**UNIVERSIDAD AUTÓNOMA GABRIEL RENÉ MORENO**

**FACULTAD DE INGENIERÍA Y CIENCIAS EN LA COMPUTACIÓN Y TELECOMUNICACIONES**

# PROYECTO 2

# WEKA I credit-g


**INTEGRANTES:**

- Mamani Rodriguez Kasandra
- Sejas Mamani Dennis
- Sabinas Brayan

**DOCENTE:** Ing.


**ASIGNATURA:** Soporte para la Toma de Decisiones

**SANTA CRUZ – BOLIVIA**

## ÍNDICE GENERAL

# 1.OBTENCIÓN DE DATOS?

Para entender cómo se obtuvieron los datos utilizados en el proyecto **Credit-G** del conjunto **German Credit Data**, es necesario analizar las características del dataset y el contexto en el que fue creado, a continuación se detalla el proceso general que puede haberse seguido para la generación de este tipo de dataset:

## 1.1.Recopilación de datos

Los datos del German Credit Dataset probablemente fueron recolectados por una institución financiera o un banco alemán. Este conjunto contiene información de clientes relacionada con su historial crediticio y datos demográficos, utilizados para evaluar la probabilidad de que un cliente cumpla con sus obligaciones de pago.

### Fuentes de datos:

- **Historial crediticio**: Los datos fueron extraídos de los registros internos del banco, como:
    - Préstamos anteriores.
    - Información de pagos atrasados o incumplidos.
    - Relación con otras instituciones crediticias.
- **Información del cliente**: Recopilada al momento de solicitar el crédito, incluyendo:
    - Edad, género, ocupación.
    - Estado civil y número de dependientes.
    - Ingresos y propiedades (p. ej., casa o automóvil).
- **Comportamiento transaccional**: Datos sobre el uso de servicios bancarios, como cuentas de ahorro o inversión.

## 1.2.Procesamiento inicial de los datos

Los datos brutos se someten a un procesamiento previo para estructurarlos adecuadamente. Esto incluye:

- **Conversión a formato tabular**: Los datos se organizan en filas (individuos) y columnas (atributos).
- **Definición de la clase objetivo**: En este caso, los registros se etiquetan como:
    - **Good (Buen riesgo)**: Clientes que cumplen con los pagos.
    - **Bad (Mal riesgo)**: Clientes que incumplen.
- **Eliminación de datos irrelevantes o confidenciales**: Información personal sensible se excluye para proteger la privacidad.

## 1.3. Normalización y codificación

Dado que los datos provienen de diferentes fuentes, se asegura la uniformidad mediante:

- **Normalización**: Atributos como montos crediticios o duración de préstamos son escalados a valores estándar.
- **Codificación de atributos categóricos**: Variables como "historial crediticio" o "estado civil" se convierten en valores discretos o numéricos. Ejemplo:
    - Historial crediticio: [1 = excelente, 2 = bueno, 3 = malo].
    - Estado civil: [1 = soltero, 2 = casado, 3 = divorciado].

## 1.4. Etiquetado manual o automático

La clase objetivo (Good/Bad) pudo haberse determinado de dos maneras:

- **Manual**: Expertos financieros evalúan el riesgo basado en políticas del banco y asignan etiquetas.
- **Automática**: Uso de reglas predefinidas, como:
    - Si un cliente tiene más de tres pagos atrasados → **Bad**.
    - Si tiene un ingreso estable y sin historial de incumplimientos → **Good**.

### 1.5. Validación y limpieza

Los datos pasan por una fase de validación para garantizar consistencia y calidad:

- **Eliminación de duplicados**: Clientes con múltiples solicitudes.
- **Manejo de valores faltantes**: Relleno con promedio, mediana o eliminación de registros incompletos.
- **Balanceo de clases**: Ajuste del número de ejemplos de "Good" y "Bad" para evitar sesgo en el análisis.

## 2. ¿QUÉ SIGNIFICA CADA DATO?

Este conjunto de datos tiene 20 atributos que representan características financieras y demográficas de los solicitantes de crédito, así como la clase objetivo:

### Clase Objetivo

- **class**:
    - **Good (buen riesgo)**: El solicitante tiene un bajo riesgo de incumplir con los pagos.
    - **Bad (mal riesgo)**: El solicitante tiene un alto riesgo de incumplir con los pagos.
    -

## Atributos del dataset

### 1. Status of existing checking account

- Estado de la cuenta corriente del solicitante:
    - **A11**: Ninguna cuenta.
    - **A12**: Saldo < 0 DM (saldo negativo).
    - **A13**: 0 ≤ Saldo < 200 DM.
    - **A14**: Saldo ≥ 200 DM.

### 2. Duration in months

- Duración del crédito solicitado (en meses).

### 3. Credit history

- Historial crediticio del solicitante:
    - **A30**: Sin créditos tomados o todos los créditos pagados a tiempo.
    - **A31**: Todos los créditos pagados a tiempo.
    - **A32**: Retrasos en los pagos previos.
    - **A33**: Problemas críticos en el historial crediticio.
    - **A34**: Otros créditos en proceso.

### 4. Purpose

- Propósito del crédito:
    - **A40**: Automóvil (nuevo).
    - **A41**: Automóvil (usado).
    - **A42**: Muebles/electrodomésticos.
    - **A43**: Radio/TV.

- **A44**: Electrodomésticos.
- **A45**: Reparaciones.
- **A46**: Educación.
- **A47**: Vacaciones/recreación.
- **A48**: Créditos en proceso.
- **A49**: Negocios.
- **A410**: Otros.

**5. Credit amount**

- Monto del crédito solicitado (en DM).

**6. Savings account/bonds**

- Ahorros o bonos del solicitante:
  - **A61**: Ninguno.
  - **A62**: < 100 DM.
  - **A63**: 100 ≤ Saldo < 500 DM.
  - **A64**: 500 ≤ Saldo < 1000 DM.
  - **A65**: ≥ 1000 DM o más.

**7. Present employment since**

- Tiempo de empleo actual:
  - **A71**: Desempleado.
  - **A72**: < 1 año.
  - **A73**: 1 ≤ Tiempo < 4 años.
  - **A74**: 4 ≤ Tiempo < 7 años.
  - **A75**: ≥ 7 años.

**8. Installment rate in percentage of disposable income**

- Porcentaje del ingreso disponible destinado a pagar el crédito (1-4).

**9. Personal status and sex**

- Estado civil y género:
    - **A91**: Hombre, soltero.
    - **A92**: Hombre, casado o viudo.
    - **A93**: Mujer, soltera.
    - **A94**: Mujer, casada o viuda.
    - **A95**: Hombre, divorciado o separado.

**10. Other debtors/guarantors**

- Otros deudores o garantes:
    - **A101**: Ninguno.
    - **A102**: Codeudor.
    - **A103**: Garante.

**11. Present residence since**

- Tiempo de residencia actual (en años).

**12. Property**

- Tipo de propiedad del solicitante:
    - **A121**: Propiedad inmobiliaria.
    - **A122**: Ahorros en seguros.
    - **A123**: Automóviles o bienes.

    - **A124**: Ninguna propiedad.

**13. Age in years**

- Edad del solicitante (en años).

**14. Other installment plans**

- Planes de crédito adicionales:
    - **A141**: Ninguno.
    - **A142**: En el banco.
    - **A143**: En las tiendas.

**15. Housing**

- Tipo de vivienda:
    - **A151**: Rentada.
    - **A152**: Propia.
    - **A153**: Viviendo con los padres.

**16. Number of existing credits at this bank**

- Número de créditos existentes en este banco.

**17. Job**

- Categoría laboral del solicitante:
    - **A171**: Desempleado o sin empleo calificado.
    - **A172**: Empleo no calificado.
    - **A173**: Empleo calificado o administrativo.
    - **A174**: Empleo altamente calificado.

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

- Número de dependientes financieros del solicitante (1-2).

**19. Telephone**

- Disponibilidad de teléfono:
    - **A191**: No tiene teléfono.
    - **A192**: Tiene teléfono registrado.

**20. Foreign worker**

- Si el solicitante es un trabajador extranjero:
    - **A201**: Sí.
    - **A202**: No.

## 3. ¿CUAL ES EL OBJETIVO DE ESOS DATOS?

El **objetivo** de los datos en el conjunto **German Credit Dataset** (credit-g.arff) es proporcionar un caso práctico para **evaluar el riesgo crediticio** de los clientes de una entidad financiera. Este dataset sirve como base para la construcción y validación de modelos de aprendizaje automático que clasifiquen a los solicitantes en dos categorías principales:

Las dos categorías principales en el conjunto de datos **German Credit Dataset** (credit-g.arff) son:

1. **Good (Buen riesgo)**
    - Representa a los solicitantes que tienen un **bajo riesgo de incumplimiento** en sus pagos.

    - Estos solicitantes son considerados como confiables por la entidad financiera y tienen alta probabilidad de cumplir con las condiciones del crédito.
2. **Bad (Mal riesgo)**
    - Representa a los solicitantes que tienen un **alto riesgo de incumplimiento** en sus pagos.
    - Estos solicitantes son considerados como no confiables o de alto riesgo por la entidad financiera, lo que podría resultar en la negativa del crédito o en la aplicación de condiciones más estrictas (como tasas de interés más altas).

## Objetivo principal

Predecir si un solicitante de crédito es un **"buen riesgo"** o un **"mal riesgo"**, basándose en sus características financieras, demográficas y comportamentales.

## Contexto práctico

En la vida real, las instituciones financieras y bancos utilizan este tipo de modelos para tomar decisiones informadas sobre:

1. **Aprobación o rechazo de créditos**.
2. **Condiciones del crédito** (tasa de interés, monto permitido, plazo de pago).
3. **Gestión del portafolio de riesgo** para minimizar pérdidas y maximizar ganancias.

## Aplicaciones del dataset

1. **Entrenamiento de modelos de clasificación**:
    - Usar algoritmos de Machine Learning (como árboles de decisión, redes neuronales, SVM, etc.) para aprender patrones que distingan entre clientes buenos y malos.

2. **Validación y comparación de algoritmos**:
    - Evaluar el rendimiento de diferentes técnicas para determinar cuál es más adecuada para este tipo de problema.
3. **Análisis de características importantes**:
    - Identificar cuáles son los atributos más relevantes que influyen en el riesgo crediticio (por ejemplo, duración del crédito, historial crediticio o monto del préstamo).
4. **Simulación y predicción**:
    - Simular escenarios de aprobación/rechazo basados en los datos de entrada de nuevos clientes.

## Beneficios del análisis

1. **Minimizar riesgos financieros**:
    - Reducir las probabilidades de otorgar créditos a clientes que no puedan pagar.
2. **Optimizar recursos**:
    - Focalizar los esfuerzos de la institución financiera en clientes con mayor probabilidad de cumplir con sus obligaciones.
3. **Toma de decisiones basadas en datos**:
    - Ofrecer un respaldo objetivo y cuantitativo a los analistas de riesgos y gerentes de crédito.