral de la economía española [16], y su alta vulnerabilidad a shocks económicos

hace que la detección temprana de la insolvencia sea una prioridad [17]. El Análisis de Supervivencia (*Survival Analysis*, SA) modela explícitamente el "tiempo hasta el evento" y maneja la censura por la derecha [8,10], siendo los métodos no paramétricos como Kaplan-Meier [12,9] fundamentales para estimar funciones de supervivencia sin supuestos distribucionales restrictivos.

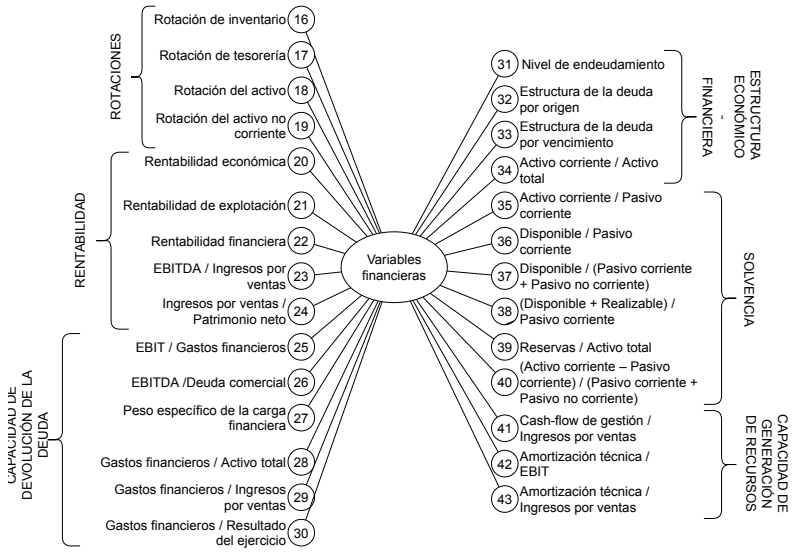
No obstante, la estratificación univariante de empresas para comparar riesgos ignora las correlaciones compensatorias entre múltiples variables financieras [20]. Para superar esta limitación, este trabajo propone un enfoque híbrido que integra K-Means para descubrir "arquetipos financieros" multidimensionales a partir de 26 ratios financieros, y valida su poder discriminatorio mediante Kaplan-Meier y la prueba de Log-Rank [14]. El objetivo es evaluar si esta estratificación proporciona una comprensión más robusta del riesgo de insolvencia en las PYMEs españolas (1999–2020) frente al análisis univariante tradicional.

2. Revisión de la Literatura

En las ciencias económicas y financieras, el Análisis de Supervivencia (*Survival Analysis*, SA) se utiliza cada vez más por su capacidad para modelar el "tiempo hasta el evento" y manejar la censura por la derecha [8,10,4]. Dentro del SA, los métodos no paramétricos como Kaplan-Meier [12,9] y la prueba de Log-Rank [14] permiten estimar funciones de supervivencia y comparar trayectorias entre grupos sin asumir supuestos distribucionales [7].

No obstante, la estratificación univariante (e.g., dicotomización por la mediana de un único ratio) [13] ignora las correlaciones compensatorias entre múltiples variables financieras [20]. Para superar esta brecha, este trabajo propone un enfoque híbrido que integra K-Means para descubrir arquetipos financieros multidimensionales y valida su poder discriminatorio mediante Kaplan-Meier estratificado y la prueba de Log-Rank multivariante.

En el ámbito financiero, diversas aplicaciones recientes han demostrado la utilidad del SA para la predicción de insolvencia. Dellana y West [6] emplearon modelos de Cox para predecir la quiebra en cadenas de suministro, y combinaron sus curvas de supervivencia con K-Means para priorizar el riesgo en tres niveles. Affes y Hentati-Kaffel [1] propusieron un enfoque híbrido K-Means–MARS que mejoró la precisión de clasificación frente a modelos individuales en la predicción de quiebra bancaria. En el contexto de PYMEs, Borges y Carvalho [2] compararon Cox con *Random Survival Forests*, encontrando que los modelos de ensemble superan a Cox en predicción a corto plazo. Mena-Siles y Gemar [15] aplicaron Kaplan-Meier y regresión *cloglog* a PYMEs españolas para identificar predictores de supervivencia. No obstante, son escasos los trabajos que integran clustering multidimensional con SA no paramétrico para la segmentación de riesgo en cohortes amplias de PYMEs, como la analizada en este estudio.



(a) Variable financieras

Figura 1: Variables financieras del conjunto de datos históricos de PYMEs.

3. Materiales y Métodos

3.1. Fuente de datos y población de estudio

La población de estudio está constituida por una cohorte histórica de pequeñas y medianas empresas (PYMEs) españolas observadas durante el periodo 1999–2020, clasificadas conforme a la definición de la Comisión Europea [3]. Los datos fueron obtenidos del Registro Mercantil Español, base que proporciona estados financieros estandarizados e información sobre el estatus legal de las firmas.

La muestra analítica final incluye **117.052 empresas**, cada una caracterizada mediante 28 ratios financieros (liquidez, endeudamiento, rentabilidad, flujo de caja operativo, entre otros) (Fig. 1). Esta amplia ventana temporal es de particular interés, ya que captura múltiples ciclos económicos: la expansión previa a 2008, la crisis financiera global, la recuperación posterior y el inicio de la pandemia por COVID-19.

3.2. Preprocesamiento de datos

El preprocesamiento se estructuró en dos fases: una fase *base* de limpieza y construcción del dataset longitudinal, y una fase *específica* de preparación para el clustering.

Fase base Los datos originales en formato *counting-process* se limpiaron eliminando columnas con una tasa de valores faltantes superior al 50 %, incluyendo

dos ratios financieros (*EBITDA/Deuda comercial* y *Estructura de la deuda por origen*), reduciendo el conjunto de variables financieras de 28 a 26. Se construyó un indicador binario de insolvencia $\delta_i \in \{0, 1\}$ y se imputaron los valores ausentes recurriendo primero a la media histórica de la empresa cuando esta disponía de observaciones suficientes, y a la media del sector en caso contrario. Para el tratamiento de outliers se aplicó *clipping* basado en el rango intercuartílico (IQR), una estrategia univariante adecuada para la limpieza inicial variable por variable. Finalmente, se colapsó el dataset a una única fila por empresa, tomando la primera observación cronológica como referencia basal. El tiempo hasta el evento se calculó como:

$$T_i = \begin{cases} Stop_i - \text{mín}(Start_i) & \text{si ocurre el evento} \\ \text{máx}(Stop_i) - \text{mín}(Start_i) & \text{si censurado} \end{cases} \tag{1}$$

Fase específica Sobre los 26 ratios resultantes tras la fase de limpieza ($F16$–$F25$, $F27$–$F31$, $F33$–$F43$), se aplicó estandarización mediante z-score. Estas variables capturan rotaciones, rentabilidad, estructura de deuda, solvencia y capacidad de generación de recursos. Para manejar los valores atípicos en el espacio multidimensional, se implementó una estrategia de filtrado basada en la distancia al centroide global: las 115.881 empresas más próximas (percentil 99) se usaron para entrenar el modelo, y las 1.171 excluidas se reasignaron al centroide más cercano, preservando la muestra total de 117.052 empresas. Este enfoque multivariante es preferible al clipping univariante en esta fase, ya que preserva la estructura conjunta de los ratios.

3.3. Descubrimiento de arquetipos financieros (clustering)

Se empleó el algoritmo K-Means para identificar patrones latentes en el espacio financiero multidimensional. El objetivo del algoritmo es minimizar la inercia (suma de las distancias al cuadrado de las muestras a su centroide más cercano).

Para determinar el número óptimo de clusters (K), se evaluó el rango $K \in [2, 10]$ sobre una muestra aleatoria estratificada de 20.000 empresas, con el fin de equilibrar la eficiencia computacional y la representatividad estadística. Se utilizaron dos métricas de validación interna:

- **Inercia** (*within-cluster sum of squares*): para observar el punto de inflexión (*elbow method*).
- **Coeficiente de Silueta**: que mide la cohesión intra-cluster y la separación inter-cluster, definido como:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}} \tag{2}$$

  donde $a(i)$ es la distancia media a otros puntos del mismo cluster y $b(i)$ la distancia media al cluster vecino más cercano.

La configuración final del modelo utilizó inicialización *k-means++*, 10 inicializaciones independientes (n_init=10) y un máximo de 300 iteraciones, con una semilla fija ($random_state$=42) para garantizar la reproducibilidad.

3.4. Análisis de supervivencia no paramétrico

Se adoptó un enfoque de tiempo-hasta-evento [8,10], donde el evento corresponde a la declaración legal de insolvencia. El tiempo de supervivencia T_i se definió como en la Ecuación (1), y las observaciones censuradas se trataron como no informativas [11], supuesto razonable en este contexto dado que la censura se debe principalmente al corte temporal de la ventana de observación (1999–2020) y no a causas relacionadas con el riesgo de insolvencia. Para estimar la función de supervivencia $S_k(t)$ de cada arquetipo k, se empleó el estimador de Kaplan-Meier [12,9,18]:

$$\hat{S}_k(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{3}$$

Los intervalos de confianza al 95 % se construyeron mediante la transformación log-log. Las diferencias entre curvas de supervivencia se evaluaron mediante la prueba de Log-Rank multivariante [14,7].

3.5. Implementación computacional y reproducibilidad

Los análisis se implementaron en Python 3.11 utilizando *pandas*, *numpy*, *scikit-learn* (`KMeans` y `StandardScaler`) y *Lifelines* [5] (`KaplanMeierFitter` y `multivariate_logrank_test`) para el análisis de supervivencia. El script completo (`archetype_clustering.py`) estará disponible en un repositorio [5] público tras la publicación.

4. Resultados

4.1. Caracterización de la muestra y supervivencia global

El conjunto de datos analizado comprende información histórica de 117.052 PYMEs españolas observadas durante el periodo 1999–2020. El tiempo de supervivencia T_i se definió según la Ecuación (1), con el evento de interés correspondiente a la declaración legal de insolvencia (concurso de acreedores) o disolución administrativa por dificultades financieras. De la muestra total, 47.752 empresas (40,8 %) experimentaron el evento, mientras que 69.300 (59,2 %) fueron censuradas por la derecha.

La estimación de la función de supervivencia global $\hat{S}(t)$, obtenida mediante el estimador no paramétrico de Kaplan-Meier [12,9], reveló una mediana de supervivencia para la cohorte completa de 18, 0 años (IC 95 %: [17, 8, 18, 2]). La probabilidad de supervivencia a distintos horizontes temporales se resume en la Tabla 1, con una probabilidad de permanencia en el mercado cercana al 48, 8 % a los 20 años.

[5] https://github.com/aavazquez-go/clustering_km_for_insolvency_prediction

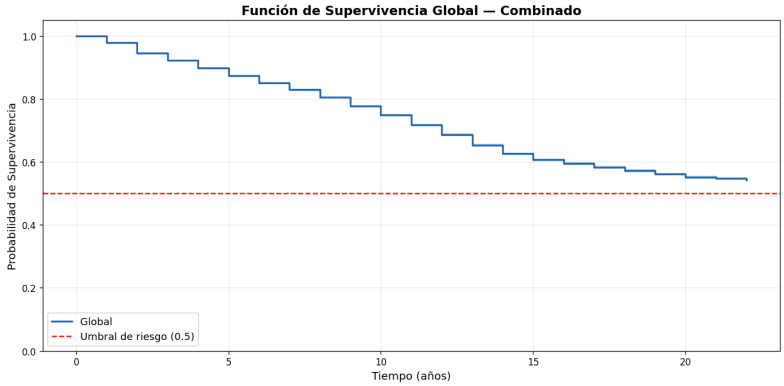


Figura 2: Estimación Kaplan-Meier de la función de supervivencia global para PYMEs españolas (1999–2020). La línea punteada roja indica el umbral de riesgo $S(t) = 0{,}5$.

Tabla 1: Probabilidades de supervivencia global estimadas mediante Kaplan-Meier

Tiempo (años)	1	3	5	10	15	20
$\hat{S}(t)$	0,9823	0,9385	0,8891	0,7542	0,6128	0,4876

4.2. Análisis estratificado univariante: línea base

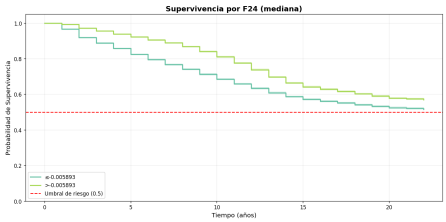
Como referencia metodológica tradicional, se evaluó la asociación entre variables predictoras individuales y el tiempo hasta la insolvencia. Las variables financieras continuas ($F16$–$F43$) fueron dicotomizadas utilizando la mediana muestral como punto de corte.

Los resultados indicaron diferencias estadísticamente significativas ($p < 0{,}05$) en la supervivencia para la mayoría de los ratios analizados. La Figura 3 muestra las curvas estratificadas. La variable F24 (relación de liquidez) presentó la mayor capacidad discriminatoria: empresas con valores superiores a la mediana alcanzaron una mediana de supervivencia no alcanzada frente a 16.0 años para el grupo inferior (log-rank $p = 5{,}35 \times 10^{-195}$). Este patrón se replicó en ratios de capacidad de devolución de deuda (F25, F29) y solvencia patrimonial (F36, F39).

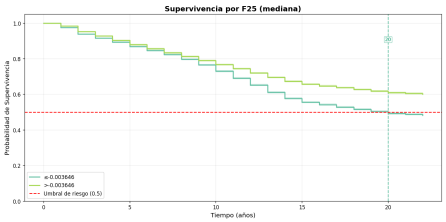
Aunque este análisis univariante proporciona una visión inicial útil, presenta limitaciones inherentes: la dicotomización por mediana de un único ratio ignora las correlaciones y compensaciones entre variables financieras, lo que puede conducir a clasificaciones erróneas del riesgo empresarial.

4.3. Determinación del número óptimo de arquetipos financieros

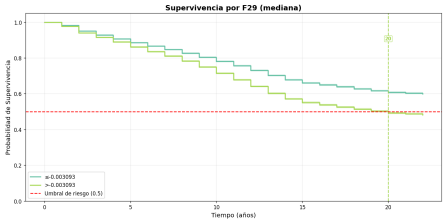
Para superar las limitaciones de la estratificación univariante, se aplicó el algoritmo K-Means sobre el espacio multidimensional de 26 variables financieras estandarizadas ($F16$–$F25$, $F27$–$F31$, $F33$–$F43$), tras filtrar temporalmente el



(a) Liquidez (F24)



(b) EBIT/Gastos financieros (F25)



(c) Gastos financieros/Ingresos (F29)

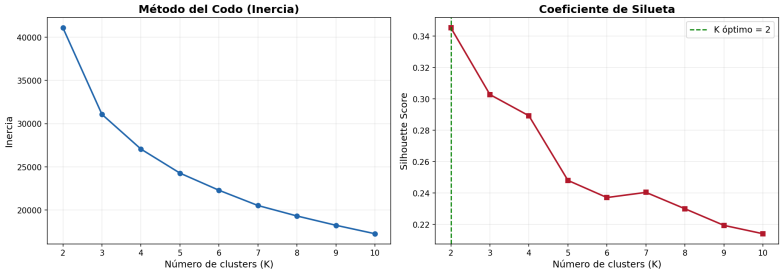
Figura 3: Curvas de supervivencia de Kaplan-Meier del análisis univariante para las tres variables con mayor capacidad discriminatoria. Las variables financieras se dicotomizaron por la mediana muestral.

1 % de *outliers* extremos (1.171 empresas) para evitar la degeneración de la partición.

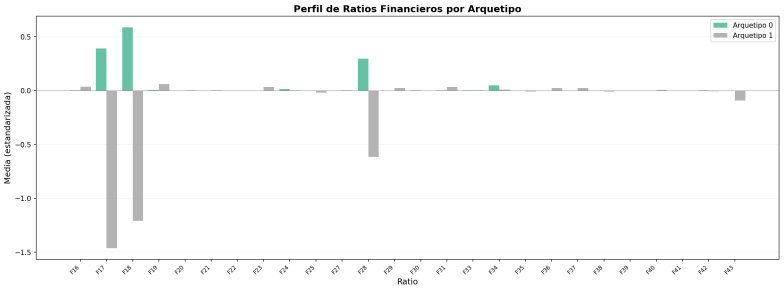
La Figura 4a presenta el método del codo y el coeficiente de silueta evaluados en el rango $K \in [2, 10]$. Aunque la inercia decrece monótonamente con K, el coeficiente de silueta —que mide simultáneamente la cohesión intra-cluster y la separación inter-cluster— alcanza su máximo en $K = 2$ ($s = 0,3453$), indicando que la partición en dos grandes arquetipos es la que maximiza la cohesión intra-cluster y la separación inter-cluster según esta métrica. Un valor de silueta de 0,3453 sugiere una separación moderada entre arquetipos, con una zona de solapamiento donde las empresas presentan perfiles financieros intermedios; no obstante, esta partición sigue siendo la más discriminante dentro del rango evaluado, como confirman los análisis de estabilidad presentados en la Subsección 4.5. Para $K \geq 3$, el coeficiente de silueta desciende progresivamente, sugiriendo que particiones más finas generan clusters menos cohesivos y con menor capacidad discriminante.

4.4. Caracterización de los arquetipos financieros

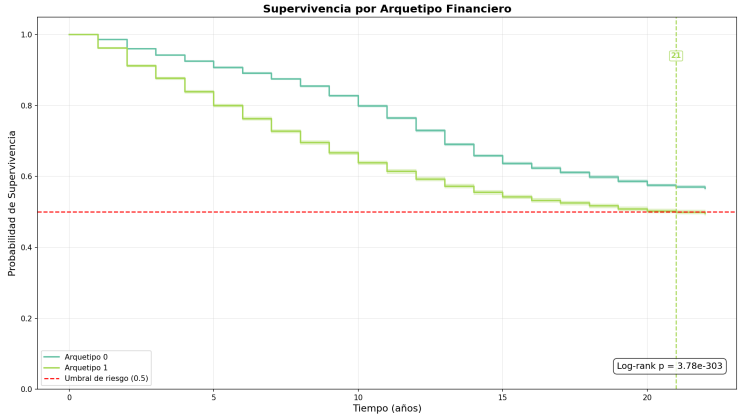
Una vez fijado $K = 2$, se ejecutó el algoritmo K-Means con inicialización *k-means++*, 10 inicializaciones independientes y un máximo de 300 iteraciones. Los *outliers* filtrados fueron reasignados al centroide más cercano. El **Arquetipo 0 (Resiliente)** agrupa el 67,7 % de la muestra (79.278 empresas), con valores positivos en F17 y F18, tasa de insolvencia del 39,3 % y mediana no alcanzada (>



(a) Selección del número óptimo de arquetipos. Izquierda: método del codo. Derecha: coeficiente de silueta, con máximo en $K = 2$.



(b) Perfil de ratios financieros por arquetipo.



(c) Curvas de Kaplan-Meier estratificadas por arquetipo.

Figura 4: Selección de arquetipos (a), perfiles financieros (b) y curvas de supervivencia estratificadas (c).

Tabla 2: Resultados de estabilidad de los arquetipos financieros

Métrica	**Valor**	**Método**
ARI promedio ($K = 5$)	0,9863	Cross-clustering
AMI promedio ($K = 5$)	0,9692	Cross-clustering
Jaccard global	0,9912	Bootstrap pairs
Estabilidad Cluster 0 (Resiliente)	0,7632	Bootstrap
Estabilidad Cluster 1 (Vulnerable)	0,2400	Bootstrap

Tabla 3: Probabilidades de supervivencia estratificadas por arquetipo financiero

Tiempo (años)	**Arquetipo 0 (Resiliente)**	**Arquetipo 1 (Vulnerable)**
1	0, 9868 [0, 9860–0, 9876]	0, 9625 [0, 9605–0, 9644]
3	0, 9425 [0, 9409–0, 9441]	0, 8770 [0, 8736–0, 8803]
5	0, 9077 [0, 9056–0, 9097]	0, 7997 [0, 7955–0, 8037]
10	0, 7987 [0, 7959–0, 8016]	0, 6389 [0, 6338–0, 6439]
15	0, 6370 [0, 6335–0, 6404]	0, 5423 [0, 5370–0, 5477]
20	0, 5752 [0, 5715–0, 5788]	0, 5024 [0, 4968–0, 5081]

20 años). El **Arquetipo 1 (Vulnerable)** concentra el 32,3 % (37.774 empresas), con valores negativos en F17 y F18, elevado F28, tasa de insolvencia del 43,9 % y mediana de 21 años.

4.5. Estabilidad de los arquetipos financieros

Para evaluar la robustez de la partición en dos arquetipos, se aplicaron tres métodos: (i) bootstrap de estabilidad con 200 iteraciones, (ii) validación cruzada K-fold ($K = 5$) con ARI y AMI, e (iii) índice de Jaccard [21] entre 50 pares de particiones bootstrap independientes.

La Tabla 2 resume los resultados. La estructura global es robusta: ARI=0,986 y Jaccard=0,991 confirman que la dicotomía Resiliente/Vulnerable emerge consistentemente. No obstante, el bootstrap revela una asimetría: el cluster Resiliente alcanza estabilidad de 0,76, mientras que el Vulnerable alcanza solo 0,24, indicando una zona de frontera difusa coherente con el coeficiente de silueta moderado ($s = 0,345$).

4.6. Validación de los arquetipos mediante análisis de supervivencia

Para validar el poder predictivo de la estratificación por arquetipos, se estimaron las funciones de supervivencia $\hat{S}_k(t)$ para cada grupo mediante el estimador de Kaplan-Meier [12,9]. La Figura 4c muestra las curvas de supervivencia estratificadas con sus intervalos de confianza al 95 %.

La Tabla 3 resume las probabilidades de supervivencia a distintos horizontes temporales para cada arquetipo:

La divergencia entre las curvas se acentúa con el tiempo: a los 5 años, la diferencia en la probabilidad de supervivencia es de aproximadamente 11 puntos porcentuales (90, 8 % vs. 80, 0 %), expandiéndose a casi 16 puntos a los 10 años

Tabla 4: Comparación de métricas para $K \in \{2, 3, 4\}$

K	Silhouette	Bootstrap	CV-ARI	Jaccard	Log-rank p
2	**0,3436**	**0,6391**	**0,9863**	**0,9912**	$3{,}78 \times 10^{-303}$
3	0,3047	0,3828	0,9833	0,9864	≈ 0
4	0,2878	0,2612	0,9254	0,9347	≈ 0

(79, 9 % vs. 63, 9 %). Esta brecha progresiva es consistente con la hipótesis de que las vulnerabilidades financieras identificadas por el clustering se asocian con un mayor riesgo acumulado de insolvencia a lo largo del tiempo.

Para determinar la significancia estadística de estas diferencias, se aplicó la prueba de Log-Rank multivariante [14,7]:

$$H_0 : S_0(t) = S_1(t) \quad \forall t \tag{4}$$

El estadístico de prueba arrojó un valor $\chi^2 = 1{,}385{,}02$ con 1 grado de libertad, y un p-valor $= 3{,}78 \times 10^{-303}$. Este valor, esencialmente cero, permite rechazar de manera contundente la hipótesis nula de que ambos arquetipos comparten la misma función de supervivencia, confirmando que la segmentación multidimensional posee un alto poder discriminatorio sobre el riesgo temporal de insolvencia.

4.7. Análisis de sensibilidad por número de arquetipos

La Tabla 4 compara $K \in \{2, 3, 4\}$ según calidad, estabilidad, separación de supervivencia y parsimonia.

$K = 2$ alcanza la máxima calidad (silueta=0,3436), estabilidad (bootstrap=0,64, ARI=0,986) y parsimonia. $K = 3$ produce una brecha de supervivencia ligeramente mayor (13,5 p.p. vs. 10,8 p.p. a 5 años) pero su estabilidad bootstrap cae un 40 % y los perfiles son variaciones de los discriminadores de $K = 2$. $K = 4$ se descarta por degradación insostenible (ARI$<$ 0,93, bootstrap$<$ 0,27).

5. Discusión y Conclusiones

El enfoque híbrido propuesto —K-Means más Kaplan-Meier— revela dos arquetipos estadísticamente distintos ($\chi^2 = 1{,}385{,}02$; $p = 3{,}78 \times 10^{-303}$) con una brecha de 16 puntos porcentuales a los 10 años, ofreciendo poder discriminante significativamente mayor que el análisis univariante.

Los hallazgos presentan implicaciones directas para la gestión financiera y el riesgo crediticio. La identificación del Arquetipo Vulnerable (32,3 % de la muestra, mediana de 21 años) frente al Resiliente (mediana no alcanzada) permite estimar la *esperanza de vida financiera* del prestatario [10,19]. La divergencia progresiva entre las curvas (11 pp a 5 años, 16 pp a 10 años) proporciona una base para calibrar tipos de interés y comisiones de riesgo [13]. La pertenencia al Arquetipo Vulnerable podría constituir una señal de alerta temprana que

oriente la priorización de revisiones de expedientes, exigencias de garantías o reestructuración preventiva de deuda. Los resultados se traducen en herramientas interpretables para gestores (*benchmarking*), entidades financieras (*scoring* con dimensión temporal) y reguladores (programas de apoyo diferenciados por perfil de riesgo).

Entre las limitaciones del estudio se encuentran el supuesto de censura no informativa —que no puede validarse formalmente en un diseño observacional retrospectivo—, la ausencia de covariables macroeconómicas dependientes del tiempo, la falta de distinción entre riesgos competitivos, la ausencia de un análisis de correlación entre variables para filtrar redundancias y la restricción geográfica y temporal de la muestra (España, 1999–2020). Como líneas futuras, se propone extender el marco con modelos semi-paramétricos de Cox y *Random Survival Forests* para evaluar si el poder predictor de los arquetipos se mantiene frente a enfoques con covariables explícitas y cuantificar la contribución individual de cada ratio financiero al riesgo temporal de insolvencia (mediante hazard ratios en Cox y medida de importancia de variables en RSF), así como realizar validación prospectiva, análisis de trayectorias financieras, incorporación de variables no financieras y filtrado basado en correlación entre ratios.

En definitiva, este trabajo realiza tres contribuciones fundamentales —metodológica, empírica y práctica— y establece una base sólida para transformar datos administrativos históricos en herramientas estratégicas para la gestión del riesgo crediticio y el diseño de políticas de apoyo al tejido empresarial.

Agradecimientos: Los autores agradecen el apoyo del Proyecto PID2023-146575NB-I00, financiado por MICIU/AEI/10.13039/501100011033, FEDER, UE, y al Registro Mercantil Español por facilitar los datos utilizados en este estudio.

Referencias

1. Affes, Z., Hentati-Kaffel, R.: Forecast bankruptcy using a blend of clustering and MARS model: Case of US banks. Annals of Operations Research **281**(1), 27–64 (Oct 2019). https://doi.org/10.1007/s10479-018-2845-8
2. Borges, A., Carvalho, M.: Short- and long-term financial distress prediction in SMEs: A survival model comparison. Journal of Applied Statistics pp. 1–30 (May 2025). https://doi.org/10.1080/02664763.2025.2501166
3. Communities, C.o.t.E.: Commission Recommendation of 6 May 2003 Concerning the Definition of Micro, Small and Medium-Sized Enterprises (Text with EEA Relevance)(Notified under Document Number C (2003) 1422),(2003/361/EC), Official Journal L 124, 20/05/2003 P. 0036–0041 (2003)
4. Cox, D.R.: Analysis of Survival Data. Chapman and Hall/CRC (2018)
5. Davidson-Pilon, C.: Lifelines: Survival analysis in Python. Journal of Open Source Software **4**(40), 1317 (2019)
6. Dellana, S., West, D.: Survival analysis of supply chain financial risk. The Journal of Risk Finance **17**, 130–151 (2016). https://doi.org/10.1108/jrf-11-2015-0112
7. Dormuth, I., Liu, T., Xu, J., Pauly, M., Ditzhaus, M.: A comparative study to alternatives to the log-rank test. Contemporary Clinical Trials **128**, 107165 (May 2023). https://doi.org/10.1016/j.cct.2023.107165

8. Emmert-Streib, F., Dehmer, M.: Introduction to survival analysis in practice. Machine Learning and Knowledge Extraction **1**(3), 1013 – 1038 (2019). https://doi.org/10.3390/make1030058, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128708055&doi=10.3390%2fmake1030058&partnerID=40&md5=b5921b7cd71bd9e76cf0ae9f8f9ce529, cited by: 35; All Open Access, Gold Open Access
9. Etikan, I., Abubakar, S., Alkassim, R.: The Kaplan-Meier estimate in survival analysis. Biometrics & biostatistics international journal **5**(2), 55–59 (2017)
10. Gao, F., He, X.: Survival analysis: Theory and application in finance. World Scientific Publishing Co. (2020). https://doi.org/10.1142/9789811202391_0120, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096248970&doi=10.1142%2f9789811202391_0120&partnerID=40&md5=4d699100b6971b4625e6f25fadc90d98, cited By 0
11. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. John Wiley & Sons (2002)
12. Kaplan, E.L., Meier, P.: Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association **53**(282), 457–481 (Jun 1958). https://doi.org/10.1080/01621459.1958.10501452
13. Kwak, W., Shi, Y., Lee, C.F.: Data Mining Applications in Accounting and Finance Context. In: Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning, pp. 823–857. WORLD SCIENTIFIC (Aug 2020). https://doi.org/10.1142/9789811202391_0021
14. Mantel, N., et al.: Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer chemotherapy reports **50**(3), 163–170 (1966)
15. Mena-Siles, D., Gemar, G.: Survival analysis for predicting bankruptcy in small and medium-sized enterprises (SMEs): A case study approach. Small Business International Review **9**(2), e773 (Dec 2025)
16. Ministerio de Industria, C.y.T.: Cifras PYME (2025)
17. Navarro-Galera, A., Lara-Rubio, J., Novoa-Hernández, P., Cruz Corona, C.A.: Using Decision Trees to Predict Insolvency in Spanish SMEs: Is Early Warning Possible? Computational Economics **65**(1), 91–116 (Mar 2025). https://doi.org/10.1007/s10614-024-10586-5
18. RichMD, J.T., NeelyMD, J.G., PanielloMD, R.C., VoelkerMD DPhil, C.C.J., NussenbaumMD, B., WangMD, E.W.: A practical guide to understanding Kaplan-Meier curves. Otolaryngology–Head and Neck Surgery **143**(3), 331–336 (Sep 2010). https://doi.org/10.1016/j.otohns.2010.05.007
19. Toledo, J.S.B., Tomazella, V.L.D., Lima, C.M.M., Felix, M.H.: Gompertz zero-inflated cure rate regression models applied to credit risk data. Applied Stochastic Models in Business and Industry **39**(2), 177–197 (2023). https://doi.org/10.1002/asmb.2732
20. Wang, P., Li, Y., Reddy, C.K.: Machine Learning for Survival Analysis: A Survey. ACM Computing Surveys **51**(6), 1–36 (Feb 2019). https://doi.org/10.1145/3214306
21. Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M., Blair, R.H.: Bootstrapping estimates of stability for clusters, observations and model selection. Computational Statistics **34**(1), 349–372 (Mar 2019). https://doi.org/10.1007/s00180-018-0830-y