|  |
| --- |
| ANÁLISE DE FRAUDE EM CARTÕES DE CRÉDITO Licenciatura em Engenharia Informática | 2024/2025 | 3º ano  Laboratório de Análise de Dados | Prof. Fátima Leal  Gabriela Almeida | 47004 João Faria | 48057 Luís Meireles | 47799 Mónica Fino | 47366 Paulo Brochado | 47406 Tiago Mesquita | 47558 |

### ÍNDICE

[**Introdução 4**](#_heading=h.3rdcrjn)

[**Título 5**](#_heading=h.ktxp7cbj5jqh)

[**Conclusão 6**](#_heading=h.927it0515gle)

# Introdução

O presente projeto, desenvolvido no âmbito da unidade curricular Laboratório de Analise de Dados, tem como principal objetivo a aplicação prática de técnicas de análise de dados e algoritmos de aprendizagem automática, utilizando a linguagem de programação Python.

A realização do projeto divide-se em duas fases complementares: Na primeira fase, é realizada uma análise estatística exploratória de um conjunto de dados escolhido, com o intuito de compreender a sua estrutura, características e padrões relevantes. Esta análise inclui o uso de medidas estatísticas, visualizações gráficas, criação de novas variáveis e normalização dos dados, culminando num painel de controlo gráfico e numa reflexão crítica dos resultados obtidos.

Na segunda fase, são aplicados e comparados diversos modelos de *Machine Learning*, deforma a resolver um problema de revisão ou de classificação. Esta etapa inclui também a análise de desempenho dos modelos, técnicas de redução de dimensionalidade, validação cruzada, e eventualmente algoritmos de recomendação e ferramentas automáticas, com o objetivo de validar os métodos escolhidos.

Este relatório descreve todo o processo de desenvolvimento do trabalho, desde a seleção do conjunto de dados até à análise final dos modelos aplicados, incluindo as decisões técnicas tomadas, os desafios enfrentados e as aprendizagens adquiridas ao longo do projeto.

# Contextualização do Dataset

O dataset utilizado neste projeto foi obtido a partir da plataforma Kaggle e é amplamente conhecido por ser utilizado em estudos sobre deteção de fraude em transações financeiras. Este conjunto de dados representa transações realizadas com cartões de crédito por titulares europeus ao longo de dois dias, em setembro de 2013. O dataset contém 284.807 registos, distribuídos por 31 variáveis. Todas as variáveis presentes no dataset classificam-se como Quantitativa Contínua, exceto a variável *class*, que é uma variável Qualitativa nominal.

O principal objetivo deste dataset é servir como base para o desenvolvimento e avaliação de modelos de Machine Learning capazes de identificar transações fraudulentas de forma automática, um problema real e crítico enfrentado por instituições financeiras em todo o mundo.

# Objetivo do Projeto

O principal objetivo deste projeto é desenvolver uma análise detalhada e um modelo preditivo para a deteção de fraudes com cartão de crédito, com base num conjunto de dados real e anonimizado. O trabalho está dividido em duas fases que se complementam.

Na primeira etapa, a análise estatística, é realizada uma exploração do dataset previamente escolhido de forma a compreender a sua estrutura, distribuição das variáveis, relações entre atributos e identificação de padrões relevantes. Inclui o cálculo de medidas estatísticas, normalização dos dados e criação de visualizações gráficas que suportem a análise crítica do fenómeno.

Na segunda etapa, a aplicação de modelos de *Machine Learning*, com as informações obtidas na fase anterior, são aplicados e comparados vários algoritmos de *Machine Learning* com o objetivo de construir um modelo que permita classificar automaticamente transações como legítimas ou fraudulentas. São exploradas técnicas como regressões, SVM, árvores de decisão, redes neurais e métodos de ensemble, com atenção à avaliação do desempenho, tempos de execução, validação cruzada e métricas como precisão, recall e F1-score.

Dado o elevado desbalanceamento da variável alvo (*class*), o projeto aborda estratégias para lidar com este desafio, garantindo que os modelos conseguem identificar com eficácia as transações fraudulentas, mesmo quando estas representam uma minoria dos dados.

# Tecnologias Utilizadas

* Python – linguagem principal do projeto
* Streamlit – responsável pela interface do utilizador
* Pandas – ferramenta utilizada para a análise e manipulação dos dados
* Numpy – ferramenta para manipulação de matrizes
* Matplotlib – responsável pela visualização dos dados
* Seaborn – complemento para a biblioteca matplotlib
* Scikit-sklearn – útil para a criação de e aplicação de algoritmos de *Machine Learning*

# Features Implementadas

(A PREENCHER)

# Visão Geral do Programa

A página principal do nosso programa mostra ao utilizador informações sobre a missão do projeto, os tipos comuns de fraude, as variáveis utilizadas no modelo e ficheiro CSV, como prevenir fraudes e tecnologias e estratégias para prevenção de fraudes.

A missão do projeto é ajudar empresas a detetarem fraudes financeiras com ajuda da Inteligência Artificial e análise de dados. O projeto também oferece soluções para a monitorização, prevenção e identificação de atividades suspeitas.

Os tipos comuns de fraudes dividem-se em:

* Fraude em Cartão de Crédito – utilização não autorizada do cartão de crédito para compras
* *Phishing* – enganar utilizadores para fornecerem informações sensíveis
* Roubo de identidade – falsificação de identidade para acesso financeiro ilícito
* Transações Falsificadas – manipulação ou falsificação de transações bancárias.

Quanto às variáveis utilizadas no modelo CSV, foram utilizadas as seguintes variáveis:

* Time – tempo decorrido desde a primeira transação no dataset
* V1-V28 – conjunto de variáveis anonimizadas resultantes de PCA (28 componentes principais)
* Amout – montante da transação
* Class – identificação do tipo de transação (0: legítima, 1: fraudulenta)
* Hour – hora do dia em que a transação ocorreu
* Rolling\_Mean\_Amount – média móvel do valor da transação (janela de 5 transações)
* Std\_Amout – desvio padrão do valor da transação (janela de 5 transações)
* Delta\_Amout – diferença entre o valor atual e o valor anterior da transação
* Amount Category – categoria do valor da transação (ex.: Muito Baixo, Baixo, Médio, etc.)
* Time\_Diff – diferença de tempo entre transações consecutivas
* Transacao\_Noturna – indica se a transação ocorreu durante a noite (1: Sim, 0: Não)
* Num\_Transacoes\_1h – número de transações realizadas na mesma hora
* Freq\_Valor\_Transacao – frequência de transações com o mesmo valor
* Delta\_Media\_Valor – Diferença entre o valor da transação e a média móvel
* Region – Região geográfica associada à transação

Focando na prevenção de fraudes, esta atividade envolve um conjunto de boas práticas e tecnologias que ajudam a proteger empresas e consumidores. São indicadas as seguintes recomendações:

* Ativar alertas de transação, utilizar a autenticação multifator e monitorização contínua
* Não partilhar informações pessoais, verificar remetentes suspeitos e utilizar autenticação de dois fatores
* Utilizar verificação biométrica, não reutilizar palavras-passe e ativar bloqueios automáticos
* Implementar monitorização de transações em tempo real e análises de comportamento.

Pensando nas empresas, o programa indica quais tecnologias e estratégias que podem ser utilizadas para reforçar a segurança:

* *Machine Learning* e Inteligência Artificial: modelos que analisam padrões e detetam anomalias
* Autenticação Multifator (MFA) – verificação em duas etapas para acessos financeiros
* Monitorização em Tempo Real - identificação de transações suspeitas à medida que ocorrem
* Criptografia Avançada – proteção de dados sensíveis contra acessos não autorizados
* Análises de Comportamento – identificação de padrões incomuns de uso do sistema

# Análise de Fraudes

Nesta secção da aplicação, foi implementado um painel de visualização interativo dedicado à análise das fraudes detetadas no dataset. Esta componente permite ao utilizador aplicar filtros personalizados, nomeadamente por intervalo de tempo e por região geográfica da Europa, proporcionando uma exploração mais dirigida dos dados. No âmbito da apresentação do projeto, optou-se por analisar os resultados considerando todas as regiões e todas as horas do dia, de forma a obter uma visão global do fenómeno.

O gráfico apresentado na Figura 1 consiste num gráfico de barras que ilustra a distribuição do número de fraudes ao longo das 24 horas do dia. A análise revela dois picos de atividade fraudulenta significativos: às 2h e às 11h da manhã. O pico registado às 2h pode ser associado a um período de baixa vigilância humana, o que pode favorecer a ocorrência de fraudes. Já o pico das 11h coincide com um horário de elevada atividade comercial, o que pode ser estrategicamente explorado pelos criminosos para disfarçar operações ilícitas entre o volume de transações legítimas.

Uma imagem com texto, captura de ecrã, Retângulo, Tipo de letra

Os conteúdos gerados por IA podem estar incorretos.Uma imagem com diagrama, Gráfico, file, Tipo de letra

Os conteúdos gerados por IA podem estar incorretos.O gráfico apresentado na Figura 2 mostra a percentagem de fraudes por região europeia, permitindo identificar possíveis disparidades geográficas. Os resultados indicam que a distribuição é relativamente equilibrada entre as regiões, embora se destaque a região Norte, onde se observa uma ligeira predominância de fraudes. Esta tendência pode estar relacionada com a alta adoção de meios de pagamento digitais nessa zona, aumentando a exposição a potenciais ataques.

Figura 2

Figura

Adicionalmente, foi incluída uma visualização sobre o valor monetário das fraudes (Figura 3). A análise indica que a maioria das fraudes envolve valores inferiores a 120€, o que pode ser uma estratégia para evitar alertas automáticos dos sistemas bancários. Contudo, foi identificada uma exceção notável, com uma transação fraudulenta superior a 2000€. Através da análise espacial-temporal com recurso a um heatmap (Figura 4), foi possível identificar que esta fraude de elevado valor ocorreu às 2h da manhã, na região Leste, reforçando a hipótese de que horários com menor controlo facilitam a concretização de fraudes mais significativas.

Uma imagem com texto, captura de ecrã, quadrado, Retângulo

Os conteúdos gerados por IA podem estar incorretos.Esta análise interativa não só facilita a compreensão da distribuição das fraudes no tempo e no espaço, como também oferece um suporte visual fundamental para a tomada de decisões em cenários reais de combate a fraudes.

Figura 4

Uma imagem com texto, captura de ecrã, file, Retângulo

Os conteúdos gerados por IA podem estar incorretos.

Figura 3

# Estatísticas Avançadas

Na interface da aplicação desenvolvida, foi implementada a secção dedicada à apresentação de Estatísticas Avançadas, permitindo ao utilizador explorar de forma aprofundada o comportamento das variáveis do dataset. Esta funcionalidade visa apoiar a análise descritiva com métricas estatísticas clássicas, fundamentais para compreender a distribuição e variabilidade dos dados.

No topo da página, são apresentadas as seguintes medidas estatísticas: média, mediana, variância e desvio padrão. Estas medidas oferecem uma perspetiva clara sobre a tendência central e a dispersão dos dados, o que é essencial na preparação e normalização dos mesmos para futuras etapas de modelação.

Uma imagem com texto, captura de ecrã, Gráfico, file

Os conteúdos gerados por IA poderão estar incorretos.Adicionalmente, foi incluída a matriz de correlação, que permite identificar relações lineares entre as variáveis. A análise dessa matriz revela que não existe uma correlação linear forte entre nenhuma das variáveis e a ocorrência de fraudes. Esta ausência de correlação direta é um comportamento comum em problemas de deteção de anomalias, nos quais as variáveis que indicam fraude não se distinguem facilmente por padrões simples.

Figura 5

Uma imagem com diagrama, captura de ecrã, file

Os conteúdos gerados por IA poderão estar incorretos.Para complementar a análise estatística, foi desenvolvido um *violin plot* que representa as fraudes em função do calor das transações e faz regiões. Este tipo de gráfico combina elementos de boxplot com a distribuição de densidade dos dados, oferecendo uma visão rica e intuitiva da variação dos valores.

Figura 6

A análise deste gráfico mostra que, em todas as regiões geográficas, a maioria das transações fraudulentas ocorreu com valores baixos, o que pode refletir uma tentativa de passar despercebido aos mecanismos de controlo. No entanto, observa-se também a ocorrência de algumas fraudes com valores elevados, especialmente na região Oeste, o que pode indicar zonas de maior risco ou fragilidade em determinados contextos de uso de cartões.



upt.pt