www.aileaders-project.eu

DEMO

- Ferramenta para auditoria de preditores

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

DEMO - Ferramenta para auditoria de preditores

01 Resumo

Tipo de REA

Demonstração utilizando AEQUITAS em um ambiente Google Colab

Objetivo/Finalidade

Fornecer aos estudantes uma ferramenta para auditar modelos de aprendizagem automática para preconceitos e imparcialidade. Ao focar num cenário de deteção de fraudes, os estudantes aprendem a avaliar criticamente o desempenho de um algorítmico de classificação, não apenas em termos de precisão, mas também em termos de tratamento equitativo entre grupos demográficos.

Resultados de aprendizagem esperados

Ao final da demonstração, os estudantes serão capazes de:

01 Aplicar a ferramenta Aequitas para auditar modelos de classificação (árvore de decisão, Random Forest, FairGBM) quanto à equidade;

02 Interpretar métricas-chave de equidade, tais como Taxa de Falsos Positivos (FPR), Taxa de Falsas Descobertas (FDR) e Paridade Estatística;

03 Identificar e explicar as disparidades ao nível do grupo nos resultados das previsões;

04 Refletir sobre o papel da auditoria de viés e das técnicas equitativas de Aprendizagem Automática em contextos de tomada de decisão de alto risco

Palavras-chave

- Aprendizagem automática
- Aequitas
- Auditoria
- Preconceitos
- Equidade

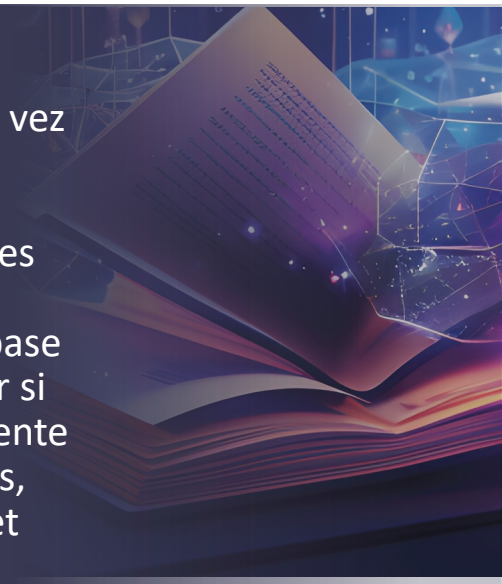
Sugerido Abordagem metodológica

Aprendizagem baseada em problemas

NOTA

É necessário ter conhecimentos intermédios de **programação na linguagem Python** para compreender e trabalhar com o conteúdo deste REA.

02 Introdução



À medida que a aprendizagem automática se torna cada vez mais integrada nos processos de tomada de decisão — particularmente em domínios de alto risco, como as finanças —, é essencial avaliar criticamente as implicações éticas da implementação destes modelos. Embora os modelos preditivos sejam normalmente avaliados com base em métricas de desempenho, como a precisão, estas por si só são insuficientes para garantir a equidade, especialmente quando os dados subjacentes incluem atributos sensíveis, como género, etnia ou estatuto socioeconómico (Jesus et al., 2024).

Esta atividade baseada em demonstração convida os estudantes a interagir com o Aequitas, um kit de ferramentas de auditoria de código aberto, por meio de uma exploração prática de um cenário realista de deteção de fraudes. O objetivo é duplo: aprofundar a compreensão dos alunos sobre o viés algorítmico e familiarizá-los com ferramentas e técnicas práticas que apoiam o desenvolvimento de sistemas de aprendizagem automática mais transparentes e responsáveis.

O Aequitas fornece uma estrutura abrangente para avaliar a equidade em modelos de classificação, examinando como os resultados preditivos são distribuídos entre os grupos demográficos. Ele oferece uma variedade de métricas de equidade — incluindo Taxa de Falsos Positivos (FPR), Taxa de Descobertas Falsas (FDR) e Paridade Estatística — que ajudam a revelar disparidades no desempenho do modelo que podem não ser visíveis através de métodos de avaliação convencionais. Mesmo quando treinados em conjuntos de dados desequilibrados ou tendenciosos, os modelos podem ser auditados para avaliar se tratam diferentes subgrupos de forma equitativa.

Ao envolverem-se diretamente com o Aequitas neste contexto aplicado, os estudantes adquirirão competências técnicas em auditoria de equidade e uma consciência ética mais ampla dos desafios inerentes à implementação da aprendizagem automática em domínios onde as consequências do viés podem ser particularmente graves.

03 Apresentação das ferramentas

O cenário envolve a auditoria de um **modelo de classificação binária** treinado para detectar **fraudes em contas bancárias.**

Os sistemas de deteção de fraudes normalmente sofrem com

- **Grave desequilíbrio de classes** (casos de fraude são raros);
- **Riscos elevados** (classificações erradas podem prejudicar indivíduos ou instituições);
- **Preconceitos ocultos** (características sensíveis podem estar correlacionadas com os resultados dos rótulos).

*A simulação usa o **conjunto de dados Bank Account Fraud (BAF)**, um conjunto de dados sintéticos em grande escala que preserva a privacidade e replica padrões reais de fraude bancária.*

As principais características incluem

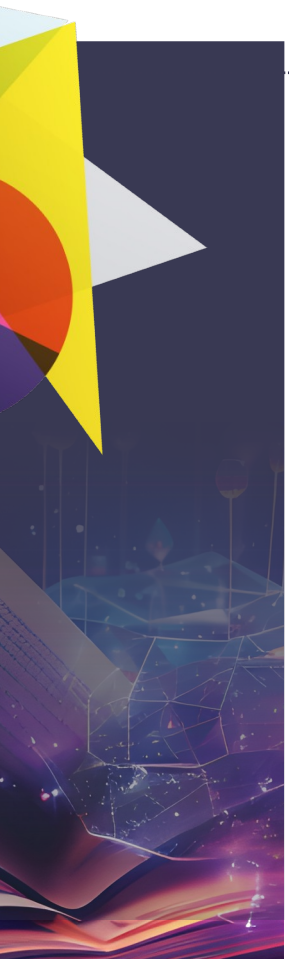
01 Variantes que simulam viés de amostragem, desvio temporal e desequilíbrio de características

02 Atributos demográficos que permitem a análise da equidade;

03 Forte desequilíbrio de classes para realismo.

Estas condições permitem um conjunto de experiências significativas sobre como as métricas de equidade se comportam sob diferentes premissas de modelo e distribuições de dados.

04 Execução da simulação



01 Acesse o Caderno de Simulação

Aceda a https://tinyurl.com/4b9u9hun

02 Execute todo o código

Execute o caderno na íntegra para carregar o conjunto de dados, treinar um modelo de classificação binária e gerar previsões. Para isso, clique em reproduzir no canto superior esquerdo de cada célula.

03 Realize uma auditoria de equidade com o Aequitas

- Use o Aequitas para gerar um relatório de equidade
- Concentre-se em métricas como FPR, FDR e paridade estatística
- Identifique quais grupos são tratados de forma injusta nas previsões do modelo

04 Compare os resultados e interprete as métricas

Analise as disparidades de equidade entre os grupos. Compare as métricas de desempenho (por exemplo, precisão) com as métricas de equidade para avaliar as compensações.

05 Reflita e discuta

- Que padrões de injustiça foram observados?
- Como é que estes resultados podem afetar indivíduos reais?
- Como essas ferramentas podem ser incorporadas ao pipeline de ML para melhorar os resultados éticos?

Detalhes do processo

Este pipeline consiste em várias etapas sequenciais projetadas para preparar dados, construir modelos preditivos e avaliá-los em termos de desempenho e justiça.
As etapas principais são as seguintes:

01 **Carregamento de dados** - O conjunto de dados é primeiro carregado a partir de uma fonte especificada (por exemplo, ficheiro CSV, base de dados). Espera-se que contenha variáveis de características e um ou mais atributos sensíveis (por exemplo, género, raça) necessários para a avaliação da equidade.

02 **Pré-processamento** - Os dados passam por várias etapas de pré-processamento para garantir a qualidade e a consistência (imputação e normalização).

03 **Modelação -** Vários modelos de aprendizagem máquina são treinados e avaliados. Esses modelos podem incluir algoritmos que não levam em consideração a equidade (padrão) e abordagens que levam em consideração a equidade, incorporando técnicas de mitigação de viés durante o treinamento ou pós-processamento.

04 **Avaliação da equidade com Aequitas** - Os modelos treinados são avaliados não apenas em termos de precisão e outras métricas de desempenho, mas também em termos de equidade, utilizando o kit de ferramentas Aequitas.

05 Conclusão

Esta demonstração mostra que mesmo modelos precisos podem levar a **resultados injustos** quando implementados sem auditoria de equidade

Utilizando o Aequitas, os estudantes descobrem como **o preconceito pode persistir** em configurações de classificação binária e como diferentes grupos demográficos podem sofrer diferentes taxas de classificação incorreta. Ao se envolverem diretamente com métricas de equidade e realizarem auditorias reais, os estudantes desenvolvem competências técnicas e consciência ética. Mais importante ainda, eles aprendem **que a deteção de preconceitos não é um complemento,** mas uma parte essencial da construção **de sistemas de IA confiáveis** — particularmente em domínios como finanças, onde as previsões dos modelos têm consequências graves.

06 Referências

- NJesus, S., Saleiro, P., e Silva, I. O., Jorge, B. M., Ribeiro, R. P., Gama, J., ... & Ghani, R. (2024). Aequitas flow: Streamlining fair ml experimentation. Journal of Machine Learning Research, 25(354), 1-7.

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.