Statlog (German Credit Data) Datasheet

i) Motivación

El conjunto de datos *Statlog (German Credit Data)* fue creado con el propósito de servir como caso de estudio para comparar distintos enfoques de clasificación: métodos estadísticos tradicionales, algoritmos de aprendizaje automático y redes neuronales aplicados a problemas reales. El conjunto de datos se utilizó como benchmark dentro del proyecto *Statlog*, cuyo objetivo fue comparar distintos métodos de clasificación sobre problemas reales, promoviendo la creación de un estándar común de evaluación.

El conjunto de datos original fue compilado por el Profesor Hans Hofmann de la Universidad de Hamburgo. Posteriormente, el grupo de la Strathclyde University (Reino Unido) adaptó el dataset para incluir únicamente variables numéricas, facilitando su uso con algoritmos que no admiten atributos categóricos.

ii) Composición

Cada instancia representa una solicitud individual de crédito realizada a una entidad bancaria alemana, evaluada según una serie de atributos personales y financieros. Entre ellos se incluyen variables como la duración del crédito solicitado, el propósito del préstamo, la edad del solicitante, el empleo y el monto del crédito. El conjunto de datos está estructurado como una colección de 1000 observaciones, cada una con 20 atributos descriptivos y una variable objetivo binaria que indica si la solicitud se considera “riesgo bueno” o “riesgo malo”.

iii) Proceso de recopilación

Las fuentes consultadas no especifican el procedimiento exacto mediante el cual se obtuvieron los datos. Sin embargo, dada la naturaleza de las variables, se infiere que los registros provienen de solicitudes reales de crédito recopiladas por un banco alemán y anonimizadas antes de su publicación. No se dispone de información sobre validaciones o verificaciones específicas realizadas durante la recopilación.

iv) Preprocesamiento, limpieza y etiquetado

El conjunto de datos original incluye tanto variables categóricas como numéricas. En el marco del proyecto *Statlog*, la Strathclyde University llevó a cabo una transformación que convirtió los atributos categóricos en variables numéricas codificadas, de forma que fuera compatible con algoritmos que requieren entradas exclusivamente numéricas.

Para el presente trabajo se utilizará la versión original del conjunto de datos, ya que mantiene la interpretabilidad de las variables categóricas, aspecto relevante para el análisis de equidad y transparencia en modelos de clasificación.

v) Usos

En la publicación original (*Machine Learning, Neural and Statistical Classification*, 1994), el conjunto fue empleado para comparar el desempeño de más de veinte algoritmos de clasificación, incluyendo modelos estadísticos y de aprendizaje automático. Se analizaron métricas de precisión, interpretabilidad y robustez, así como el impacto de costos de clasificación asimétricos, reflejando la importancia del contexto financiero del problema.

Por otro lado, existen numerosos trabajos científicos y de investigación que utilizan el conjunto de datos. El *UCI Machine Learning Repository* mantiene un registro de los documentos que lo citan. La versión disponible puede consultarse en:
Statlog (German Credit Data) - UCI Machine Learning Repository

Referencias

- UCI Machine Learning Repository. (1994). *Statlog (German Credit Data)*. Recuperado de https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

- Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (Eds.). (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood. https://www.researchgate.net/publication/2335004_Machine_Learning_Neural_and_Statistical_Classification

- Hofmann, H. (Universidad de Hamburgo). Comunicación y contribución original de los datos, citada en UCI Machine Learning Repository.