Ferramenta 4: lista de verificação de IA robusta e responsável

Esta ferramenta consolida os principais receios quanto à dimensão de risco do ciclo de vida da IA. A lista de verificação deve ser continuamente revisada pela equipe técnica, acompanhada pelo tomador de decisões (Fritzler, 2015; drivendata, 2019).

Conceitualização e formulação

☐ **Definição correta do problema e a resposta da política pública:**

- (Qualitativo) O problema de política pública está claramente definido?
- (Qualitativo) Descrever como esse problema é abordado atualmente, considerando as respostas das instituições relacionadas, e como o uso da IA melhoraria a resposta do governo a esse problema.
- (Qualitativo) Os grupos ou atributos protegidos dentro do projeto foram identificados (por exemplo, idade, sexo, escolaridade, raça, nível de marginalização etc.)?
- (Qualitativo) As ações ou intervenções a serem realizadas foram definidas com base no resultado do sistema de IA?

☐ **Princípios de IA**

- (Quantitativo) A necessidade de um sistema de IA é justificada, levando-se em conta outras soluções possíveis que não requerem a utilização de dados pessoais e decisões automatizadas?
- (Quantitativo) Há evidências de que tanto a ação das políticas públicas quanto a indicação do sistema de IA beneficiarão as pessoas e o planeta ao impulsionar o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar?
- (Qualitativo) Projetos semelhantes foram identificados e analisados em busca de aprendizados e erros comuns?
- (Quantitativo) A possibilidade de minimizar a exposição de informações pessoais identificáveis foi considerada (por exemplo, anonimizando ou não coletando informações não relevantes para a análise)?

Ciclo de vida

Coleta e processamento de dados

☐ **Qualidade e relevância dos dados disponíveis**

Estados indesejáveis ou subótimos nos dados coletados

- (Qualitativo) Discutir possíveis desigualdades sociais históricas no caso de uso com especialistas no assunto.
- (Quantitativo) Fazer uma análise exploratória dos dados disponíveis com base nos quais o modelo será treinado para identificar possíveis vieses históricos ou estados indesejáveis.

Falta de correspondência entre variáveis disponíveis e variáveis ideais

- (Qualitativo) As variáveis-alvo ideais devem ser claramente estabelecidas. As variáveis coletadas/disponíveis devem ser analisadas para entender até que ponto são adequadas para substituir a variável-alvo. Os vieses sistemáticos ou a validade da métrica substituta devem ser identificados.
- (Qualitativo) O uso da variável de resposta selecionada foi claramente justificado para os propósitos da intervenção?

☐ **Qualificação e exaustividade dos dados para a população-alvo**

Amostras probabilísticas e naturais

- (Qualitativo) Foram analisadas possíveis diferenças entre o banco de dados e a população para a qual o sistema de IA está sendo desenvolvido? (Utilizar a bibliografia relacionada ao tema e as informações dos especialistas. Estudar, em particular, os vieses de seleção não medidos).
- (Quantitativo) Embora os modelos possam ser construídos com diversas fontes de dados, projetadas ou naturais, o ideal é que a validação seja realizada com uma amostra que permita uma inferência estatística da população. A amostra de validação deve abranger adequadamente a população-alvo e as subpopulações de interesse.

Atributos ausentes ou incompletos

- (Qualitativa) Foi realizada uma análise de valores ausentes e variáveis omitidas?
- (Qualitativa) Foi identificado se existem variáveis omitidas importantes para as quais não há medidas associadas (se houver)?
- (Qualitativa) Foram identificadas as razões pelas quais existem observações ausentes (se houver)?
- (Quantitativa) Os processos de imputação precisam ser avaliados quanto à sua sensibilidade a suposições e dados. De preferência, devem ser usados métodos de imputação múltipla que permitam avaliar incertezas na imputação (Little & Rubin, 2002; Buuren & Groothuis-Oudshoorn, 2011).

☐ **Comparação causal**

- (Qualitativo) Compreender e descrever as razões pelas quais a variável de resposta está correlacionada às variáveis conhecidas e desconhecidas. Descrever possíveis vieses com base no conhecimento e na análise de especialistas.
- (Qualitativo) Caso não se tenha trabalhado para garantir a causalidade dos resultados, as limitações dos resultados foram comunicadas explicitamente ao responsável pelas políticas públicas?
- (Quantitativa) No caso de tentativa de inferência causal com modelos, as hipóteses, considerações ou métodos usados para apoiar uma interpretação causal devem ser descritos. As verificações de robustez devem ser realizadas e documentadas.

Desenvolvimento do modelo e validação

☐ **Ausência ou uso inadequado de amostras de validação**

- (Quantitativa) As amostras de validação e teste foram construídas adequadamente, considerando-se um tamanho apropriado, abrangendo-se subgrupos de interesse e protegidos, e evitando-se vazamentos de informações durante sua implementação?

☐ **Vazamentos de informações**

Contaminação treinamento-validação

- (Quantitativa) Qualquer processamento e preparação de dados de treinamento deve evitar o uso de dados de validação ou teste. Uma barreira sólida deve ser mantida entre treinamento, de um lado, e validação e teste, do outro. Isso inclui recodificação de dados, normalizações, seleção de variáveis, identificação de dados atípicos e qualquer outro tipo de preparação de qualquer variável a ser incluída nos modelos, o que também inclui ponderações ou equilíbrio de amostras com base em sobre/subamostragem.

Vazamentos de dados não disponíveis na previsão

- O esquema de validação **deve replicar com a maior precisão possível** o esquema com base no qual as previsões serão aplicadas. Isso inclui a necessidade de replicar:
    - Janelas temporárias de observação e registro de variáveis, bem como janelas de previsão.
    - Se houver grupos nos dados, considerar se as informações estarão disponíveis para cada grupo quando a previsão for feita ou se é necessário realizar previsões para novos grupos.

☐ **Probabilidades e classes**

Dados não balanceados

- (Quantitativa) Realizar **previsões de probabilidade** em vez de classe. Essas probabilidades podem ser incorporadas ao processo de decisão posterior como tais. Evitar pontos de corte de probabilidade padrão, como 0,5, ou previsões realizadas com base na probabilidade máxima.
- (Quantitativa) Quando o número absoluto de casos minoritários é muito pequeno, pode ser muito difícil encontrar informações apropriadas para discriminar essa classe. Mais dados precisam ser **coletados** da classe minoritária.
- Quantitativa) A subamostragem da classe dominante (ponderando-se os casos para cima a fim de não perder a calibração) pode ser uma estratégia bem-sucedida para reduzir o tamanho dos dados e o tempo de treinamento sem afetar o desempenho preditivo.
- (Quantitativa) Replicar a classe minoritária a fim de melhor balancear as classes (sobremostragem).
- Quantitativa) Algumas técnicas de aprendizado de máquina permitem que cada classe seja ponderada por um peso diferente, para que o peso total de cada classe fique equilibrado. Se isso for possível, é preferível em relação à subamostragem ou à sobreamostragem.

Pontos de corte arbitrários

- (Quantitativo) O uso de algoritmos de classificação probabilística é mais adequado para que a tomada de decisões incorpore a incerteza quanto à classificação.
- (Quantitativo) Evitar pontos de corte de probabilidade padrão, como 0,5. Escolher uma interpretação ideal das probabilidades previstas, analisando-se as métricas de erro.

Idoneidade das métricas de avaliação

- (Qualitativa) As implicações dos diferentes tipos de erro para o caso de uso específico foram questionadas, bem como a forma correta de avaliá-los?
- (Qualitativa) As limitações do modelo foram claramente explicadas, identificando-se tanto falsos positivos quanto falsos negativos, e as implicações que uma decisão do sistema teria na vida da população beneficiária?
- (Quantitativa) Foi implementada uma análise de custo-benefício do sistema em relação ao *status quo* ou outras estratégias de suporte/tomada de decisões (quando possível)?

Sub e sobreajuste

- (Quantitativa) Sobreajuste: modelos com uma grande lacuna entre validação e treinamento (indícios de sobreajuste) devem ser evitados. Se necessário, métodos para moderar o sobreajuste, como regularização, restrição do espaço funcional de modelos possíveis, utilização de mais dados de treinamento ou disrupção os dados de treinamento, entre outros, devem ser refinados (Hastie, Tibshirani & Friedman, 2017).
- (Quantitativa) Subajuste: subconjuntos importantes de casos (por exemplo, grupos protegidos) devem ser revisados para confirmar que não existem erros sistemáticos indesejáveis.

☐ **Erros não quantificados e avaliação humana**

Falhas não medidas pelo modelo

- (Qualitativa) Foi realizada uma avaliação com especialistas do caso de uso para procurar vieses ou erros conhecidos? (Por exemplo, podem ser usados painéis de revisores para examinar previsões específicas e considerar se são ou não razoáveis. Esses painéis devem ser equilibrados em termos dos tipos de usuário previstos, inclusive tomadores de decisão, se necessário).

☐ **Equidade e desempenho diferencial dos preditores**

Definição de justiça e equidade algorítmicas

- (Qualitativa) A medida de justiça algorítmica a ser utilizada no projeto foi definida com especialistas e tomadores de decisões?
- (Quantitativa) Quando há atributos protegidos, deve-se avaliar até que ponto as previsões se afastam da definição de justiça algorítmica escolhida.
- (Quantitativa) No caso de classificação, os pontos de corte podem ser ajustados para diferentes subgrupos a fim de atingir a igualdade de oportunidades.

Uso e monitoramento

☐ **Degradação do desempenho**

- (Qualitativo) Existe um plano para monitorar o desempenho do modelo e a coleta de informações ao longo do tempo?
- (Quantitativo) Monitorar várias métricas associadas a previsões em subgrupos predefinidos (inclusive variáveis protegidas).
- (Quantitativo) Monitorar o desvio nas distribuições de características em relação ao conjunto de treinamento.

- (Quantitativo) Monitorar mudanças na metodologia de coleta e processamento de dados que possam reduzir a qualidade das previsões.
- (Quantitativo) Sempre que possível, planejar atribuir, por meio de projetos experimentais, tratamentos aleatórios (ou conforme o status quo) a algumas unidades. Fazer comparações de desempenho e comportamento entre essa amostra e os resultados de acordo com o regime algorítmico.
- (Quantitativo) Identificar as variáveis não observadas e procurar a forma de medi-las. Se possível, reajustar o modelo e avaliar seu desempenho com base nessas novas informações.

Prestação de contas

☐ Interpretabilidade e explicação das previsões

Explicabilidade de previsões individuais

- (Qualitativo) Foram analisados os requisitos legais e éticos de explicabilidade e interpretabilidade necessários para o caso de uso?
- (Qualitativo) Existe algum processo para explicar a um determinado indivíduo por que uma determinada decisão foi tomada?
- (Qualitativo) Os prós e contras dos algoritmos, de acordo com seu nível de interpretabilidade e explicabilidade, foram discutidos para escolher o mais apropriado?

Modelos parcimoniosos

- (Qualitativa) Incluir todas as características disponíveis para construir modelos aumenta o risco de vieses. As variáveis, por serem incluídas no processo de aprendizagem, devem ter alguma base teórica ou explicação do motivo pelo qual podem ajudar na tarefa de previsão.
- (Quantitativa) Métodos mais parcimoniosos, que usam menos características, são preferíveis a modelos que usam muitas características.
- (Quantitativa) Métodos como gráficos de dependência parcial (Friedman, 2001) ou importância baseada em permutações (Breiman, 2001; Molnar, 2019) podem apontar variáveis problemáticas que recebem muito peso na previsão, em contraste com observações anteriores ou conhecimento especializado.

☐ Rastreabilidade

- (Quantitativa) O processo de ingestão, transformação, modelagem e tomada de decisões (inclusive fonte de dados, infraestrutura e dependências, código, métricas e interpretação de resultados) está bem documentado?
- (Qualitativo) As deficiências, limitações e vieses do modelo foram comunicados às partes interessadas para que possam ser levados em consideração na tomada e no suporte às decisões?
- (Qualitativo) A equipe técnica completou o “Perfil de dados” (ver Ferramenta 2) e o “Perfil do modelo” (ver Ferramenta 3), e foi definido algum processo de atualização contínua dessas ferramentas?