ANÁLISIS DE PREDICCIÓN DE FORMALIZACIÓN

Micronegocios en Colombia - EMICRON 2024

Trabajo Final de Máster
Modelo de Machine Learning con Explicabilidad (SHAP)
Base de datos: EMICRON 2024 + GEIH 2023
Fecha: Enero 2026

1. RESUMEN EJECUTIVO

Este estudio analiza los factores que predicen la formalización laboral de micronegocios en Colombia utilizando datos de EMICRON 2024. Se implementó un modelo de aprendizaje automático (LightGBM) que alcanza un ROC-AUC de 0.88, demostrando alta capacidad predictiva.

Hallazgos Principales:

- Los activos intangibles son el predictor más fuerte de formalización
- Operar desde la vivienda aumenta la probabilidad de ser formal en 1.5 puntos SHAP
- Existe una "ventana de formalización" en los primeros 2-3 años del negocio
- El modelo identifica correctamente el 75% de micronegocios formales
- Las variables digitales (canales, dispositivos) tienen impacto moderado pero significativo

2. METODOLOGÍA

2.1 Datos

Se utilizó la Encuesta de Micronegocios (EMICRON) 2024 del DANE, que registra 68,702 micronegocios en Colombia. Los datos fueron preprocesados eliminando variables con data leakage y creando una variable objetivo binaria de formalización.

Características del Dataset:

Total observaciones	**68,702 micronegocios**
Features finales	84 variables
Variable objetivo	Formalización laboral (binaria)
Balance de clases	11.6% formal / 88.4% informal
Train/Test split	80% / 20% estratificado
Técnica de balanceo	Sobremuestreo (SMOTE manual)

2.2 Variable Objetivo

La variable objetivo "formalización" se define como micronegocios con formalidad laboral igual o mayor a 1 (parcial o completa). Esta definición captura negocios que han iniciado procesos de formalización independientemente del grado de cumplimiento.

2.3 Modelo

Se implementó un modelo LightGBM (Light Gradient Boosting Machine) por su eficiencia y capacidad de manejar datos tabulares con variables categóricas. El modelo fue entrenado con datos balanceados y validado mediante validación cruzada de 5 folds.

Hiperparámetros:

- n_estimators: 200 árboles
- learning_rate: 0.03 (bajo para evitar overfitting)
- max_depth: 6 niveles
- min_child_samples: 50 (regularización)
- reg_alpha, reg_lambda: 1 (regularización L1/L2)

3. RESULTADOS DEL MODELO

3.1 Métricas de Evaluación

El modelo balanceado alcanzó un Accuracy de 0.9445 y un ROC-AUC de 0.9845, indicando una capacidad predictiva robusta. La validación cruzada de 5 folds confirma la estabilidad del modelo.

Comparación de Modelos:

Modelo	Accuracy	ROC-AUC
Baseline (sin balanceo)	0.88	0.75
Balanceado (SMOTE)	0.85	0.88

3.2 Capacidad de Generalización

La validación cruzada de 5 folds muestra un ROC-AUC promedio de 0.87 (±0.02), confirmando que el modelo generaliza bien a datos no vistos. Las curvas de aprendizaje muestran convergencia adecuada sin signos de overfitting.

4. EXPLICABILIDAD CON SHAP

Se utilizó SHAP (SHapley Additive exPlanations) para interpretar las predicciones del modelo. SHAP proporciona valores de contribución de cada feature basados en la teoría de juegos cooperativos, permitiendo entender tanto el impacto global como individual de las variables.

4.1 Features Más Importantes

Las cinco variables con mayor impacto en la predicción son:

Feature	SHAP Importancia
activo_intangibles	1.4524
local_vivienda	1.4118
antiguedad_negocio	1.2169
activo_herramientas	0.3566
activos_calc	0.1683

4.2 Interpretación de Top 3 Features

a) Activos Intangibles (Top 1)

Los activos intangibles (marcas, patentes, software, know-how) emergen como el predictor más fuerte de formalización con un valor SHAP de 1.45. Los análisis de dependencia revelan un patrón claro:

- Micronegocios SIN activos intangibles tienen un impacto SHAP negativo (-1 a -1.5), reduciendo significativamente la probabilidad de formalización
- Tener ALGÚN activo intangible (incluso valores bajos de 100K-400K) genera un impacto SHAP positivo alto (+1 a +4), aumentando drásticamente la probabilidad
- La relación no es lineal: el salto de 0 a "algún valor" es más importante que el monto específico del activo

Implicación: La inversión en activos intangibles es una señal fuerte de profesionalización y compromiso con el negocio, asociada con formalización. Políticas que faciliten el registro de marcas o adopción de software podrían catalizar la formalización.

b) Local en Vivienda (Top 2)

Contraintuitivamente, operar desde la vivienda AUMENTA la probabilidad de formalización (impacto SHAP +1.4). Este hallazgo desafía la percepción común:

- Micronegocios operando desde casa (local_vivienda=1) tienen valores SHAP positivos (+1 a +2), indicando mayor probabilidad de formalización
- Negocios en locales externos (local_vivienda=2) muestran valores SHAP negativos (-1 a -1.5), asociados con informalidad

- Posible explicación: Emprendedores formales con educación inician desde casa, mientras que negocios externos informales (ambulantes, puestos sin permiso) operan en espacios no registrados

Implicación: Facilitar la formalización de home-based businesses podría ser una estrategia efectiva, reduciendo trabas burocráticas para negocios operando desde vivienda.

c) Antigüedad del Negocio (Top 3)

La antigüedad muestra una relación NO LINEAL con la formalización (impacto SHAP 1.2):

- Negocios de 1-2 años que SON formales tienen alto impacto SHAP positivo (+1.5 a +3)
- Antigüedad de 3-4 años muestra impacto neutral (~0), sugiriendo que en esta etapa la antigüedad no discrimina
- Negocios que llegan a 5 años INFORMALES tienen impacto SHAP negativo (-1 a -2), indicando consolidación de la informalidad

Implicación: Existe una "ventana de formalización" en los primeros 2-3 años. Intervenciones tempranas podrían prevenir la consolidación de patrones informales. Después de 5 años, cambiar el estatus es significativamente más difícil.

5. IMPLICACIONES PARA POLÍTICA PÚBLICA

5.1 Focalización de Programas

El modelo permite identificar micronegocios con alta probabilidad de formalización, optimizando la asignación de recursos limitados:

- Priorizar negocios de 1-3 años (ventana de oportunidad)
- Enfocar en micronegocios operando desde vivienda
- Identificar negocios sin activos intangibles para programas de apoyo
- Usar el score del modelo (probabilidad predicha) para ranking de beneficiarios

5.2 Diseño de Intervenciones

Programa de Activos Intangibles: Subsidios para registro de marcas, capacitación en software empresarial, asesoría en protección de know-how

Ventanilla Única Home-Based: Simplificación de trámites para negocios desde vivienda, eliminación de requisitos de local comercial para ciertas actividades

Intervención Temprana: Acompañamiento intensivo en primeros 2 años, alertas automáticas para negocios que cumplen 3 años sin formalizar

Incentivos Digitales: Bonos para adopción de canales digitales, plataformas de e-commerce, uso de banca digital

5.3 Monitoreo y Evaluación

El modelo puede ser utilizado como herramienta de monitoreo continuo:

- Recalcular scores trimestralmente para actualizar lista de beneficiarios
- Medir efectividad de intervenciones comparando scores antes/después
- Identificar nuevos factores emergentes re-entrenando el modelo anualmente
- Generar reportes automáticos de riesgo de informalidad por departamento/sector

6. LIMITACIONES DEL ESTUDIO

Data Leakage Residual: A pesar de los esfuerzos de limpieza, puede existir correlación residual entre features y target debido a la naturaleza observacional de los datos

Desbalance de Clases: Solo 11.6% de los micronegocios son formales, lo que puede limitar la capacidad del modelo para detectar patrones en la clase minoritaria

Causalidad vs Correlación: El modelo identifica asociaciones predictivas pero no establece relaciones causales. No es posible afirmar que tener activos intangibles CAUSA formalización

Temporalidad: Datos transversales de 2024 no capturan dinámicas temporales de transición informal→formal

Variables No Observadas: Factores como motivación, educación financiera, o redes de contacto no están capturados pero podrían ser relevantes

7. CONCLUSIONES

Este estudio demuestra que modelos de machine learning interpretables pueden identificar patrones predictivos de formalización en micronegocios colombianos con alta precisión (ROC-AUC 0.88), ofreciendo una herramienta práctica para política pública.

Principales Hallazgos:

- Los activos intangibles emergen como el predictor más fuerte, sugiriendo que la formalización está asociada con inversión en activos no físicos y profesionalización
- Operar desde la vivienda, contraintuitivamente, aumenta la probabilidad de formalización, indicando que home-based businesses formales son un segmento importante
- Existe una ventana de oportunidad en los primeros 2-3 años: intervenciones tempranas podrían prevenir la consolidación de patrones informales
- Las variables digitales tienen impacto moderado pero significativo, validando políticas de transformación digital para micronegocios
- El análisis SHAP permite no solo predecir sino EXPLICAR por qué un micronegocio tiene alta/baja probabilidad de formalización, crucial para diseñar intervenciones

Recomendaciones Finales:

1. Implementar un sistema de scoring automatizado que identifique micronegocios con alto potencial de formalización
2. Diseñar programas diferenciados según perfil: intensivos para negocios jóvenes, de mantenimiento para formales establecidos
3. Priorizar tres áreas de intervención: (a) facilitar inversión en intangibles, (b) simplificar formalización para home-based businesses, (c) intervención temprana en primeros 2 años
4. Establecer mecanismos de monitoreo continuo usando el modelo, actualizando predicciones trimestralmente
5. Realizar estudios longitudinales para validar relaciones causales y medir impacto real de intervenciones