UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Aprendizado de Máquina

Mini Trabalho 2

Aquisição de dados

Grupo - 1

Ana Clara Barbosa Borges
André Emanuel Bispo da Silva
Artur Handow Krauspenhar
Gabriel Moura dos Santos
João Artur Leles Ferreira Pinheiro
João Pedro Anacleto Ferreira Machado

Brasília, DF
2025

Documentação:

O dataset escolhido, *Credit Card Fraud Detection*, tem como fonte o site de comunidade de *machine learning* e ciência de dados *Kaggle* e é uma base de dados de transações de cartão de crédito anonimizadas de titulares Europeus em setembro de 2013 (ANDREA; MACHINE LEARNING GROUP, 2018). O dataset consiste em 284.807 transações, com 492 fraudes ao decorrer de 2 dias, e tem suas *features* numéricas transformadas por meio de uma **Análise de Componentes Principais**. A maioria das features foram anonimizadas (colunas V1, V2, … V28) com exceção de Tempo (*Time*) que descreve o tempo desde a primeira transação em segundos em vez de um tempo absoluto, e quantidade (*Amount*) presumidamente em euros, e finalmente classe (*Class*) que indica se houve fraude (1) ou não (0) na transação.

Conjuntos de dados de transações são raros de encontrar devido a confidencialidade da informação, assim nosso grupo teve a escolha entre dados sintéticos e completos e dados reais mas anonimizados e acabamos por optar por dados reais pois podem proporcionar uma aproximação maior ao domínio do que dados sintéticos. Algo que encontramos relevante para o aprendizado na disciplina.

Quanto ao armazenamento e ambiente, os dados, pelo fato dos dados serem relativamente pequenos (≅100Mb) e estarem hospedados no kaggle não há necessidades especiais de armazenamento, sendo que é possível que os integrantes da equipe baixem os dados a qualquer momento. Dito isso, teremos uma cópia dos dados nas máquinas “chococcino” do campus Gama, onde existe *hardware* com maior capacidade de processamento para nossos fins.

Qualidade da aquisição de dados

Este dataset é altamente relevante para o problema proposto, que é a detecção de fraudes em transações. Ele contém um total de 284.807 transações, das quais 492 são fraudulentas. Essa desproporcionalidade entre as classes reflete com precisão o cenário real do problema, onde fraudes são eventos que não ocorrem com tanta frequência. A presença desse desequilíbrio é importante, pois permite o desenvolvimento e avaliação de técnicas específicas para lidar com classificações desbalanceadas, como reamostragem ou uso de métricas mais adequadas.

Além disso, as variáveis do conjunto de dados foram previamente transformadas por Análise de Componentes Principais (PCA), o que garante a confidencialidade dos dados originais ao mesmo tempo em que preserva as estruturas latentes mais relevantes para a discriminação entre fraudes e transações legítimas. Isso é particularmente vantajoso para aplicações práticas, pois evita o vazamento de informações sensíveis ao mesmo tempo em que reduz a dimensionalidade, melhorando o desempenho computacional e a capacidade de generalização do modelo. A variável alvo ‘Class’ é binária, o que facilita a aplicação direta de algoritmos de classificação supervisionada.

Dessa forma, o conjunto de dados que será utilizado é plenamente adequado ao objetivo de identificar transações fraudulentas. Sua estrutura, composição e características estatísticas são coerentes com o problema real, permitindo a criação e avaliação de um modelo preditivo aplicado em uma situação real.

Conformidade Legal e Ética:

A conformidade legal e ética na coleta e uso de dados é uma prioridade neste projeto. Embora o dataset "Credit Card Fraud Detection" do Kaggle apresente dados anonimizados, é fundamental considerar as implicações da Lei Geral de Proteção de Dados (LGPD - Lei nº 13.709/2018) e os princípios do sigilo bancário.

LGPD e Dados Anonimizados:

A LGPD define "dado pessoal" como qualquer informação relacionada a uma pessoa natural identificada ou identificável (Art. 5º, I). Dados anonimizados, por sua vez, são aqueles relativos a um titular que não pode ser identificado (Art. 5º, III). A lei não se aplica a dados totalmente anonimizados, a menos que o processo de anonimização seja reversível com meios razoáveis (Art. 12).

No contexto deste projeto, é essencial garantir que a anonimização dos dados do Kaggle seja robusta e irreversível, dentro de um limite razoável. Ou seja, a equipe não deve tentar reverter a anonimização, nem utilizar técnicas que visem identificar os titulares dos dados. Caso contrário, a LGPD passaria a ser aplicável, exigindo bases legais para o tratamento (como o consentimento), além de outras obrigações.

Sigilo Bancário:

Mesmo com os dados anonimizados, é importante reconhecer que as transações financeiras estão tradicionalmente protegidas pelo sigilo bancário. Embora a LGPD não trate especificamente do sigilo bancário, o princípio da minimização de dados (Art. 6º, III da LGPD) e o respeito à privacidade (Art. 2º, I da LGPD) exigem que a equipe utilize apenas os dados estritamente necessários para o desenvolvimento do projeto de detecção de fraudes, evitando qualquer tentativa de inferir informações sensíveis sobre os titulares dos cartões.

Referências Bibliográficas

ANDREA; MACHINE LEARNING GROUP. **Credit Card Fraud Detection**. Disponível em: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>.

Brasil. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União,** Brasília, DF, 15 ago. 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm> Acesso em: 13 de abril de 2025.