

**Tarea sobre análisis de datos y minería de datos**

**Facultad de ciencias de la computación y telecomunicaciones**

**Nombre**: Jose Carlos Girón Solano

**Registro**: 216161924

**Docente**: Miguel Jesús Peinado

**Materia**: Sistemas para el soporte a la toma de decisiones

**Contenido**

[**Pasos para obtener los datos** 4](#_Toc183738303)

[**1. Recolección de datos 4**](#_Toc183738304)

[**2. Limpieza de datos 4**](#_Toc183738305)

[**3. Codificación de atributos cualitativos 4**](#_Toc183738306)

[**4. Estandarización de atributos numéricos 5**](#_Toc183738307)

[**5. Definición de las etiquetas o clases 5**](#_Toc183738308)

[**6. Generación del archivo ARFF 5**](#_Toc183738309)

[**Descripción de los atributos 6**](#_Toc183738310)

[**1. Status of existing checking account (Cualitativo) 6**](#_Toc183738311)

[**2. Duration in month (Numérico) 6**](#_Toc183738312)

[**3. Credit history (Cualitativo) 7**](#_Toc183738313)

[**4. Purpose (Cualitativo) 7**](#_Toc183738314)

[**5. Credit amount (Numérico) 8**](#_Toc183738315)

[**6. Savings account/bonds (Cualitativo) 8**](#_Toc183738316)

[**7. Present employment since (Cualitativo) 9**](#_Toc183738317)

[**8. Installment rate in percentage of disposable income (Numérico) 10**](#_Toc183738318)

[**9. Personal status and sex (Cualitativo) 10**](#_Toc183738319)

[**10. Other debtors / guarantors (Cualitativo) 11**](#_Toc183738320)

[**11. Present residence since (Numérico) 11**](#_Toc183738321)

[**12. Property (Cualitativo) 12**](#_Toc183738322)

[**13. Age in years (Numérico) 12**](#_Toc183738323)

[**14. Other installment plans (Cualitativo) 13**](#_Toc183738324)

[**15. Housing (Cualitativo) 13**](#_Toc183738325)

[**16. Number of existing credits at this bank (Numérico) 14**](#_Toc183738326)

[**17. Job (Cualitativo) 14**](#_Toc183738327)

[**18. Number of people being liable to provide maintenance for (Numérico) 15**](#_Toc183738328)

[**19. Telephone (Cualitativo) 15**](#_Toc183738329)

[**20. Foreign worker (Cualitativo) 16**](#_Toc183738330)

[**Ejemplos de uso 17**](#_Toc183738331)

[**1. Segmentación de clientes 17**](#_Toc183738332)

[**2. Evaluación de riesgo crediticio 17**](#_Toc183738333)

[**3. Análisis de la estabilidad financiera 17**](#_Toc183738334)

[**4. Impacto de las características demográficas 18**](#_Toc183738335)

[**5. Análisis de productos financieros 18**](#_Toc183738336)

[**6. Optimización de políticas de crédito 19**](#_Toc183738337)

[**7. Predicción del comportamiento futuro 19**](#_Toc183738338)

[**8. Detección de fraude 19**](#_Toc183738339)

[**9. Análisis de carga financiera 20**](#_Toc183738340)

[**10. Tendencias de mercado 20**](#_Toc183738341)

[**Técnicas y Algoritmos 21**](#_Toc183738342)

[** Arboles de decisión (J48): 21**](#_Toc183738343)

[** Regresión Logística: 28**](#_Toc183738344)

# **Pasos para obtener los datos**

### ****1. Recolección de datos****

#### **Acción:**

* Recopilar datos de diversas fuentes, como bases de datos, sistemas de gestión de clientes (CRM), archivos CSV o informes financieros.
* Asegurarse de incluir variables relevantes como atributos cualitativos (categorías) y cuantitativos (valores numéricos).

#### **Ejemplo:**

Se obtiene información de clientes que han solicitado créditos: ingresos, duración del empleo, historial crediticio, edad, etc.

### ****2. Limpieza de datos****

#### **Acción:**

* **Eliminar registros incompletos:** Filtrar datos con valores faltantes o incoherentes.
* **Corrección de errores:** Normalizar categorías mal escritas o convertir formatos.
* **Imputación de valores faltantes:** Asignar valores por defecto o promedio para datos faltantes, si es necesario.
* **Discretización de datos:** Transformar atributos numéricos continuos en categorías discretas.

#### **Ejemplo:**

* Clientes sin historial (no checking account) se categorizan como A14.
* Créditos con montos negativos o valores imposibles se corrigen o eliminan.

### ****3. Codificación de atributos cualitativos****

#### **Acción:**

* Transformar valores categóricos en identificadores únicos para que puedan ser representados en formato ARFF.

#### **Ejemplo:**

* Convertir Status of existing checking account:
  + < 0 DM → A11
  + 0 <= ... < 200 DM → A12
* Reemplazar valores como male: single con A93.

### ****4. Estandarización de atributos numéricos****

#### **Acción:**

* Escalar valores numéricos a rangos uniformes, si es necesario (e.g., normalización entre 0 y 1).
* Asegurar consistencia en las unidades (por ejemplo, duración siempre en meses).

#### **Ejemplo:**

* Convertir todas las duraciones de años a meses.
* Normalizar montos de crédito a miles de DM, si es necesario.

### ****5. Definición de las etiquetas o clases****

#### **Acción:**

* Si se trata de un problema supervisado, definir claramente la etiqueta o clase objetivo para predicción.

#### **Ejemplo:**

* Clasificar clientes como:
  + 1 (crédito otorgado sin problemas).
  + 0 (crédito denegado o incumplido).

### ****6. Generación del archivo ARFF****

#### **Acción:**

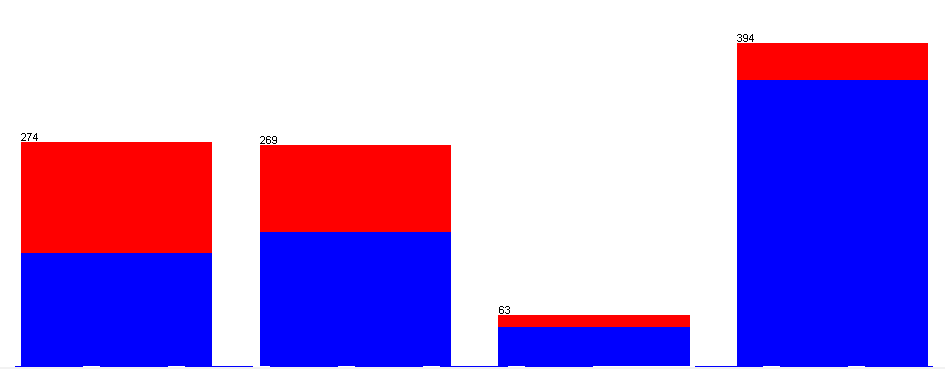
* Crear un archivo de texto que siga la estructura específica de ARFF:
  1. Encabezado con nombre de la relación (@relation).
  2. Definición de atributos (@attribute).
  3. Sección de datos (@data).

# **Descripción de los atributos**

**1. Status of existing checking account (Cualitativo)**

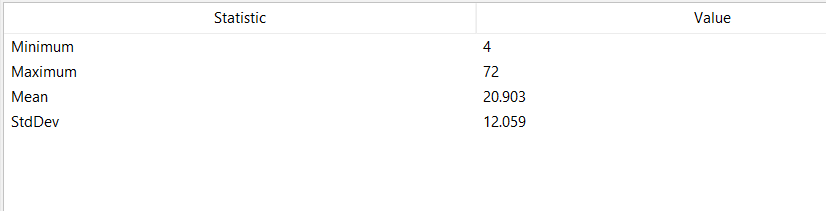
Representa el estado de la cuenta corriente existente. Las categorías reflejan la cantidad de dinero en la cuenta:

* **A11**: Saldo negativo (< 0 DM).
* **A12**: Saldo entre 0 y menos de 200 DM.
* **A13**: Saldo >= 200 DM o salario depositado regularmente por al menos un año.
* **A14**: No tiene cuenta corriente.



**2. Duration in month (Numérico)**

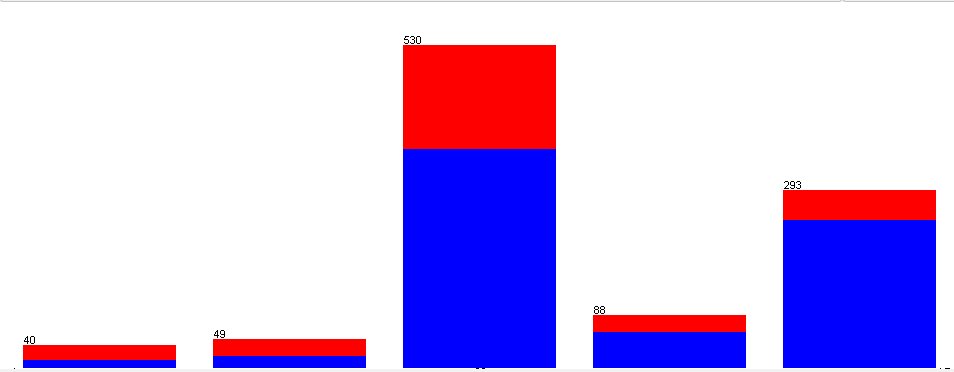
Duración del préstamo en meses.



**3. Credit history (Cualitativo)**

Indica el historial crediticio del cliente:

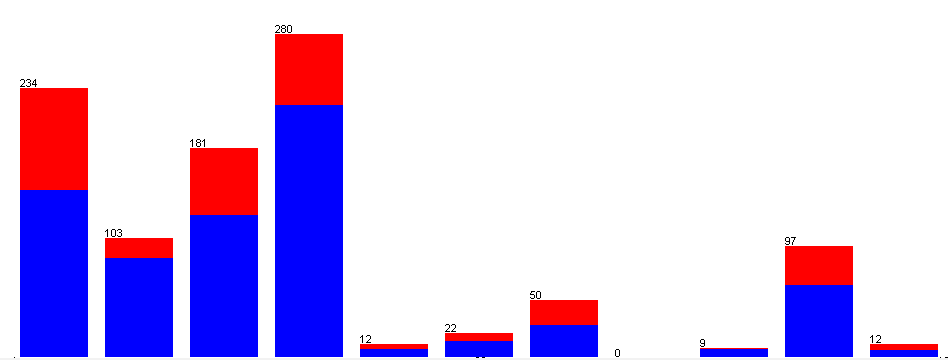
* **A30**: Nunca tuvo créditos / Todos los créditos anteriores fueron pagados puntualmente.
* **A31**: Todos los créditos en este banco fueron pagados puntualmente.
* **A32**: Créditos existentes pagados puntualmente hasta ahora.
* **A33**: Retrasos en pagos en el pasado.
* **A34**: Cuenta crítica / Créditos existentes en otras instituciones.



**4. Purpose (Cualitativo)**

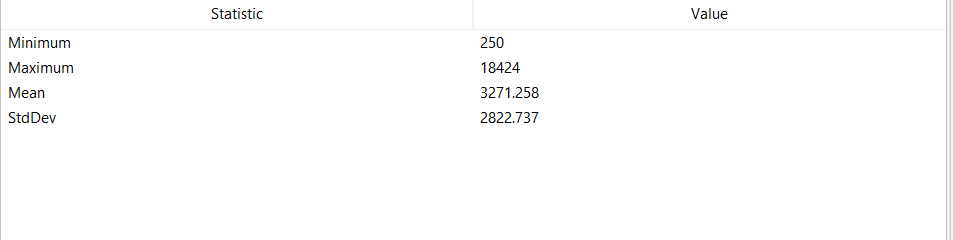
El propósito del crédito solicitado:

* **A40**: Automóvil nuevo.
* **A41**: Automóvil usado.
* **A42**: Muebles/equipamiento.
* **A43**: Radio/televisión.
* **A44**: Electrodomésticos.
* **A45**: Reparaciones.
* **A46**: Educación.
* **A47**: Vacaciones (categoría posiblemente inexistente).
* **A48**: Reentrenamiento.
* **A49**: Negocios.
* **A410**: Otros.



**5. Credit amount (Numérico)**

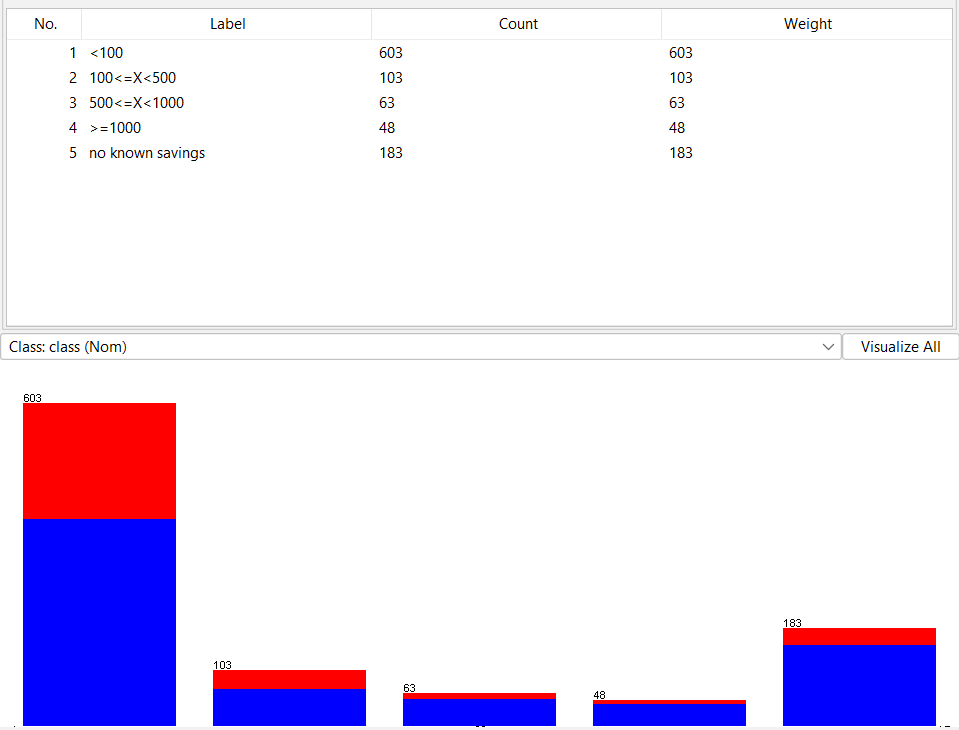
Monto del crédito solicitado en DM.



**6. Savings account/bonds (Cualitativo)**

Estado de cuentas de ahorros o bonos:

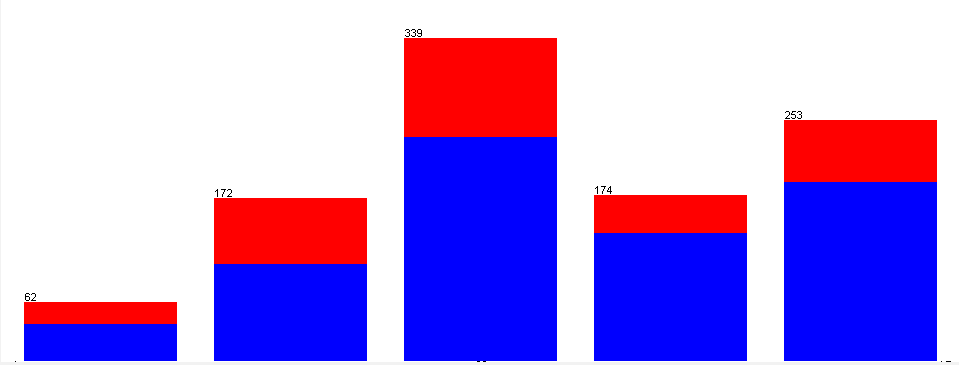
* **A61**: Saldo < 100 DM.
* **A62**: Saldo entre 100 y menos de 500 DM.
* **A63**: Saldo entre 500 y menos de 1000 DM.
* **A64**: Saldo >= 1000 DM.
* **A65**: Desconocido / Sin cuenta de ahorros.



**7. Present employment since (Cualitativo)**

Duración del empleo actual:

* **A71**: Desempleado.
* **A72**: Menos de 1 año.
* **A73**: Entre 1 y menos de 4 años.
* **A74**: Entre 4 y menos de 7 años.
* **A75**: 7 años o más.



**8. Installment rate in percentage of disposable income (Numérico)**

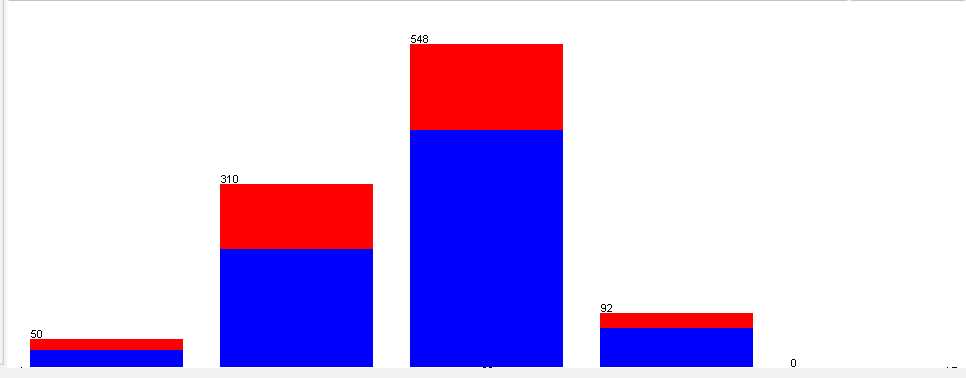
Porcentaje del ingreso disponible destinado a pagar la cuota mensual del crédito.



**9. Personal status and sex (Cualitativo)**

Estado civil y género del cliente:

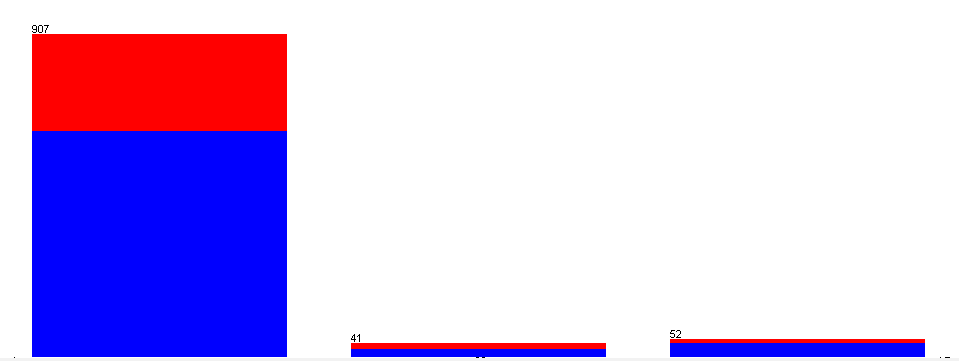
* **A91**: Hombre divorciado/separado.
* **A92**: Mujer divorciada/separada/casada.
* **A93**: Hombre soltero.
* **A94**: Hombre casado/viudo.
* **A95**: Mujer soltera.



**10. Other debtors / guarantors (Cualitativo)**

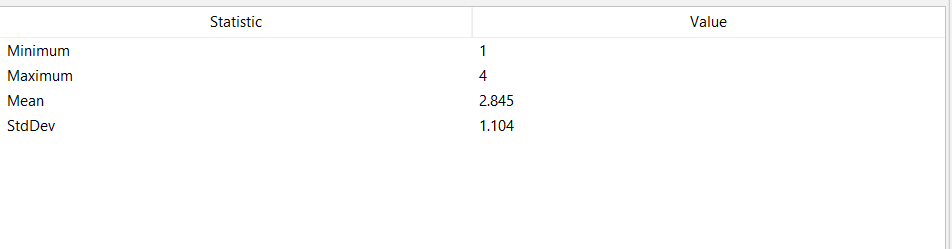
Otros responsables del crédito:

* **A101**: Ninguno.
* **A102**: Co-solicitante.
* **A103**: Avalista.



**11. Present residence since (Numérico)**

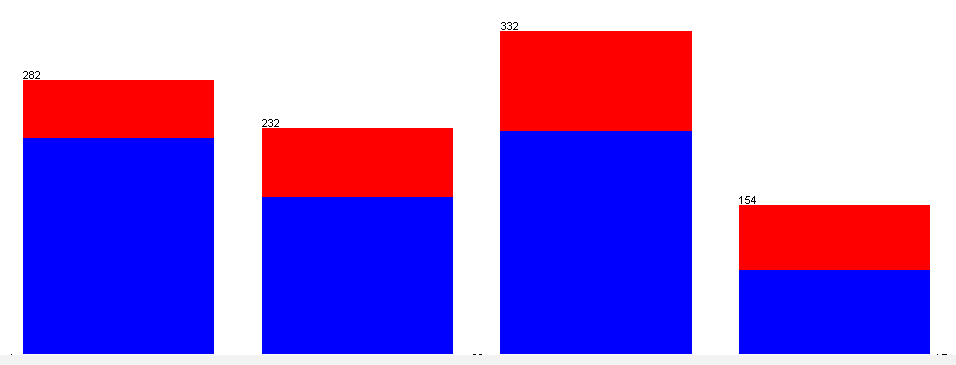
Número de años viviendo en la residencia actual.



**12. Property (Cualitativo)**

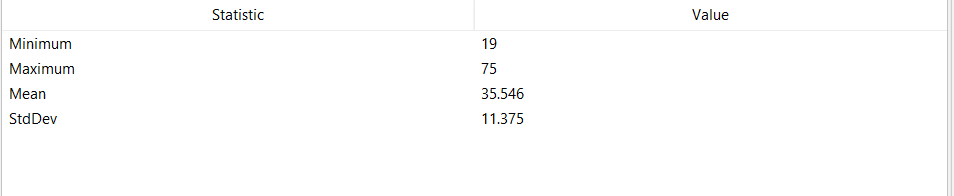
Tipo de propiedad del cliente:

* **A121**: Bienes raíces.
* **A122**: Ahorros en una sociedad de construcción / seguro de vida.
* **A123**: Coche u otro activo (no en el atributo 6).
* **A124**: Desconocido / Sin propiedad.



**13. Age in years (Numérico)**

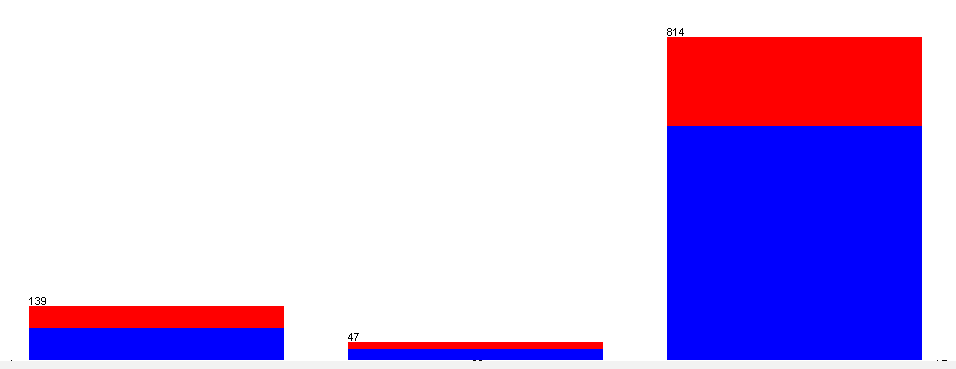
Edad del cliente en años.



**14. Other installment plans (Cualitativo)**

Planes de pago adicionales:

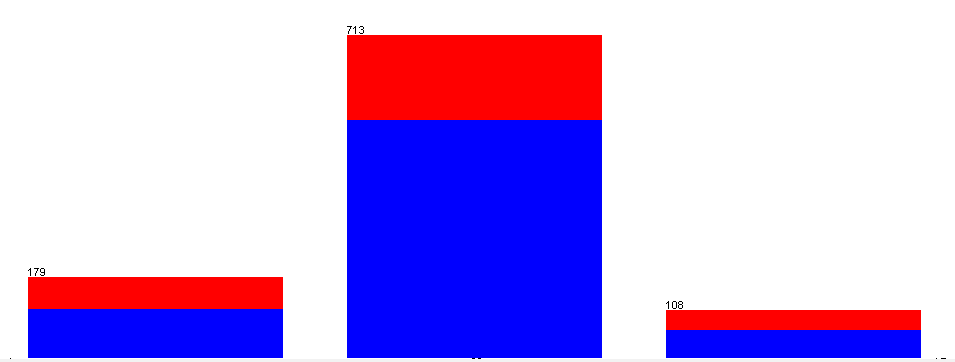
* **A141**: Banco.
* **A142**: Tiendas.
* **A143**: Ninguno.



**15. Housing (Cualitativo)**

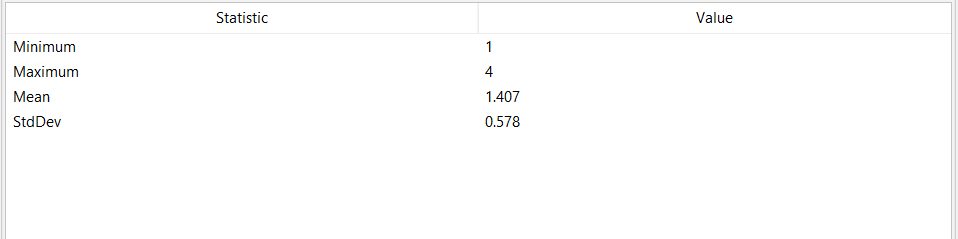
Tipo de vivienda:

* **A151**: Renta.
* **A152**: Propia.
* **A153**: Sin costo (por ejemplo, prestada).



**16. Number of existing credits at this bank (Numérico)**

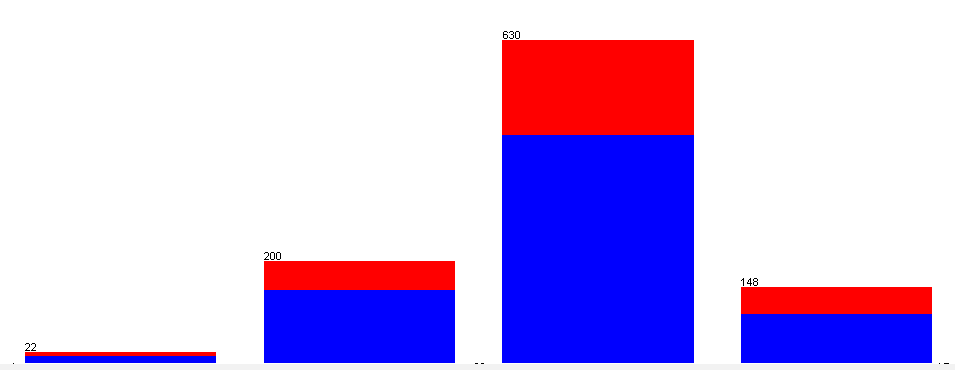
Cantidad de créditos existentes con este banco.



**17. Job (Cualitativo)**

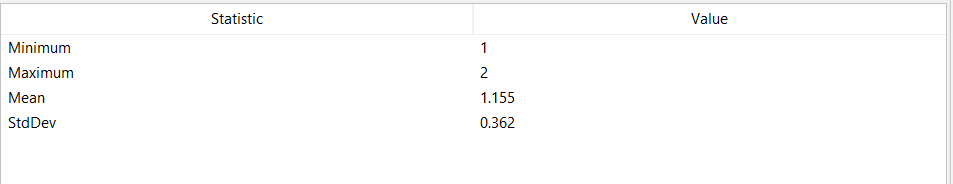
Categoría laboral del cliente:

* **A171**: Desempleado / No calificado (no residente).
* **A172**: No calificado (residente).
* **A173**: Empleado calificado / Oficial.
* **A174**: Alta gerencia / Autónomo / Profesional altamente calificado.



**18. Number of people being liable to provide maintenance for (Numérico)**

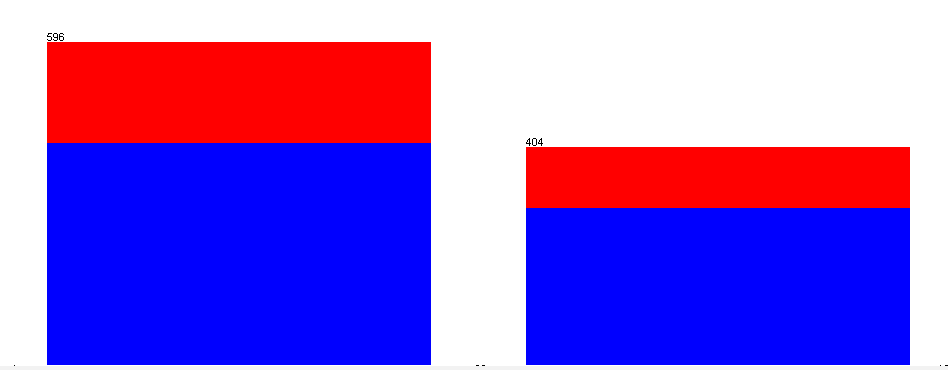
Número de personas dependientes del cliente.



**19. Telephone (Cualitativo)**

Disponibilidad de teléfono:

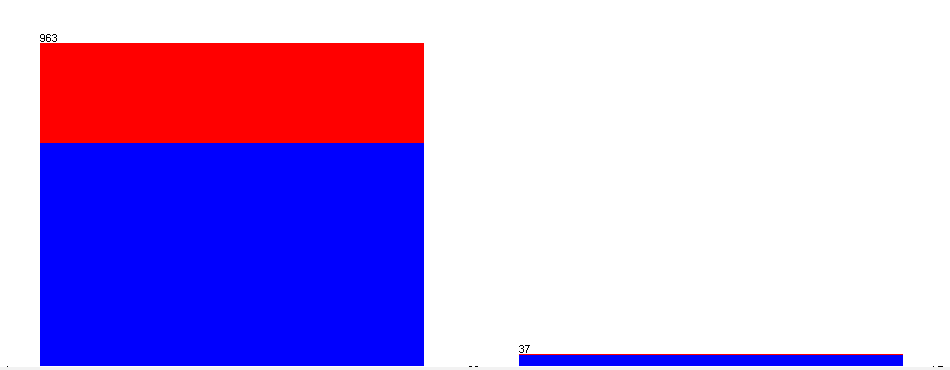
* **A191**: No tiene teléfono.
* **A192**: Sí, registrado a nombre del cliente.



**20. Foreign worker (Cualitativo)**

Si el cliente es un trabajador extranjero:

* **A201**: Sí.
* **A202**: No.



# **Ejemplos de uso**

### ****1. Segmentación de clientes****

#### **Objetivo:**

Identificar grupos de clientes con características similares para ofrecer productos financieros personalizados.

#### **Ejemplo:**

* **Segmentos según riesgo:**
  + Clientes con historial impecable (Credit history = A30) y altos ingresos (Savings >= 1000 DM).
  + Clientes con pagos atrasados (Credit history = A33) y poco tiempo en el empleo actual (Present employment < 1 año).
* **Segmentos según monto de crédito solicitado:**
  + Créditos pequeños (< 5000 DM): Pueden ofrecerse con menos requisitos.
  + Créditos grandes (≥ 10,000 DM): Requieren más evaluaciones.

### ****2. Evaluación de riesgo crediticio****

#### **Objetivo:**

Determinar las probabilidades de incumplimiento de pago basadas en patrones históricos.

#### **Ejemplo:**

* **Factores que más influyen en el incumplimiento:**
  + Historial crítico (Credit history = A34) combinado con ingresos bajos (Savings < 100 DM).
  + Altos montos de crédito (Credit amount > 15,000 DM) con tasa de pago alta (Installment rate > 30%).

#### **Resultados posibles:**

* Identificar características comunes de los clientes que incumplen.
* Crear reglas para denegar créditos automáticamente si ciertos criterios se cumplen.

### ****3. Análisis de la estabilidad financiera****

#### **Objetivo:**

Medir la estabilidad financiera de los clientes y su capacidad para manejar deudas.

#### **Ejemplo:**

* **Relación entre tiempo en el empleo y cumplimiento:**
  + Clientes con empleo estable (> 4 años) tienen menor probabilidad de incumplir.
  + Clientes desempleados tienen mayor probabilidad de incumplir.
* **Impacto de la propiedad:**
  + Clientes con propiedad (Property = A121) son menos riesgosos que aquellos sin ahorros ni bienes (Savings = A65 y Property = A124).

### ****4. Impacto de las características demográficas****

#### **Objetivo:**

Comprender cómo edad, estado civil, género y otros factores afectan el comportamiento financiero.

#### **Ejemplo:**

* **Edad:**
  + Jóvenes (< 25 años) tienden a solicitar créditos más pequeños y para propósitos como vacaciones o muebles.
  + Adultos (> 45 años) solicitan créditos más grandes para negocios o propiedades.
* **Estado civil y género:**
  + Solteros (Personal status = A93 o A95) tienen un historial más irregular en comparación con personas casadas (A94).
  + Mujeres (A92, A95) pueden ser estadísticamente más responsables en pagos.

### ****5. Análisis de productos financieros****

#### **Objetivo:**

Determinar cuáles productos financieros son más populares y rentables.

#### **Ejemplo:**

* **Propósito de los créditos:**
  + Los créditos para automóviles (Purpose = A40, A41) son los más comunes.
  + Los créditos para educación (Purpose = A46) tienen menor incidencia.
* **Relación entre propósito y cumplimiento:**
  + Créditos para negocios (Purpose = A49) tienen mayor tasa de incumplimiento.
  + Créditos para educación tienen la tasa de incumplimiento más baja.

### ****6. Optimización de políticas de crédito****

#### **Objetivo:**

Mejorar las reglas para la concesión de créditos y maximizar los beneficios minimizando el riesgo.

#### **Ejemplo:**

* **Simulación de tasas de interés:**
  + Clientes con alto riesgo (Credit history = A34, Savings < 500 DM) podrían recibir tasas de interés más altas.
* **Límites de crédito basados en ingresos:**
  + Créditos aprobados no deben exceder el 30% del ingreso disponible (Installment rate).

### ****7. Predicción del comportamiento futuro****

#### **Objetivo:**

Anticipar el comportamiento de los clientes en base a patrones históricos.

#### **Ejemplo:**

* Utilizar algoritmos de machine learning para predecir:
  + Probabilidad de incumplimiento.
  + Monto máximo de crédito que un cliente puede manejar.
  + Tiempo estimado para el reembolso completo de un crédito.

### ****8. Detección de fraude****

#### **Objetivo:**

Identificar patrones anómalos que puedan indicar intentos de fraude.

#### **Ejemplo:**

* Clientes con muchas solicitudes rechazadas pero cambiando atributos como propósito (Purpose) o monto solicitado.
* Combinaciones inusuales como altos créditos sin historial previo (Credit history = A30).

### ****9. Análisis de carga financiera****

#### **Objetivo:**

Calcular qué proporción del ingreso de los clientes está destinada al pago de deudas.

#### **Ejemplo:**

* Clientes con tasas de pago (Installment rate) superiores al 40% de sus ingresos pueden tener mayor probabilidad de incumplir.

### ****10. Tendencias de mercado****

#### **Objetivo:**

Analizar patrones en las solicitudes para entender la demanda de productos financieros.

#### **Ejemplo:**

* Incremento en créditos solicitados para automóviles (Purpose = A40, A41) puede reflejar una tendencia en la compra de vehículos.

# **Técnicas y Algoritmos**

## Arboles de decisión (J48):

* + Basado en el algoritmo C4.5.
  + Interpretable y muestra cómo se toman las decisiones.
  + Ventaja: Adecuado para datos cualitativos y numéricos.
  + Uso: Ideal para explicar cuáles son los factores más relevantes en la evaluación de riesgos.

*=== Run information ===*

*Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2*

*Relation: german\_credit*

*Instances: 1000*

*Attributes: 21*

*checking\_status*

*duration*

*credit\_history*

*purpose*

*credit\_amount*

*savings\_status*

*employment*

*installment\_commitment*

*personal\_status*

*other\_parties*

*residence\_since*

*property\_magnitude*

*age*

*other\_payment\_plans*

*housing*

*existing\_credits*

*job*

*num\_dependents*

*own\_telephone*

*foreign\_worker*

*class*

*Test mode: evaluate on training data*

*=== Classifier model (full training set) ===*

*J48 pruned tree*

*------------------*

*checking\_status = <0*

*| foreign\_worker = yes*

*| | duration <= 11*

*| | | existing\_credits <= 1*

*| | | | property\_magnitude = real estate: good (8.0/1.0)*

*| | | | property\_magnitude = life insurance*

*| | | | | own\_telephone = none: bad (2.0)*

*| | | | | own\_telephone = yes: good (4.0)*

*| | | | property\_magnitude = car: good (2.0/1.0)*

*| | | | property\_magnitude = no known property: bad (3.0)*

*| | | existing\_credits > 1: good (14.0)*

*| | duration > 11*

*| | | job = unemp/unskilled non res: bad (5.0/1.0)*

*| | | job = unskilled resident*

*| | | | purpose = new car*

*| | | | | own\_telephone = none: bad (10.0/2.0)*

*| | | | | own\_telephone = yes: good (2.0)*

*| | | | purpose = used car: bad (1.0)*

*| | | | purpose = furniture/equipment*

*| | | | | employment = unemployed: good (0.0)*

*| | | | | employment = <1: bad (3.0)*

*| | | | | employment = 1<=X<4: good (4.0)*

*| | | | | employment = 4<=X<7: good (1.0)*

*| | | | | employment = >=7: good (2.0)*

*| | | | purpose = radio/tv*

*| | | | | existing\_credits <= 1: bad (10.0/3.0)*

*| | | | | existing\_credits > 1: good (2.0)*

*| | | | purpose = domestic appliance: bad (1.0)*

*| | | | purpose = repairs: bad (1.0)*

*| | | | purpose = education: bad (1.0)*

*| | | | purpose = vacation: bad (0.0)*

*| | | | purpose = retraining: good (1.0)*

*| | | | purpose = business: good (3.0)*

*| | | | purpose = other: good (1.0)*

*| | | job = skilled*

*| | | | other\_parties = none*

*| | | | | duration <= 30*

*| | | | | | savings\_status = <100*

*| | | | | | | credit\_history = no credits/all paid: bad (8.0/1.0)*

*| | | | | | | credit\_history = all paid: bad (6.0)*

*| | | | | | | credit\_history = existing paid*

*| | | | | | | | own\_telephone = none*

*| | | | | | | | | existing\_credits <= 1*

*| | | | | | | | | | property\_magnitude = real estate*

*| | | | | | | | | | | age <= 26: bad (5.0)*

*| | | | | | | | | | | age > 26: good (2.0)*

*| | | | | | | | | | property\_magnitude = life insurance: bad (7.0/2.0)*

*| | | | | | | | | | property\_magnitude = car*

*| | | | | | | | | | | credit\_amount <= 1386: bad (3.0)*

*| | | | | | | | | | | credit\_amount > 1386: good (11.0/1.0)*

*| | | | | | | | | | property\_magnitude = no known property: good (2.0)*

*| | | | | | | | | existing\_credits > 1: bad (3.0)*

*| | | | | | | | own\_telephone = yes: bad (5.0)*

*| | | | | | | credit\_history = delayed previously: bad (4.0)*

*| | | | | | | credit\_history = critical/other existing credit: good (14.0/4.0)*

*| | | | | | savings\_status = 100<=X<500*

*| | | | | | | credit\_history = no credits/all paid: good (0.0)*

*| | | | | | | credit\_history = all paid: good (1.0)*

*| | | | | | | credit\_history = existing paid: bad (3.0)*

*| | | | | | | credit\_history = delayed previously: good (0.0)*

*| | | | | | | credit\_history = critical/other existing credit: good (2.0)*

*| | | | | | savings\_status = 500<=X<1000: good (4.0/1.0)*

*| | | | | | savings\_status = >=1000: good (4.0)*

*| | | | | | savings\_status = no known savings*

*| | | | | | | existing\_credits <= 1*

*| | | | | | | | own\_telephone = none: bad (9.0/1.0)*

*| | | | | | | | own\_telephone = yes: good (4.0/1.0)*

*| | | | | | | existing\_credits > 1: good (2.0)*

*| | | | | duration > 30: bad (30.0/3.0)*

*| | | | other\_parties = co applicant: bad (7.0/1.0)*

*| | | | other\_parties = guarantor: good (12.0/3.0)*

*| | | job = high qualif/self emp/mgmt: good (30.0/8.0)*

*| foreign\_worker = no: good (15.0/2.0)*

*checking\_status = 0<=X<200*

*| credit\_amount <= 9857*

*| | savings\_status = <100*

*| | | other\_parties = none*

*| | | | duration <= 42*

*| | | | | personal\_status = male div/sep: bad (8.0/2.0)*

*| | | | | personal\_status = female div/dep/mar*

*| | | | | | purpose = new car: bad (5.0/1.0)*

*| | | | | | purpose = used car: bad (1.0)*

*| | | | | | purpose = furniture/equipment*

*| | | | | | | duration <= 10: bad (3.0)*

*| | | | | | | duration > 10*

*| | | | | | | | duration <= 21: good (6.0/1.0)*

*| | | | | | | | duration > 21: bad (2.0)*

*| | | | | | purpose = radio/tv: good (8.0/2.0)*

*| | | | | | purpose = domestic appliance: good (0.0)*

*| | | | | | purpose = repairs: good (1.0)*

*| | | | | | purpose = education: good (4.0/2.0)*

*| | | | | | purpose = vacation: good (0.0)*

*| | | | | | purpose = retraining: good (0.0)*

*| | | | | | purpose = business*

*| | | | | | | residence\_since <= 2: good (3.0)*

*| | | | | | | residence\_since > 2: bad (2.0)*

*| | | | | | purpose = other: good (0.0)*

*| | | | | personal\_status = male single: good (52.0/15.0)*

*| | | | | personal\_status = male mar/wid*

*| | | | | | duration <= 10: good (6.0)*

*| | | | | | duration > 10: bad (10.0/3.0)*

*| | | | | personal\_status = female single: good (0.0)*

*| | | | duration > 42: bad (7.0)*

*| | | other\_parties = co applicant: good (2.0)*

*| | | other\_parties = guarantor*

*| | | | purpose = new car: bad (2.0)*

*| | | | purpose = used car: good (0.0)*

*| | | | purpose = furniture/equipment: good (0.0)*

*| | | | purpose = radio/tv: good (18.0/1.0)*

*| | | | purpose = domestic appliance: good (0.0)*

*| | | | purpose = repairs: good (0.0)*

*| | | | purpose = education: good (0.0)*

*| | | | purpose = vacation: good (0.0)*

*| | | | purpose = retraining: good (0.0)*

*| | | | purpose = business: good (0.0)*

*| | | | purpose = other: good (0.0)*

*| | savings\_status = 100<=X<500*

*| | | purpose = new car: bad (15.0/5.0)*

*| | | purpose = used car: good (3.0)*

*| | | purpose = furniture/equipment: bad (4.0/1.0)*

*| | | purpose = radio/tv: bad (8.0/2.0)*

*| | | purpose = domestic appliance: good (0.0)*

*| | | purpose = repairs: good (2.0)*

*| | | purpose = education: good (0.0)*

*| | | purpose = vacation: good (0.0)*

*| | | purpose = retraining: good (0.0)*

*| | | purpose = business*

*| | | | housing = rent*

*| | | | | existing\_credits <= 1: good (2.0)*

*| | | | | existing\_credits > 1: bad (2.0)*

*| | | | housing = own: good (6.0)*

*| | | | housing = for free: bad (1.0)*

*| | | purpose = other: good (1.0)*

*| | savings\_status = 500<=X<1000: good (11.0/3.0)*

*| | savings\_status = >=1000: good (13.0/3.0)*

*| | savings\_status = no known savings: good (41.0/5.0)*

*| credit\_amount > 9857: bad (20.0/3.0)*

*checking\_status = >=200: good (63.0/14.0)*

*checking\_status = no checking: good (394.0/46.0)*

*Number of Leaves : 103*

*Size of the tree : 140*

*Time taken to build model: 0.07 seconds*

*=== Evaluation on training set ===*

*Time taken to test model on training data: 0.02 seconds*

*=== Summary ===*

*Correctly Classified Instances 855 85.5 %*

*Incorrectly Classified Instances 145 14.5 %*

*Kappa statistic 0.6251*

*Mean absolute error 0.2312*

*Root mean squared error 0.34*

*Relative absolute error 55.0377 %*

*Root relative squared error 74.2015 %*

*Total Number of Instances 1000*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0,956 0,380 0,854 0,956 0,902 0,640 0,857 0,905 good*

*0,620 0,044 0,857 0,620 0,720 0,640 0,857 0,783 bad*

*Weighted Avg. 0,855 0,279 0,855 0,855 0,847 0,640 0,857 0,869*

*=== Confusion Matrix ===*

*a b <-- classified as*

*669 31 | a = good*

*114 186 | b = bad*

**Resumen del Modelo**

1. **Esquema del Modelo**:
   * Clasificador: weka.classifiers.trees.J48
   * Parámetros:
     + Confianza (-C): 0.25 (nivel de poda)
     + Número mínimo de instancias por hoja (-M): 2
2. **Datos de Entrenamiento**:
   * Número de instancias: 1000
   * Número de atributos: 21, incluido el atributo objetivo (class)
3. **Atributos Utilizados**:
   * **checking\_status**, **duration**, **foreign\_worker**, **job**, **purpose**, entre otros.
   * El atributo objetivo (class) toma valores "good" o "bad".
4. **Profundidad del Árbol**:
   * El árbol es bastante detallado, reflejando una segmentación compleja basada en reglas específicas derivadas de los atributos.
5. **Evaluación**:
   * El modelo fue evaluado sobre el **conjunto completo de entrenamiento**, lo que significa que la precisión reportada será probablemente más alta debido al sobreajuste.

**Análisis del Árbol de Decisión**

1. **División Raíz**:
   * El atributo **checking\_status** es la raíz del árbol, lo que indica que el estado de la cuenta bancaria del cliente es el predictor más importante.
2. **Niveles de Anidación**:
   * El árbol tiene múltiples niveles de condiciones anidadas, donde se toman decisiones basadas en atributos como foreign\_worker, duration, job, y otros.
3. **Poda**:
   * La poda se realiza para evitar el sobreajuste. Esto se refleja en las fracciones como (8.0/1.0), donde el clasificador ajusta reglas generales y permite algunas excepciones.
4. **Ejemplo de Reglas Extraídas**:
   * Si checking\_status = <0 y foreign\_worker = yes y duration <= 11 y existing\_credits <= 1 y property\_magnitude = real estate, entonces la predicción es **good** con (8.0/1.0) casos.
5. **Casos Ambiguos**:
   * Algunas ramas muestran excepciones o incertidumbre, como (14.0/4.0), donde 14 instancias cumplen la regla, pero 4 son clasificaciones incorrectas.

## Regresión Logística:

*=== Run information ===*

*Scheme: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4*

*Relation: german\_credit*

*Instances: 1000*

*Attributes: 21*

*checking\_status*

*duration*

*credit\_history*

*purpose*

*credit\_amount*

*savings\_status*

*employment*

*installment\_commitment*

*personal\_status*

*other\_parties*

*residence\_since*

*property\_magnitude*

*age*

*other\_payment\_plans*

*housing*

*existing\_credits*

*job*

*num\_dependents*

*own\_telephone*

*foreign\_worker*

*class*

*Test mode: evaluate on training data*

*=== Classifier model (full training set) ===*

*Logistic Regression with ridge parameter of 1.0E-8*

*Coefficients...*

*Class*

*Variable good*

*========================================================*

*checking\_status=<0 -0.778*

*checking\_status=0<=X<200 -0.4032*

*checking\_status=>=200 0.1877*

*checking\_status=no checking 0.9339*

*duration -0.0279*

*credit\_history=no credits/all paid -0.8129*

*credit\_history=all paid -0.9562*

*credit\_history=existing paid -0.2268*

*credit\_history=delayed previously 0.0403*

*credit\_history=critical/other existing credit 0.6229*

*purpose=new car -0.692*

*purpose=used car 0.9744*

*purpose=furniture/equipment 0.0996*

*purpose=radio/tv 0.1996*

*purpose=domestic appliance -0.1692*

*purpose=repairs -0.4756*

*purpose=education -0.7283*

*purpose=vacation 0*

*purpose=retraining 1.3674*

*purpose=business 0.0481*

*purpose=other 0.7967*

*credit\_amount -0.0001*

*savings\_status=<100 -0.4402*

*savings\_status=100<=X<500 -0.0825*

*savings\_status=500<=X<1000 -0.0641*

*savings\_status=>=1000 0.8989*

*savings\_status=no known savings 0.5064*

*employment=unemployed -0.2934*

*employment=<1 -0.2265*

*employment=1<=X<4 -0.1106*

*employment=4<=X<7 0.5376*

*employment=>=7 -0.0168*

*installment\_commitment -0.3301*

*personal\_status=male div/sep -0.4922*

*personal\_status=female div/dep/mar -0.2168*

*personal\_status=male single 0.3238*

*personal\_status=male mar/wid -0.1252*

*personal\_status=female single 0*

*other\_parties=none -0.1798*

*other\_parties=co applicant -0.6158*

*other\_parties=guarantor 0.7988*

*residence\_since -0.0048*

*property\_magnitude=real estate 0.2572*

*property\_magnitude=life insurance -0.0242*

*property\_magnitude=car 0.0627*

*property\_magnitude=no known property -0.4732*

*age 0.0145*

*other\_payment\_plans=bank -0.3273*

*other\_payment\_plans=stores -0.2041*

*other\_payment\_plans=none 0.3191*

*housing=rent -0.3497*

*housing=own 0.0939*

*housing=for free 0.3341*

*existing\_credits -0.2721*

*job=unemp/unskilled non res 0.5096*

*job=unskilled resident -0.0265*

*job=skilled -0.0451*

*job=high qualif/self emp/mgmt 0.0301*

*num\_dependents -0.2647*

*own\_telephone=yes 0.3*

*foreign\_worker=no 1.3922*

*Intercept 3.1983*

*Odds Ratios...*

*Class*

*Variable good*

*========================================================*

*checking\_status=<0 0.4593*

*checking\_status=0<=X<200 0.6682*

*checking\_status=>=200 1.2064*

*checking\_status=no checking 2.5443*

*duration 0.9725*

*credit\_history=no credits/all paid 0.4436*

*credit\_history=all paid 0.3843*

*credit\_history=existing paid 0.7971*

*credit\_history=delayed previously 1.0411*

*credit\_history=critical/other existing credit 1.8643*

*purpose=new car 0.5006*

*purpose=used car 2.6496*

*purpose=furniture/equipment 1.1047*

*purpose=radio/tv 1.2209*

*purpose=domestic appliance 0.8443*

*purpose=repairs 0.6215*

*purpose=education 0.4827*

*purpose=vacation 1*

*purpose=retraining 3.925*

*purpose=business 1.0492*

*purpose=other 2.2182*

*credit\_amount 0.9999*

*savings\_status=<100 0.6439*

*savings\_status=100<=X<500 0.9208*

*savings\_status=500<=X<1000 0.9379*

*savings\_status=>=1000 2.457*

*savings\_status=no known savings 1.6593*

*employment=unemployed 0.7457*

*employment=<1 0.7973*

*employment=1<=X<4 0.8953*

*employment=4<=X<7 1.7119*

*employment=>=7 0.9834*

*installment\_commitment 0.7189*

*personal\_status=male div/sep 0.6113*

*personal\_status=female div/dep/mar 0.8051*

*personal\_status=male single 1.3824*

*personal\_status=male mar/wid 0.8824*

*personal\_status=female single 1*

*other\_parties=none 0.8354*

*other\_parties=co applicant 0.5402*

*other\_parties=guarantor 2.2229*

*residence\_since 0.9952*

*property\_magnitude=real estate 1.2933*

*property\_magnitude=life insurance 0.9761*

*property\_magnitude=car 1.0647*

*property\_magnitude=no known property 0.623*

*age 1.0146*

*other\_payment\_plans=bank 0.7209*

*other\_payment\_plans=stores 0.8154*

*other\_payment\_plans=none 1.3758*

*housing=rent 0.7049*

*housing=own 1.0984*

*housing=for free 1.3967*

*existing\_credits 0.7618*

*job=unemp/unskilled non res 1.6647*

*job=unskilled resident 0.9738*

*job=skilled 0.9559*

*job=high qualif/self emp/mgmt 1.0306*

*num\_dependents 0.7674*

*own\_telephone=yes 1.3499*

*foreign\_worker=no 4.0237*

*Time taken to build model: 0.05 seconds*

*=== Evaluation on training set ===*

*Time taken to test model on training data: 0 seconds*

*=== Summary ===*

*Correctly Classified Instances 786 78.6 %*

*Incorrectly Classified Instances 214 21.4 %*

*Kappa statistic 0.4563*

*Mean absolute error 0.2921*

*Root mean squared error 0.3823*

*Relative absolute error 69.5095 %*

*Root relative squared error 83.4247 %*

*Total Number of Instances 1000*

*=== Detailed Accuracy By Class ===*

*TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class*

*0,894 0,467 0,817 0,894 0,854 0,463 0,834 0,918 good*

*0,533 0,106 0,684 0,533 0,599 0,463 0,834 0,688 bad*

*Weighted Avg. 0,786 0,358 0,777 0,786 0,778 0,463 0,834 0,849*

*=== Confusion Matrix ===*

*a b <-- classified as*

*626 74 | a = good*

*140 160 | b = bad*

**Análisis del modelo de regresión logística**

**1. Modelo y Coeficientes**

La regresión logística predice la probabilidad de pertenecer a una clase (good o bad) en función de las características del cliente. Los **coeficientes** determinan cómo cada variable afecta esa probabilidad.

* **Coeficientes (Class: good):**
  + Los valores negativos indican que la variable disminuye la probabilidad de que el cliente sea clasificado como good.
  + Los valores positivos indican que la variable aumenta la probabilidad de que el cliente sea clasificado como good.

Por ejemplo:

* checking\_status=<0 tiene un coeficiente de -0.778, lo que significa que tener un saldo menor a 0 reduce la probabilidad de ser clasificado como good.
* foreign\_worker=no tiene un coeficiente de 1.3922, lo que significa que no ser un trabajador extranjero aumenta considerablemente la probabilidad de ser clasificado como good.

**2. Odds Ratios**

El modelo también calcula los **odds ratios**, que son la exponencial de los coeficientes. Los odds ratios muestran cuánto se multiplican las probabilidades de pertenecer a la clase good al cambiar el valor de una variable.

* **Valores > 1:** Aumentan la probabilidad de que el cliente sea good.
* **Valores < 1:** Reducen la probabilidad de que el cliente sea good.

Ejemplo:

* checking\_status=no checking tiene un odds ratio de 2.5443, lo que significa que los clientes sin cuenta corriente tienen 2.54 veces más probabilidades de ser clasificados como good.
* credit\_history=all paid tiene un odds ratio de 0.3843, lo que indica que haber pagado todos los créditos anteriores disminuye la probabilidad de ser good.

**3. Evaluación del Modelo**

* **Precisión General:**  
  El modelo clasifica correctamente el 78.6% de las instancias en el conjunto de entrenamiento. Esto sugiere un rendimiento decente, aunque con margen de mejora.
* **Métricas por Clase:**
  + Para la clase good:
    - **TP Rate (Recall):** 0.894 → El modelo identifica correctamente el 89.4% de los casos good.
    - **Precision:** 0.817 → El 81.7% de las predicciones good son correctas.
  + Para la clase bad:
    - **TP Rate (Recall):** 0.533 → El modelo identifica correctamente el 53.3% de los casos bad.
    - **Precision:** 0.684 → El 68.4% de las predicciones bad son correctas. Esto indica que el modelo es más efectivo identificando casos good que bad.
* **Kappa:**  
  Un valor de 0.4563 indica un nivel de acuerdo moderado entre las predicciones del modelo y las clases reales, más allá del azar.
* **Errores:**
  + **Mean Absolute Error (MAE):** 0.2921 → El promedio de error absoluto en las probabilidades predichas es del 29.21%.
  + **Root Mean Squared Error (RMSE):** 0.3823 → Indica la magnitud promedio del error, siendo más sensible a valores extremos.
* **Curva ROC (Area bajo la curva):**  
  El área bajo la curva ROC es 0.834, lo que sugiere un buen desempeño en la separación entre las clases good y bad.