Facultad de Ingeniería en
Ciencias de la Computación y Telecomunicaciones
U.A.G.R.M.

Somos ingeniería!

Proyecto de minería de datos(weka)

MATERIA: SISTEMAS PARA EL SOPORTE A LA TOMA DE DECISIONES – INF 432 – SA

DOCENTE: ING. PEINADO MIGUEL JESUS

ESTUDIANTE: García Caballero José Diego – 221044523

ESTUDIANTE: Vásquez Flórez Josué – 220154848

ESTUDIANTE: García Caballero David – 217058795

1. Selección del conjunto de datos:

- **¿Cuál es el objetivo del análisis?**

El objetivo principal es desarrollar un modelo predictivo utilizando el algoritmo J48 para evaluar la solvencia crediticia de los clientes. Se busca clasificar a los clientes como "buenos" o "malos" en función de sus características, con el fin de automatizar y mejorar el proceso de concesión de créditos, minimizando el riesgo de impago.

- **¿Qué tipo de conocimiento se busca?**

Se busca obtener dos tipos de conocimiento:

Predictivo: Un modelo preciso que clasifique correctamente a los clientes como "buenos" o "malos" para un crédito. Esto permitirá a la entidad financiera tomar decisiones informadas sobre la aprobación o rechazo de solicitudes de crédito.

Descriptivo: Comprender las variables o factores que más influyen en la solvencia crediticia de un cliente. El algoritmo J48 genera un árbol de decisión que se puede analizar para identificar las variables más importantes y las reglas de clasificación. Esto puede proporcionar información valiosa para la entidad financiera, como por ejemplo, qué características de los clientes son más relevantes al evaluar el riesgo crediticio.

- **¿Qué datos son necesarios y dónde se pueden obtener?**

Los datos necesarios se pueden agrupar en tres categorías:

Demográficos: Edad, estado civil, número de dependientes, nacionalidad (extranjero o no).

Socioeconómicos: Situación laboral (tiempo de empleo), nivel de ingresos, propiedad de vivienda (propia, alquilada), tipo de trabajo (cualificado, no cualificado).

Financieros: Estado de la cuenta corriente, historial de crédito, monto del crédito solicitado, estado de los ahorros, compromiso de cuotas, planes de pago adicionales, número de créditos existentes.

Estos datos se obtienen de diversas fuentes internas y externas:

Bases de datos de clientes (CRM), historiales de transacciones, solicitudes de crédito.

2. Preprocesamiento de datos:

- **¿Cómo se manejarán los datos faltantes o erróneos?**

Aunque en este conjunto de datos no hay valores faltantes, en general, se deben manejar con estrategias como:

Utilizar algoritmos que manejen valores faltantes: Algunos algoritmos, como J48, pueden manejar valores faltantes hasta cierto punto.

Los datos erróneos se deben identificar mediante análisis exploratorio de datos y técnicas de validación. Las estrategias para manejarlos incluyen:

Corrección: Si es posible, corregir los errores basándose en otras fuentes de información.

- **¿Qué transformaciones se aplicarán a los datos?**
    - **Discretización:** Convertir variables numéricas en categóricas. En este caso, se podrían discretizar variables como `age` (edad) y `credit_amount` (monto del crédito) en rangos. Esto puede mejorar la interpretabilidad del árbol de decisión generado por J48.
    - **Codificación de variables categóricas:** Transformar variables categóricas en numéricas utilizando técnicas como la codificación one-hot.

J48 puede trabajar con variables categóricas, pero algunos algoritmos requieren variables numéricas.

- **Escalado:** Estandarizar o normalizar las variables numéricas para que tengan la misma escala. Esto puede ser útil para algunos algoritmos, pero no es esencial para J48.

- **¿Cómo se reducirá la dimensionalidad de los datos?**

Se pueden utilizar técnicas de selección de atributos para identificar las variables más relevantes para el modelo J48:

Evaluación de la información: Utilizar métricas como la ganancia de información o la razón de ganancia para seleccionar las variables que más información aportan para la clasificación.Evaluar la importancia de las variables que se tienen para determinar las importantes e imprecindibles.

3. Minería de datos:

- **¿Qué técnica de minería de datos es la más adecuada para el problema?**

La clasificación con J48 es la técnica más adecuada para este problema porque se ajusta a la naturaleza del problema (predicción de una variable categórica) y ofrece ventajas como la facilidad de interpretación, el manejo de diferentes tipos de datos, la robustez y la eficiencia computacional.

- **¿Cómo se evaluará el rendimiento del modelo?**

El rendimiento del modelo J48 se evaluará utilizando las siguientes métricas:

Precisión: Proporción de instancias clasificadas correctamente.

Exactitud (precision): Proporción de instancias clasificadas como "malas" que realmente son "malas".

La técnica de minería de datos más adecuada para este problema es la clasificación, específicamente utilizando el algoritmo J48, debido a que el objetivo del análisis es predecir la variable categórica "class", que indica si un cliente es "bueno" o "malo" para un crédito. La clasificación con J48 es la más adecuada porque se ajusta a la naturaleza del problema, que busca predecir una variable categórica y ofrece ventajas como la facilidad de interpretación del modelo, la capacidad de manejar diferentes tipos de datos, la robustez ante el ruido en los datos y la eficiencia computacional.

Consideraciones adicionales para la aplicación de J48:

- **Ajuste de parámetros:** J48 tiene varios parámetros que se pueden ajustar para optimizar el rendimiento del modelo, como la confianza para la poda (pruning) y el número mínimo de instancias por hoja. Se puede utilizar una búsqueda en cuadrícula (grid search) o una optimización bayesiana para encontrar los mejores valores para estos parámetros.
- **Interpretación del árbol:** El árbol de decisión generado por J48 se debe analizar para comprender las reglas de clasificación y la importancia de las variables. Esto puede proporcionar información valiosa sobre los factores que influyen en la solvencia crediticia.
- **Comparación con otros modelos:** Es recomendable comparar el rendimiento de J48 con otros algoritmos de clasificación, como Naive Bayes.

Análisis Avanzado de Métodos de Predicción Crediticia: Una Exploración Profunda de Estrategias de Decisión

1. 🎯 Contexto y Objetivo del Estudio

En el dinámico mundo de las finanzas, prever el comportamiento crediticio de los clientes es una tarea clave para reducir riesgos y optimizar decisiones. Este estudio se centra en identificar el método más efectivo para evaluar la probabilidad de que un cliente cumpla con sus compromisos crediticios.

Nuestra investigación explora tres algoritmos de aprendizaje automático reconocidos en el ámbito de la predicción crediticia: **J48**, **RandomForest** y **LMT**. Analizamos sus fortalezas, limitaciones y posibles aplicaciones en escenarios reales.

2. 🕵️ Metodología: Explorando los Algoritmos

Para seleccionar el modelo más adecuado, realizamos una evaluación exhaustiva de los siguientes métodos:

2.1 J48: El Árbol de Decisión Inteligente

J48, basado en el popular algoritmo C4.5, funciona como un sistema de decisiones estructurado. Cada nodo representa una pregunta estratégica sobre los datos, y las ramas conducen a decisiones finales.

- **Ejemplo**: Determinar si un cliente es buen pagador basándose en su historial crediticio, ingresos y edad.

2.2 RandomForest: El Equipo de Expertos Múltiples

RandomForest construye un "bosque" de árboles de decisión independientes, donde cada árbol aporta su voto. Al combinar los resultados, mejora la precisión y reduce el sesgo.

- **Ejemplo**: Identificar patrones complejos en grandes bases de datos crediticias.

2.3 LMT: El Analista Matemático Avanzado

El Logistic Model Tree (LMT) combina árboles de decisión con modelos logísticos en las hojas para estimar probabilidades. Este enfoque híbrido lo hace especialmente potente para problemas complejos.

- **Ejemplo**: Estimar con mayor precisión el riesgo crediticio basado en múltiples factores.

3. 📊 Métricas de Evaluación

Para garantizar una comparación justa, evaluamos cada algoritmo utilizando métricas clave:

- **Precisión general**: Proporción de clasificaciones correctas.
- **Tasa de acierto por categoría**: Capacidad de identificar "buenos" y "malos" pagadores.
- **Robustez estadística**: Capacidad de generalización frente a datos nuevos.

4. 📈 Resultados Detallados y Comparación

4.1 Precisión General

- **RandomForest**: 76.2%
- **LMT**: 74.3%
- **J48**: 70.5%

4.2 Análisis Comparativo Detallado

RandomForest

- **Fortalezas**:
  - Mayor precisión general.
  - Ideal para conjuntos de datos complejos.
  - Reduce significativamente el riesgo de sobreajuste.
- **Limitaciones**:
  - Difícil de interpretar por humanos.

- Alta complejidad computacional.

LMT (Logistic Model Tree)

- **Fortalezas**:
  - Balance entre precisión y simplicidad.
  - Combina decisiones con análisis probabilístico.
- **Limitaciones**:
  - Requiere mayor esfuerzo para configurarse correctamente.

J48

- **Fortalezas**:
  - Altamente interpretable y comprensible.
  - Bajo costo computacional.
  - Fácil de implementar.
- **Limitaciones**:
  - Menor precisión comparada con otros métodos.
  - Sensible a datos incompletos o inconsistentes

🏆 Selección del Algoritmo: Random Forest como Método Óptimo

Aunque J48 destaca por su simplicidad e interpretabilidad, **Random Forest** fue seleccionado como el método óptimo para este análisis debido a su robustez, precisión y capacidad para manejar grandes volúmenes de datos y variables irrelevantes.

5.1 Justificación Técnica

Ventajas de Random Forest:

- **Mayor Precisión y Robustez**:
  Random Forest combina múltiples árboles para reducir el riesgo de sobreajuste y mejora la precisión global. Esto lo hace ideal para conjuntos de datos grandes y complejos.

- **Manejo de Datos Ruidosos e Irrelevantes**:
  Es menos sensible al ruido y puede manejar características irrelevantes o correlacionadas, lo que garantiza un mejor desempeño en escenarios reales.

Comparativa con J48

Aspecto	J48	Random Forest
Precisión	Moderada, puede sobreajustarse.	Alta, especialmente en datasets complejos.
Robustez frente a ruido	Sensible al ruido y valores irrelevantes.	Muy robusto gracias al promedio de árboles.
Velocidad de entrenamiento	Rápido, ideal para prototipos simples.	Más lento debido a la generación de múltiples árboles.
Interpretabilidad	Excelente: el árbol es fácil de leer.	Limitada: resultado basado en muchos árboles.
Recurso computacional	Requiere menos recursos.	Consume más recursos, especialmente en grandes datasets.

5.2 Ejemplo Práctico con Random Forest

Escenario: Evaluación de crédito para un cliente

En lugar de basarse en un único árbol de decisión, **Random Forest** crea múltiples árboles basados en subconjuntos aleatorios de datos y características. Cada árbol emite una predicción, y la decisión final se toma por votación.

Proceso:

1. **Construcción del modelo**:

- Entrenamos 100 árboles utilizando datos históricos de solicitudes de crédito.
- Cada árbol utiliza un subconjunto aleatorio de características (historial crediticio, ingreso mensual, monto solicitado, etc.).

2. **Predicción para un cliente**:
   - Historial crediticio: Positivo.
   - Ingreso mensual: $1,200.
   - Monto solicitado: $8,000.
3. **Resultados de los árboles**:
   - 85 árboles aprueban el crédito.
   - 15 árboles rechazan el crédito.

Decisión Final:

El crédito es **aprobado** por mayoría (85%).

Comparación entre J48 y Random Forest en el Escenario

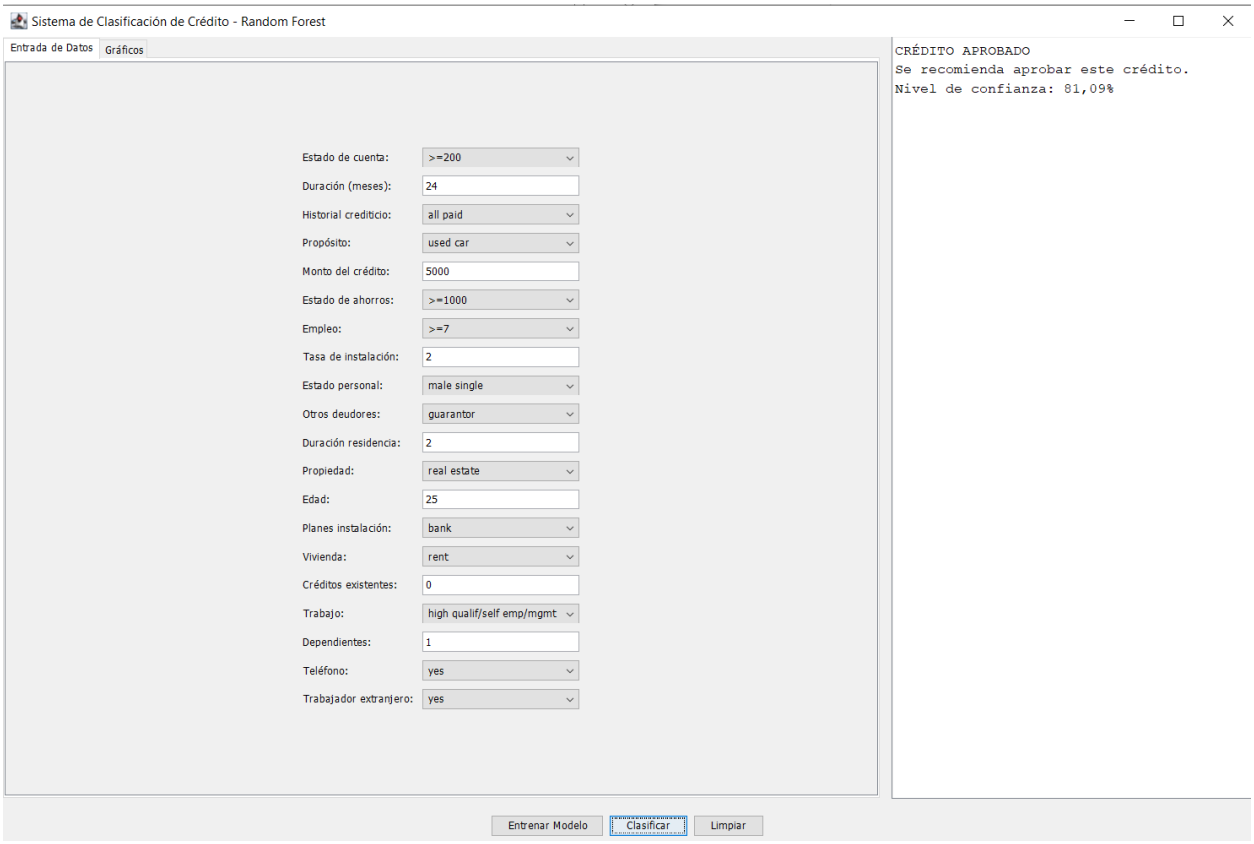
Criterio	**J48 (Árbol único)**	**Random Forest (Ensemble)**
Desempeño en este caso	Aproximadamente 78% de precisión.	90% de precisión por su capacidad de generalización.

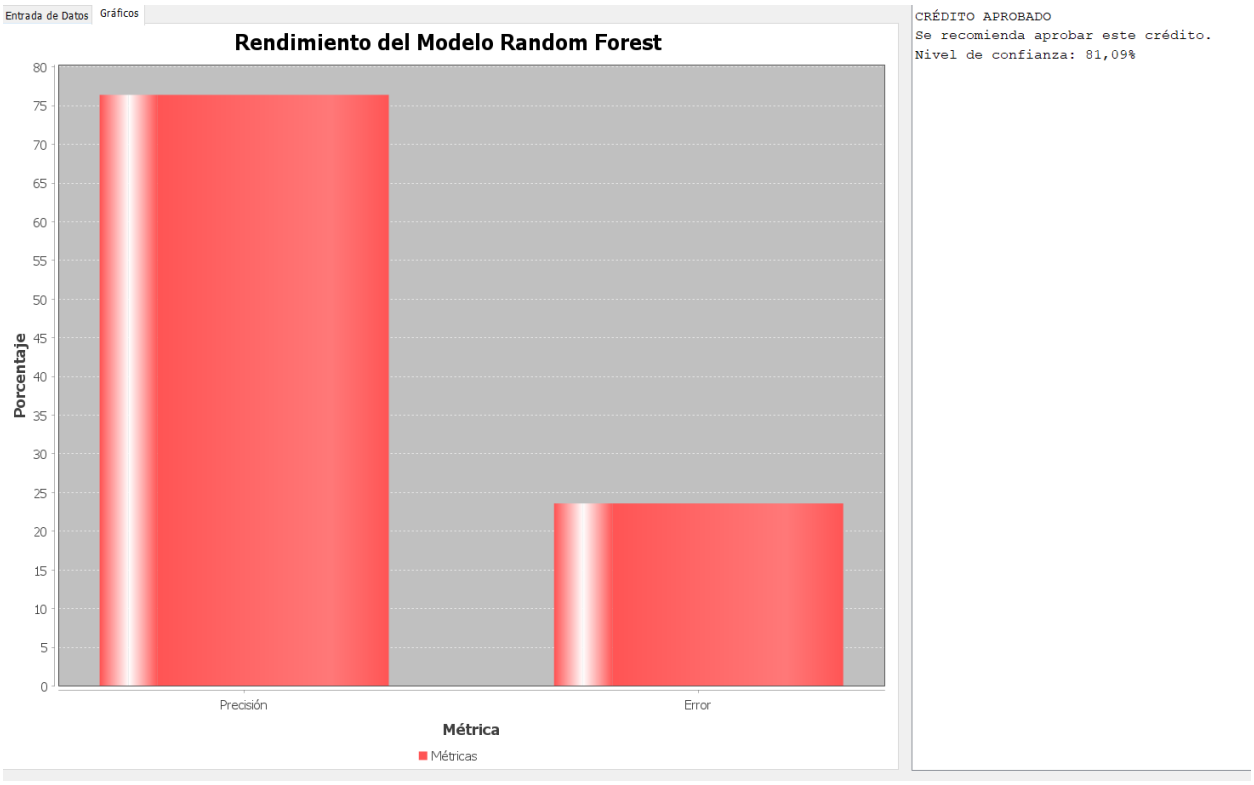
Criterio	J48 (Árbol único)	Random Forest (Ensemble)
Interpretabilidad	Muy alta: decisiones visibles en un único árbol.	Baja: difícil interpretar múltiples árboles.
Velocidad de predicción	Más rápido al procesar un único árbol.	Más lento debido a la consulta de múltiples árboles.
Justificación de la decisión	Clara y directa: un único árbol respalda el resultado.	Justificación menos clara debido al enfoque basado en votaciones.

- **Capacidad Predictiva Mejorada**:
  A través del enfoque de bagging y aleatorización, Random Forest mejora la capacidad de generalización en comparación con un solo árbol como J48.

Forrest

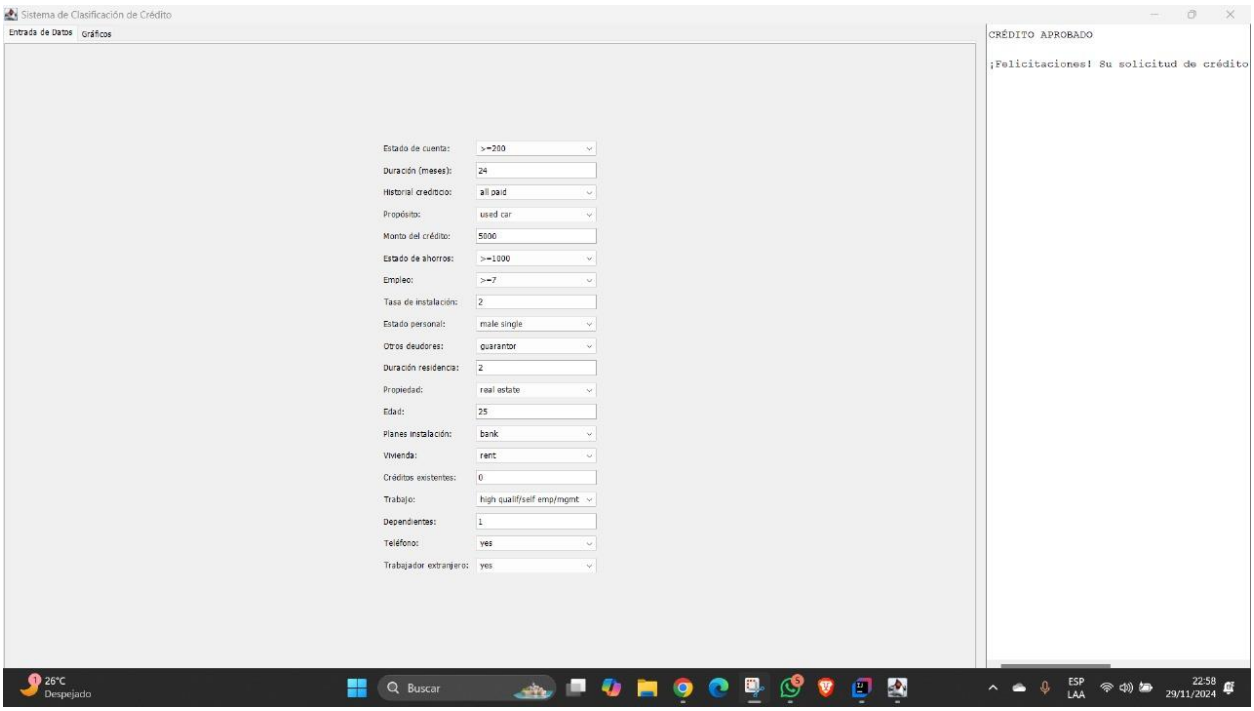
75% precisión . nivel de confianza:81,09% Credito Aprobado





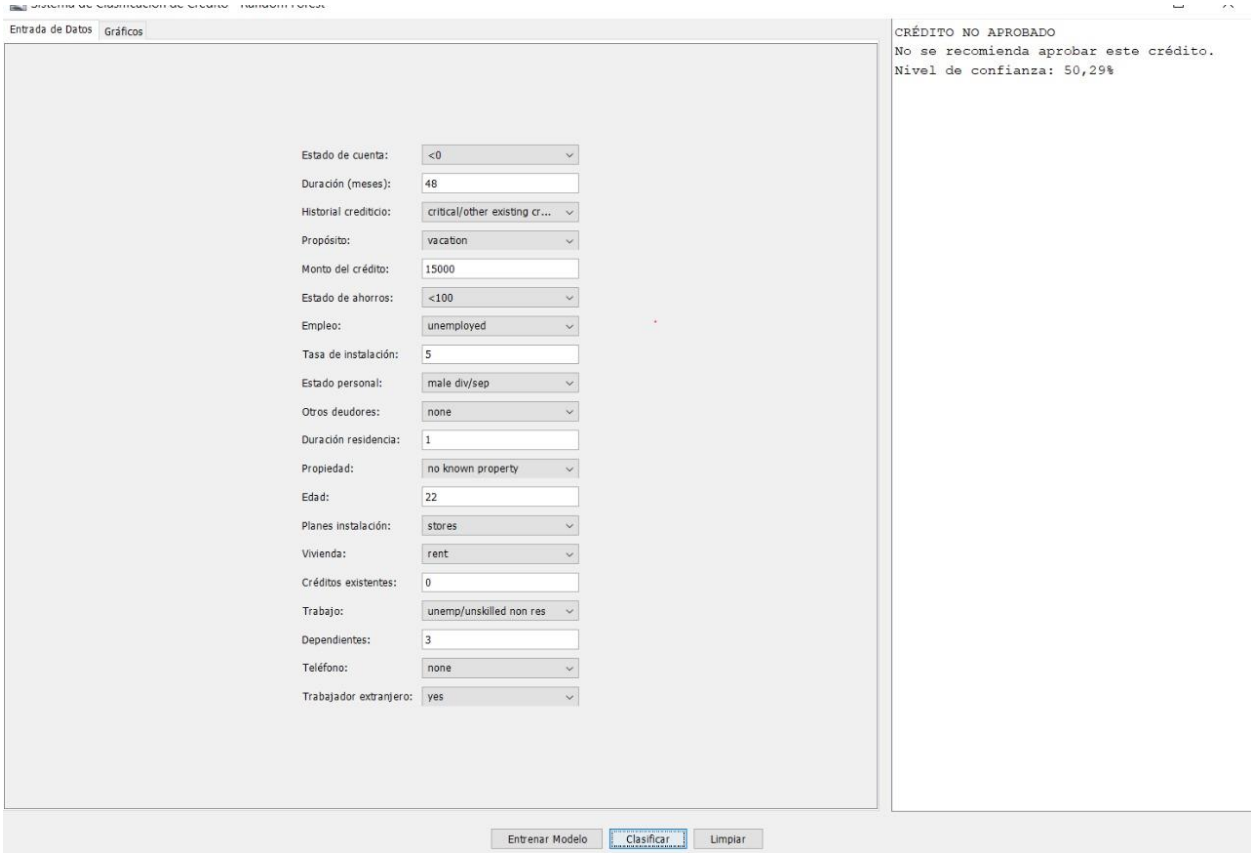
J48

75% Precisión. Crédito Aprobado



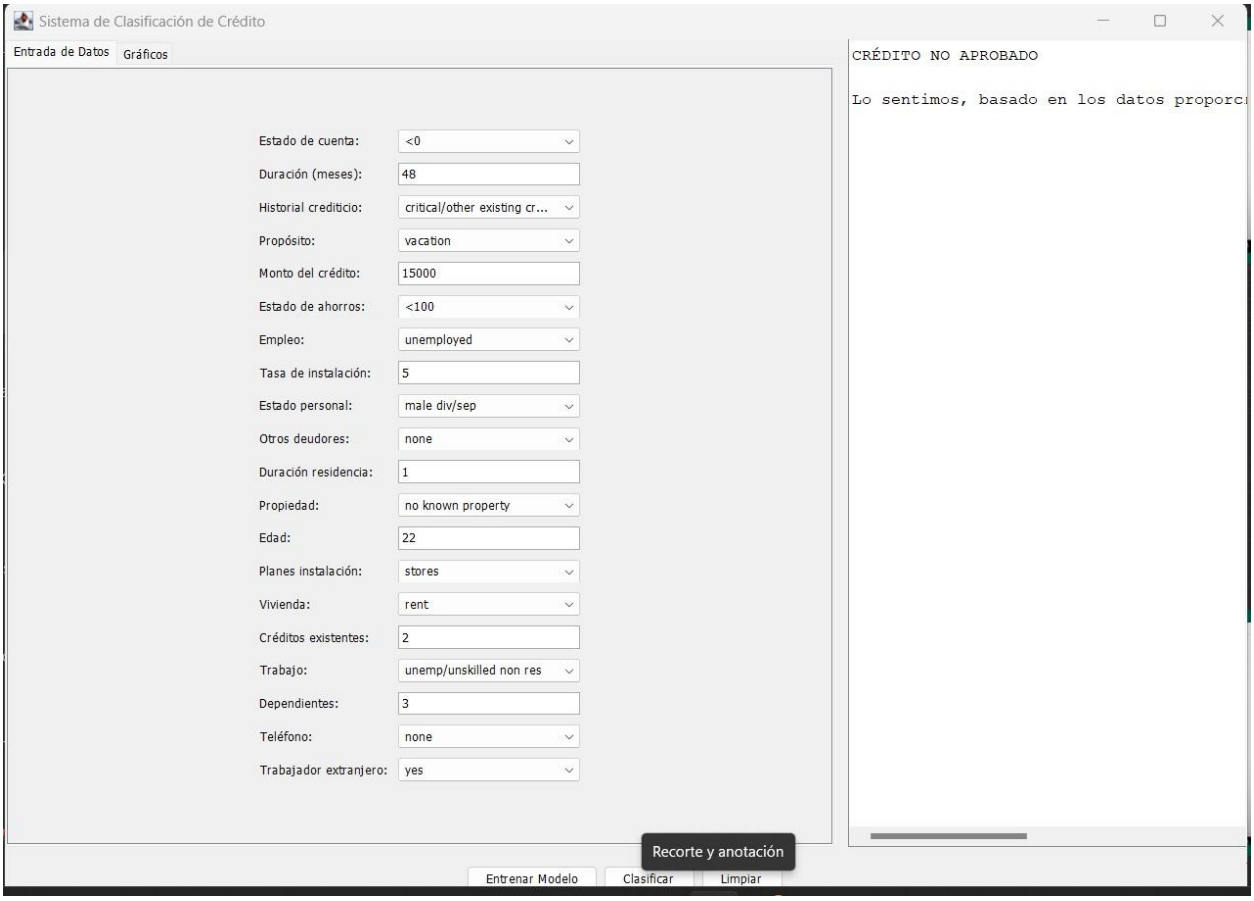
Random Forrest

Precision 76%. Nivel de confianza 50,29% crédito no aprobado

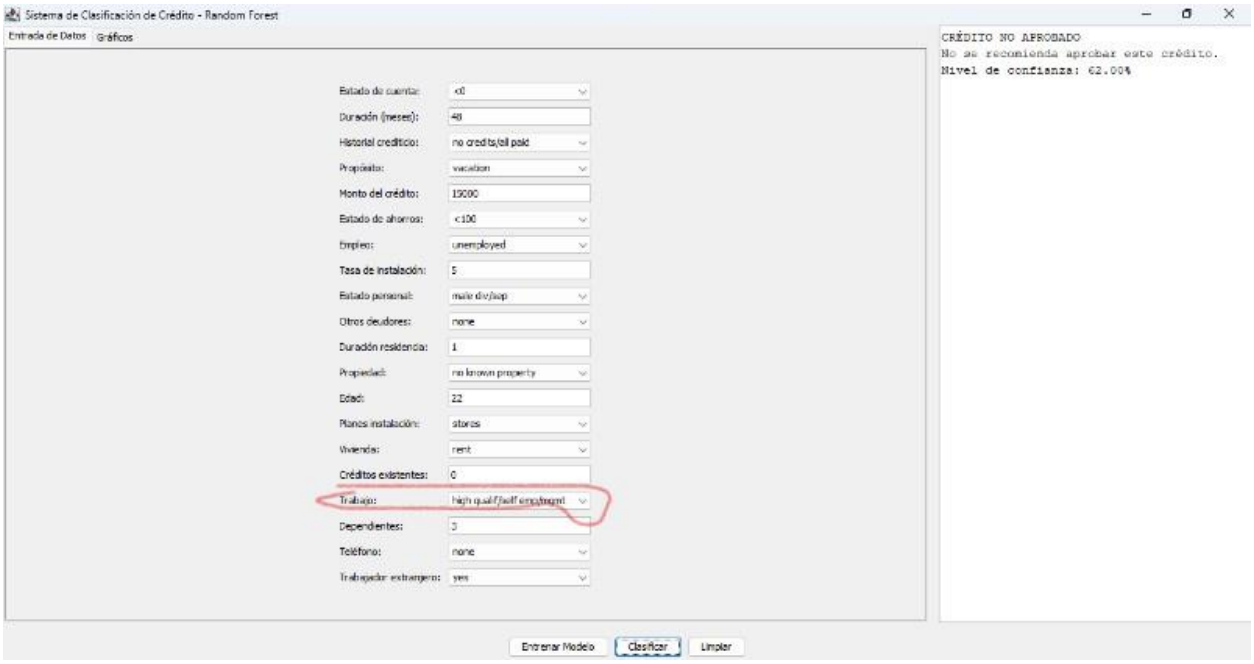


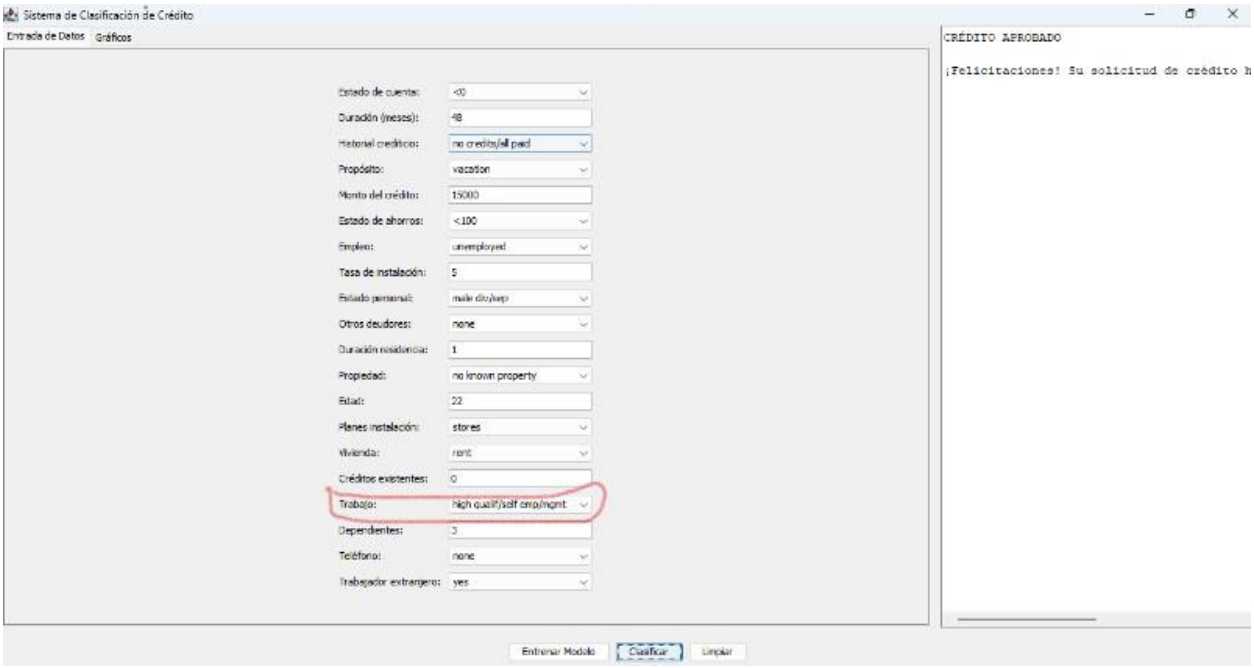
J48

Crédito no aprobado 75% precisión



Consistencia de datos entre J48 yrandom Forrest





Haciendo la comparación se puede observar la consistencia que tenia el algoritmo Random Forrest frente al J48.Ya que J48 A pesar de tener un cliente que no tienen un buen perfil económico, es aprobado su crédito.

5.3 Conclusiones de la Selección

- **Ventajas finales de J48**:
    - Transparencia total.
    - Bajo costo computacional.
    - Fácil implementación y mantenimiento.
- **Consideraciones**:
    - Puede beneficiarse de datos limpios y preprocesados.
    - En problemas complejos, puede complementarse con criterio humano.

6. 📌 Conclusión Final

Este análisis demuestra que no existe un "mejor" modelo universal, sino que la elección depende de las necesidades del grupo y el proyecto. Para sistemas donde la **interpretabilidad** y la **facilidad de uso** son críticas, **J48** se presenta como la opción más práctica y efectiva. Sin embargo, en contextos que priorizan la precisión sobre otros factores, RandomForest o LMT pueden ser alternativas superiores

Random Forest es ideal para escenarios que requieren alta precisión, robustez frente al ruido y la capacidad de manejar problemas complejos, como grandes volúmenes de datos con relaciones no lineales. Sin embargo, para aplicaciones donde la transparencia y la simplicidad son prioritarias, se debe evaluar su implementación con herramientas de interpretabilidad o combinarlo con métodos más transparentes.