Herramienta 1: Lista de verificación de IA robusta y responsable

Esta herramienta consolida las principales preocupaciones referidas a la dimensión de riesgo del ciclo de vida de IA. La lista de verificación debe revisarla continuamente el equipo técnico, acompañado por el tomador de decisiones (Fritzler, 2015; drivendata, 2019).

Conceptualización y diseño

☐ **Definición correcta del problema y de la respuesta de la política pública:**

- (Cualitativo) ¿Está claramente definido el problema de política pública?
- (Cualitativo) Describir cómo se aborda actualmente este problema –considerando las respuestas de las instituciones relacionadas– y cómo el uso de la IA mejoraría la respuesta gubernamental a este problema.
- (Cualitativo) ¿Se identificaron los grupos o los atributos protegidos dentro del proyecto (por ejemplo, edad, género, nivel educativo, raza, nivel de marginación, etc.)?
- (Cualitativo) ¿Se definieron las acciones o intervenciones que se llevarán a cabo en función del resultado del sistema de IA?

☐ **Principios de la IA**

- (Cuantitativo) ¿Se ha justificado la necesidad de un sistema de IA, teniendo en cuenta otras posibles soluciones que no requieran el uso de datos personales y decisiones automatizadas?
- (Cuantitativo) ¿Existen pruebas de que tanto la acción de las políticas públicas como la recomendación del sistema de IA supondrán un beneficio para las personas y el planeta al impulsar el crecimiento inclusivo, el desarrollo sostenible y el bienestar?
- (Cualitativo) ¿Se han identificado y analizado proyectos similares para obtener aprendizajes y errores comunes?
- (Cuantitativo) ¿Se ha considerado la posibilidad de minimizar la exposición de información personal identificable (por ejemplo, anonimizando o no recogiendo información no relevante para el análisis)?

Ciclo de vida

☐ **Recolección y procesamiento de datos**

Calidad y relevancia de los datos disponibles
Estados indeseables o subóptimos en datos recolectados

- (Cualitativo) Discutir las posibles desigualdades sociales históricas en el caso de uso con especialistas en la materia.
- (Cuantitativo) Realizar un análisis exploratorio de los datos disponibles con los que se entrenará el modelo para identificar posibles sesgos históricos o estados indeseables.

Mala correspondencia entre variables disponibles y variables ideales

- (Cualitativo) Las variables objetivo ideales deben estar claramente establecidas. Las variables recogidas/disponibles deben analizarse para comprender hasta qué punto son adecuadas para sustituir a la variable objetivo. Deben identificarse los sesgos sistemáticos o la validez de la métrica sustituto.
- (Cualitativo) ¿Se ha justificado claramente el uso de la variable de respuesta seleccionada para los fines de la intervención?

☐ **Cualificación y exhaustividad de los datos para la población objetivo**

Muestras probabilísticas y naturales

- (Cualitativo) ¿Se han analizado las posibles diferencias entre la base de datos y la población para la que se está desarrollando el sistema de IA? (Utilizar la bibliografía relacionada con el tema y la información de los expertos. Estudiar en particular los sesgos de selección no medidos).
- (Cuantitativo) Aunque los modelos pueden construirse con diversas fuentes de datos, diseñadas o naturales, lo ideal es que la validación se realice con una muestra que permita la inferencia estadística a la población. La muestra de validación debe cubrir adecuadamente la población objetivo y las subpoblaciones de interés.

Atributos faltantes o incompletos

- (Cualitativa) ¿Se ha realizado un análisis de valores faltantes y de variables omitidas?
- (Cualitativa) ¿Se ha identificado si existen variables omitidas importantes para las cuales no se tiene mediciones asociadas (en caso de existir)?
- (Cualitativa) ¿Se ha identificado las razones por las que existen observaciones faltantes (en caso de existir)?
- (Cuantitativa) Los procesos de imputación tienen que evaluarse en cuanto a su sensibilidad a supuestos y datos. De preferencia, deben utilizarse métodos de imputación múltiple que permitan evaluar incertidumbre en la imputación (Little & Rubin, 2002) (Buuren & Groothuis-Oudshoorn, 2011).

☐ **Comparación causal**

- (Cualitativo) Comprender y describir las razones por las que la variable de respuesta está correlacionada con variables conocidas y desconocidas. Describir los posibles sesgos basados en el conocimiento y el análisis de los expertos.
- (Cualitativo) En caso de que no se haya trabajado para asegurar la causalidad en los resultados, ¿se comunicaron explícitamente las limitaciones de los resultados al responsable de las políticas públicas?
- (Cuantitativo) En el caso de que se intente la inferencia causal con modelos, deben describirse las hipótesis, consideraciones o métodos utilizados para apoyar una interpretación causal. Deben realizarse y documentarse las comprobaciones de robustez.

Desarrollo del modelo y validación

☐ Ausencia o uso inadecuado de muestras de validación

- (Cuantitativa) ¿Se construyeron adecuadamente las muestras de validación y prueba, considerando un tamaño apropiado, cubriendo a subgrupos de interés y protegidos y evitando fugas de información durante su implementación?

☐ Fugas de información

Contaminación entrenamiento-validación

- (Cuantitativa) Cualquier procesamiento y preparación de datos de entrenamiento debe evitar de cualquier manera usar los datos de validación o prueba. Debe mantenerse una barrera sólida entre entrenamiento vs. validación y prueba. Esto comprende recodificación de datos, normalizaciones, selección de variables, identificación de datos atípicos y cualquier otro tipo de preparación de cualquier variable que va a incluirse en los modelos, lo que también abarca ponderaciones o balance de muestras basadas en sobre/sub muestreo.

Fugas de datos no disponibles en la predicción

- El esquema de validación debe replicar tan cerca como sea posible el esquema con el cual se aplicarán las predicciones. Esto incluye que hay que replicar:
- Ventanas temporales de observación y registro de variables y ventanas de predicción.
- Si existen grupos en los datos, considerar si se tendrá información disponible de cada grupo cuando se haga la predicción, o si es necesario predecir para nuevos grupos.

☐ Probabilidades y clases

Datos desbalanceados

- (Cuantitativa) Hacer predicciones de probabilidad en lugar de clase. Estas probabilidades pueden incorporarse al proceso de decisión posterior como tales. Evitar puntos de corte estándar de probabilidad como 0.5, o predecir según máxima probabilidad.
- (Cuantitativa) Cuando el número absoluto de casos minoritarios es muy reducido, puede ser muy difícil encontrar información apropiada para discriminar esa clase. Se requiere recolectar más datos de la clase minoritaria.
- (Cuantitativa) Submuestrear la clase dominante (ponderando hacia arriba los casos para no perder calibración) puede ser una estrategia exitosa para reducir el tamaño de los datos y el tiempo de entrenamiento sin afectar el desempeño predictivo.
- (Cuantitativa) Replicar la clase minoritaria, para balancear mejor las clases (sobremuestreo).
- (Cuantitativa) Algunas técnicas de aprendizaje automático permiten ponderar cada clase por un peso distinto para que el peso total de cada clase quede balanceado. Si esto es posible, es preferible a sub o sobremuestreo.

Puntos de corte arbitrarios

- (Cuantitativo) El uso de algoritmos de clasificación probabilística es más adecuado para que la toma de decisiones incorpore la incertidumbre con respecto a la clasificación.
- (Cuantitativo) Evitar los puntos de corte de probabilidad estándar, como el 0,5. Elegir una interpretación óptima de las probabilidades predichas analizando las métricas de error.

Idoneidad de las métricas de evaluación

- (Cualitativa) ¿Se cuestionaron las implicaciones de los diferentes tipos de errores para el caso de uso específico, así como la forma correcta de evaluarlos?
- (Cualitativa) ¿Se explicó en forma clara las limitantes del modelo, identificando tanto los falsos positivos como los falsos negativos y las implicaciones que una decisión del sistema tendría en la vida de la población beneficiaria?
- (Cuantitativa) ¿Se implementó un análisis costo-beneficio del sistema contra el statu quo u otras estrategias de toma/soporte de decisión (cuando es posible)?

Sub y sobreajuste

- (Cuantitativa) Sobreajuste: debe evitarse modelos cuya brecha validación - entrenamiento sea grande (indicios de sobreajuste). De ser necesario, deben afinarse métodos para moderar el sobreajuste como regularización, restricción del espacio funcional de modelos posibles, usar más datos de entrenamiento o perturbar los datos de entrenamiento, entre otros (Hastie, Tibshirani, & Friedman, 2017).
- (Cuantitativa) Subajuste: deben revisarse subconjuntos importantes de casos (por ejemplo, grupos protegidos) para verificar que no existen errores sistemáticos indeseables.

☐ **Errores no cuantificados y evaluación humana**

Fallas no medidas por el modelo

- (Cualitativa) ¿Se realizó una evaluación con expertos del caso de uso para buscar sesgos o errores conocidos? (Por ejemplo, pueden usarse paneles de revisores que examinen predicciones particulares y consideren si son razonables o no. Estos paneles deben ser balanceados en cuanto al tipo de usuarios que se prevén, incluyendo tomadores de decisiones, si es necesario).

☐ **Equidad y desempeño diferencial de predictores**

Definición de justicia y equidad algorítmicas

- (Cualitativa) ¿Se definió con expertos y tomadores de decisiones la medida de justicia algorítmica que va a usarse en el proyecto?
- (Cuantitativa) Cuando existen atributos protegidos, debe evaluarse qué tanto se alejan las predicciones de la definición de justicia algorítmica elegida.
- (Cuantitativa) En el caso de clasificación, pueden ajustarse puntos de corte para distintos subgrupos con el fin de lograr equidad de oportunidad.

Uso y monitoreo

☐ **Degradación de desempeño**

- (Cualitativo) ¿Existe un plan para monitorear el desempeño del modelo y la recolección de información a lo largo del tiempo?
- (Cuantitativo) Monitorear varias métricas asociadas a las predicciones en subgrupos definidos con antelación (incluyendo variables protegidas).
- (Cuantitativo) Monitorear deriva en distribuciones de características con respecto al conjunto de entrenamiento.
- (Cuantitativo) Monitorear cambios en la metodología de levantamiento y procesamiento de datos que pueden reducir calidad de las predicciones.
- (Cuantitativo) Cuando sea posible, planear asignar bajo diseños experimentales tratamientos aleatorios (o según el statu quo) a algunas unidades. Hacer comparaciones de desempeño y comportamiento entre esta muestra y los resultados de acuerdo con el régimen algorítmico.
- (Cuantitativo) Identificar las variables no observadas y buscar la forma de medirlas. Si es posible, volver a ajustar el modelo y evaluar su rendimiento utilizando esta nueva información.

Rendición de cuentas

☐ **Interpretabilidad y explicación de predicciones**

Explicabilidad de predicciones individuales

- (Cualitativo) ¿Se analizaron los requerimientos legales y éticos de explicabilidad e interpretabilidad necesarios para el caso de uso?
- (Cualitativo) ¿Existe algún proceso para dar explicaciones a un individuo en particular sobre por qué se tomó una determinada decisión?
- (Cualitativo) ¿Se discutieron los pros y contras de los algoritmos según su nivel de interpretabilidad y explicabilidad para elegir el más apropiado?

Modelos parsimoniosos

- (Cualitativa) Incluir todas las características disponibles para construir modelos aumenta el riesgo de que se generen sesgos. Las variables por incluirse en el proceso de aprendizaje deben tener algún sustento teórico o explicación de por qué pueden ayudar en la tarea de predicción.
- (Cuantitativa) Métodos más parsimoniosos, que usan menos características, son preferibles a modelos que utilizan muchas características.
- (Cuantitativa) Métodos como gráficas de dependencia parcial (Friedman, 2001) o importancia basada en permutaciones (Breiman, 2001) (Molnar, 2019) pueden señalar variables problemáticas que reciben mucho peso en la predicción, en contra de observaciones pasadas o conocimiento experto.

☐ **Trazabilidad**

- (Cuantitativa) ¿Está bien documentado el proceso de ingesta, transformación, modelado y toma de decisión (incluyendo fuente de datos, infraestructura y dependencias, código, métricas e interpretación de resultados)?
- (Cualitativo) ¿Se han comunicado las deficiencias, limitaciones y sesgos del modelo a las partes interesadas para que se tengan en cuenta en la toma de decisiones y el apoyo a las mismas?
- (Cualitativo) ¿Ha completado el equipo técnico el Perfil de datos (ver la Herramienta 2) y el Perfil del modelo (ver la Herramienta 3), y se ha definido un proceso para la actualización continua de estas herramientas?