[🎯 Objetivo 1 – Identificación de los Criterios de Otorgamiento de Créditos 2](#_Toc47417967)

[Modelo: Árbol de Decisión (Decision Tree Learner en KNIME) 2](#_Toc156271193)

[🌳 Análisis del Árbol de Decisión 2](#_Toc863601206)

[📌 Variable raíz 2](#_Toc1857871706)

[🧭 Principales caminos del árbol 3](#_Toc2002593812)

[📊 Otras variables relevantes 3](#_Toc1106276535)

[✅ Conclusiones 3](#_Toc275596239)

[🤖 Comparación con otros algoritmos de clasificación 4](#_Toc847516391)

[✅ Recomendación 4](#_Toc890964666)

[📌 Conclusión 5](#_Toc383066125)

[🎯 Objetivo 2 – Identificación y Caracterización de Grupos de Clientes 5](#_Toc1368918317)

[🧩 Introducción 5](#_Toc472685697)

[🔍 Algoritmos Aplicados 5](#_Toc274183134)

[📦 WEKA 6](#_Toc669697487)

[🧠 KNIME 6](#_Toc2097714013)

[📈 Evaluación de la Segmentación 6](#_Toc520796545)

[👥 Análisis Comparativo 6](#_Toc242941072)

[🔹 1. Con todos los datos 6](#_Toc1472289073)

[🔹 2. Solo con clientes que obtuvieron créditos 7](#_Toc516829754)

[🔎 Calidad de los Clústeres 7](#_Toc368105355)

[✅ Conclusiones 7](#_Toc642677437)

[🎯 Objetivo 3 – Identificación de los Factores de Incidencia en Clientes con Ingresos Superiores a $15.000 7](#_Toc652874432)

[🧠 Introducción 8](#_Toc841896102)

[🧪 Metodología 8](#_Toc96704222)

[📊 Resultados 8](#_Toc193224086)

[✅ Conclusiones 9](#_Toc1870299770)

[🎯 Objetivo 4 – Predicción para Nuevos Casos Usando Redes Bayesianas 9](#_Toc1577474613)

[🧠 Introducción 9](#_Toc856121946)

[🧪 Metodología 9](#_Toc1573903700)

[🔍 Análisis de Resultados 10](#_Toc1031988727)

[✅ Predicción de Casos Nuevos 11](#_Toc2067921491)

[📌 Conclusión 11](#_Toc1812528452)

[🎯 Objetivo 5 – Selección de los Atributos Más Significativos del Caso de Estudio 11](#_Toc1911936009)

[🧠 Introducción 12](#_Toc1137171265)

[🧪 Métodos Aplicados y Resultados 12](#_Toc1767749691)

[🔹 1. Feature Selection Loop Start (1:1) 12](#_Toc1766954611)

[🔹 2. AttributeSelectedClassifier con CfsSubsetEval 12](#_Toc1452855289)

[🔹 3. Apriori (Reglas de Asociación) 12](#_Toc1949438061)

[📊 Conclusiones del Análisis 13](#_Toc885956377)

[🏠 Vivienda 13](#_Toc1908674353)

[👨‍👩‍👧‍👦 Composición Familiar 13](#_Toc1898062527)

[💳 Otros Créditos 13](#_Toc1562059486)

[📡 Servicios 13](#_Toc1843190709)

[✅ Recomendaciones 13](#_Toc129173297)

Trabajo Practico - Inteligencia de Negocios

# 🎯 Objetivo 1 – Identificación d[[1]](#footnote-1)e los Criterios de Otorgamiento de Créditos

# Modelo: Árbol de Decisión (Decision Tree Learner en KNIME)

### **🌳 Análisis del Árbol de Decisión**

Para identificar los criterios utilizados por el banco al momento de otorgar créditos, se entrenó un modelo de árbol de decisión sobre el dataset histórico de clientes. Este modelo permite visualizar de forma clara las reglas que definen la aprobación o

rechazo de un préstamo.

Como se detalló previamente, el árbol de decisión permitió obtener reglas claras que explican las decisiones de otorgamiento de crédito, utilizando principalmente las variables:

* **Otros Créditos**
* **Vivienda**
* **Servicios contratados**
* **Composición Familiar**

Estas reglas hacen del árbol una herramienta altamente **interpretable**, ideal para comunicar decisiones a responsables de negocio no técnicos.

#### **📌 Variable raíz**

El árbol comienza dividiendo los datos según la variable **“Otros Créditos”**, indicando que esta es la característica más influyente. Se observa una clara tendencia a aprobar créditos cuando el cliente tiene uno o dos créditos previos, y a rechazarlos cuando posee tres.

#### **🧭 Principales caminos del árbol**

A continuación, se describen algunas de las **reglas más relevantes** inferidas por el árbol:

* ✅ **Regla 1:**  
   Si el cliente tiene **1 o 2 créditos previos**  
   y tiene **vivienda propia**  
   → **Se aprueba el préstamo**
* ✅ **Regla 2:**  
   Si el cliente tiene **1 o 2 créditos previos**,  
   **alquila**, y posee **servicios básicos**  
   → **Se aprueba el préstamo**
* ❌ **Regla 3:**  
   Si el cliente tiene **3 créditos previos**  
   y **servicios completos** (básicos, TV y celular)  
   → **Se rechaza el préstamo**
* ❌ **Regla 4:**  
   Si el cliente es **casado con 2 hijos**,  
   tiene **vivienda alquilada** y **3 créditos previos**  
   → **Se rechaza el préstamo**

#### **📊 Otras variables relevantes**

Aunque aparecen en niveles más profundos del árbol, otras variables como **“Servicios”**, **“Composición Familiar”** e **“Ingresos”** también inciden en la decisión:

* Los clientes con **servicios básicos** tienden a tener más aprobaciones.
* Las familias **casadas con hijos** tienen más rechazos que los **solteros** o **casados sin hijos**.
* Los ingresos, si bien son importantes, no son determinantes principales según el árbol.

### **✅ Conclusiones**

* El **número de créditos previos** es el factor más decisivo.
* La **tenencia de vivienda** (propia) y la **simplicidad de los servicios contratados** también aumentan la probabilidad de aprobación.
* El árbol de decisión es un modelo altamente **interpretable**, ideal para explicar decisiones a perfiles no técnicos del banco.
* Se recomienda complementarlo con otros modelos (como Random Forest o Naive Bayes) para verificar la consistencia de los patrones hallados y mejorar la precisión predictiva.

### **🤖 Comparación con otros algoritmos de clasificación**

Además del árbol de decisión, se implementaron otros algoritmos de clasificación disponibles en KNIME, incluyendo:

|  |  |  |
| --- | --- | --- |
| **Algoritmo** | **Tipo** | **Ventajas principales** |
| Naive Bayes | Probabilístico | Rápido, efectivo con pocos datos, útil en Objetivo 3 y 4 |
| Random Forest | Ensamble de árboles | Alta precisión, buena generalización |
| JRip / ID3 / OneR | Árboles simples / reglas | Fáciles de interpretar, pero menos precisos |
| SVM / KNN | Numéricos (requieren OHE) | Potentes en datasets complejos, menos interpretables |

Las métricas de evaluación (precisión, recall, accuracy, etc.) fueron obtenidas con los nodos **Scorer** y **ROC Curve**, utilizando particiones 70/30 y validación cruzada con **X-Partitioner**.

Se observó que:

* **Random Forest** obtuvo el mejor rendimiento general en términos de precisión y exactitud.
* **Naive Bayes** mostró buen desempeño y fue particularmente útil para analizar segmentos de clientes.
* **Árbol de decisión** ofreció un buen balance entre interpretabilidad y rendimiento.

### **✅ Recomendación**

Se recomienda al banco **utilizar el modelo de Árbol de Decisión como base para la toma de decisiones comerciales**, ya que:

* Proporciona **reglas simples y comprensibles**.
* Permite entender fácilmente qué factores impulsan la aprobación o rechazo.
* Es adecuado para generar políticas de crédito claras.

Sin embargo, para casos donde se requiera **mayor precisión predictiva**, como en simulaciones de riesgo o nuevos clientes, se sugiere **complementarlo con Random Forest o Naive Bayes**, que ofrecen mayor rendimiento, aunque con menor interpretabilidad.

### **📌 Conclusión**

Este análisis permitió identificar patrones sólidos de comportamiento crediticio en los clientes del banco. El conocimiento extraído con árboles de decisión puede utilizarse como base para políticas de aprobación y, combinado con modelos más complejos, puede brindar una solución robusta y estratégica.

## **🎯 Objetivo 2 – Identificación y Caracterización de Grupos de Clientes**

### **🧩 Introducción**

El presente análisis tiene como finalidad segmentar a los clientes del banco mediante algoritmos de agrupamiento (clustering), con el objetivo de **identificar grupos homogéneos de personas** que compartan características similares. Esta segmentación permite diseñar líneas de crédito personalizadas, alineadas a las necesidades específicas de cada grupo.

### **🔍 Algoritmos Aplicados**

Se utilizaron tanto algoritmos del paquete **WEKA** como algoritmos nativos de **KNIME** para realizar la segmentación:

#### **📦 WEKA**

* **EM (Expectation-Maximization):** asigna probabilidades de pertenencia a distintos clústeres.
* **FarthestFirst:** selecciona puntos de partida lejanos entre sí, optimizando la separación.
* **SimpleKMeans:** agrupa en torno a centroides cercanos.

#### **🧠 KNIME**

* **K-Means Clustering:** particiona el dataset en clústeres según cercanía a centroides.
* **Hierarchical Clustering:** construye una jerarquía de agrupamiento ascendente o descendente.

### **📈 Evaluación de la Segmentación**

Para evaluar la calidad de los clústeres generados, se utilizaron las siguientes métricas:

* **Silhouette Coefficient:** mide qué tan bien se agrupan los elementos dentro de un clúster.
* **Entropy Scorer:** evalúa la pureza de cada clúster respecto a las categorías internas de los datos.

### **👥 Análisis Comparativo**

Se realizaron dos segmentaciones distintas:

#### **🔹 1. Con todos los datos**

* Permitió observar **diversidad de perfiles** entre clientes, con variabilidad en características como ingresos, composición familiar, servicios y tipo de vivienda.
* Los clústeres mostraron patrones amplios, útiles para definir **líneas de crédito generales**.

#### **🔹 2. Solo con clientes que obtuvieron créditos**

* Se filtró el dataset para incluir únicamente a quienes se les otorgó un préstamo.
* Esta segmentación destacó **patrones específicos de los clientes aprobados**, como:
  + Mayor presencia de vivienda propia
  + Menor número de créditos simultáneos
  + Preferencia por servicios básicos o intermedios

### **🔎 Calidad de los Clústeres**

* En ambos casos, los **coeficientes de Silhouette fueron altos**, indicando buena cohesión interna y separación entre grupos.
* La **entropía fue baja**, reflejando una alta homogeneidad dentro de los clústeres.

### **✅ Conclusiones**

* La segmentación permite al banco **entender mejor los perfiles de sus clientes**, adaptando sus productos financieros según cada grupo.
* El análisis global brinda una visión panorámica útil para estrategias generales, mientras que el filtrado por aprobaciones revela **criterios ocultos** detrás de la concesión de créditos.
* Esta información es clave para **diseñar campañas de marketing específicas**, reducir riesgos crediticios y mejorar la experiencia del cliente.

## **🎯 Objetivo 3 – Identificación de los Factores de Incidencia en Clientes con Ingresos Superiores a $15.000**

### **🧠 Introducción**

El objetivo de este análisis fue identificar qué características tienen mayor peso en la decisión de otorgar un crédito a clientes con **ingresos superiores a $15.000**. Este conocimiento resulta clave para definir políticas diferenciadas y ofertas personalizadas para un segmento considerado de alto valor.

### **🧪 Metodología**

1. **Revisión previa:** Se analizaron los resultados del algoritmo **Naive Bayes** aplicado al conjunto total de datos.
2. **Filtrado del dataset:** Se utilizó el nodo **Row Filter** para seleccionar únicamente los registros con ingresos **“Más de $15.000”**.
3. **División del conjunto:** Se aplicó **Partitioning** (70% entrenamiento, 30% prueba) para evaluar el modelo de forma controlada.
4. **Modelo aplicado:** Se volvió a entrenar un clasificador **Naive Bayes** sobre el subconjunto filtrado.
5. **Evaluación:** Se compararon las métricas de desempeño (precisión, recall, área bajo la curva ROC) con las obtenidas previamente sobre el dataset completo.

### **📊 Resultados**

* Se evidenció una **mejora en la precisión del modelo** aplicado al segmento de ingresos altos, lo que sugiere que sus características presentan **mayor regularidad** en la aprobación de créditos.
* El área bajo la curva ROC también se incrementó, lo que indica una mejor capacidad del modelo para discriminar entre aprobaciones y rechazos en este grupo.
* A través de la visualización de probabilidades condicionales, se identificaron variables **determinantes para este segmento**, entre ellas:
  + ✔️ **Vivienda propia**: se asocia fuertemente con aprobación del préstamo.
  + ✔️ **Menor cantidad de créditos previos**: mejora la probabilidad de aceptación.
  + ✔️ **Servicios contratados**: los perfiles con servicios más simples tienden a tener más aprobaciones.

### **✅ Conclusiones**

El análisis reveló que los clientes con ingresos superiores a $15.000 presentan **patrones de comportamiento más estables** y previsibles. Variables como la **tenencia de vivienda**, la **cantidad de créditos** y los **servicios contratados** permiten anticipar con mayor certeza si obtendrán un préstamo.

Esta información es fundamental para:

* **Optimizar políticas crediticias**
* **Diseñar productos específicos para este grupo**
* **Reducir el riesgo financiero** a través de decisiones basadas en datos

El uso del algoritmo Naive Bayes demostró ser eficaz para este tipo de segmentación, ya que proporciona tanto **predicciones precisas** como **explicaciones interpretables**.

## **🎯 Objetivo 4 – Predicción para Nuevos Casos Usando Redes Bayesianas**

### **🧠 Introducción**

El banco desea validar la efectividad de los modelos de minería de datos aplicados en el TP para **predecir el otorgamiento de créditos en nuevos casos**, utilizando **redes bayesianas**. Para ello, se analiza una muestra reducida de clientes con información similar a la del dataset histórico, y se busca estimar si obtendrán o no el préstamo.

### **🧪 Metodología**

Se aplicó el algoritmo **Naive Bayes** sobre el conjunto completo de datos históricos. Luego, se utilizaron tres nuevos casos provistos por el banco para realizar una predicción basada en:

* **Probabilidades a priori**: proporción general de aprobaciones y rechazos en el dataset.
* **Probabilidades condicionales**: probabilidad de observar ciertos atributos dentro de cada clase (préstamo otorgado o rechazado).
* **Normalización**: ajuste de probabilidades para comparar escenarios y definir la clase más probable.

Además, se compararon los resultados automáticos del algoritmo con los cálculos realizados de forma manual, como forma de validar el modelo y comprender su lógica de funcionamiento.

### **🔍 Análisis de Resultados**

A continuación, se describen los factores con mayor incidencia en la predicción:

* **Composición Familiar:**  
   Los grupos "Soltero", "Casado sin hijos" y "Casado con 1 hijo" concentran la mayoría de los créditos aprobados (~75%), en contraste con "Casado con 2 hijos", que tiene mayor proporción de rechazos.
* **Otros Créditos:**  
   Tener **tres créditos previos** incrementa notablemente la probabilidad de **rechazo**. En cambio, tener **uno o dos créditos** se relaciona con mayor tasa de aprobación.
* **Servicios:**  
   Los clientes con **servicios básicos o intermedios** tienen más probabilidades de ser aprobados. A medida que se incrementa el nivel de servicios contratados, la tasa de aprobación tiende a descender.
* **Vivienda:**  
   Poseer vivienda **propia** se asocia con mayor estabilidad financiera y, por ende, con más aprobaciones de crédito.
* **Ingreso:**  
   En el modelo, los clientes con ingresos **"Entre $8.000 y $15.000"** y los de **"Más de $15.000"** presentan tasas de aprobación relativamente similares, lo cual indica que el ingreso por sí solo no es un factor decisivo.

### **✅ Predicción de Casos Nuevos**

Se analizaron los siguientes casos:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Caso** | **Ingreso** | **Composición Familiar** | **Vivienda** | **Servicios** | **Otros Créditos** | **Predicción** |
| 1 | Más de $15.000 | Casado con 2 hijos | Propia | Básicos | Un crédito | Sí |
| 2 | Entre $8.000 y $15.000 | Casado con 1 hijo | Alquila | Básicos | Tres créditos | No |
| 3 | Entre $8.000 y $15.000 | Soltero | Propia | Básicos, TV por cable y teléfono celular | Un crédito | Sí |

Estas predicciones se determinaron en base a los cálculos de probabilidad condicional y posterior normalización, y fueron consistentes con los resultados obtenidos por KNIME.

### **📌 Conclusión**

El modelo basado en redes bayesianas permitió realizar predicciones confiables sobre nuevos casos, validando su utilidad para escenarios reales. A través de este análisis se logró:

* Visualizar qué factores tienen mayor peso en la aprobación de créditos.
* Confirmar la consistencia entre cálculos manuales y automáticos.
* Demostrar que el modelo es **intuitivo, explicable y aplicable** a casos prácticos.

Esta técnica puede ser utilizada como **herramienta de apoyo en decisiones crediticias**, complementando las políticas actuales del banco.

## **🎯 Objetivo 5 – Selección de los Atributos Más Significativos del Caso de Estudio**

### **🧠 Introducción**

El objetivo de este análisis fue identificar cuáles son los atributos más relevantes al momento de predecir si un cliente obtendrá o no un crédito. Para ello, se aplicaron distintos algoritmos de **selección de características** disponibles en KNIME, que permiten reducir la dimensionalidad del problema y mejorar tanto la interpretabilidad como el rendimiento de los modelos.

### **🧪 Métodos Aplicados y Resultados**

#### **🔹 1. Feature Selection Loop Start (1:1)**

Este método automático de selección iterativa identificó a la variable **“Vivienda”** como uno de los atributos más influyentes. Tanto la categoría "Alquila" como "Propia" aparecen frecuentemente asociadas a patrones de otorgamiento de crédito.

#### **🔹 2. AttributeSelectedClassifier con CfsSubsetEval**

Mediante la evaluación de subconjuntos, este método destacó los atributos:

* **“Composición Familiar”**
* **“Otorga\_Credito”**

Esto indica que la estructura familiar y la clase objetivo tienen fuerte relación con el comportamiento de las variables predictoras.

#### **🔹 3. Apriori (Reglas de Asociación)**

Este método permitió extraer reglas con alto nivel de confianza. Entre ellas, se destaca:

**Ingreso = Más de $15.000** ∧ **Otros\_Créditos = Un crédito** → **Otorga\_Credito = Sí**  
 **Confianza: 100%**

Además, los atributos más recurrentes en las reglas fueron:

* **Ingreso**
* **Vivienda**
* **Otros Créditos**

### **📊 Conclusiones del Análisis**

#### **🏠 Vivienda**

El atributo "Vivienda" es altamente significativo. La categoría **“Propia”** se asocia consistentemente con aprobaciones de crédito, sugiriendo mayor confianza por parte del banco hacia quienes poseen un activo inmobiliario. Sin embargo, "Alquila" también aparece con relevancia, por lo que no debe descartarse.

#### **👨‍👩‍👧‍👦 Composición Familiar**

Las familias **“Casado sin hijos”** presentan mayores probabilidades de aprobación, posiblemente por su equilibrio entre estabilidad y nivel de gasto. Este grupo podría representar un segmento objetivo estratégico.

#### **💳 Otros Créditos**

Tanto **“Un crédito”** como **“Tres créditos”** se identifican como relevantes en varias reglas de asociación. Esto refuerza la importancia del historial crediticio como indicador clave de solvencia.

#### **📡 Servicios**

Aunque de menor peso relativo, los tipos de servicios contratados también influyen. Los clientes con **servicios básicos** tienden a tener mejores tasas de aprobación, aunque ciertos niveles adicionales no parecen ser determinantes negativos por sí solos.

### **✅ Recomendaciones**

Con base en los hallazgos, se recomienda al banco enfocar su análisis de riesgo en los siguientes atributos:

* **Vivienda**
* **Composición Familiar**
* **Historial de Otros Créditos**

Estos factores deberían ser considerados con prioridad en futuros modelos predictivos o políticas de análisis crediticio. Documentar y aplicar estos criterios basados en evidencia empírica puede aumentar la efectividad de las decisiones de otorgamiento, reducir riesgos y fomentar relaciones financieras más sostenibles con los clientes.

1. [↑](#footnote-ref-1)