Ferramenta 5: perfil de dados

O perfil de dados é uma análise exploratória que fornece informações para avaliar a qualidade, integridade, temporalidade, consistência e possível viés de um conjunto de dados que será usado para treinar um modelo de aprendizado de máquina (Gebru *et al*., 2018).

Fonte de coleta e origem dos dados	
Nome do conjunto de dados usado	
Que instituição criou o conjunto de dados?	
Com que finalidade a instituição criou o conjunto de dados utilizado?	
Que mecanismos ou procedimentos foram usados para coletar os dados (por exemplo, pesquisa domiciliar, sensor, software, API)? Eles estão em conformidade com a regulamentação atual de proteção de dados?	
Qual é a escala do conjunto de dados?	
Obter documentação para cada variável do conjunto de dados. Fornecer uma breve descrição, inclusive seu nome e tipo, o que representa, como é medido etc	

Governança de dados	
Qual é o domínio dos dados (por exemplo, proprietário, público, pessoal)?	
Se forem pessoais, os dados são identificados, têm pseudônimos, têm pseudônimos não vinculados, são anônimos ou agregados?	
Se forem privados, os direitos de propriedade intelectual e a proteção de dados pessoais são levados em conta?	

Estrutura dos dados	
Os dados são estáticos ou dinâmicos? Se forem dinâmicos, com que frequência serão atualizados?	
Captar a frequência (diária, semanal, mensal) ou o número médio de observações por indivíduo. Que versão do conjunto de dados está sendo usada?	
É o conjunto de dados mais adequado disponível, levando-se em conta o problema em questão?	

Qualidade dos dados	
Como os dados foram obtidos (observados, derivados, sintéticos ou fornecidos por indivíduos ou organizações)?	
Os dados são representativos da população de interesse?	
Descrever o tipo de amostragem usada.	
Analisar a cobertura espacial e temporal dos dados.	
Analisar a cobertura dos grupos protegidos (sexo, raça, idade etc.).	

Descrever as dimensões importantes nas quais a amostra de dados pode diferir da população, particularmente os vieses de seleção não medidos. Use a literatura relacionada ao assunto e informações de especialistas	
Identificar possíveis “estados indesejáveis” nos dados que poderiam dar origem a vieses e inequidade prejudiciais para um determinado subgrupo ou qualquer outro padrão considerado subótimo ou indesejável do ponto de vista da política social	
Está faltando algum valor? Em caso afirmativo, explicar os motivos pelos quais essas informações não estão disponíveis (inclui informações removidas intencionalmente). Identificar as razões para a falta desses dados e considerar se os dados ausentes estão associados à variável a ser prevista. Documentar qualquer processo de imputação usado para substituir os dados que faltam.	