# Análisis del Dataset South German Credit y Documentación con ML Canvas

## 1. Análisis del problema y del dataset “South German Credit (Update)”

El dataset South German Credit contiene 1,000 ejemplos (700 créditos buenos, 300 malos) con 20 variables predictoras más la variable objetivo ('credit\_risk': bueno/malo). Es un problema clásico de clasificación binaria orientado a predecir si un solicitante de crédito será buen o mal pagador, con aplicaciones directas en el ámbito financiero.

### 1.1 Naturaleza del problema

El objetivo es construir un modelo de clasificación binaria que prediga la probabilidad de impago de un solicitante. Esto permite a una entidad financiera tomar decisiones más informadas sobre aprobación o rechazo de crédito, reduciendo pérdidas por morosidad.

### 1.2 Variables principales

El dataset contiene variables relacionadas con historial crediticio, empleo, ahorros, propiedades, estado civil, edad y más. No hay valores faltantes y las variables están discretizadas o codificadas.

### 1.3 Principales retos, sesgos y riesgos

- Desbalance entre clases reales y dataset.  
- Posibles sesgos demográficos (e.g., trabajador extranjero).  
- Datos antiguos (años 70s) que limitan la vigencia.  
- Necesidad de explicabilidad en modelos financieros.  
- Costos de error asimétricos (falsos negativos más costosos).  
- Riesgo de multicolinealidad entre variables.

### 1.4 Oportunidades

El dataset es ideal para experimentación, prototipado y benchmarking en proyectos de scoring crediticio. Permite comparar distintos algoritmos y técnicas de ingeniería de características.

## 2. Documentación con Machine Learning Canvas

El ML Canvas describe de forma estructurada los elementos de un proyecto de machine learning: objetivo, fuentes de datos, características, predicción, decisiones y evaluación. A continuación, se presenta la versión adaptada para el dataset South German Credit.

### Machine Learning Canvas - South German Credit

* Objetivo / Propuesta de valor:

Construir un modelo de scoring de riesgo crediticio para clasificar solicitantes entre riesgo bajo y alto, reduciendo pérdidas y aumentando eficiencia en otorgamientos.

* Fuentes de datos:

Datos históricos de créditos, historial de pagos, burós de crédito, datos macroeconómicos.

* Recolección de datos:

Información obtenida al solicitar un crédito, más registros internos y externos de comportamiento de pago.

* Características (feature engineering):

Variables base del dataset más variables derivadas (ratios, transformaciones, interacciones).

* Construcción / actualización de modelos:

Entrenamiento inicial con datos históricos y reentrenamiento periódico con nuevos datos.

* Tarea ML / Predicción:

Clasificación binaria (default sí/no) usando regresión logística, árboles, ensemble o redes neuronales.

* Decisiones:

Usar la probabilidad estimada para aprobar, rechazar o ajustar condiciones del crédito.

* Momento de predicción:

En el momento de la solicitud o para monitorear cartera existente.

* Evaluación offline (validación):

Métricas: AUC-ROC, sensibilidad, especificidad, matriz de costos.

* Evaluación en producción / monitoreo:

Comparar predicciones con defaults reales, detectar drift y degradación del modelo.

* Feedback / retroalimentación:

Incorporar nuevos resultados de pago como etiquetas para reentrenar periódicamente.

* Limitaciones / riesgos / supuestos:

Sesgos de datos, falta de representatividad, cambios regulatorios, posibles variables discriminatorias.

## 3. Propuesta de valor del proyecto

La implementación de un modelo de scoring crediticio basado en aprendizaje automático ofrece valor tangible a una institución financiera al mejorar la toma de decisiones, reducir pérdidas, aumentar la rentabilidad y permitir procesos más ágiles y auditables.

### Propuesta de valor general

- Incrementar la tasa de aprobación segura de créditos.  
- Reducir pérdidas por impagos.  
- Automatizar la decisión crediticia.  
- Mejorar la transparencia mediante modelos explicables.  
- Adaptación continua con datos actualizados.

### Indicadores de éxito (KPI)

- Reducción en tasa de impago.  
- Mejora en AUC/ROC del modelo.  
- Aumento de créditos aprobados sin elevar riesgo.  
- Ahorro operativo.  
- ROI positivo del proyecto.

### Riesgos asociados

- Falta de generalización del modelo.  
- Variables no disponibles en producción.  
- Cambios regulatorios.  
- Resistencia organizacional al cambio.