Una guía de

desafíos y recomendaciones para la implementación de una IA más justa

Fundación Vía Libre

Índice

Resumen ejecutivo

Proponemos recomendaciones técnicas para acompañar en los procesos de creación de sistemas de Inteligencia Artificial, más concretamente, en sistemas basados en Aprendizaje Automático. Estas recomendaciones surgen y se elaboran a partir de entrevistas estructuradas con personas que trabajan en aplicaciones prácticas de sistemas basados en datos, en diferentes roles y diferentes organizaciones del ecosistema tecnológico argentino.

Las principales recomendaciones que proponemos son:

1. Como paso previo a cualquier desarrollo, cuestionar la necesidad y valor social del mismo, teniendo en cuenta potenciales beneficios y perjuicios.
2. Definir el problema que se quiere resolver en términos no técnicos, desde la reflexión conjunta entre las partes interesadas en desarrollar una aplicación de inteligencia artificial. Tener en cuenta posibles usos secundarios de la aplicación, y sus potenciales efectos perniciosos.
3. Seleccionar las fuentes de datos con las que se va a entrenar el modelo preservando la protección de datos personales y propiedad intelectual, y con atención a posibles sesgos, de forma que todos los grupos queden representados de forma adecuada y no discriminatoria, con especial atención a las minorías.
4. En el entrenamiento de los modelos, incorporar metodología de comprobación de limitaciones habituales de los modelos que repercuten en resultados perniciosos: clases poco representadas, clases mayoritarias, tendencia al sobreajuste, etc.
5. En la evaluación de modelos, incorporar métricas para la detección temprana y sistemática de posibles sesgos en las predicciones.
6. Puesta en producción por fases y acompañada con los correspondientes mecanismos de monitoreo para detección temprana de efectos perniciosos del funcionamiento del algoritmo, incorporando diferentes perspectivas.

Se sostiene que el análisis de equidad debe ser transversal, es decir, un recurso reflexivo presente a lo largo todo el proceso de desarrollo y construcción de los sistemas, y no solo ser aplicado a la hora de evaluar los impactos de discriminación de un sistema una vez que ha sido desplegado.

El documento pretende contribuir al entendimiento de los procesos de desarrollo y los problemas asociados al mismo y servir de guía para quienes estén interesados en generar productos tecnológicos más respetuosos con los derechos humanos.

1. Introducción

Los sistemas basados en Inteligencia Artificial han demostrado tener el potencial de replicar y amplificar desigualdades sociales. Los modelos que se infieren de datos indiscriminados pueden inferir y amplificar patrones nocivos, como por ejemplo lenguaje abusivo o estereotipos.[1] Por ejemplo, se ha observado a menudo que buscadores de internet como Google search refuerzan estereotipos de género o raza. En 2018, para la consulta mujeres negras el buscador de Google ofrecía mucho más contenido pornográfico que para hombres blancos.[2] . Por esta razón, en el último tiempo desde algunos sectores tanto gubernamentales como académicos, desde el área de informática, y desde el activismo por los derechos humanos en el entorno digital, se está reclamando la necesidad de establecer reglas claras para el uso de Inteligencia Artificial respetuosa de los principios fundamentales de organización de las sociedades, un área conocida generalmente como ética de la Inteligencia Artificial.

Desde la Comisión Europea se diseñaron propuestas de reglamentos para el desarrollo y el uso de la inteligencia artificial[3], así como la OCDE (Organización para la Cooperación y el Desarrollo Económicos) presentó iniciativas de regulación de la IA[4]. Sin embargo, para garantizar derechos fundamentales como no discriminación o respeto a la privacidad, y en general principios de equidad en la recolección, representación, procesamiento y disponibilización de datos y en el desarrollo y despliegue de algoritmos de aprendizaje automático, es fundamental crear sistemas de IA que incluyan, traduzcan o implementen estos lineamientos en todo su ciclo de vida y no solo a la hora de evaluar impactos de discriminación del sistema una vez puesto en funcionamiento, cuando ya puede haber perjudicado a una gran cantidad de personas[5].

En lugar de pensar en la equidad como una iniciativa separada de la construcción del sistema o un requerimiento más que se debe cumplir, resulta más orgánico mantener una perspectiva centrada en los derechos de las personas a lo largo de todo el proceso de construcción y desarrollo de estas aplicaciones. Esto es especialmente importante cuando la IA se utiliza en procesos críticos o aspectos sociales sensibles, como es el caso de sistemas que preparan la redacción de fallos

---

[1] Emily M. Bender et al.: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 21), 2021. Disponible en https://dl.acm.org/doi/10.1145/3442188.3445922 (Consultado: 4 julio 2021)

[2] Safiya Noble. "Google Has a Striking History of Bias Against Black Girls", Time, March 26th 2018, https://time.com/5209144/google-search-engine-algorithm-bias-racism/ (Consultado: 4 Julio 2021)

[3] digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (Consultado: 4 Diciembre 2022)

[4] www.oecd.org/gov/innovative-government/hola-mundo-la-inteligencia-artificial-y-su-uso-en-el-sector-publico (Consultado: 4 Diciembre 2022)

[5] news.bloombergtax.com/tax-insights-and-commentary/we-can-all-learn-a-thing-or-two-from-the-dutch-ai-tax-scandal (Consultado: 4 Diciembre 2022)

judiciales[6], sugieren posibles diagnósticos para un paciente[7] o proponen acciones para evitar la deserción escolar[8] que afectan directamente la vida de muchos usuarios finales.

En este documento se proponen una serie de recomendaciones técnicas que surgen a partir del diálogo con personas que trabajan en aplicaciones prácticas de sistemas basados en datos, en diferentes roles y diferentes organizaciones del ecosistema tecnológico argentino. Este diálogo tuvo como objetivo comprender e identificar las necesidades, desafíos y posibles consecuencias que surgen a lo largo de la aplicación de los procesos de desarrollo de inteligencia artificial.

Metodología y perfil de los entrevistados

Condujimos entrevistas semiestructuradas, cada una con una duración aproximada de 40 minutos, mediante videoconferencias. El guión para las entrevistas fue adaptado del propuesto en el artículo "Data Cascades in High Stakes AI".[9] En este trabajo las autoras introducen y desarrollan el concepto de "Data Cascades": Eventos que combinados causan efectos negativos tales como daños a las comunidades, el desgaste de las relaciones con diferentes actores, el descarte de conjuntos de datos enteros y la duplicación de trabajo al reescribir partes del software. Dichos efectos se manifiestan en etapas posteriores en aplicaciones de Machine Learning. Estos eventos no tienen que ver con malas intenciones conscientes por parte de quienes programan, sino más bien con malas prácticas, malos incentivos y una serie de motivos que las autoras identifican mediante entrevistas a practicantes de distintas áreas (salud, economía, universidades, etc).

Se entrevistó a 6 desarrolladores de IA, quienes se desenvuelven en distintos roles y que cuentan con diferente experiencia en empresas con características distintas en cuanto a la cantidad de colaboradores, antigüedad, y áreas en los cuales aplican Inteligencia Artificial. Entre los entrevistados se encuentran un líder de datos, un científico de datos, un líder técnico, dos fundadores de start-ups y un investigador que desarrolla sus tareas en una empresa de software. Todas las empresas (excepto

---

[6] Estevez, Elsa; Fillotrani, Pablo; Linares Lejarraga, Sebastián (2020-06). «PROMETEA: Transformando la administración de justicia con herramientas de inteligencia artificial». Banco Interamericano de Desarrollo

[7] Layes, María Elisabeth Silva, Marcelo Alejandro Falappa and G. Simari. "Sistemas de soporte a las decisiones clínicas." 4to Congreso Argentino de Informática y Salud, CAIS 2013.

[8] Ignacio Urteaga, Laura Siri, Guillermo Garófalo (2020). "Predicción temprana de deserción mediante aprendizaje automático en cursos profesionales en línea". RIED Revista Iberoamericana de Educación a Distancia, 23 (2), pp. 147-167.

[9] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 39, 1–15.

una) son latinoamericanas, . Algunas son pequeñas empresas y otras grandes ya que no es lo mismo desarrollar IA en una empresa pequeña con presupuesto, empleados y clientes limitados, que en una empresa multinacional donde abundan estos recursos.

Entrevistamos a 2 trabajadores de Mercado Libre, empresa multinacional de origen Argentino líder en e-commerce, con más de 8 mil empleados y oficinas en diferentes países de Latinoamérica; 1 trabajador de Rappi, empresa multinacional de origen Colombiano, muy similar a Mercado Libre en cuanto a tamaño y modelo de negocio; 1 Data Scientist de Medalia, empresa multinacional de origen Estadounidense cuyo modelo de negocio se basa en extraer información de encuestas de satisfacción de clientes; y 2 fundadores de startups de Argentina, una de las cuales desarrolla chatbots para asignación de turnos en clínicas (sector salud), y la otra desarrolla modelos predictivos en el sector educativo.
Las entrevistas fueron grabadas y luego desgrabadas. Se generaron cuadros con índices temáticos de cada una para poder identificar nodos en común. A partir del visionado de las entrevistas se identificaron problemáticas comunes en los procesos de desarrollo de inteligencia artificial que afectan el desempeño del producto final y las técnicas y/o acciones que aplican los entrevistados en sus respectivos espacios de trabajo para abordar aquellos desafíos.

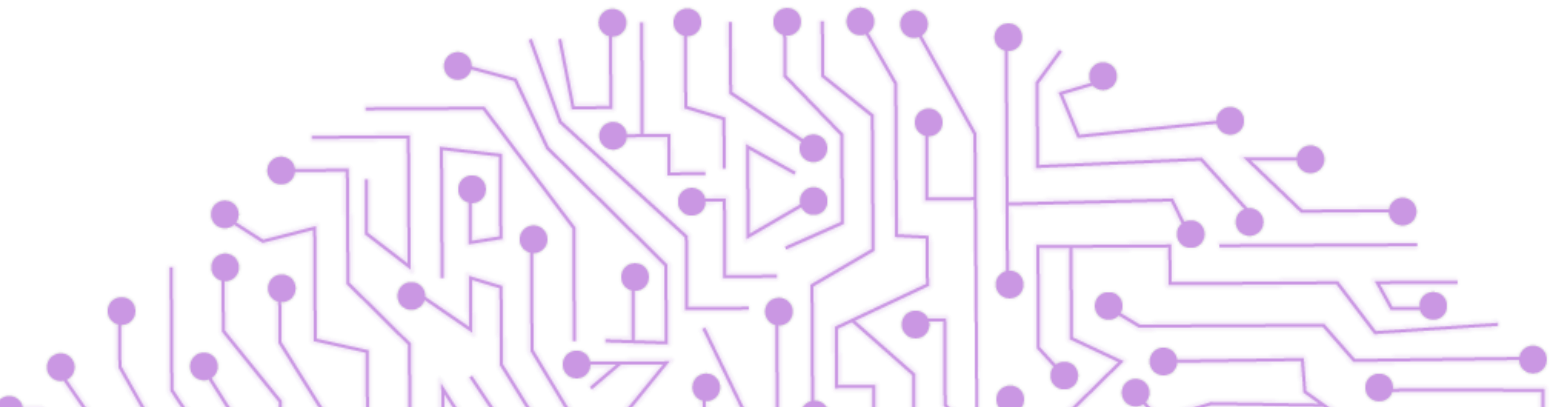
Estructura de este documento

El presente informe organiza esta información en función de un flujo de trabajo estándar dentro de los procesos de desarrollo e implementación basados en aprendizaje automático. En este proceso, desagregado en cinco partes, se agruparon posibles problemáticas y recomendaciones para cada una de ellas entendiendo que en cada momento del proceso pueden surgir diferentes situaciones. Este abordaje parte de la premisa de que cada una de las partes del proceso de desarrollo de IA tiene sus particularidades en cuanto a los actores involucrados y las herramientas técnicas en las cuales se trabaja: por ejemplo, en el proceso de recolección de los datos el foco está puesto en los datos y los propietarios de los mismos, ya sean bases de datos propias de la empresa, usuarios finales, o clientes externos.

Las etapas que distinguimos en el proceso de desarrollo de IA son las siguientes. Es importante resaltar que estas etapas no son secuenciales, pero cualquier paso puede indicar que es necesario volver a un paso anterior y revisar su resultado.

a) *Consideraciones previas:* momento reflexivo para preguntarnos si realmente es necesario construir el sistema, si es necesario hacerlo mediante Inteligencia Artificial, teniendo en cuenta a quién beneficia y a quién puede llegar a perjudicar.

b) *Definición del problema*: momento de diálogo entre las partes interesadas en desarrollar una aplicación de inteligencia artificial. Se define el problema que se quiere resolver en términos no técnicos y por qué sería necesario desarrollar dicha aplicación.
c) *Compilación y Curación de datos*: Esta es la etapa del proceso en donde se seleccionan las fuentes de datos que va a utilizar el modelo en su entrenamiento y se evalúan si son acordes para el problema.
d) *Entrenamiento de modelos*. Este es el proceso en donde se utilizan los datos seleccionados en la etapa anterior para entrenar o crear modelos que aborden el problema definido anteriormente.
e) *Evaluación y selección de modelos*. En esta etapa se definen métricas a partir de las cuales se medirá el rendimiento del modelo. Distintas métricas miden y priorizan características distintas del mismo y permiten decidir su mejora o puesta en funcionamiento.
f) *Puesta en producción*. En este momento el modelo comienza a interactuar con el mundo real, y es donde pueden surgir resultados inesperados



2. Desafíos y recomendaciones

A continuación, se agruparán desafíos y recomendaciones organizados en función del flujo del trabajo estándar dentro de la industria de la inteligencia artificial. Este flujo de trabajo consiste, de manera generalizada, en cinco etapas que definimos en la introducción. En estas etapas dialogan diferentes actores y surgen diversos problemas. Las etapas no son taxativas sino que se superponen cronológicamente entre sí, no es un proceso lineal sino más bien un ciclo en el cual los resultados del desarrollo de las etapas pueden requerir redefiniciones en etapas anteriores o afectar etapas subsiguientes. Por ejemplo, en la etapa de selección de modelos si modificamos los datos de entrenamiento es necesario procesar de nuevo los datos para adaptarlos a los diferentes modelos que sean tenidos en cuenta.

2.a. Consideraciones previas

Una primera consideración, previa a todo desarrollo, es preguntarnos si realmente es necesario construir un sistema de IA ya que como dicen Paz Peña y Joana Varón en "Inteligencia Artificial opresiva: categorías feministas para entender sus efectos políticos", en lugar de preguntarnos cómo desarrollar y desplegar un sistema de inteligencia artificial, ¿no deberíamos preguntarnos primero «por qué construirlo», «si es realmente necesario», «a petición de quién», «quién se beneficia», «quién pierde» con el despliegue de un determinado sistema de inteligencia artificial? ¿Debería incluso desarrollarse y desplegarse?"[10].

Una vez decidido que es necesaria la construcción del sistema de IA podemos avanzar en los siguientes puntos.

Por otro lado, algunos elementos transversales aplican a todo el proceso de producción. En principio, se recomienda la conformación de equipos diversos que incluyan a comunidades subrepresentadas en el desarrollo de tecnologías, desde una perspectiva interdisciplinaria y de derechos humanos, para evitar reproducir los enfoques hegemónicos que actualmente guían el desarrollo de sistemas de IA.

Desde un punto de partida hay que tener en cuenta que los datos son un recorte de la realidad. En muchas ocasiones se toman los datos como una fuente de verdad universal y absoluta, con lo cual resulta muy difícil ejercer una perspectiva crítica sobre el funcionamiento de estos sistemas. Para evitar esta dinámica, puede resultar beneficioso partir de la premisa, y tener presente en todo momento de que

[10] Paz Peña y Joana Varón, "Inteligencia Artificial opresiva: categorías feministas para entender sus efectos políticos", Not my A.I., octubre 10, 2021, https://notmy.ai/es/noticias-es/inteligencia-artificial-opresiva-categorias-feministas-para-entendersus-efectos-politicos/

los datos son un recorte de la realidad. Esta premisa debería contribuir a habilitar mecanismos que faciliten una perspectiva crítica, en lugar de naturalizar a los sistemas de IA como infalibles. Cada vez que se toma una problemática del mundo real y se la traduce en datos, hay que considerar que se toma solo una muestra del fenómeno que se quiere representar. Además de ser una muestra, la manera en la cual se modela el problema es también un recorte de la realidad. Por ejemplo, la elección de qué características incluir para describir a una persona como cliente de un banco para un modelo de segmentación de clientes, va a definir qué características de los clientes son relevantes para el modelo: género, edad, conducta crediticia, podrían ser variables que a priori tiene sentido incluir, aunque no se nos ocurriría por ejemplo incluir la altura de la persona en cambio sí estoy desarrollando un sistema para predecir el stock de ropa que voy a vender en verano si necesito esta consideración.

2.b Definición del problema

¿En qué consiste esta parte del proceso?

En este primer momento se delimitan los fenómenos que se quieren tratar de forma automática y el tipo de tratamiento que se les quiere dar. Se decide

1) **delimitación del problema**: qué problemas son interesantes para dedicarles el esfuerzo de desarrollo, implementación y seguimiento. El interés de un problema puede radicar en que permita hacer cosas imposibles hasta el momento (revisar una gran cantidad de documentos para ver cuáles tienen una determinada palabra), en la cantidad de trabajo que permite ahorrar, en la cantidad de gente a la que beneficia, en la cantidad de tiempo o dinero que permite ahorrar, en lo graves que son los casos que permite tratar... y la factibilidad de que sean tratados satisfactoriamente de forma automática.
2) **definición de los resultados**: las soluciones que se pretende obtener para cada tipo de problema a tratar.
3) **elegir y crear métricas de rendimiento**, que permitan comparar la bondad de diferentes aproximaciones y detectar problemas en el comportamiento del sistema.

El proceso de definición del problema debe incluir no solamente al equipo técnico que desarrollará el sistema, sino también expertos de dominio, que son quienes definen el problema en términos de una problemática del mundo real o una métrica empresarial a optimizar, como el número de ventas.

En cada uno de los pasos de definición, es crítico contar con miradas que puedan detectar potenciales impactos en diferentes grupos de población o contextos de

aplicación. Por esta razón es crítico que los equipos incorporen diversidades en etapa de diseño y no solamente en etapa de evaluación. Por ejemplo, una persona con movilidad reducida puede detectar que se requiere una categoría extra en los resultados de un sistema de reconocimiento de imágenes, que identifique si un objeto constituye un obstáculo.

Tomemos como ejemplo el problema de elegir qué publicación mostrar en una red social. En esta etapa el problema se puede definir como la optimización de diferentes medidas: por ejemplo, podemos querer maximizar la probabilidad de que un usuario interactúe con la publicación. Otra opción puede ser la que optimiza la probabilidad de que un usuario otorgue una valoración positiva a la publicación (carita feliz, like, etc.).
Estas diferentes formas de modelar el problema derivan en el desarrollo de productos muy distintos, que impactan de diversa manera en el comportamiento de los usuarios. Por ejemplo, si se optimiza la probabilidad de interacción se favorecen discusiones que, sin mecanismos de moderación o reglas de interacción muy claras, a menudo terminan en crispación, polarización y contenido violento.

Problemas posibles

La representación del problema a resolver, y de cuál es un resultado positivo (a optimizar) determina el funcionamiento del sistema. Si creo que las bananas son sólo amarillas, al definir el problema voy a introducir este sesgo y al crear un sistema que detecte bananas, el sistema va a detectar cosas ovaladas amarillas. En el mundo existen bananas de otros colores pero como no estaban en la concepción de banana inicial, las va a dejar afuera. (Es decir que la concepción inicial del problema no se corresponde con la realidad, muchas veces esto ocurre de manera inconsciente.) Aunque muchas veces esto ocurre de manera inconsciente, cada vez que la concepción inicial del problema deja afuera parte de la realidad puede haber impactos negativos en ciertas comunidades que no estén representadas por el sistema.

Recomendaciones

- **Incluir conocimiento existente en el área del problema** que se quiere abordar durante la ideación del enfoque, especialmente al determinar qué constituye un buen resultado o un mal resultado. Es una forma de integrar en el diseño del sistema conocimiento sobre el problema que podría no estar representado por los datos, pero sí ser conocido por expertos de dominio.
- **Hacer explícitas todas las hipótesis.** Ej: Asumimos que nuestros datos son representativos.
- **Incluir diversidad, usuarios finales y comunidad de ciencias sociales** especialmente en la definición del problema, para entender mejor el impacto de diferentes opciones en la representación del problema y evitar la llamada “discriminación por diseño”.
- **Favorecer la intervención humana**, y concretamente, de sistemas analíticos, de reporting o de BI (Business Intelligence). Podemos encontrar distintas situaciones en las cuales la intervención humana es deseable:
    - **Recomendación para decisiones:** En algunos casos es posible utilizar los modelos de IA para mostrar información a las personas que ayudan a tomar decisiones.
    - **Emergencias:** En el caso del papel higiénico, el sistema no detectaba a tiempo el cambio de comportamiento dado por la pandemia, en este caso se debería poder modificar parámetros por un especialista. Un caso similar es el del autoPilot donde se puede pedir al vehículo que delegue el manejo al piloto.
    - **Necesidad de interacción humana:** En los casos donde el agente automático no funciona como el usuario desearía, es deseable que exista la opción de hacer algo a mano, por ejemplo, si quiero sacar un turno médico con un programa de chat con IA y la IA no me entiende repetidas veces, que exista la opción de hablar con un humano (Human handoff).

2.c Compilación y curación de datos

¿En qué consiste esta parte del proceso?

Esta parte del proceso consiste en primer lugar en recopilar y seleccionar las fuentes de datos con los que va a ser entrenado el algoritmo de aprendizaje automático. Esto puede incluir el diseño del dispositivo de recolección de datos, o determinar cuáles de los conjuntos de datos ya existentes resultan pertinentes para el enfoque planteado en la fase de diseño. A la hora de definir cuáles van a ser estas fuentes de datos, se tienen en cuenta distintos factores, principalmente disponibilidad y costo de recolección, aunque también utilidad para abordar el problema.

En esta sección incluimos también el análisis preliminar de datos o análisis exploratorio de datos. Mediante este análisis se pueden detectar propiedades interesantes de los datos que se pueden explotar en el proceso de aprendizaje automático, como patrones desconocidos previamente. También se pueden detectar errores en el proceso de recolección, en el diseño de la muestra, etc. Finalmente, el análisis exploratorio es una buen momento para la detección temprana de sesgos, ya que evita desarrollos posteriores basados en datos sesgados.. Consideramos la etapa de análisis preliminar como una buena práctica en sí misma, porque enfoca el trabajo de fases posteriores y evita problemas mayores.

Problemas posibles:

- Errores en el proceso de adquisición
- Sesgos que discriminen a grupos históricamente marginalizados
- Asumir que tener una gran cantidad de datos equivale a tener datos de calidad, representativos del dominio del problema, sin preguntarse si tienen en cuenta diferentes perspectivas, otras fuentes, si cubren todos los casos, o los casos de interés?
- Falta de información sobre la captura y etiquetado. Las etiquetas que se asignan a los datos definen el problema, ya que el modelo va a inferir patrones a partir de estas etiquetas.
- Utilizar todos los datos disponibles en el modelo, sin refinar variables correlacionadas o sensibles, variables que contienen errores, variables irrelevantes para la representación del problema, o comparando variables numéricas que están en distinta escala, sin interpretar màs en profundidad su significado, por ejemplo comparar años con tamaño de pupilas.

Recomendaciones

- **Análisis preliminar de los datos.** Consiste en revisar las características de los datos que se disponen, para esto pueden realizarse 3 actividades:
    - Descripción de los datos: qué valores pueden tomar los datos, de qué tipo son, qué representan de mi problema, etc.
    - Inspección anecdótica: tomar una muestra de los datos que se tienen y analizar caso por caso.
    - Estadísticas descriptivas: Mostrar características agregadas : valores más frecuentes y más escasos, valores mínimos y máximos, correlaciones entre los valores, análisis de varianza, outliers, etc.
- **Verificar que datos representen adecuadamente a las poblaciones.** Muchas veces ocurre que las muestras están desbalanceadas, y representan de forma inadecuada a las poblaciones. La configuración más común es que algunos valores o conglomerados de valores sean mucho más frecuentes que otros. Esto sucede de forma natural porque muchos datos tienen distribución de Pareto, pero el fenómeno se agudiza porque se aplican de forma poco cuidada los mecanismos de recolección de datos. Además, muchos algoritmos de aprendizaje automático son susceptibles de exagerar una distribución desbalanceada (con clase mayoritaria, valores mayoritarios para características). En este momento es importante verificar que ningún grupo de individuos, de características, de fenómenos, quede subrepresentado o simplemente representado de forma que los sistemas automáticos basados en esos datos tengan efectos discriminatorios sobre alguna población. También es muy importante documentar la distribución de clases, y tener en cuenta las medidas necesarias para que en fases posteriores no se produzcan efectos discriminatorios debidos a las falencias de algunos algoritmos en un contexto de desbalanceo de clases.
- **Evaluar si un conjunto de datos es representativo del universo a tratar.** Por ejemplo, una muestra de estudiantes universitarios podría contener muchos más datos sobre jóvenes entre 20-30 años que de adultos mayores, si usamos estos datos para predecir edad de graduación, los resultados del modelo podrían no tener sentido para adultos mayores ya que no se cuentan con datos sobre esa población.
- **Si se usan datos sintéticos, compararlos con datos del mundo real o validarlos con un experto de dominio.** En el caso de que no existan datos, o estos no sean suficientes para el entrenamiento del modelo, a menudo se recurre a la generación de datos artificiales. Por ejemplo, un software utilizado para predecir patologías en estudios médicos, podría estar entrenado con imágenes de alta calidad, generadas específicamente para el entrenamiento del modelo, mientras que las utilizadas por los médicos en los consultorios podrían ser de menor calidad, o estar desenfocadas, capturadas por cámaras de celular. Si bien muchas veces no es posible

conseguir datos no sintéticos antes de la puesta en producción del modelo, recomendamos tener en cuenta las diferencias que podrían existir.

- **Documentar fuentes de datos y procesos de selección y recopilación.** Una tentación ante la carencia de datos es utilizar datos recopilados para otro problema. La falta de contexto sobre la producción de datos realizada lleva en ciertos casos a malinterpretarlos o usarlos mal. Frameworks como Datasheets[11] tienen como objetivo guiar a quienes construyen fuentes de datos para analizar las suposiciones involucradas, riesgos e implicancias de su uso. Esta metodología también permite transparentar quién captura, etiqueta y define las reglas para procesar los datos.
- **Documentar los procesos de curación.** Es muy común realizar transformaciones sobre los datos para eliminar errores en el proceso de adquisición de datos, representar mejor algún fenómeno de interés, o simplemente evitar limitaciones conocidas y maximizar los resultados de los sistemas de aprendizaje automático. Los procesos de curación suelen tener una componente muy importante de conocimiento basado en la experiencia, aplicando "sentido común", ad hoc para cada caso particular, que solamente se puede conocer si se explicita. Por esta razón resulta especialmente importante documentarlo, para que cualquier equipo que use los datos más adelante pueda detectar posibles problemas.
- **Crear una buena interfaz de usuario para la recolección de la información** (bien explicada y accesible). En algunas ocasiones, los datos ingresados por los usuarios o por las personas contratadas para la recolección de los mismos, contienen información contradictoria que podría evitarse desde el diseño de la interfaz, como por ejemplo: aclarar las unidades de medida que pueden no estar unificadas, o si se pregunta por la provincia de origen, no dejar que el usuario la escriba sino que elija de una lista.
- **Definir la menor cantidad de características o información necesaria para el problema acerca del usuario**. A la hora de recopilar o compilar los datos que utilizará el modelo para representar el problema, muchas veces se intenta recopilar la mayor cantidad de datos posibles, en una estrategia avariciosa por la posible utilidad de estos datos en potenciales aplicaciones futuras. Sin embargo, las políticas de recolección avariciosas suponen una potencial amenaza a los usuarios, por la vulnerabilidad de que estén de alguna forma disponibles datos que pueden llegar a ser sensibles de alguna forma. Aunque se establezcan mecanismos de seguridad, ningún mecanismo es infalible. Por otro lado, también suponen un coste ecológico evitable, por el costo de adquisición y almacenado de datos, y también suponen es recomendable evitar incluir datos que no aporten información estrictamente necesaria para la definición del problema. Por ejemplo, para construir un modelo que prediga atrasos en contribuyentes a un sistema

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723

tributario, no sería necesario para el modelo información sobre el género, o el número de documento del contribuyente.

- **Asegurar la conformidad del autor para el uso de datos con propiedad intelectual.** Se debe verificar que la licencia de los datos permita el uso que se le va a dar o conseguir una excepción en caso contrario.
- **Asegurar la conformidad del usuario para el uso de datos personales.** Si la aplicación que se va a desarrollar utiliza datos personales, es importante estar informado y actualizado sobre la legislación en términos de protección de datos personales. Por ejemplo, la ley Argentina de protección de datos personales indica requisitos a cumplir al tomar y almacenar este tipo de datos, como el consentimiento explícito del usuario.
- **Desplegar mecanismos de seguridad para proteger datos sensibles.** **"***Se consideran datos sensibles aquellos datos personales que revelan origen racial, étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical e información referente a la salud o vida sexual."* (Ley de protección de datos personales 25.326). En el caso de proyectos que incluyen este tipo de datos, lo recomendable es almacenar la información en bases de datos que implementen todos los mecanismos de seguridad posibles. Pero incluso de esa forma, cuando esta información sensible queda guardada en bases de datos, existe la posibilidad de que los mecanismos de seguridad sean vulnerados y quede expuesta la privacidad de los titulares de los datos. Se recomienda como medida de seguridad adicional utilizar herramientas de anonimización de datos personales y que luego dicha anonimización sea revisada manualmente por expertos del dominio. También es importante tener en cuenta que algunos datos que no son considerados sensibles en el contexto actual, pueden resultar sensibles en contextos futuros o en usos secundarios.

2.d Entrenamiento de modelos

¿En qué consiste esta parte del proceso?

Esta parte del proceso, conducida en gran parte por el equipo técnico, incluye la inferencia, evaluación y selección del modelo o modelos de aprendizaje automático que se utilizará para las predicciones. Como parte de este proceso se pueden realizar transformaciones sobre los datos para sacar el mayor partido de las características de cada método de aprendizaje automático, o de . Es importante dividir los datos de entrenamiento en conjuntos de entrenamiento y validación.

Problemas posibles:

- Inferir modelos sobre datos que no han pasado por un proceso de curación, y pueden traer problemas de representación inadecuada del problema y sesgos de diferente tipo.
- No incorporar las conclusiones del análisis de resultados a la representación de los datos, es decir, al proceso de curación de datos, para mejorar el modelado.
- No tener en cuenta los efectos de una clase mayoritaria en métricas agregadas como el acierto promedio (*accuracy*).
- Reproducir o amplificar sesgos presentes en los datos, especialmente frecuente en modelos de tipo lineal, que tienden a favorecer a la clase mayoritaria.
- Utilizar modelos que asuman distribuciones sobre los datos y no verificar que se cumplan estas restricciones, por ejemplo, usar un modelo que asume una distribución gaussiana de los datos sobre datos con distribución geométrica.
- Aplicar modelos inferidos en una población sobre una población distinta, por ejemplo, usar un modelo inferido sobre población de piel clara a una población de piel oscura.

Recomendaciones:

- **Utilizar técnicas que reduzcan la tendencia a favorecer la clase mayoritaria** de los métodos lineales, como regresiones o máquinas de vectores de soporte, como por ejemplo técnicas de balanceo de clases:
  - **sobremuestreo** de clases minoritarias
  - **submuestreo** de la clase mayoritaria.
- Al aplicar técnicas de suavizado (*smoothing*) para evitar sobreajuste (*overfitting*), verificar que no se estén sobrerepresentando los valores mayoritarios o la clase mayoritaria en general.
- En el caso de métodos más expresivos, como árboles de decisión o redes neuronales, se pueden **aplicar métodos para mejorar la representación de los casos minoritarios:**
  - Usar algoritmos de **Boosting**: estos modelos por definición suelen centrarse en mejorar los errores que cometen. Por ejemplo, en un XGBoost, aumentando el número de árboles, podemos ir corrigiendo los errores de los árboles anteriores.
  - Usar algoritmos de **Stacking** y algoritmos de **aprendizaje por refuerzo:** del mismo modo que los Boosting, estos algoritmos permiten ir mejorando los aciertos de la clase minoritaria.
- **Entrenar modelos con datos lo más parecidos posibles al lugar donde se va a utilizar.** Es común entrenar modelos con un conjunto de datos y aplicarlos sobre contextos diferentes a los de la recolección de los datos. Esta práctica puede ocasionar errores muy grandes en las predicciones. Por

ejemplo, supongamos, estamos desarrollando un sistema que prenda un aire acondicionado a una temperatura y usamos los datos de temperatura ambiente a la cual los usuarios del norte de México prenden el aire. Al entrenar ese modelo y aplicarlo, por ejemplo, en la Patagonia Argentina, las temperaturas para los usuarios de la zona de Argentina pueden ser muy diferentes.

- **Correlación no implica causalidad**. La correlación no implica causalidad es un concepto muy importante en estadística y en la investigación en general. Básicamente, se refiere a la idea de que solo porque dos cosas están relacionadas, no necesariamente significa que una causa la otra. Por ejemplo, si observamos que hay una correlación entre el hecho de comer manzanas y tener una piel saludable, no podemos concluir que comer manzanas causa una piel saludable. Podría ser que ambas cosas estén relacionadas por algún otro factor, como por ejemplo el hecho de llevar un estilo de vida saludable en general. Por lo tanto, es importante tener en cuenta este concepto cuando se realizan estudios y se interpretan los resultados.

2.e Evaluación y selección de modelos

¿En qué consiste esta parte del proceso?

En esta etapa se analizan los resultados de diferentes modelos entrenados para seleccionar las opciones que mejor se adecuan a las necesidades del problema. La principal pregunta que surge en esta etapa es de qué manera se va a evaluar el modelo, y esto depende del uso que se le vaya a dar. Existen distintas métricas para evaluar el rendimiento de los modelos, y cada una de estas métricas valora cualidades distintas.

Problemas posibles:

- Un problema que puede aparecer en esta etapa es encontrar modelos que tengan características generales muy buenas, pero no se comportan bien para mi problema particular. Por ejemplo, un clasificador diseñado para predecir el medio de locomoción utilizado por estudiantes para asistir a un colegio (a pie, transporte público, bicicleta) tenía muy buen desempeño en precisión y exhaustividad (recall) globalmente, pero cuando se analizó más en profundidad, se descubrió que la mayoría de los errores se encontraban en la clase bicicleta: el modelo predecía que los estudiantes que efectivamente usaban bicicletas, no iban a usarla. En otras palabras, predecía que menos estudiantes iban a usar bicicletas. Este sistema iba a

ser utilizado para estimar cuántas bicicletas sería necesario comprar para los estudiantes.

- Otra consideración puede ser la explicabilidad de los modelos, es decir muchas veces “da buenos resultados” pero no sabemos por qué y en casos particulares no sabemos por qué tomó ciertas decisiones. Por ejemplo, en el ámbito médico, la toma de decisiones se trata de un punto crítico, pues una decisión puede influir directamente en la vida y salud de las personas. Por ello, si se utilizan estos métodos de IA como ayuda en la toma de decisiones, es necesario saber algo más sobre cómo afecta cada variable a la predicción emitida por el modelo.

Recomendaciones:

- **Favorecer el uso de modelos cuyo comportamiento sea fácil de explicar.** Algunos modelos de Inteligencia Artificial son interpretables *per se*. Modelos sencillos como regresiones, que de por sí nos ofrecen la importancia de cada variable en las decisiones que se toman, o árboles de decisión, que por su propia estructura nos indican el camino de decisiones sobre las distintas variables que conducen a la predicción o decisión final.
- **Utilizar métodos de explicabilidad para entender el comportamiento del modelo.** Los métodos de explicabilidad ayudan a entender el comportamiento de modelos “caja negra”. Es importante utilizar estas técnicas para detectar debilidades y posibles comportamientos discriminatorios del modelo, pero no como una garantía de su correcto funcionamiento. Algunas técnicas existentes son, por ejemplo, los índices SHAP (existe una herramienta del mismo nombre que lo aplica) y los mapas de saliencia.
- **Análisis de error, en la etapa de desarrollo del modelo.** Por análisis de error no sólo nos referimos a evaluar los modelos sino describir cómo los errores afectan a las diferentes clases o en general cómo es que se da ese error y si afecta más a ciertos grupos que a otros. Es muy importante asegurarse de que el error esté distribuido aleatoriamente entre las distintas clases. Después de todo, achicar estas brechas de performance es mejorar el producto. Ejemplo: ¿El modelo se equivoca más cuando tiene que identificar a una doctora que a un doctor?
- **Utilizar métricas resistentes al desbalance de clases.** Si bien el desbalance de clases se vio como un problema en el entrenamiento de modelos, en la etapa de evaluación de modelos podemos centrarnos en las métricas resistentes a dicho problema.
    - f1, la sensitividad o la precisión.
    - Roc auc. threshold independiente.
    - Average precision score. treshold independiente.
- **Data Ablations para medir impacto previo.** Es una técnica que consiste en modificar los datos para probar qué pasaría con modificaciones en las

predicciones. No es necesario esperar a que algo pase para saber cómo se va a comportar el modelo en esos casos.

2.f Puesta en producción

¿En qué consiste esta parte del proceso?

Esta etapa consiste en el despliegue del sistema desarrollado para su uso final. En etapas anteriores, nuestro modelo interactúa con datos previamente curados y seleccionados para optimizar su rendimiento. En esta etapa es crítica la capacidad de generalización del modelo, ya que va a interactuar con datos previamente no vistos, y que pueden tener diferencias con los datos de entrenamiento, a veces diferencias inesperadas.

Problemas posibles

- El principal problema que podemos tener en esta etapa es que los datos nuevos del mundo real sean muy distintos de los usados para entrenar y que por lo tanto el modelo no sepa cómo comportarse con los mismos, o se comporte de formas inesperadas, produciendo posibles problemas éticos. El caso más paradigmático de este problema es el conocido como “Data Drift”, o la paradoja de predecir el pasado: Los modelos predicen tendencias y si hay un cambio brusco en "la realidad" esos datos demoran en ser incorporados y afianzados para el reentrenamiento de modelos. Un ejemplo es el “desabastecimiento” de papel higiénico en los supermercados españoles al principio de la pandemia. Evidentemente, los modelos predictivos utilizados para las reposiciones de stock, tardaron demasiado tiempo en detectar ese cambio de comportamiento, hasta el punto de que llegaron tarde.
- Un segundo problema que emerge es el de los usos secundarios de una aplicación, que pueden llevar a problemas que no estaban previstos en el diseño de la aplicación. Un caso ilustrativo es el uso de sistemas que socializan el comportamiento de usuarios, por ejemplo, que comparten la información de dónde y cuándo practican un cierto deporte, que tiene como uso primario la formación de grupos de ejercitación en deporte, pero como uso secundario se ha usado para acosar a personas.

Recomendaciones

- **Evitar los feedback loops** o circuitos de retroalimentación, es decir, evitar que el modelo solo utilice sus propias predicciones como fuente para re-entrenarse, pues esto hace que cada vez tenga mayor divergencia de las predicciones de los datos reales. Para ello es importante que el sistema pueda actualizarse re-entrenando su modelo con feedback del funcionamiento del sistema desplegado.
- **Monitoreo de métricas del modelo en producción.** Las métricas pre-productivas reflejan la calidad del modelo sobre los datos de entrenamiento, que pueden ser muy distintos a los datos reales. Es recomendable llevar un registro de las predicciones y contrastarlas con la realidad para saber cuando es necesario reentrenar los modelos. Es una decisión de esta etapa establecer las métricas que indican la necesidad de reentrenamiento.
- **Poner un canal de reporte para fallas en el modelo (de sesgo o discriminación).** El modelo va a tener fallas, y estas pueden impactar en los usuarios finales, sometiéndolos por ejemplo a una auditoría. Es necesario que los usuarios tengan herramientas para reportar estos casos, y poder rectificar las predicciones que resulten perjudiciales. Por ejemplo, las redes sociales suelen incluir un método de reporting , para decir si algo es spam, no me interesa, tiene contenido de odio, etc. Otro caso es el de traducciones de Google que te permite decir que una traducción no te parece buena y sugerir otra. Un caso muy representativo es el de hacer disclaimer ante el caso de una baja de un video de Youtube por falsa predicción de infracción de copyright. Si el algoritmo me da de baja el video porque predice que estoy violando las leyes de copyright puedo presentar una nota mediante la interfaz de Youtube.
- **No hardcodear enfoques a problemas del modelo en la interfaz, sino integrar como feedback para un reentrenamiento.** En algunos casos cuando una empresa recibe una queja por un mal funcionamiento, en lugar de incorporarlo como entrenamiento o hacer un análisis integral de la falla, se parcha el caso específico. Como desarrollador y desde un punto de vista pragmático, esto sería como resolver un bug para el caso particular, o forzar un resultado para un caso particular.
- **No poner en producción modelos entrenados para otros contextos sin verificar su desempeño.** En muchos casos se cree que los modelos sirven para distintos contextos, por ejemplo usar un modelo de recomendación de películas entrenado con datos chilenos para recomendaciones en argentina puede ser muy distinto, en todo caso antes de hacerlo, hay que hacer un análisis de desempeño.

Palabras finales

Cuando los desarrolladores nos dejamos llevar por “el sentido común", es decir, lo que nos parece más probable, lo estándar, no chequear condiciones, no pensar en las diversidades y representaciones, lo que hacemos es reproducir prácticas y formas de pensar hegemónicas. Necesitamos estar muy atentos a poner en tela de juicio las creencias que tenemos más naturalizadas para generar una IA anti-hegemónica, para no producir una IA eurocentrista, blanca, patriarcal, capacitista, etc. En este sentido, lo que podemos ver a lo largo del documento, si lo podemos resumir en una frase, es “no dar nada por sentado”.

A lo largo del documento estuvimos revisando los procesos principales en el proceso de desarrollo de un producto basado en datos, identificando problemáticas que podrían surgir y recomendaciones para la mitigación de estas problemáticas. Bajo una perspectiva de calidad de producto y Derechos Humanos uno no puede asegurar en este tipo de sistemas la ausencia de “fallas” o la ausencia de comportamientos indeseados. Estas prácticas apuntan a mejorar la calidad que incide fuertemente en una Ética de la Inteligencia Artificial.

Creemos fuertemente que la mirada técnica y social no están disociadas. En este sentido, como técnicos, hablar de una IA que discrimina es hablar de una IA de baja calidad, una IA que funciona mal. De la misma manera, evitar comportamientos discriminadores, evitar cualquier tipo de sesgo indeseado, es producir un software de mayor calidad. Buena parte de los sesgos que encontramos en sistemas basados en aprendizaje automático son causados por restricciones técnicas, por lo tanto, resolver este problema puede considerarse un verdadero desafío técnico. Como dicen Friedman y Nissenbaum [12](1996) “la ausencia de sesgos debe formar parte del selecto conjunto de criterios según los cuales debe juzgarse la calidad de los sistemas vigentes en la sociedad” (*ibid*.: p. 345f).

[12] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

eticaenia@vialibre.org.ar
www.vialibre.org.ar

Con el apoyo de: