Avaliação de Faithfulness em Machine Learning: Análise das Métricas ROAD e Sensitivity-N

Trabalho Prático de Interpretabilidade e Causalidade do curso
Inteligência Artificial e Ciência de Dados
(1º ciclo de estudos)

Rafael Antunes nº55336

Professor: João Carlos Raposo Neves

Dezembro de 2025

Prefácio

Este trabalho prático foi realizado no âmbito da Unidade Curricular (UC) de Interpretabilidade e Causalidade, lecionada no primeiro semestre do terceiro ano da Licenciatura em Inteligência Artificial e Ciência de Dados. O principal objetivo deste trabalho é perceber, implementar e aplicar métricas de avaliação, métodos de interpretabilidade e comparar na explicação de modelos de Machine Learning.

Tudo o referido e implementado neste trabalho está disponível no Github do trabalho: Link Github

Resumo

Este trabalho analisa as métricas propostas (ROAD e Sensitivity-n) e a sua implementação, juntamente com métodos de explicação, aplicados a modelos de Machine Learning em conjuntos de dados de imagem e tabulares.

Palavras-chave

Explainable AI, ROAD, Sensitivity-N

Índice

Capítulo 1

Introdução

1.1 Contexto na Explainable Artificial Intelligence

A Explainable Artificial Intelligence (XAI) é um ramo da Inteligência Artificial que tem vindo a ganhar destaque nos últimos anos, impulsionado pela necessidade de compreender o funcionamento de modelos considerados *black box*. Nestes modelos, não é possível interpretar diretamente o processo de decisão apenas através da análise das suas camadas ou parâmetros internos.
Como resposta a esta limitação, têm sido desenvolvidos métodos específicos no âmbito da XAI que permitem explicar e interpretar as decisões tomadas pelos modelos, fornecendo uma melhor compreensão dos fatores e padrões nos quais essas decisões se baseiam.

1.2 Métodos de Explicação em XAI

De forma a compreender melhor as decisões tomadas pelos modelos caixa negra, recorrem-se a métodos de explicação. Estes métodos podem ser utilizados para caracterizar o comportamento do modelo a nível local, explicando decisões individuais, ou a nível global, fornecendo uma visão geral do processo de decisão do modelo. Cada abordagem interage de forma distinta com o modelo, permitindo extrair diferentes tipos de informação sobre os fatores que influenciam as suas predições.

1.3 Métricas de Avaliação em XAI

Associados aos métodos de explicação existem métricas que permitem avaliar a qualidade das explicações geradas e retirar conclusões mais fiáveis sobre o com-

portamento dos modelos. Estas métricas ajudam a verificar se uma explicação reflete corretamente o processo de decisão do modelo.

As métricas em XAI avaliam diferentes propriedades das explicações, como a fidelidade, a estabilidade ou a suficiência, sendo utilizadas para comparar métodos de explicação e evitar interpretações incorretas das decisões do modelo.

Neste trabalho, é dada especial atenção às métricas de fidelidade (*faithfulness*), que medem até que ponto as características identificadas como relevantes pela explicação têm impacto efetivo na predição do modelo. Estas métricas baseiam-se geralmente na alteração ou remoção dessas características e na observação da variação da saída do modelo.

Capítulo 2

Métrica Sensitivity-n

2.1 Introdução

Para uma compreensão aprofundada desta métrica, recorreu-se ao estudo original de Marco Ancona et al., intitulado *"Towards better understanding of gradient--based attribution methods for Deep Neural Networks"*.

Neste trabalho, a métrica é proposta como uma alternativa quantitativa às avaliações puramente visuais de relevância de características (como *heatmaps*), que são frequentemente subjetivas e passíveis de interpretações erróneas. Através da *Sensitivity-*n, é possível realizar uma análise de fidelidade (*faithfulness*) mais objetiva, permitindo a comparação direta entre diferentes métodos de explicação. No artigo, a métrica foi utilizada para comparar métodos baseados em gradiente, nomeadamente: *Gradient* $\times$ *Input*, *Integrated Gradients*, ε-LRP e *DeepLIFT*.

2.2 Explicação

A *Sensitivity-*n quantifica se a relevância atribuída a um conjunto de características de entrada se correlaciona linearmente com a variação na predição do modelo quando essas mesmas características são removidas ou neutralizadas.

2.2.1 Procedimento de Avaliação

Visto ser computacionalmente inviável testar todos os subconjuntos combinatórios possíveis para valores elevados de n, a métrica é estimada estatisticamente. O procedimento consiste em:

1. Amostrar aleatoriamente uma quantidade significativa de subconjuntos S de tamanho n.
2. Para cada subconjunto, calcular a soma das atribuições teóricas e a variação real observada na saída do modelo (comparando com a *baseline*).

3. Calcular o **Coeficiente de Correlação de Pearson** (ρ) entre os pares de valores obtidos.

2.2.2 Interpretação dos Resultados

O coeficiente resultante ρ situa-se no intervalo $[-1, 1]$, permitindo as seguintes interpretações:

- $\rho \to 1$ **(Alta Fidelidade)**: Indica que o método de explicação é um reflexo fiel do comportamento do modelo. As características identificadas como importantes são as que, de facto, mais influenciam a decisão.
- $\rho \to 0$ **(Ausência de Correlação)**: Sugere que as atribuições não possuem valor explicativo real, assemelhando-se a uma atribuição aleatória.
- $\rho < 0$ **(Anti-correlação)**: Indica um método enganoso, onde as características apontadas como relevantes têm um impacto oposto ao esperado na predição.

Um aspeto crucial desta métrica é a variação do parâmetro n. Testar valores baixos de n avalia a importância de características isoladas, enquanto valores elevados testam a robustez da explicação perante interações complexas entre os dados.

2.3 Aplicação

O objetivo desta fase do trabalho consistiu na implementação e aplicação desta métrica recorrendo a diferentes métodos de explicação para fins demonstrativos, tendo sido usada a biblioteca *tf-explain*. Embora tenha sido recomendado o uso da biblioteca *Quantus* para a métrica, optou-se por uma implementação própria. Esta decisão permitiu uma análise mais precisa e personalizada, como a neutralização progressiva de características até ao valor n impacta a confiança e a exatidão das decisões do modelo.

2.3.1 Configuração

Como base para a aplicação, utilizou-se o conjunto de dados *MNIST* (dígitos manuscritos). Foi treinado um modelo personalizado, cujo código está presente no

ficheiro *sensn_framework.py* dentro da classe *CNN_MNIST*, alcançando uma exatidão de 99%. Posteriormente, realizaram-se testes visuais com os métodos de explicação para validar a coerência das atribuições e descartar indícios de *overfitting*.

Os métodos de explicação avaliados foram o *Gradient* $\times$ *Input* e o *Integrated Gradients* (presentes no artigo original), complementados pelo *SmoothGrad*, usando uma função wrapper personalizada no ficheiro *xai_funcs.py* chamada de *display_explanations()*.

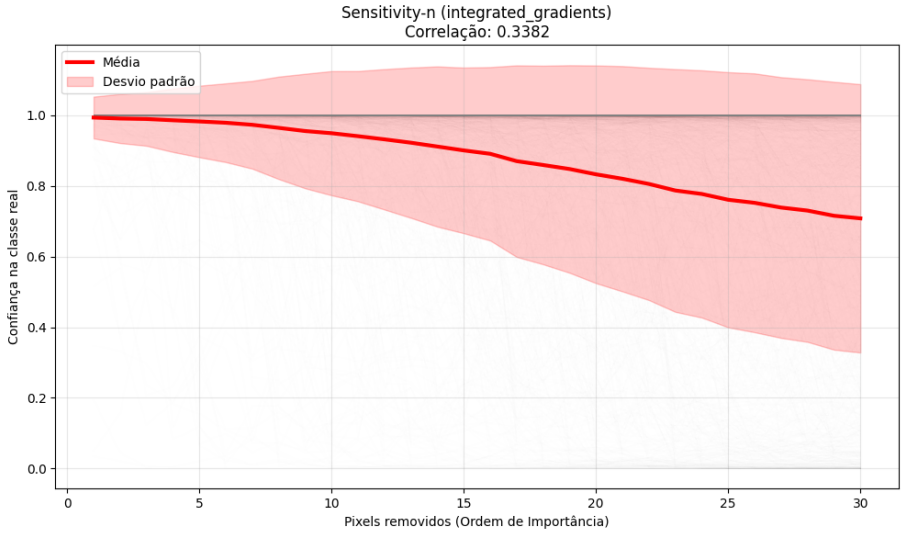
Numa análise puramente visual, o método *Gradient* $\times$ *Input* apresentou, neste contexto específico, resultados mais nítidos.

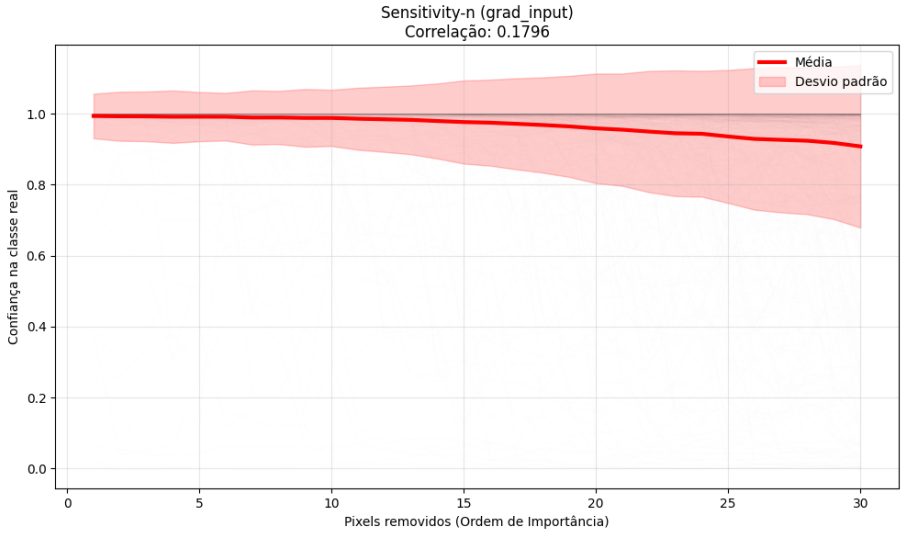
2.3.2 Implementação e Metodologia

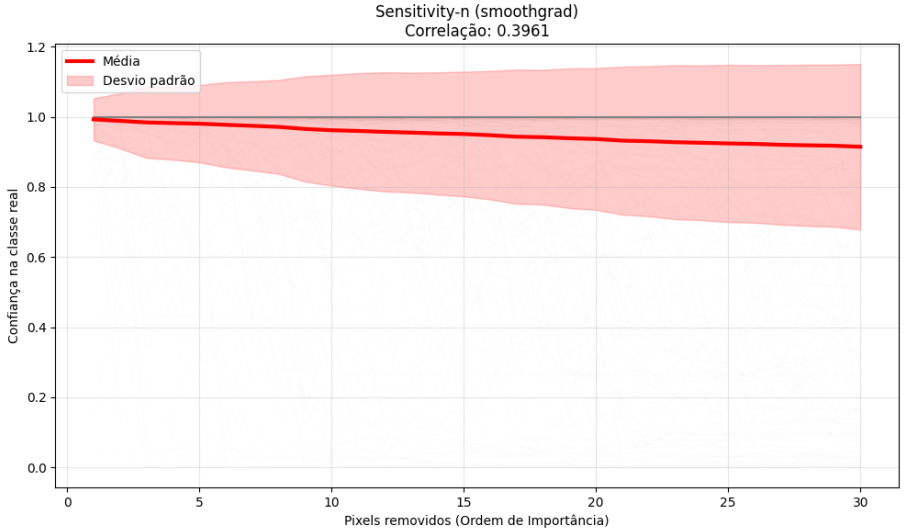
A implementação da métrica também localiza-se na *xai_funcs.py* e é uma função denominada de *analisar_sensitivity_n()*, esta seguiu um processo iterativo: para cada imagem do conjunto de teste, as características (pixéis) são neutralizadas progressivamente com base na sua importância atribuída. A cada remoção, o modelo é reavaliado para registar a variação na confiança da resposta. Este processo é repetido para diferentes valores de n, resultando num mapeamento da correlação entre a soma da importância acumulada e a perda de exatidão (ou variação do *score*). O resultado final permite observar, através de representações gráficas, quais os métodos que mantêm uma maior fidelidade ao comportamento real do modelo à medida que a perturbação no input aumenta.

2.3.3 Avaliação dos Resultados

Para a validação desta métrica, foram utilizados 100 exemplos de cada dígito, totalizando 1000 amostras. Este conjunto de dados foi mantido constante em todos os testes realizados para garantir a comparabilidade dos resultados apresentados a seguir:







Através da análise destes gráficos, observa-se que o método que obteve a melhor correlação foi o *SmoothGrad*. No entanto, a confiança do modelo permanece re-

lativamente elevada durante grande parte do processo de neutralização.

A análise destes gráficos permite identificar o ponto crítico (número de píxeis n) a partir do qual o modelo começa a sentir dificuldade em classificar o dígito corretamente. Estes resultados demonstram a robustez do modelo; tratando-se de uma imagem, uma oclusão parcial ou pontual não impede necessariamente a classificação correta. Conclui-se que o impacto na decisão só é significativo quando uma parte substancial da estrutura do dígito é removida. Este comportamento foi particularmente visível no método *Integrated Gradients*, que apresentou o maior impacto final, terminando com um nível de confiança médio inferior a 80%.

2.4 Análise Extra: Aplicação em Dados Tabulares

Após a aplicação em dados de imagem, surgiu a hipótese de que a lógica da métrica *Sensitivity-*n seria ainda mais transparente em dados tabulares. Neste domínio, as decisões dos modelos tendem a basear-se num conjunto finito e explícito de variáveis, permitindo verificar com maior clareza se a remoção das características consideradas ”importantes” afeta, de facto, a confiança do modelo.

2.4.1 Metodologia e Dataset

Para esta análise, utilizou-se o *Wine Quality Dataset*. A variável alvo original (qualidade do vinho, numa escala de 3 a 8) foi transformada num problema de classificação binária para simplificar a interpretação da explicação:

- **Classe 1 (Vinho de Qualidade):** Classificação original ≥ 6.
- **Classe 0 (Vinho Normal/Baixo):** Classificação original < 6.
- **Modelo:** Foi treinado um classificador *XGBoost*, obtendo uma exatidão de aproximadamente 85.6%.

2.4.2 Análise da Importância das Características

Para extrair a importância das características, utilizaram-se os *SHAP values*, pela biblioteca *shap*, permitindo uma interpretação local consistente. Observaram-se, inicialmente, três instâncias de vinhos de qualidade superior e três de qualidade inferior.

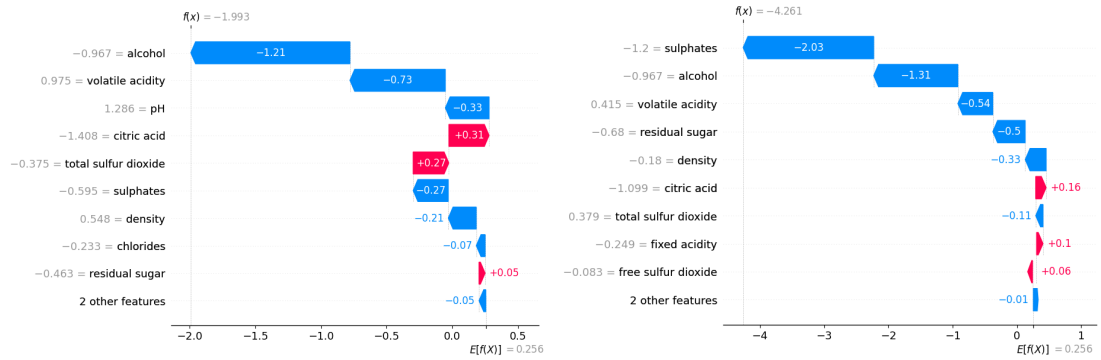


Figura 2.1: Análise SHAP para vinhos de qualidade Normal/Baixa.

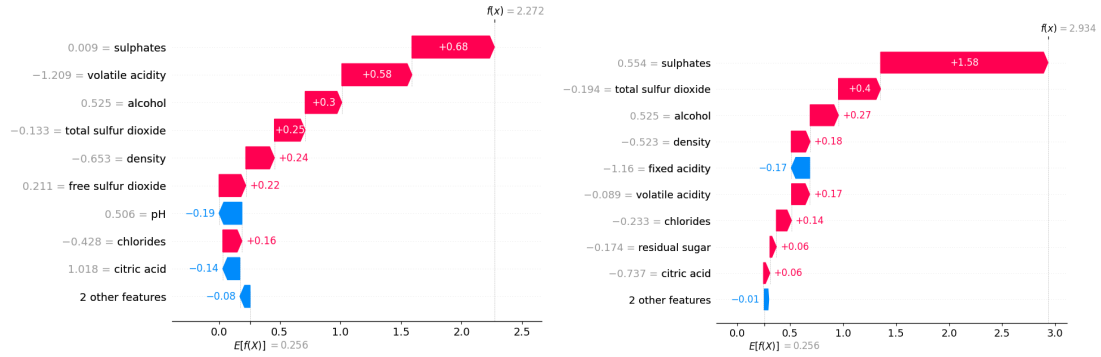
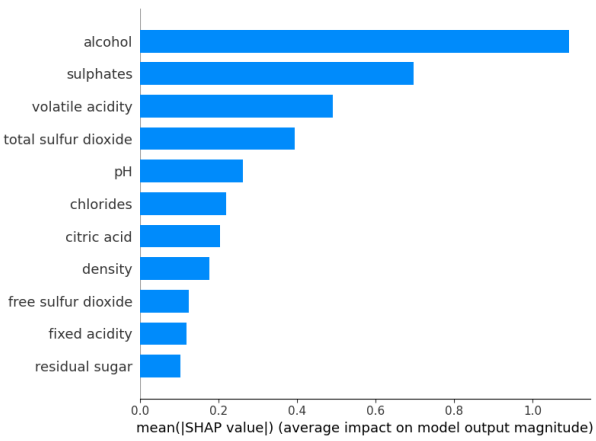


Figura 2.2: Análise SHAP para vinhos de Qualidade.

Através destas análises locais, é possível identificar um padrão na tomada de decisão do modelo, onde um subconjunto de aproximadamente duas a três características parece ser predominante, embora a diferenciação visual entre as classes não seja imediatamente evidente apenas pelos *gráficos em cascata*.

Para uma visão global, gerou-se um gráfico de barras com a média dos valores SHAP absolutos para todas as instâncias, comparando-o com os valores de correlação das variáveis. Observou-se que, embora a importância mude entre as várias características, não existe um único ”causador” isolado, mas sim uma combinação de características.



quality	1.000000
alcohol	0.476166
volatile acidity	0.390558
sulphates	0.251397
citric acid	0.226373
total sulfur dioxide	0.185100
density	0.174919
chlorides	0.128907
fixed acidity	0.124052
pH	0.057731
free sulfur dioxide	0.050656
residual sugar	0.013732

Figura 2.3: Comparação entre a média dos SHAP values e a matriz de correlação.

2.4.3 Resultados da Fidelidade

Utilizou-se a métrica *Sensitivity-*n, presente na função *run_sensitivity_n()* dentro da classe criada para englobar o modelo *Modelo_XGB_Classifier* no ficheiro *wines_sensn.py*, para validar a fiabilidade das explicações fornecidas pelo SHAP e confirmar se o modelo baseia as suas decisões nas características identificadas como mais importantes.

Os resultados confirmam a alta fidelidade do método: observa-se uma queda acentuada na confiança do modelo (para níveis próximos dos 50%) logo após a neutralização das três características principais. Além disso, obteve-se um **Coeficiente de Correlação de Pearson de 0.90**, um valor extremamente elevado que demonstra que as características selecionadas são, de facto, cruciais para a tomada de decisão do modelo.

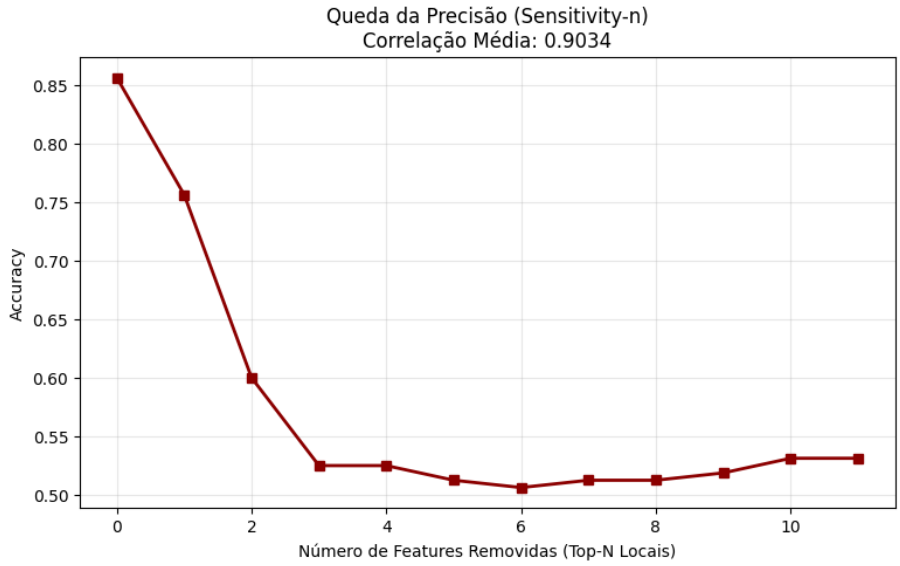


Figura 2.4: Gráfico da métrica Sensitivity-n aplicada ao dataset de vinhos.

2.4.4 Conclusão da Análise Extra

O valor de correlação de 0.90 demonstra uma fidelidade muito superior à observada nos dados de imagem. Ao contrário das imagens, onde a informação é espacial e difusa, nos dados tabulares a métrica conseguiu capturar uma relação quase linear. Ao neutralizar as características vitais (como o teor alcoólico ou a acidez), o modelo perde a sua capacidade preditiva quase instantaneamente. Isto valida a *Sensitivity-n* como uma ferramenta robusta para distinguir explicações genuínas de ruído estatístico, confirmando os verdadeiros motores de decisão do algoritmo.

Capítulo 3

Métrica ROAD

3.1 Introdução

A avaliação quantitativa de métodos de *Explainable AI* (XAI) é fundamental para garantir que as explicações geradas são fiéis ao comportamento do modelo e não apenas visualmente apelativas para o ser humano. Para este efeito, recorreu-se ao estudo de Yao Rong et al., intitulado *"A Consistent and Efficient Evaluation Strategy for Attribution Methods"*.

Neste trabalho, é apresentada a métrica ROAD (*Remove And Debias*), proposta como uma evolução eficiente e consistente face às estratégias de avaliação baseadas em perturbação (como o ROAR - *Remove and Retrain*). O ROAD aborda dois problemas fundamentais: o elevado custo computacional do re-treino e a "Fuga de Informação de Classe" (*Class Information Leakage*), onde o modelo consegue adivinhar a classe apenas pela forma da máscara de remoção, e não pelo conteúdo restante.

3.2 Fundamentação Teórica

O princípio base das métricas de remoção é que, se as características (ex: pixéis) identificadas como importantes forem removidas, a performance do modelo deve degradar-se. No entanto, a forma como essa remoção é feita é crítica. Preencher os pixéis removidos com uma cor sólida (cinzento ou preto) ou com a média do dataset cria artefactos artificiais que enviesam o modelo (problema de *Out-of-Distribution*).

A inovação do ROAD reside na introdução da **Imputação Linear Ruidosa** (*Noisy Linear Imputation*). Em vez de remover pixéis e re-treinar o modelo para se adaptar à nova distribuição (como no ROAR), o ROAD substitui os pixéis removidos por valores que são estatisticamente plausíveis baseados nos seus vizinhos, impedindo que o modelo utilize a "ausência de informação" como uma pista visual.

3.2.1 Procedimento de Avaliação

O procedimento do ROAD evita o re-treino, permitindo avaliações na ordem dos segundos em vez de horas, mantendo uma alta correlação com métricas mais dispendiosas. O processo consiste em:

1. **Ordenação (Ranking):** Com base no método de explicação (ex: *Integrated Gradients*), ordena-se a importância de cada pixel.
2. **Mascaramento (Masking):** Seleciona-se uma percentagem k dos pixéis para remoção. Esta seleção pode seguir duas ordens:
    - **MoRF (*Most Relevant First*):** Removem-se primeiro os pixéis mais importantes. Espera-se uma queda rápida na precisão do modelo.
    - **LeRF (*Least Relevant First*):** Removem-se primeiro os pixéis menos importantes. Espera-se que a precisão se mantenha estável.
3. **Imputação (*Debiasing*):** Os pixéis selecionados são substituídos utilizando a *Noisy Linear Imputation*. O valor do pixel $x_{i,j}$ é estimado pela média ponderada dos seus vizinhos diretos e diagonais, adicionando-se um ruído Gaussiano ($\sigma = 0.1$).
4. **Inferência:** A imagem modificada é submetida ao classificador original para medir a confiança na classe correta.

3.3 Aplicação e Implementação

Nesta fase do trabalho, procurou-se implementar a métrica ROAD para validar a fidelidade de diferentes métodos de explicação num cenário prático de classificação de imagem.

3.3.1 Configuração do Modelo e Dados

Utilizou-se o conjunto de dados *Dogs vs Cats*, composto por imagens RGB de cães e gatos. Para a tarefa de classificação, foi instanciado e treinado um modelo de arquitetura **EfficientNetBo** (pré-treinado na ImageNet e afinado para este dataset), que atingiu uma exatidão de validação superior a 98%, garantindo uma base sólida para a geração de explicações.

3.3.2 Métodos de Explicação Avaliados

Foram gerados mapas de saliência para as imagens de teste utilizando três métodos distintos, disponíveis na biblioteca `tf-explain`, de forma a comparar a sua granularidade e foco:

- **Integrated Gradients:** Um método axiomático que acumula gradientes ao longo de um caminho linear.
- **Gradient Inputs:** Gradiente da saída em relação à entrada, ponderado pela própria entrada.
- **Saliency (Vanilla Gradients):** O método base de backpropagation de gradientes.

Visualmente, o método *Gradient Inputs* apresentou, para este modelo e dataset específicos, resultados com menos ruído de fundo em comparação com os *Vanilla Gradients*.

3.3.3 Limitações na Implementação da Métrica

Para a avaliação quantitativa via ROAD, selecionou-se a biblioteca **Quantus**, uma ferramenta de referência para avaliação de XAI que, teoricamente, suporta a métrica ROAD.

O objetivo era configurar a métrica para avaliar a fidelidade (*Faithfulness*) utilizando a estratégia de perturbação baseada na *Noisy Linear Imputation*, iterando sobre um subconjunto de 100 imagens de teste.

No entanto, durante a fase de desenvolvimento, não foi possível concluir a execução da métrica ROAD através do `quantus`.

Devido à escassez de documentação detalhada e exemplos práticos funcionais da implementação específica do ROAD nesta biblioteca (visto ser uma métrica recente, proposta em 2022), e à complexidade de reimplementar manualmente a lógica de imputação linear ruidosa de forma eficiente, não foi possível obter os gráficos de curva de degradação (MoRF/LeRF) para esta métrica específica neste trabalho.

Capítulo 4

Conclusão

O presente trabalho teve como objetivo aprofundar o estudo sobre a avaliação da fidelidade (*Faithfulness*) em Inteligência Artificial Explicável (XAI), explorando duas métricas fundamentais: ROAD e Sensitivity-n. Através da implementação prática em dados não estruturados (imagens) e estruturados (tabular), foi possível compreender os desafios inerentes à validação de explicações de modelos ”caixa negra”.

No domínio da visão computacional, a exploração da métrica ROAD destacou a complexidade de avaliar atribuições sem introduzir viéses. Embora a fundamentação teórica do ROAD — especificamente a *Noisy Linear Imputation* — se mostre promissora para evitar o vazamento de informação, a implementação prática recorrendo à biblioteca *Quantus* revelou-se um obstáculo técnico. As limitações de compatibilidade encontradas impediram a geração das curvas de degradação MoRF/LeRF, evidenciando que, embora a teoria de XAI esteja a avançar rapidamente, as ferramentas de software disponíveis para a sua avaliação ainda carecem, em certos casos, de maturidade e estabilidade.

Por outro lado, a aplicação da métrica *Sensitivity-n* ao conjunto de dados tabular (*Wine Quality*) provou ser um sucesso demonstrável. A obtenção de um coeficiente de correlação de Pearson de **0.90** entre a importância das características (via SHAP) e a queda de desempenho do modelo validou, de forma quantitativa, que as explicações geradas são fiéis ao funcionamento interno do modelo. Este resultado contrasta com a ambiguidade visual muitas vezes sentida na análise de imagens, demonstrando que, em dados tabulares, a relação entre remoção de características e perda de confiança é mais linear e direta.

Em suma, este trabalho permitiu concluir que a escolha da métrica de avaliação é tão crítica quanto a escolha do método de explicação. Enquanto métodos visuais oferecem intuição humana, apenas métricas quantitativas rigorosas como o *Sensitivity-n* (e teoricamente o ROAD) podem garantir que o modelo está, de facto, a ”olhar” para onde a explicação indica. O estudo sublinha a necessidade contínua de desenvolvimento de *frameworks* de avaliação mais robustos e acessíveis para padronizar a confiança em sistemas de Aprendizagem Computacional.