SUMMARY OF THE PROJECT IN DESIGN * (*)

ThemisAI - Auditoría Algorítmica

PITCH ELIGIBILITY DATE	COUNTRY(IES)
02/08/2023	Colombia

ALIGNED WITH COUNTRY STRATEGY?
Yes

PARTNER(S)
Quantil S.A.S.

PRELIMINARY CLASSIFICATION ENVIRONMENTAL AND SOCIAL IMPACT
C (**)

TOTAL BUDGET	IDB Lab	LOCAL COUNTERPART AND COFINANCING
US 194,000	US 135,000	US 59,000

DESCRIPTION

El problema Actualmente, empresas y gobiernos de todo el mundo utilizan sistemas o modelos de inteligencia artificial (IA) o aprendizaje automatizado (AA) en diversas áreas para la toma de decisiones, pero a pesar del incremento en el uso de este conjunto de tecnologías, aún hace falta un mayor esfuerzo para desarrollar herramientas a nivel técnico que permitan mitigar riesgos comúnmente asociados con el despliegue de la IA, tales como los sesgos, la discriminación, la exclusión, la vigilancia e invasión a la privacidad, la falta de transparencia y la falta de rendición de cuentas. Estos riesgos, si no se mitigan, pueden tener importantes repercusiones directas para las comunidades en situación de pobreza y/o históricamente vulneradas, que suelen ser tanto usuarios, como clientes e incluso trabajadores de estos sistemas, plataformas y modelos.

En el sector privado, el aumento en el uso de la inteligencia artificial, a nivel global, en los últimos años y en el lustro por venir indica un crecimiento importante y sostenido. De acuerdo con el IDC Worldwide Semiannual Artificial Intelligence Tracker, se estima que para el periodo comprendido entre 2022 y 2026 el mercado global de esta tecnología lograría un crecimiento anual compuesto de 18.6% alcanzando los $900 mil millones de dólares hacia 2026. En materia de adopción, una investigación realizada por el Consejo Nacional de Ciencia y Tecnología de los Estados Unidos (NSTC por sus siglas en inglés) concluyó que el 72% de las grandes compañías, es decir, aquellas con más de 500 empleados, utilizan la IA o el AA en alguna medida, y una encuesta realizada por el MIT Technology Review reveló que el 85% de las empresas con más de 1000 empleados están utilizando alguna tecnología tanto de IA como de AA. La adopción de la IA en el sector privado es también una realidad a nivel regional: un estudio de NTT Data y el MIT Technology Review indicó que en América Latina y el Caribe cerca del 40% de las grandes empresas utilizan este tipo de tecnología para mejorar la experiencia del cliente, implementar la analítica predictiva, optimizar cadenas de suministro, detectar fraudes y desplegar diversas modalidades de personalización de productos y servicios para clientes y usuarios. En la práctica, desde la experiencia de BID Lab, a través de la iniciativa fAIr LAC, todas estas aplicaciones se utilizan en ámbitos particularmente sensibles, como la preselección de currículums para una vacante, la predicción del crimen o para la asignación de créditos financieros. En el sector público, si bien, los avances no han sido tan significativos y menos aún en América Latina y el Caribe (ALC), el incremento en el uso de Sistemas de Decisión Automatizada (SDAs) tiene implicaciones cada vez mayores para la provisión de servicios públicos a los ciudadanos, tales como mejoras en la formulación de políticas públicas, en el diseño y entrega de servicios a los ciudadanos, y en la gestión interna de las instituciones del Estado. Un

*The information mentioned in this document is indicative and may be altered throughout the project cycle prior to approval. This document does not guarantee approval of the project.

**The IDB categorizes all projects into one of six E/S impact categories. Category A projects are those with the most significant and mostly permanent E/S impacts, category B those that cause mostly local and short-term impacts, and category C those with minimal or no negative impacts. A fourth category, FI-1 (high risk) Financial Intermediary (FI)'s portfolio includes exposure to business activities with potential significant adverse environmental or social risks or impacts that are diverse, mostly irreversible or unprecedented, FI-2 (medium risk) FI's portfolio consists of business activities that have potential limited adverse environmental or social risks or impacts, FI-3 (low risk) FI's portfolio consists of financial exposure to business activities that predominantly have minimal or no adverse environmental and social impacts.

estudio reciente de OCDE y CAF indica que la importancia en la adopción de la IA por parte del sector público es reconocida por la mayoría de las estrategias nacionales de IA desarrolladas en la región (Argentina, Brasil, Chile, Colombia, México, Perú y Uruguay). No obstante, solo seis de estos países (excluyendo a Uruguay), más Costa Rica, se han adherido a los principios de IA de la OCDE, los cuales promueven una serie de principios para la adopción responsable de la IA en el ámbito público. Ante el auge de la IA, tanto a nivel público como privado, diversos estudios, como el “State of AI Report 2022”, advierten que el uso de sistemas de IA puede conllevar serios riesgos para los seres humanos si no existen mecanismos de supervisión y control humano. Una de las razones es que, para entrenar dichos modelos, se hace uso de datos históricos que pueden contener, de forma implícita o explícita, sesgos hacia cualquier característica de la unidad de análisis (por ejemplo, el sexo o la raza de una persona). En efecto, la tecnología asociada a los modelos no está diseñada para comprender la dimensión de los problemas que puede generar, que son de índole contextual. De este modo, las decisiones que podrían tomar las organizaciones que usan la IA en base a los resultados (buenos o malos) del modelo, tienen el riesgo de perjudicar a diversos grupos de personas, en particular a las minorías o a los grupos que han sido históricamente discriminados, incluyendo aquellos en condición de pobreza y desigualdad. Por lo anterior, el uso de estas tecnologías requiere herramientas de control que permitan determinar cuál es el mejor modelo según determinados parámetros de “justicia algorítmica”, concepto que hace referencia al grado de robustez y fiabilidad de la tecnología en cuestión, y que permita a su vez establecer medidas de corrección o mitigación dentro del algoritmo. Estas herramientas son las llamadas auditorías algorítmicas, solución propuesta en este proyecto.

La solución Las auditorías algorítmicas son procesos que permiten examinar el desempeño y diseño de un algoritmo en función de su justicia, transparencia y rendición de cuentas. Es una forma de asegurar que los algoritmos funcionan como se pretende y que no conllevan sesgos severos en contra de determinados grupos de personas. Las auditorías algorítmicas se pueden llevar a cabo por organizaciones o individuos dependiendo de la metodología que implementen, y pueden incluir la revisión de los datos y el código del algoritmo, analizar los resultados del modelo y hacer estudios de usuario para entender el impacto de los algoritmos en determinados grupos.

QuantilAI es un aplicativo desarrollado por Quantil S.A.S, en Colombia, que permite que cualquier usuario que desarrolle modelos de inteligencia artificial de aprendizaje supervisado pueda evaluar su desempeño, en función de determinados parámetros de justicia algorítmica.

La solución propuesta permite generar escenarios comparativos donde, según el contexto y las métricas de justicia que el usuario defina, se puedan observar las mejoras en las predicciones tales que eviten la toma de decisiones en escenarios discriminatorios. Todo esto adaptado a una interfaz de fácil manipulación para que cualquier usuario con conocimientos de conceptos elementales de aprendizaje supervisado pueda tomar decisiones informadas.

Con esta solución, es posible estandarizar y automatizar el proceso de auditoría algorítmica, o algunas partes de este. Además, se puede escalar a cualquier industria que use modelos de IA para tomar decisiones que afecten el bienestar de grupos poblacionales. Teniendo en cuenta el nivel de madurez tecnológica del sector financiero, se priorizará está industria.

Los beneficiarios, Aunque la población objetivo para el uso de la herramienta son entidades que utilizan aprendizaje automatizado para la toma de decisiones, la población que se beneficiará de la misma puede ser muy variada. En términos generales, se podría beneficiar a las poblaciones minoritarias, históricamente discriminadas al obtener una medida de igualdad en los modelos predictivos. Esto puede impactar, por ejemplo, desde la primera aplicación a un trabajo (primer filtro), hasta la asignación de un crédito hipotecario.

En el marco de este proyecto los beneficiarios dependerán de las empresas con las que se haga la validación de la herramienta. Algunos ejemplos de aplicación son:

- Caso de uso bancario: usar herramienta para descubrir sesgos en los modelos de scoring (construidos a partir de datos históricos) en la financiación de mujeres.
- Caso de uso transferencias condicionadas: usar herramienta para descubrir sesgos en la transferencia de recursos a las EPS a partir de fórmulas de redistribución de recursos y ajuste de riesgo, que potencialmente pueden discriminar a las poblaciones con menor acceso a los servicios de salud.
- Caso de uso de la fiscalía: usar herramienta para descubrir sesgos en la población afro / LGBT+ de cara al uso de casa por cárcel y otras penas

El socio El proyecto sería ejecutado por Quantil S.A.S, una empresa colombiana con más de 14 años de experiencia enfocada en la implementación de metodologías matemáticas, estadísticas y de inteligencia artificial para responder a las necesidades de instituciones del sector real, financiero, y gubernamental.

La contribución del BID Lab será un financiamiento de Recuperación Contingente de US $135,000.