UNIVERSIDADE DE PERNAMBUCO

ESCOLA POLITÉCNICA DE PERNAMBUCO

Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Inteligência Computacional

Título: Avaliação de Técnicas de Explicabilidade (XAI) em Modelos de Classificação Aplicados a Dados Não Estruturados e Desbalanceados

Orientador: Cleyton Mário de Oliveira Rodrigues (cleyton.rodrigues@upe.br)

Contexto

O crescimento exponencial de dados gerados diariamente, em especial os não estruturados, como textos e postagens em redes sociais, trazem desafios para garantir transparência e interpretabilidade nos modelos de Machine Learning (ML). Modelos avançados, como Redes Neurais e Transformers, frequentemente funcionam como "caixas-pretas", dificultando a compreensão das previsões realizadas. Em cenários críticos, como análises sociais ou tomadas de decisão automatizadas, essa falta de explicabilidade pode comprometer a confiança nos resultados.

Problema

O desbalanceamento de dados é um desafio significativo, especialmente em bases textuais, onde algumas classes são super-representadas em relação a outras. Esse desbalanceamento pode distorcer previsões e comprometer a capacidade explicativa dos modelos, amplificando vieses e limitando a transparência. A falta de diretrizes práticas sobre como lidar com esses cenários reforça a relevância de estudos que combinem explicabilidade com métodos robustos para dados desbalanceados.

Objetivos

O projeto visa avaliar o impacto do desbalanceamento de dados nas métricas de explicabilidade de diferentes algoritmos de classificação e identificar técnicas de XAI (e.g., LIME, SHAP) mais eficazes para dados textuais desbalanceados. Especificamente, busca:

- Revisar sistematicamente a literatura sobre XAI em dados desbalanceados.
- Comparar técnicas de XAI em algoritmos clássicos, recorrentes, convolucionais e Transformers.
- Propor métricas claras para balanceamento e explicabilidade.
- Desenvolver diretrizes práticas para cenários reais.

Método

O projeto será conduzido em etapas:

1. **Coleta e Pré-processamento de Dados**: Textos de redes sociais serão utilizados, aplicando limpeza, tokenização e padronização.

2. **Construção de Modelos**: Algoritmos clássicos (SVM, Random Forest), recorrentes (LSTM), convolucionais (CNN) e Transformers serão treinados.
3. **Técnicas de Balanceamento**: Aplicação de oversampling e undersampling para criar bases com diferentes proporções de classes.
4. **Aplicação de Técnicas de XAI**: Uso de LIME e SHAP para explicar os modelos.
5. **Análise e Avaliação**: Avaliação de desempenho (e.g., F1 Score, AUC) e explicabilidade (consistência, robustez).

Resultados Esperados

Espera-se identificar as técnicas de XAI que mantêm a melhor consistência e interpretabilidade em cenários com dados desbalanceados, bem como entender o impacto do desbalanceamento nas explicações fornecidas pelos modelos. A pesquisa deve resultar em:

- Desenvolvimento de um banco de dados público para testes futuros.
- Diretrizes práticas para tratar dados desbalanceados em tarefas de classificação.
- Aplicação do método em cenários reais, como análise de sentimentos em redes sociais.

Referências Bibliográficas:

- Araújo, J. P. B. (2020). *Interpretabilidade de modelos de machine learning: aplicação no mercado de crédito*. Universidade Federal do Ceará.
- Pereira, J. R. L. (2021). *Transparência pela cooperação: como a regulação responsiva pode auxiliar na promoção de sistemas de machine-learning inteligíveis*. *Journal of Law and Regulation*.
- Pocebon, M. (2023). *Ética na IA: Desafios e considerações éticas em aplicativos de IA e Machine Learning*.
- Rodas, C. M. et al. (2022). *Análise de sentimentos sobre vacinas contra Covid-19*. *Rev. Saúde Digital Tec. Educ.*
- Sánchez-Hernández, F. et al. (2019). *Predictive modeling of ICU healthcare-associated infections from imbalanced data*. *Applied Sciences*.