Neutrosophic Computing and Machine Learning

, Vol. 45, 2026

University of New Mexico

Predicción del riesgo crediticio mediante aprendizaje automático y análisis neutrosófico de la incertidumbre.

Credit Risk Prediction Through Machine Learning and Neutrosophic Analysis of Uncertainty.

Kleber Overdan Cedeño Pinto[1], Richard Manolo Calero Villareal[2], Tania Jeesenia Peralta Guaraca[3], Tatiana Mabel Alcívar Maldonado [4], Alfonso Aníbal Guijarro Rodríguez[5,6]

[1] Universidad de Guayaquil, Guayas, Ecuador, kleber.cedenop@ug.edu.ec, https://orcid.org/0009-0005-4857-4403
[2] Universidad de Guayaquil, Guayas, Ecuador, richard.calerovi@ug.edu.ec, https://orcid.org/0009-0001-9634-2093
[3] Universidad de Guayaquil, Guayas, Ecuador, tania.peraltag@ug.edu.ec, https://orcid.org/0000-0002-4879-6824
[4] Universidad de Guayaquil, Guayas, Ecuador, tatiana.alcivarm@ug.edu.ec, https://orcid.org/0009-0002-0993-7185
[5] Universidad de Guayaquil, Guayas, Ecuador, alfonso.guijarror@ug.edu.ec, https://orcid.org/0000-0001-6046-426X
[6] Grupo de Investigación de Inteligencia Artificial, Universidad de Guayaquil, Guayas, Ecuador.

Autor por correspondencia: alfonso.guijarror@ug.edu.ec

Resumen. Los modelos de aprendizaje automático empleados en la evaluación del riesgo crediticio suelen producir una probabilidad puntual que no distingue predicciones respaldadas por evidencia clara de aquellas hechas bajo alta ambigüedad de los datos. Este estudio propone un esquema híbrido que combina modelos de aprendizaje automático con un componente de análisis neutrosófico, capaz de cuantificar la indeterminación asociada a cada predicción individual. Se entrenaron y compararon tres modelos regresión logística, Random Forest y XGBoost sobre el conjunto German Credit Data (1000 solicitantes), y sobre sus salidas se formalizó una tripleta neutrosófica (verdad, indeterminación, falsedad), combinando indeterminación de datos e indeterminación de desacuerdo entre modelos. Random Forest obtuvo el mejor desempeño (AUC-ROC = 0.771, F1 = 0.789), aunque clasificó erróneamente a la mitad de los solicitantes de mal riesgo reales. El hallazgo principal fue que la indeterminación del modelo se correlacionó con el error en la dirección esperada, mientras que la de datos mostró una correlación negativa y significativa (r = -0.156, p = 0.007) una categoría marcada a priori como ausencia de información resultó ser una señal predictiva válida. Se concluye que la indeterminación en esquemas neutrosóficos de riesgo crediticio requiere validación empírica previa de cada fuente candidata, no solo criterios semánticos.

Palabras clave: riesgo crediticio, aprendizaje automático, lógica neutrosófica, incertidumbre, conjuntos neutrosóficos, scoring crediticio.

Abstract. Machine learning models used for credit risk assessment typically output a single point probability that fails to distinguish predictions backed by clear evidence from predictions made under highly ambiguous input data. This study proposes a hybrid scheme combining machine learning with a neutrosophic analysis component capable of explicitly quantifying the indeterminacy associated with each individual prediction. Three models logistic regression, Random Forest, and XGBoost were trained and compared on the German Credit Data dataset (1000 applicants), and a neutrosophic triplet (truth, indeterminacy, falsity) was formalized over their outputs, combining data-level indeterminacy with model-disagreement indeterminacy. Random Forest achieved the best performance (AUC-ROC = 0.771, F1 = 0.789), although it still misclassified half of the true bad risk applicants. The main finding was that model-disagreement indeterminacy correlated with classification error in the expected direction, whereas data-level indeterminacy showed a negative, statistically significant correlation (r = -0.156, p = 0.007), revealing that a category flagged a priori as missing information was in fact a valid predictive signal. It is concluded that building indeterminacy within neutrosophic schemes for credit risk requires prior empirical validation of each candidate source, rather than relying solely on semantic criteria.

Keywords: credit risk, machine learning, neutrosophic logic, uncertainty, neutrosophic sets, credit scoring.

1. Introducción

La evaluación del riesgo crediticio constituye uno de los procesos de decisión más críticos dentro del sistema financiero, ya que determina el acceso de personas y empresas al crédito y condiciona, a la vez, la estabilidad de las instituciones que lo otorgan [8]. Un error en esta evaluación sería clasificar como "buen pagador" a un solicitante que incumplirá (falso negativo) o rechazar a un solicitante solvente (falso positivo), lo cual genera pérdidas económicas directas y, en el segundo caso, exclusión financiera de agentes que sí merecían el crédito. Por esta razón, la búsqueda de modelos de predicción más precisos y más conscientes de sus propios márgenes de error ha sido una constante en la literatura de riesgo financiero [9, 10].

Durante las últimas dos décadas, los modelos de aprendizaje automático (Machine Learning, ML) regresión logística, árboles de decisión, Random Forest, máquinas de soporte vectorial y, más recientemente, métodos de boosting como XGBoost, han desplazado progresivamente a los modelos estadísticos clásicos en la construcción de sistemas de credit scoring, debido a su capacidad para capturar relaciones no lineales entre variables financieras, demográficas y de comportamiento [5, 8]. Sin embargo, estos modelos comparten una limitación conceptual, donde su salida es, en la inmensa mayoría de los casos, una probabilidad puntual o una clasificación binaria que no distingue entre una predicción respaldada por evidencia clara y una predicción hecha bajo condiciones de alta ambigüedad en los datos de entrada [14].

Esta limitación es particularmente relevante en el dominio crediticio, donde la información disponible sobre un solicitante suele ser incompleta, contradictoria o directamente ausente (por ejemplo, la falta de historial crediticio previo, o el estado "desconocido" de una cuenta corriente). Los modelos probabilísticos tradicionales absorben esta incertidumbre en un único valor numérico, lo que produce una falsa sensación de certeza: un solicitante con datos ambiguos y otro con datos claramente favorables pueden recibir un puntaje de riesgo similar, aunque la confianza que debería depositarse en cada predicción sea muy distinta [14].

La teoría de conjuntos neutrosóficos, propuesta originalmente por Smarandache [11] como generalización de la lógica difusa, ofrece un marco formal para abordar precisamente este problema. A diferencia de un conjunto difuso clásico, que solo captura el grado de pertenencia (verdad), un conjunto neutrosófico caracteriza cada elemento mediante tres componentes independientes: el grado de verdad (T), el grado de indeterminación (I) y el grado de falsedad (F) [12]. Aplicado a la predicción de riesgo crediticio, este marco permite expresar no solo si un solicitante es clasificado como en buen o mal riesgo, sino también cuánta incertidumbre rodea esa clasificación en distinción con los enfoques puramente probabilísticos no pueden representarse de forma explícita, y que trabajos recientes en el dominio financiero ya han comenzado a operacionalizar a través de puntuaciones neutrosóficas escalares [4].

El presente estudio parte de este vacío: los modelos de ML utilizados actualmente para scoring crediticio no incorporan un tratamiento explícito de la indeterminación inherente a los datos de entrada ni a la confianza de la predicción, lo cual limita su utilidad en escenarios donde la transparencia sobre el grado de certeza es tan importante como la exactitud de la clasificación misma, por ejemplo, ante requerimientos regulatorios de explicabilidad [13].

Por tanto, el objetivo de este trabajo es evaluar un modelo híbrido que combine algoritmos de aprendizaje automático con un componente de análisis neutrosófico, capaz de cuantificar la indeterminación asociada a cada predicción de riesgo crediticio, utilizando como caso de estudio el conjunto de datos German Credit Data (UCI Machine Learning Repository).

Finalmente, el resto de este artículo se articula de la siguiente manera: la Sección 2 presenta el marco teórico y los trabajos relacionados; la Sección 3 describe la metodología, incluyendo el dataset, el preprocesamiento y la formalización del componente neutrosófico; la Sección 4 reporta los resultados obtenidos; la Sección 5 discute sus implicaciones y limitaciones; y la Sección 6 presenta las conclusiones y líneas futuras de investigación.

2. Marco teórico y trabajos relacionados

2.1 Modelos de aprendizaje automático en riesgo crediticio

La aplicación de modelos de aprendizaje automático a los riesgo crediticio ha sido objeto de una intensa producción científica en los últimos años. Noriega et al. [9], en una revisión sistemática de la literatura, documentan cómo algoritmos como Random Forest, XGBoost y redes neuronales han superado consistentemente a la regresión logística clásica en términos de capacidad predictiva, aunque a costa de una menor interpretabilidad. En la misma línea, Shi et al. [10] señalan que el uso combinado de técnicas de ensemble learning con procesos de selección de variables ha permitido mejorar de forma sostenida las métricas de discriminación (AUC, KS) en distintos contextos institucionales.

Montevechi et al. [8] ofrecen una revisión exhaustiva del estado del arte en modelado de riesgo crediticio mediante ML, destacando que, si bien algoritmos como los árboles de decisión, las máquinas de soporte vectorial y las redes neuronales han incrementado la capacidad predictiva de los modelos, su adopción regulatoria aún enfrenta barreras relacionadas con la interpretabilidad y la necesidad de herramientas de explicabilidad. En un estudio aplicado sobre clientes de tarjetas de crédito, Chang et al. [5] compararon redes neuronales, regresión logística, AdaBoost, XGBoost y LightGBM, encontrando que XGBoost obtuvo el mejor desempeño (exactitud del 99.4%), lo que confirma la tendencia hacia algoritmos de boosting como estándar de facto en la industria.

No obstante, esta línea de investigación comparte una limitación señalada de forma creciente en la literatura reciente: la optimización casi exclusiva de la exactitud predictiva deja en segundo plano la cuantificación de la incertidumbre asociada a cada predicción individual. Xu et al. [14] abordan directamente este vacío, proponiendo métricas de incertidumbre basadas en el beneficio esperado para credit scoring, y demuestran que un mecanismo de clasificación con opción de rechazo identifica los casos de mayor incertidumbre predictiva y mejora la rentabilidad del proceso de decisión crediticia frente a los modelos que ignoran esta dimensión. Esta línea de trabajo es especialmente relevante para el presente estudio, ya que evidencia que la comunidad científica reconoce la necesidad de ir más allá de la clasificación binaria, aunque las soluciones propuestas hasta ahora provienen predominantemente del campo probabilístico (funciones de pérdida, redes bayesianas, ensembles de incertidumbre) y no del campo de los conjuntos neutrosóficos.

2.2 Fundamentos de conjuntos neutrosóficos

La teoría de conjuntos neutrosóficos fue formulada por Smarandache [11] como una generalización de la teoría de conjuntos difusos [15] y de los conjuntos difusos intuicionistas [3]. Mientras que un conjunto difuso clásico asocia a cada elemento un único grado de pertenencia, y un conjunto difuso intuicionista añade un grado de no pertenencia, un conjunto neutrosófico de valor único (single-valued neutrosophic set, SVNS) caracteriza a cada elemento mediante tres funciones independientes y no necesariamente complementarias: verdad (T), indeterminación (I) y falsedad (F), cada una definida en el intervalo [0, 1].

Esta independencia entre los tres componentes constituye la principal ventaja teórica de la neutrosofía frente a otras extensiones de la lógica difusa: permite representar explícitamente la ignorancia o la contradicción en la información, en lugar de forzar su resolución en un único valor numérico [12]. Por esta razón, los conjuntos neutrosóficos han sido aplicados con éxito en problemas de toma de decisiones multicriterio bajo incertidumbre, como la evaluación de estrategias de sostenibilidad [7], la selección de tecnologías de bioenergía [6], y la representación de conocimiento experto en dominios con información incompleta [1].

En el ámbito específico de la modelación financiera, Bustos-Chiliquinga et al. [4] desarrollan un enfoque de puntuación neutrosófica (neutrosophic score) que transforma cada observación financiera en una tripleta (T, I, F), donde la indeterminación se construye a partir de la volatilidad temporal y la desviación sectorial de cada indicador. Este enfoque, aplicado al análisis estructural de cooperativas de ahorro y crédito ecuatorianas, demuestra que la incorporación de indeterminación modelada no distorsiona las relaciones financieras fundamentales, pero sí modifica sustancialmente la ordenación de similitud temporal y la configuración factorial de los datos con lo cual se obtiene evidencia empírica de que la transformación neutrosófica aporta información adicional relevante que los modelos puramente numéricos no capturan.

2.3 Estado del arte: ML y neutrosofía en el dominio financiero

La integración explícita de modelos de aprendizaje automático con conjuntos neutrosóficos en el dominio financiero es todavía un campo emergente. Bustos-Chiliquinga et al. [4] constituyen, hasta donde alcanza la revisión de este estudio, uno de los antecedentes más directos y recientes de aplicación de puntuaciones neutrosóficas a variables de riesgo financiero en instituciones del sector de cooperativas ecuatorianas, aunque su enfoque se centra en el análisis estructural multivariante (Dual STATIS) y no en la predicción supervisada del incumplimiento a nivel de solicitante individual.

Fuera del dominio financiero, Al-Masri [2] propone NeutroSENSE, un esquema que integra un ensamble de Random Forest, XGBoost y regresión logística con una capa de razonamiento neutrosófico que descompone cada predicción en T, I y F para habilitar la abstención en sistemas de detección de intrusiones en IoT. Este antecedente es arquitectónicamente el más cercano al esquema propuesto en el presente estudio (mismo trío de algoritmos, misma lógica de descomposición T-I-F sobre las salidas de un ensamble), aunque aplicado a un dominio distinto (ciberseguridad en lugar de riesgo crediticio) y sin un componente de indeterminación derivado de los datos de entrada, ya que su indeterminación proviene exclusivamente del desacuerdo entre modelos.

Por su parte, la literatura sobre incertidumbre en credit scoring basada en ML como la de Xu et al. [14] ha avanzado en la cuantificación de la confianza de las predicciones, pero lo hace desde marcos estadísticos (funciones de pérdida sensibles al beneficio, estimación bayesiana) que no incorporan una representación explícita de la indeterminación como componente independiente de la verdad y la falsedad, tal como sí lo permite el marco neutrosófico [12].

En conjunto, la revisión de la literatura evidencia un vacío de investigación claro: si bien existen (a) modelos de ML robustos para la predicción de riesgo crediticio, (b) marcos neutrosóficos consolidados para la toma de decisiones bajo incertidumbre, (c) una primera aplicación de puntuaciones neutrosóficas al análisis financiero estructural en el contexto ecuatoriano, y (d) un antecedente arquitectónico de ensamble ML + descomposición T-I-F fuera del dominio financiero, no se ha identificado un estudio que integre explícitamente un modelo de clasificación supervisada de riesgo crediticio a nivel de solicitante individual con un componente neutrosófico que combine indeterminación de datos e indeterminación de modelo capaz de cuantificar la indeterminación de cada predicción. Es precisamente este vacío el enfoque principal de este estudio que busca atender mediante el caso de estudio del conjunto de datos German Credit Data.

3. Metodología

3.1 Descripción del dataset

Se utiliza el conjunto de datos Statlog (German Credit Data), disponible en el UCI Machine Learning Repository, donado por el Prof. Hans Hofmann. Está compuesto por 1000 instancias correspondientes a solicitantes de crédito, sin valores faltantes, y una variable objetivo binaria que clasifica a cada solicitante como “buen riesgo” (700 casos) o “mal riesgo” (300 casos), lo que evidencia un desbalance de clases (70/30) relevante para la selección de métricas de evaluación.

El dataset incluye 20 variables predictoras, agrupables en cinco categorías como muestra la tabla 1:

Categoría	Variables
Historial financiero	Estado de cuenta corriente, historial crediticio, monto del crédito, cuentas de ahorro, duración del crédito
Empleo e ingresos	Antiguedad laboral, tasa de cuota sobre ingreso disponible, otros deudores/garantes
Datos demográficos	Edad, estado civil/sexo, número de dependientes, residencia actual
Propiedad y otros créditos	Propiedad, otros planes de pago, número de créditos existentes
Propósito y vivienda	Propósito del crédito, tipo de vivienda, calificación laboral, teléfono, trabajador extranjero

Tabla 1. Categorías de variables predictoras del dataset.

Se emplea la versión categórica original (german.data) en lugar de la versión pre-numerizada (german.data-numeric), a fin de realizar la codificación de variables de forma explícita y metodológicamente controlada.

Cabe destacar que varias categorías del dataset (por ejemplo, estado de cuenta corriente “desconocido” o ausencia de historial crediticio previo) constituyen, en sí mismas, una forma de indeterminación presente ya en los datos de entrada, motiva y justifica la incorporación del componente neutrosófico desde la etapa de preprocesamiento, y no únicamente como una capa posterior sobre la salida del modelo.

3.2 Preprocesamiento

Sobre la versión categórica original (german.data) se aplican los siguientes pasos:

1. **Codificación de variables categóricas.** Las variables nominales sin orden natural (propósito del crédito, tipo de vivienda, sexo/estado civil) se codifican mediante *one-hot encoding*; las variables categóricas con orden implícito (estado de cuenta corriente, cuentas de ahorro, antiguedad laboral) se codifican de forma ordinal, preservando la jerarquía presente en los códigos originales del dataset (por ejemplo, A11 < A12 < A13 < A14 para el estado de cuenta corriente).

2. **Marcado explícito de categorías indeterminadas.** Antes de codificar, se identifican y se registran en una variable auxiliar binaria las categorías que representan ausencia de información más que un valor sustantivo (por ejemplo, "sin cuenta corriente", "sin historial crediticio", "desconocido/sin cuentas de ahorro"). Este registro no se descarta tras el encoding, se conserva como insumo directo para la construcción del componente de indeterminación (I) en la Sección 3.4.
3. **Partición de datos.** Se emplea una partición estratificada 70/30 (entrenamiento/prueba) respecto de la variable objetivo, con el fin de preservar la proporción original 700/300 en ambos subconjuntos. Adicionalmente, se aplica validación cruzada estratificada de 5 particiones (5-fold stratified cross-validation) sobre el conjunto de entrenamiento para la selección de hiperparámetros.
4. **Tratamiento del desbalance de clases.** Dado el desbalance 70/30 documentado en la Sección 3.1, se opta por ponderación de clases (class_weight inversamente proporcional a la frecuencia) en lugar de sobremuestreo sintético (por ejemplo, SMOTE). Esta decisión metodológica es deliberada: técnicas de sobremuestreo generan observaciones sintéticas que podrían distorsionar artificialmente el componente de indeterminación construido en la Sección 3.4, el cual depende de la presencia real de categorías ausentes o ambiguas en los datos originales.
5. **Escalado de variables.** Las variables numéricas continuas (monto del crédito, duración, edad) se estandarizan (media 0, desviación estándar 1) únicamente para los modelos sensibles a la escala (por ejemplo, regresión logística); los modelos basados en árboles no requieren este paso.

3.3 Modelos de aprendizaje automático empleados

Siguiendo el esquema comparativo habitual en la literatura de credit scoring [5, 8, 9], se entrenan y comparan tres modelos:

- Regresión logística (LR), como modelo de referencia clásico, interpretable y ampliamente utilizado como línea base regulatoria en la industria financiera.
- Random Forest (RF), como representante de los métodos de ensemble por bagging, robusto frente a variables categóricas codificadas y a relaciones no lineales.
- XGBoost, como representante de los métodos de boosting, que en estudios recientes ha mostrado el mejor desempeño predictivo individual en tareas comparables de clasificación de riesgo crediticio [5].

Los hiperparámetros de cada modelo (p. ej., profundidad máxima y número de estimadores en RF y XGBoost; parámetro de regularización C en LR) se seleccionan mediante búsqueda en cuadrícula (grid search) sobre la validación cruzada descrita en la Sección 3.2, optimizando el área bajo la curva ROC (AUC-ROC) como criterio principal, dado el desbalance de clases del dataset.

Los tres modelos entrenados cumplen una doble función en este estudio: generar la predicción base de riesgo crediticio, y proveer, mediante el grado de acuerdo o desacuerdo entre sus predicciones individuales, una de las dos fuentes de indeterminación que alimentan el componente neutrosófico descrito a continuación.

3.4 Formalización del componente neutrosófico (T, I, F)

Siguiendo la lógica de representación neutrosófica aplicada recientemente a variables financieras por Bustos-Chiliquinga et al. (2026) donde cada observación se representa mediante una tripleta (T, I, F) y se agrega posteriormente en una puntuación escalar, se propone una adaptación de este esquema al problema de clasificación supervisada de riesgo crediticio a nivel de solicitante individual.

Verdad (T). Para cada solicitante *i*, el componente de verdad se define como la probabilidad predicha de "buen riesgo" otorgada por el modelo de mejor desempeño individual (determinado en la Sección 4):

$T_i = P(\text{buen riesgo} \mid x_i)$

Indeterminación (I). A diferencia de los enfoques puramente probabilísticos (Xu et al., 2025), que derivan la incertidumbre exclusivamente de la dispersión de la salida del modelo, se propone construir la indeterminación a partir de dos fuentes independientes, análogas en su lógica de ponderación a la volatilidad temporal y la desviación sectorial empleadas por Bustos-Chiliquinga et al. (2026) en el dominio financiero longitudinal:

1. *Indeterminación de datos ($I_{dato,i}$)*: proporción ponderada de variables del registro del solicitante *i* marcadas como categorías ausentes o ambiguas en el paso 2 del preprocesamiento (Sección 3.2), normalizada al intervalo [0, 1].

2. *Indeterminación de modelo ($I_{modelo,i}$)*: grado de desacuerdo entre los tres modelos de la Sección 3.3, calculado como la entropía (o, alternativamente, la desviación estándar) de las tres probabilidades predichas para el solicitante *i*, normalizada al intervalo [0, 1].

Ambas fuentes se combinan mediante una ponderación fijada a priori como parámetro de modelado (sujeta a análisis de sensibilidad, siguiendo la práctica de Bustos-Chiliquinga et al., 2026):

$I_i = w_1 \cdot I_{dato,i} + w_2 \cdot I_{modelo,i}$, con $w_1 + w_2 = 1$

Falsedad (F). Consistente con la implementación de Bustos-Chiliquinga et al. (2026), donde la falsedad se define como complemento determinístico de la verdad una vez que ambos componentes han sido restringidos a [0, 1], se adopta:

$F_i = 1 - T_i$

Puntuación neutrosófica (S). Los tres componentes se agregan en una puntuación escalar mediante una función de deneutrosofización ponderada:

$S_i = \alpha \cdot T_i + \beta \cdot (1 - I_i) + \gamma \cdot (1 - F_i)$, con $\alpha + \beta + \gamma = 1$

Dado que $F_i = 1 - T_i$, esta expresión es algebraicamente equivalente a $S_i = (\alpha + \gamma) \cdot T_i + \beta \cdot (1 - I_i)$. Los pesos ($\alpha$, β, γ) se fijan a priori y su influencia sobre los resultados se evalúa mediante un análisis de sensibilidad (Sección 4).

Uso de la indeterminación en la regla de decisión. Además de la puntuación S_i, se define un umbral de indeterminación τ que permite distinguir dos regímenes de decisión: para solicitantes con $I_i \leq \tau$, la clasificación final se basa directamente en S_i; para solicitantes con $I_i > \tau$, el caso se marca como de "alta indeterminación" y se recomienda revisión manual adicional, en lugar de forzar una clasificación automática de baja confiabilidad.

3.5 Métricas de evaluación

Se emplean dos grupos de métricas:

Métricas de clasificación estándar, calculadas sobre las predicciones binarias de cada modelo de la Sección 3.3: exactitud (*accuracy*), precisión, sensibilidad (*recall*), puntuación F1 y área bajo la curva ROC (AUC-ROC). Dado el desbalance de clases (70/30), se reporta con especial atención el F1 y el AUC-ROC por sobre la exactitud simple, siguiendo la recomendación estándar en la literatura de *credit scoring* desbalanceado (Chang et al., 2024).

Métricas específicas de incertidumbre, orientadas a validar si el componente neutrosófico aporta información útil más allá de la clasificación binaria:

1. *Correlación indeterminación-error*: correlación entre I_i y la ocurrencia de un error de clasificación, para verificar si los casos de mayor indeterminación concentran efectivamente una mayor tasa de error.
2. *Curva de cobertura-exactitud* (*accuracy-coverage curve*): exactitud del modelo calculada únicamente sobre el subconjunto de solicitantes con $I_i \leq \tau$ (los casos "ciertos"), en función de distintos valores del umbral τ, en la línea de los esquemas de clasificación con opción de rechazo propuestos por Xu et al. (2025).
3. *Indeterminación media por clase real y por clase predicha*, para caracterizar si los "mal riesgo" y los casos mal clasificados presentan sistemáticamente mayor indeterminación que los casos correctamente clasificados.

4. Resultados

Los resultados que se presentan a continuación corresponden a la ejecución completa del pipeline descrito en la Sección 3 sobre el conjunto de prueba (300 solicitantes: 90 de "mal riesgo" y 210 de "buen riesgo"), con los tres modelos entrenados sobre 700 solicitantes mediante validación cruzada estratificada de 5 particiones y búsqueda en cuadrícula de hiperparámetros.

4.1 Resultados de los modelos base

La Tabla 2 resume el desempeño de los tres modelos sobre el conjunto de prueba.

Modelo	Accuracy	Precisión	Recall	F1	AUC-ROC	Hiperparámetros seleccionados
Regresión Logística (LR)	0.663	0.819	0.667	0.735	0.728	C = 1
Random Forest (RF)	0.703	0.787	0.790	0.789	**0.771**	n_estimators = 200, max_depth = 8
XGBoost	0.687	0.805	0.729	0.765	0.726	n_estimators = 200, max_depth = 3, learning_rate = 0.1

Tabla 2. Comparación de desempeño de los modelos de aprendizaje automático (conjunto de prueba, n = 300).

Random Forest obtiene el mejor AUC-ROC (0.771) y el mejor F1 (0.789), por lo que se selecciona como modelo base para la construcción del componente de verdad (T) del esquema neutrosófico. Su matriz de confusión es la siguiente:

	Predicho: mal riesgo	Predicho: buen riesgo
Real: mal riesgo	45	45
Real: buen riesgo	44	166

Tabla 3. Matriz de confusión de Random Forest sobre el conjunto de prueba.

Es de notar que, pese a tener el mejor F1 y AUC entre los tres modelos, Random Forest clasifica erróneamente a la mitad de los solicitantes de "mal riesgo" del conjunto de prueba (45 de 90) como "buen riesgo" el tipo de error más costoso desde la perspectiva del prestamista, lo que refuerza la pertinencia de un mecanismo adicional (el componente de indeterminación) capaz de señalar qué predicciones ameritan menor confianza.

4.2 Resultados del tratamiento neutrosófico

Aplicando la formalización de la Sección 3.4 sobre las salidas de los tres modelos, se obtienen los siguientes valores promedio del conjunto de prueba, tal como muestra la tabla 4:

Componente	Media
Verdad (T)	0.616
Indeterminación (I)	0.304
Falsedad (F)	0.384
Puntuación neutrosófica (S)	0.640

Tabla 4. Estadísticos descriptivos de los componentes neutrosóficos (n = 300).

Al desagregar la indeterminación en sus dos fuentes constitutivas (Sección 3.4), se observa una asimetría relevante: la indeterminación de datos (I_{dato}) construida a partir de las categorías "sin cuenta corriente", "sin cuentas de ahorro conocidas", "sin propiedad conocida" y "sin historial crediticio previo" presenta una media de 0.270, mientras que la indeterminación de modelo (I_{modelo}) el desacuerdo entre LR, RF y XGBoost presenta una media de 0.338, ligeramente superior. En el conjunto de prueba, 124 solicitantes no presentan ninguna categoría marcada como indeterminada, 117 presentan una, 51 presentan dos y solo 8 presentan tres.

4.3 Análisis de casos con alta indeterminación

Fijando el umbral de indeterminación en $\tau = 0.5$ (Sección 3.4), el 13.7% de los solicitantes del conjunto de prueba (41 de 300) queda clasificado como de "alta indeterminación" y sería candidato a revisión manual antes de una decisión automática.

Un hallazgo que merece discusión explícita (retomado en la Sección 5) es que, contrario a lo hipotetizado inicialmente, la exactitud del modelo sobre el subconjunto de "alta indeterminación" ($I > 0.5$) resultó superior (0.829) a la exactitud sobre el subconjunto "cierto" ($I \leq 0.5$, exactitud = 0.683). La correlación puntual-biserial entre I y la ocurrencia de un error de clasificación fue negativa y estadísticamente significativa ($r = -0.156$, $p = 0.007$), es decir, en sentido contrario al esperado bajo la hipótesis de que una mayor indeterminación debería asociarse con mayor probabilidad de error.

El desglose por componente aclara el origen de este resultado: la indeterminación de modelo (I_{modelo}) sí se correlaciona en la dirección esperada, aunque de forma débil y no significativa ($r = 0.081$, $p = 0.163$) a mayor desacuerdo entre los tres algoritmos, ligeramente mayor probabilidad de error, mientras que la indeterminación de datos (I_{dato}) se correlaciona negativa y significativamente con el error ($r = -0.263$, $p < 0.001$). Esto se debe, muy probablemente, a que la categoría "sin cuenta corriente" (A14), marcada como indeterminada en el preprocesamiento, está fuertemente asociada en este dataset específico con la clase "buen riesgo" una particularidad ya documentada informalmente en la literatura aplicada sobre el German Credit Data y por tanto aporta señal predictiva real en lugar de mera ausencia de información.

La Tabla 5 presenta la curva de cobertura-exactitud completa.

τ	Cobertura	Exactitud en el subconjunto
0.1	0.107	0.625
0.2	0.313	0.585
0.3	0.540	0.654

0.4	0.727	0.688
0.5	0.863	0.683
0.6	0.940	0.691
0.7	0.977	0.696
0.8	0.983	0.698
0.9	0.993	0.701
1.0	1.000	0.703

Tabla 5. Curva de cobertura-exactitud según el umbral de indeterminación τ.

Finalmente, la indeterminación media difiere entre clases: los solicitantes reales de "buen riesgo" presentan una I media de 0.317, frente a 0.273 en los de "mal riesgo"; y los casos correctamente clasificados presentan una I media (0.322) mayor que los casos con error de clasificación (0.261). Este patrón es consistente con el hallazgo anterior y confirma que, en la configuración actual del componente I_{dato}, la indeterminación construida no está capturando incertidumbre predictiva genuina para este dataset, sino en parte una señal de clase disfrazada de ausencia de información.

5. Discusión

5.1 El desempeño predictivo en perspectiva

El desempeño de los tres modelos (AUC entre 0.726 y 0.771) se ubica en un rango consistente con lo reportado en la literatura reciente para el German Credit Data, un dataset reconocido por su tamaño reducido (1000 instancias) y su techo de desempeño relativamente bajo en comparación con conjuntos de datos crediticios de mayor escala (Chang et al., 2024; Noriega et al., 2023). Que Random Forest supere a XGBoost en este caso contrario a lo observado por Chang et al. (2024) en un dataset de tarjetas de crédito de mayor tamaño es coherente con la literatura que señala que los métodos de *boosting* tienden a requerir volúmenes de datos más amplios para explotar plenamente su capacidad de modelado; con solo 700 observaciones de entrenamiento, la ventaja teórica de XGBoost sobre RF se diluye. Este resultado refuerza una idea ya señalada por Montevechi et al. (2024) la superioridad de un algoritmo sobre otro en *credit scoring* depende fuertemente del contexto de datos, y no puede asumirse de forma genérica.

Más relevante para los objetivos de este estudio es que, incluso el mejor modelo, clasifica erróneamente al 50% de los solicitantes de "mal riesgo" reales como "buen riesgo" (Tabla 3). Esta cifra por sí sola justifica la pregunta de fondo del artículo: si el modelo se equivoca con esa frecuencia en la clase de mayor costo para el prestamista, es posible identificar, mediante algún mecanismo adicional, ¿cuáles de esas predicciones merecen menor confianza? Es exactamente aquí donde el aporte neutrosófico muestra su valor.

5.2 La paradoja de la indeterminación de datos

El hallazgo más significativo de este estudio y el que amerita la discusión más detenida es que el componente de indeterminación (I), tal como fue formalizado en la Sección 3.4, no cumplió la función para la que fue diseñado. En lugar de identificar los casos de mayor riesgo de error, el subconjunto de "alta indeterminación" ($I > 0.5$) presentó una exactitud *superior* (0.829) a la del subconjunto "cierto" (0.683), y la correlación global entre I y el error fue negativa ($r = -0.156$, $p = 0.007$).

El desglose por componente (Sección 4.3) permite descartar que se trate de un error de implementación del marco neutrosófico como tal: la indeterminación de modelo (I_{modelo}), construida a partir del desacuerdo entre los tres algoritmos de forma análoga a como Xu et al. (2025) derivan incertidumbre de la dispersión de las predicciones, sí se orienta en la dirección teóricamente esperada ($r = 0.081$), aunque de forma débil y no significativa en este tamaño muestral. El problema recae específicamente en la indeterminación de datos (I_{dato}), con una correlación negativa y fuertemente significativa ($r = -0.263$, $p < 0.001$).

La explicación más plausible es de naturaleza sustantiva y no metodológica: la categoría "sin cuenta corriente" (atributo A14), que en la Sección 3.2 se marcó como indeterminada bajo el supuesto de que representa ausencia de información, es en realidad en la idiosincrasia particular del German Credit Data una categoría con fuerte poder predictivo hacia la clase "buen riesgo". Este comportamiento, documentado de forma recurrente en análisis exploratorios aplicados a este dataset, probablemente refleja un sesgo de selección en la recolección original de los datos (por ejemplo, si los solicitantes sin cuenta corriente en el banco evaluado correspondían predominantemente a un segmento de clientes ya preseleccionado como de bajo riesgo). En otras palabras: lo que en el diseño teórico del componente neutrosófico se interpretó como "ausencia de información" resultó ser, empíricamente, información válida y fuertemente informativa.

Este hallazgo tiene una implicación metodológica de fondo que trasciende el caso específico del German Credit Data: la construcción de la indeterminación de datos no puede basarse únicamente en un criterio semántico a priori. Es necesario, como paso adicional al preprocesamiento propuesto en la Sección 3.2, validar empíricamente mediante un análisis exploratorio previo, por ejemplo la asociación bivariada entre cada categoría candidata y la variable objetivo que las categorías seleccionadas como "indeterminadas" no posean, en el dataset específico bajo estudio, una capacidad discriminante propia que contradiga su tratamiento como ausencia de información. Esta es una diferencia relevante respecto del enfoque de Bustos-Chiliquinga et al. (2026), donde la indeterminación se deriva de la volatilidad temporal de series financieras una fuente más directamente vinculada a la incertidumbre real de la medición y no de categorías declarativas cuya relación con la variable objetivo puede variar según el contexto institucional de origen de los datos.

Resulta particularmente ilustrativo contrastar este hallazgo con el de Al-Masri (2025), quien propone un esquema arquitectónicamente análogo al de este estudio un ensamble de Random Forest, XGBoost y regresión logística cuyas salidas se descomponen en una tripleta neutrosófica (T, I, F) para habilitar la abstención de decisiones en un sistema de detección de intrusiones en IoT. En ese trabajo, la indeterminación (derivada de la entropía de las probabilidades del ensamble, es decir, un análogo directo de nuestro I_{modelo}) sí se comportó exactamente en la dirección esperada, las muestras mal clasificadas presentaron una indeterminación media sustancialmente mayor ($I = 0.622$) que las correctamente clasificadas ($I = 0.241$). La coincidencia de arquitectura entre ambos estudios permite aislar la variable relevante, cuando la indeterminación se construye únicamente a partir del desacuerdo entre modelos como en Al-Masri (2025) y como en nuestro propio I_{modelo}, se comporta de forma consistente con la teoría; el comportamiento anómalo únicamente aparece al introducir la indeterminación de datos basada en categorías declarativas no validadas empíricamente. Esto refuerza que la paradoja identificada en la Sección 5.2 es atribuible a la fuente específica de indeterminación empleada, y no al marco neutrosófico en general.

5.3 Implicaciones para el diseño del componente neutrosófico

Este resultado no invalida el enfoque neutrosófico en sí, pero sí exige matizar su aporte. La indeterminación de modelo (I_{modelo}) el desacuerdo entre algoritmos se comportó de manera consistente con la teoría, lo cual sugiere que el valor del marco neutrosófico en este dominio puede residir más en capturar la incertidumbre *epistémica* del propio proceso de modelado (cuán de acuerdo están los distintos algoritmos entrenados sobre los mismos datos) que en capturar la incertidumbre *aleatoria* atribuible a huecos declarativos en el registro de un solicitante individual, al menos sin una validación previa de estos últimos.

Esto es consistente, en un sentido más amplio, con la observación de Xu et al. (2025) de que los esquemas de clasificación con opción de rechazo son más confiables cuando la incertidumbre se deriva directamente del comportamiento del modelo (dispersión de probabilidades, pérdida esperada) que son señales indirectas sobre la calidad del dato de entrada. El aporte específico de la neutrosofía frente a estos enfoques puramente probabilísticos seguiría residiendo en la posibilidad de mantener T, I y F como dimensiones independientes y auditable por separado una ventaja de transparencia regulatoria señalada por Weber et al. (2024) pero su utilidad práctica depende críticamente de que cada fuente de indeterminación esté bien especificada.

5.4 Limitaciones del estudio

Se reconocen las siguientes limitaciones:

1. **Tamaño y representatividad del dataset.** El German Credit Data (1000 instancias, recolectadas en Alemania en la década de 1990) es un conjunto de datos reducido y desactualizado respecto a los patrones de crédito contemporáneos, lo cual limita la generalización directa de los resultados a contextos crediticios actuales.
2. **Especificidad de la paradoja del I_{dato}.** El hallazgo de la Sección 5.2 fue identificado en este dataset particular; no puede descartarse que la construcción de la indeterminación de datos funcione de forma más consistente con la teoría en otros conjuntos de datos donde las categorías "desconocido/sin X" reflejen efectivamente ausencia de información y no un artefacto de muestreo.
3. **Pesos fijados a priori.** Tanto los pesos de combinación de la indeterminación (w_1, w_2) como los de la deneutrosofización (α, β, γ) se fijaron de forma a priori (0.5/0.5 y 0.5/0.3/0.2, respectivamente) y no fueron optimizados sobre los datos; un análisis de sensibilidad más exhaustivo sobre estos parámetros queda pendiente.
4. **Umbral de indeterminación único.** El umbral $\tau = 0.5$ utilizado para distinguir casos "ciertos" de casos de "alta indeterminación" es también una elección a priori; la curva de cobertura-exactitud (Tabla 5) sugiere que el comportamiento del sistema es sensible a esta elección, y una calibración específica por institución sería necesaria en una aplicación real.
5. **Alcance del componente de indeterminación de modelo.** Al basarse en solo tres algoritmos (LR, RF, XGBoost), la indeterminación de modelo puede estar subestimada; un ensamble con más algoritmos heterogéneos podría producir una señal de desacuerdo más robusta.

5.5 Implicaciones prácticas

Para instituciones financieras, el hallazgo de la Sección 5.2 tiene una lectura práctica directa: antes de implementar cualquier esquema de indeterminación basado en categorías "desconocidas" del historial crediticio, es indispensable validar empíricamente, con los datos propios de la institución, si dichas categorías realmente representan falta de información o si como ocurrió aquí constituyen en sí mismas una señal de riesgo (positiva o negativa). Ignorar esta validación podría llevar a un sistema de alerta de indeterminación que, en la práctica, dirija la revisión manual precisamente hacia los casos que el modelo clasifica con mayor acierto, y no hacia los que efectivamente lo requieren.

6. Conclusiones

El presente estudio propuso y evaluó un esquema híbrido de aprendizaje automático y análisis neutrosófico para la predicción del riesgo crediticio, utilizando el conjunto de datos *German Credit Data* como caso de estudio. Se entrenaron y compararon tres modelos (regresión logística, Random Forest y XGBoost), y sobre sus salidas se formalizó una tripleta neutrosófica (T, I, F) y una puntuación agregada (S), con el objetivo de cuantificar no solo la clasificación de riesgo, sino también el grado de indeterminación asociado a cada predicción individual.

Random Forest fue el modelo con mejor desempeño (AUC-ROC = 0.771, F1 = 0.789), aunque incluso este modelo clasificó erróneamente a la mitad de los solicitantes de "mal riesgo" reales del conjunto de prueba, lo que confirma la necesidad de mecanismos complementarios de identificación de predicciones poco confiables.

El hallazgo central del estudio fue de naturaleza inesperada: mientras que la indeterminación derivada del desacuerdo entre modelos (I_{modelo}) se comportó en la dirección teóricamente prevista mayor desacuerdo, mayor probabilidad de error, la indeterminación derivada de categorías declarativas ausentes en los datos de entrada (I_{dato}) mostró una correlación negativa y significativa con el error de clasificación ($r = -0.156$, $p = 0.007$), revelando que, en este dataset particular, una categoría marcada a priori como "ausencia de información" (la falta de cuenta corriente) constituye en realidad una señal predictiva válida hacia la clase de menor riesgo.

Esta paradoja constituye el aporte metodológico más relevante del artículo ya que evidencia que la formalización de la indeterminación de datos en esquemas neutrosóficos no puede basarse únicamente en criterios semánticos a priori, sino que requiere una validación empírica previa de la relación entre cada categoría candidata y la variable objetivo en el contexto específico de los datos utilizados. Sin esta validación, un sistema de alerta de indeterminación podría, paradójicamente, señalar como más confiables los casos que en realidad concentran mayor riesgo de error, y viceversa.

Como líneas futuras de investigación se plantea: (a) replicar el esquema propuesto sobre conjuntos de datos crediticios de mayor tamaño y vigencia, idealmente provenientes de instituciones financieras ecuatorianas, para evaluar si la paradoja identificada es específica del German Credit Data o se generaliza a otros contextos; (b) incorporar un paso de validación empírica automatizado que descarte, antes de su uso, categorías candidatas a "indeterminadas" cuya asociación con la variable objetivo sea estadísticamente significativa; (c) realizar un análisis de sensibilidad sistemático sobre los pesos de combinación (w_1, w_2) y de deneutrosofización (α, β, γ); y (d) explorar ensambles de mayor tamaño y heterogeneidad algorítmica para robustecer el componente de indeterminación de modelo, que en este estudio mostró un comportamiento más alineado con la teoría neutrosófica que su contraparte basada en datos.

Referencias

[1]. Al-Hijjawi, S., Ahmad, A. G., & Alkhazaleh, S. (2024). Effective neutrosophic soft expert set and its application. International Journal of Neutrosophic Science, 23(1), 27–50. https://doi.org/10.54216/IJNS.230103

[2]. Al-Masri, E. (2025). Deciding when not to decide: Indeterminacy-aware intrusion detection with NeutroSENSE (arXiv:2507.00003). arXiv. https://arxiv.org/abs/2507.00003

[3]. Atanassov, K. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

[4]. Bustos-Chiliquinga, G. V., Galindo-Villardón, P., Vicente-Galindo, P., & Cornejo Gaete, C. (2026). Modeling financial structural dynamics under uncertainty: A neutrosophic-score dual STATIS approach for cooperative financial systems. Mathematics, 14(18), Artículo 3278. https://doi.org/10.3390/math14183278

[5]. Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. Risks, 12(11), Artículo 174. https://doi.org/10.3390/risks12110174

[6]. Hezam, I. M., Mishra, A. R., Rani, P., Saha, A., Smarandache, F., & Pamucar, D. (2023). An integrated decision support framework using single-valued neutrosophic-MASWIP-COPRAS for sustainability assessment of bioenergy production technologies. Expert Systems with Applications, 211, Artículo 118674. https://doi.org/10.1016/j.eswa.2022.118674

[7]. Mishra, A. R., Saha, A., Rani, P., Hezam, I. M., Shrivastava, R., & Smarandache, F. (2022). An integrated decision support framework using single-valued-MEREC-MULTIMOORA for low carbon tourism strategy assessment. IEEE Access, 10, 24411–24432. https://doi.org/10.1109/ACCESS.2022.3155171

[8]. Montevechi, A. A., Miranda, R. de C., Medeiros, A. L., & Montevechi, J. A. B. (2024). Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art. Engineering Applications of Artificial Intelligence, 138, Artículo 109082. https://doi.org/10.1016/j.engappai.2024.109082

[9]. Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. Data, 8(11), Artículo 169. https://doi.org/10.3390/data8110169

[10]. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. Neural Computing and Applications, 34(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

[11]. Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

[12]. Voskoglou, M. Gr. (2023). An application of neutrosophic sets to decision making (arXiv:2306.01746). arXiv. https://arxiv.org/abs/2306.01746

[13]. Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. Management Review Quarterly, 74(2), 867–907. https://doi.org/10.1007/s11301-023-00320-0

[14]. Xu, Y., Kou, G., & Ergu, D. (2025). Profit-based uncertainty estimation with application to credit scoring. European Journal of Operational Research, 325(2), 303–316. https://doi.org/10.1016/j.ejor.2025.03.007

[15]. Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(5), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X