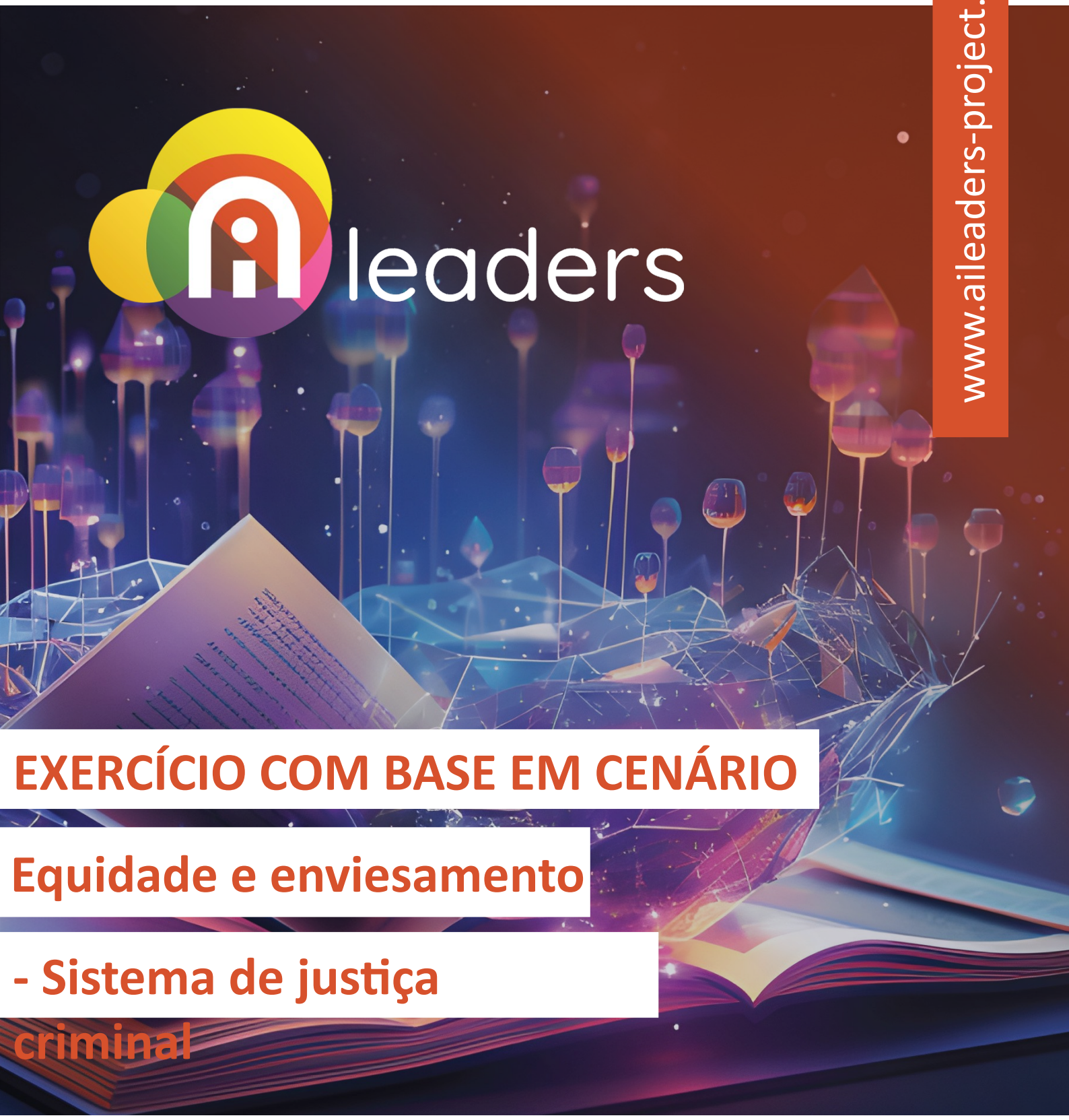


www.aileaders-project.eu

EXERCÍCIO COM BASE EM CENÁRIO

Equidade e enviesamento

- Sistema de justiça criminal

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.

EXERCÍCIO COM BASE EM CENÁRIO
Equidade e enviesamento - Sistema de justiça criminal

01 Resumo

Tipo de Recurso de Aprendizagem Online

Exercício com base em cenário

Objetivo/Finalidade

Análise do viés num algoritmo de aprendizagem automática para o sistema de justiça criminal

Resultados de aprendizagem esperados

- Consciência do viés nas previsões dos algoritmos de aprendizagem automática
- Deteção de preconceitos / vieses nas previsões

Palavras-chave

- Viés
- Equidade
- Aprendizagem automática
- Sistema de justiça criminal

Abordagem Metodológica Sugerida:

Aprendizagem baseada em casos

02 Introdução

01 De forma específica, este exercício aborda o viés no algoritmo de aprendizagem automática COMPAS utilizado no sistema de justiça criminal dos Estados Unidos.

02 O algoritmo de aprendizagem automática COMPAS calcula uma pontuação de risco que é utilizada para ajudar o juiz a decidir sobre a prisão preventiva e a definição da fiança.

03 O juiz decide com base na sua avaliação do risco de um arguido libertado não comparecer no julgamento ou causar danos ao público (reincidência).

04 O algoritmo COMPAS calcula uma pontuação de risco, que mede o risco de um arguido reincidir (previsão de reincidência).

05 Essa pontuação de risco é utilizada para ajudar o juiz a tomar decisões sobre a prisão preventiva e a fiança.

ProPublica

01 **A ProPublica** publicou um **artigo** sobre o viés no algoritmo COMPAS.

02 Os dados do COMPAS também foram obtidos e divulgados publicamente pela ProPublica.

03 Apresentação das ferramentas

Este exercício com base em cenário requer conhecimentos sobre:

01 **A linguagem de programação Python;**

02 **O software Jupyter notebook.**

- Um computador com Python e Jupyter é também necessário para este exercício com base em cenário.
- Além dos módulos Python básicos / integrados, são também necessárias as bibliotecas pandas, NumPy, Matplotlib e scikit-learn.

01 Este exercício com base em cenário utiliza conteúdos que foram desenvolvidos pela NOS (uma operadora de telecomunicações portuguesa), mas que estão disponíveis publicamente na plataforma de desenvolvimento GitHub.

02 Mais especificamente, neste exercício com base em cenário, utilizaremos o conteúdo do módulo SLU17 – Ethics and Fairness.

03 Este módulo faz parte do curso Intro to Data Science da NOS.

04 Por sua vez, este curso faz parte do percurso de aprendizagem mais amplo FAAST Advance Data Science.

05 Todos estes conteúdos são utilizados pela NOS para integrar novos funcionários, mas estão disponíveis publicamente no GitHub.

06 Além disso, é possível ver todos os repositórios públicos da NOS na sua página principal do GitHub.

07 Todos os ficheiros necessários para este exercício de cenário são fornecidos na pasta «oer_files».

08 Portanto, não há necessidade de descarregar nenhum ficheiro da página do GitHub correspondente ao módulo SLU17 – Ethics and Fairness .

09 De facto, todos esses ficheiros já estão disponíveis na pasta “oer_files”.

10 Naturalmente, todo o crédito desses ficheiros pertence à NOS.

11 Existem algumas diferenças mínimas entre alguns ficheiros na pasta «oer_files» e os ficheiros correspondentes no GitHub.

12 No ficheiro «README.md» fornecido, foi eliminada uma ligação para um ficheiro do Google Docs que está disponível apenas para os funcionários da NOS. No entanto, este ficheiro não é necessário para este exercício com base em cenário.

No ficheiro «Exercise notebook.ipynb foram feitas as seguintes alterações:

01 A primeira célula de código foi tornada editável e uma linha de código que remetia para um estilo Matplotlib obsoleto foi eliminada;

02 Foi adicionada uma célula de código logo abaixo do cálculo da FPR para réus negros, na qual é feito o output de «fpr_b», para torná-la idêntica ao que é feito acima no caso de réus brancos;

03 A última célula de código estava vazia e, como tal, foi eliminada.

A pasta “oer_files” também inclui um ficheiro que não está incluído na página GitHub, nomeadamente o ficheiro **“Exercise notebook solved.ipynb”.** Por conveniência e referência, este ficheiro contém o código/soluções para todos os exercícios no ficheiro “Exercise notebook.ipynb”.

04 Atividades práticas

O primeiro passo é ler o conteúdo do ficheiro “Learning notebook.ipynb”.

Este ficheiro aborda conceitos básicos relativos:

01 **aos componentes de um sistema de aprendizagem**

02 **a privacidade por predefinição e;**

03 **preconceito e imparcialidade.**

- Em seguida, deverão ser feitos os exercícios no ficheiro “Exercise notebook.ipynb”.
- Este ficheiro começa com uma descrição básica do cenário: viés nas pontuações de risco calculadas pelo algoritmo COMPAS usado no sistema de justiça criminal dos EUA.
- Para referência adicional, o caderno tem links para um livro sobre justiça e aprendizagem automática e para o artigo da ProPublica sobre enviesamento de algoritmos de aprendizagem automática na determinação de penas criminais.

EXERCÍCIO 01

- O exercício 1 consiste em desenhar distribuições para a pontuação de risco calculada pelo algoritmo COMPAS.
- O objetivo é desenhar a distribuição geral, bem como as distribuições por raça.
- Em particular, devem ser representadas as distribuições da pontuação de risco para réus brancos e negros.

EXERCÍCIO 02

- O exercício 2 consiste em desenhar distribuições para as pontuações de risco atribuídas à classe positiva (reincidentes).
- O objetivo é desenhar a distribuição geral, bem como as distribuições por raça.
- Mais uma vez, as distribuições da pontuação de risco para réus reincidentes, brancos e negros, devem ser desenhadas.

EXERCÍCIO 03

- O exercício 3 considera os réus que foram classificados como de alto risco de reincidência (ou seja, que tiveram um valor alto na pontuação de risco).
- O objetivo é calcular a Taxa de Falsos Positivos (FPR) para esses réus de alto risco.
- A FPR deve ser calculada para réus de alto risco, tanto brancos quanto negros.

A FPR (taxa de falsos positivos) também é conhecida como taxa de falsos alarmes ou probabilidade de falsos alarmes. A FPR é descrita no ficheiro. No entanto, como se trata de uma medida crucial e fundamental para compreender o viés neste exercício com base em cenário, fornecemos também aqui uma explicação mais detalhada sobre a FPR.

A FPR é calculada como

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

01 FP é o número de falsos positivos, ou seja, no nosso caso, o número de não reincidentes reais que foram incorretamente classificados como reincidentes;

02 TN é o número de verdadeiros negativos, ou seja, no nosso caso, o número de não reincidentes reais que foram corretamente classificados como não reincidentes;

03 N é o número total de negativos reais, ou seja, no nosso caso, o número total de não reincidentes reais.

01 Portanto, a FPR é a taxa dos negativos reais que foram incorretamente classificados como positivos.

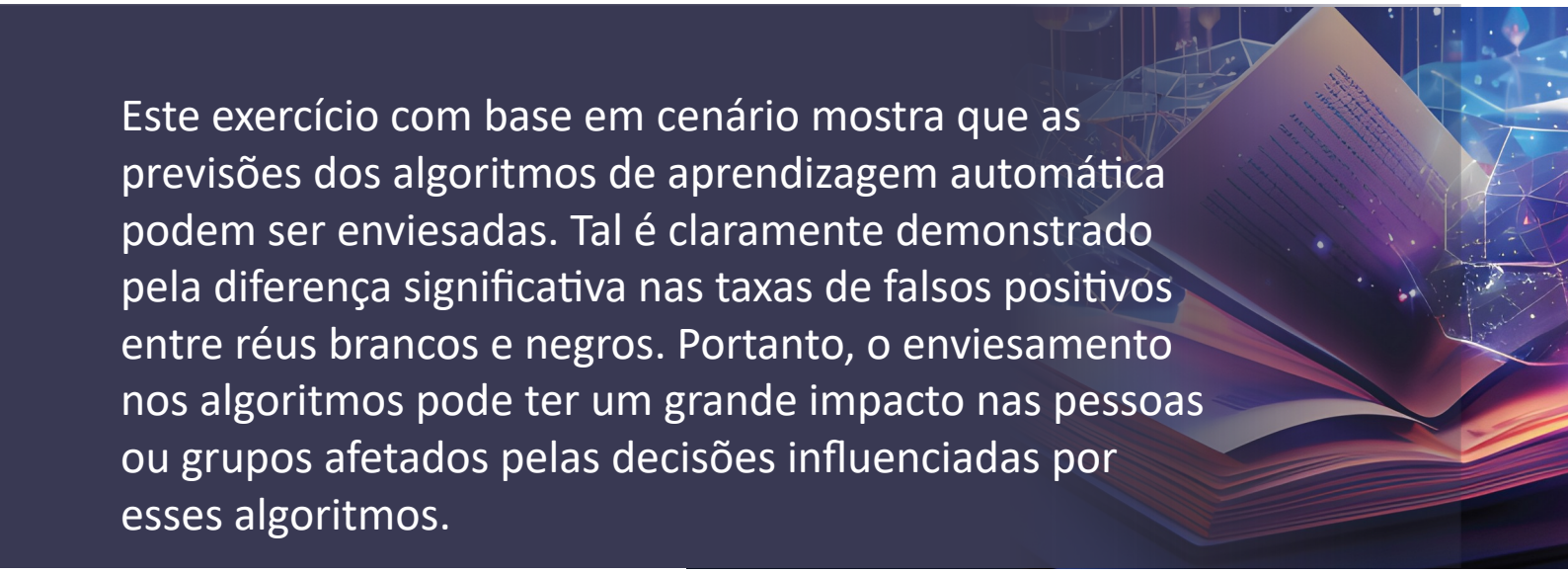
02 No nosso contexto, a FPR é a taxa de não reincidentes reais que foram incorretamente classificados como reincidentes.

03 Valores mais altos são, portanto, mais indesejáveis, uma vez que uma FPR mais alta significa que uma proporção maior de não reincidentes reais foi incorretamente classificada como reincidente.

04 Como mencionado, o objetivo é escrever o código para fazer todos os exercícios no ficheiro **"Exercise notebook.ipynb".**

05 Como mencionado anteriormente, por conveniência e referência, também é fornecido um ficheiro que já está preenchido com **o código** apropriado **("Exercise notebook solved.ipynb").**

05 Conclusões



Este exercício com base em cenário mostra que as previsões dos algoritmos de aprendizagem automática podem ser enviesadas. Tal é claramente demonstrado pela diferença significativa nas taxas de falsos positivos entre réus brancos e negros. Portanto, o enviesamento nos algoritmos pode ter um grande impacto nas pessoas ou grupos afetados pelas decisões influenciadas por esses algoritmos.

06 Referências

- Fairness and Machine Learning – Book
- ProPublica Article on Machine Bias in Criminal Sentencing
- NOS Ethics and Fairness Learning Unit on GitHub
- NOS Intro to Data Science Course on GitHub
- NOS FAAST Learning Path on GitHub
- Main NOS Page on GitHub

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.