
ESCOLA SUPERIOR DE

TECNOLOGIA DA INFORMAÇÃO

INSTITUTO INFNET

LEONARDO DA CONCEIÇÃO MUNIZ

# Projeto de Bloco:
# Inteligência Artificial e Machine Learning

## TP 3

Docente: Tiago Cariolano de Souza Xavier

Rio de Janeiro

04/08/2025

**LINK DO NOTEBOOK DO GOOGLE COLAB REFERENTE AO TP 3 DE PB DE IA:**
HTTPS://COLAB.RESEARCH.GOOGLE.COM/DRIVE/14WJAZXPV6C_D2G6F9TABC5EYYHFDU2MF

***Relatório Final***

O classificador binário criado demonstra alta eficiência e capacidade discriminatória para a tarefa proposta, evidenciada por um **AUC de 0.89** e um bom equilíbrio entre precisão e recall, especialmente para a classe positiva ("Mina").

**1. Preparação dos Dados e Otimização do Modelo**
O projeto iniciou com um dataset de **60 features**, que foram eficientemente reduzidas para **29 componentes principais** utilizando **Análise de Componentes Principais (PCA)**. Esta etapa foi crucial para diminuir a dimensionalidade, o que pode mitigar a maldição da dimensionalidade e melhorar o desempenho e interpretabilidade do modelo. A **visualização PCA 2D** das classes ("Rocha" e "Mina") revelou que, embora as classes apresentem alguma sobreposição visual em duas dimensões, existe uma tendência de agrupamento, sugerindo que a separação é possível em um espaço de maior dimensão.
O modelo de árvore de decisão foi otimizado utilizando **GridSearch** com validação cruzada (**5 folds**). Os melhores hiperparâmetros encontrados foram **ccp_alpha=0.0**, **criterion='entropy'**, **max_depth=5** e **min_samples_leaf=5**. A melhor pontuação (acurácia média da validação cruzada) de **0.8251** indica que o modelo otimizado apresenta boa generalização para dados não vistos durante o treinamento. A poda da árvore (ccp_alpha) contribui para prevenir o **overfitting** e melhorar a capacidade de **generalização** do modelo.

**2. Avaliação de Desempenho do Classificador**
A avaliação do classificador foi realizada utilizando diversas figuras de mérito, conforme solicitado:

**2.1. Métricas Globais**
- **Acurácia:** 0.8333 (ou 83.33%). Indica que a maioria das previsões do modelo está correta. É uma boa métrica geral, mas não captura a performance em classes desbalanceadas.

**2.2. Matriz de Confusão e Métricas por Classe**
A Matriz de Confusão forneceu uma visão detalhada dos erros e acertos do modelo:

| | Predito Rocha (0) | Predito Mina (1) |
|---|---|---|
| **Real Rocha (0)** | **14 (TN)** | **6 (FP)** |
| **Real Mina (1)** | **1 (FN)** | **21 (TP)** |

A partir dela, calculamos as métricas essenciais para a classe "Mina" (considerada a classe positiva ou '1') e "Rocha" (considerada a classe negativa ou '0'):
- **Precisão (Mina - Positiva):** 0.7778
  - Dos casos previstos como "Mina", 77.78% estavam corretos.
- **Recall (Mina - Positiva / Sensibilidade):** 0.9545
  - Das amostras que *realmente* eram "Mina", o modelo identificou 95.45% corretamente. Este é um **ponto forte significativo**, indicando que o modelo é excelente em detectar a classe positiva.
- **F1-Score (Mina - Positiva):** 0.8571

    - Representa a média harmônica entre precisão e recall para a classe "Mina", indicando um bom equilíbrio.
- **Especificidade (Rocha - Negativa / Recall da Classe 0):** 0.7000
    - Das amostras que *realmente* eram "Rocha", o modelo identificou 70% corretamente.
- **Precisão (Rocha - Negativa):** 0.9333
    - Dos casos previstos como "Rocha", 93.33% estavam corretos.

**2.3. Curva ROC e AUC**

A Curva ROC plotou a Taxa de Verdadeiros Positivos (TPR) versus a Taxa de Falsos Positivos (FPR) para diferentes limiares.

- **AUC (Area Under the Curve): 0.89**
    - Um valor de AUC de 0.89 é **excelente** e demonstra a **alta capacidade discriminatória** do classificador. Isso significa que há uma probabilidade de 89% de que o modelo classificará uma instância positiva aleatória com uma pontuação mais alta do que uma instância negativa aleatória. A curva se afasta significativamente da linha diagonal (classificador aleatório), indicando que o modelo é muito mais eficaz do que o acaso na distinção entre "Rocha" e "Mina".

**3. Eficiência do Classificador Criado**

O classificador criado é **altamente eficiente e robusto** para a tarefa de classificação binária. Ele demonstra uma forte capacidade de generalização e um excelente poder discriminatório, conforme evidenciado pelo alto AUC.

- **Pontos Fortes:**
    - **Alta Detecção da Classe Positiva:** O recall de 0.9545 para a classe "Mina" é notável, indicando que o modelo é muito eficaz em identificar amostras verdadeiramente "Mina".
    - **Alta Confiança na Previsão Negativa:** A precisão de 0.9333 para a classe "Rocha" significa que, quando o modelo prevê uma amostra como "Rocha", essa previsão é altamente confiável.
    - **Excelente Poder Discriminatório:** O AUC de 0.89 valida que o modelo é ótimo em distinguir entre as duas classes em um nível de probabilidade.
    - **Generalização:** A acurácia de validação cruzada e o uso de pruning sugerem que o modelo não está superajustado e é capaz de performar bem em novos dados.
- **Considerações:**
    - Há um **trade-off** notável onde 6 amostras de "Rocha" foram erroneamente classificadas como "Mina" (Falsos Positivos). Dependendo do custo associado a esse tipo de erro (por exemplo, o custo de investigar uma falsa "Mina" em vez de uma "Rocha"), isso pode ser um ponto a ser refinado. No entanto, o erro oposto (classificar uma "Mina" como "Rocha") é muito raro (apenas 1 Falso Negativo), o que é ideal se a detecção de "Mina" for a prioridade principal.

Em conclusão, o classificador desenvolvido é um sucesso em seu objetivo de classificação binária, demonstrando a eficácia das técnicas de PCA, árvores de decisão e otimização de hiperparâmetros para lidar com problemas de aprendizado de máquina no mundo real.