El conjunto de datos contiene transacciones realizadas por tarjetas de crédito en septiembre de 2013 por titulares de tarjetas europeos. Este conjunto de datos presenta transacciones que ocurrieron en dos días, con 492 fraudes de un total de 284,807 transacciones. El conjunto de datos está altamente desequilibrado, ya que la clase positiva (fraudes) representa el 0.172% de todas las transacciones.

Solo contiene variables de entrada numéricas, que son el resultado de una transformación PCA. Debido a problemas de confidencialidad, no se pueden proporcionar las características originales y más información de fondo sobre los datos. Las características V1, V2, ... V28 son los componentes principales obtenidos con PCA; las únicas características que no han sido transformadas con PCA son 'Time' y 'Amount'. La característica 'Time' contiene los segundos transcurridos entre cada transacción y la primera transacción en el conjunto de datos. La característica 'Amount' es el monto de la transacción, y esta característica se puede utilizar, por ejemplo, para el aprendizaje dependiente del costo. La característica 'Class' es la variable de respuesta y toma el valor 1 en caso de fraude y 0 en caso contrario.

Dada la proporción desequilibrada de clases, se recomienda medir la precisión utilizando el Área Bajo la Curva de Precisión-Recall (AUPRC). La precisión de la matriz de confusión no es significativa para la clasificación desequilibrada.

Actualización (03/05/2021): Se ha lanzado un simulador de datos de transacciones como parte del manual práctico sobre Machine Learning para la detección de fraudes con tarjetas de crédito. Invitamos a todos los profesionales interesados en conjuntos de datos de detección de fraudes a consultar también este simulador de datos y las metodologías presentadas en el libro.

Agradecimientos: El conjunto de datos ha sido recopilado y analizado durante una colaboración de investigación entre Worldline y el Grupo de Aprendizaje Automático (http://mlg.ulb.ac.be) de ULB (Université Libre de Bruxelles) sobre minería de grandes datos y detección de fraudes. Se proporcionan detalles adicionales sobre proyectos actuales y pasados en temas relacionados en https://www.researchgate.net/project/Fraud-detection-5 y en la página del proyecto DefeatFraud.