Nombre del evento: "Diálogos y experiencias sobre IA en la evaluación: El presente y la perspectiva futura"

Fecha: 1 de junio de 2026

Organizadores: Econometría y Natura Software.

1. Apertura del Evento y Perfil de los Participantes

El espacio fue moderado por **Carolina Suárez** (Gerente Técnica de Econometría Consultores). El panel contó con la participación de dos expertos en la materia:

- **Adriana Cárdenas (Directora de Proyectos en Econometría):** Ingeniera de Sistemas y Magíster en Análisis de Big Data, con más de 20 años de trayectoria en el sector social y evaluación de políticas públicas.

- **Jorge Posada (Director Natura Software):** Ingeniero Electrónico y Magíster en Ingeniería Industrial, especialista en Inteligencia Artificial (IA), machine learning y procesamiento de lenguaje natural (NLP).

Al iniciar el evento, se consultó al público mediante la plataforma *Mentimeter* sobre las expectativas que les generaba la IA en la evaluación, destacándose los conceptos de **optimización e innovación**, acompañados por nociones de **incertidumbre, angustia y oportunidad**.

2. Presentación de Experiencias y Casos de Éxito

Los panelistas expusieron la trayectoria conjunta que han consolidado desde 2019, acumulando más de 20 proyectos en los que se aplican tecnologías como *web scraping*, análisis de redes sociales, *topic modeling*, modelos predictivos y el desarrollo de **DOKU**, una herramienta propia orientada a la explotación y análisis de repositorios documentales.

Caso 1: Econometría (Procesamiento masivo en tiempo récord)

- **El Reto:** Evaluar un repositorio documental de **7.500 documentos** para verificar el cumplimiento de **230 requisitos** específicos de calidad. El cronograma asignado era de un mes inicial, pero operativamente se redujo a **dos semanas**, con una fecha de entrega inamovible debido a un evento de reconocimiento público dependiente de los puntajes.

- **La Solución:** Implementación de IA generativa para el análisis y la puntuación de evidencias, supeditada enteramente a la supervisión de un panel de expertos.

- **Resultado:** Cumplimiento exitoso del objetivo con un alto estándar de calidad, equidad en

las calificaciones y plena satisfacción del cliente y de las entidades evaluadas.

Caso 2: Natura Software (Análisis de Sentencias del sector justicia)

- **El Reto:** Medir el goce efectivo del derecho a la verdad judicial analizando **200 sentencias judiciales** masivas y complejas (algunas de hasta 4.000 páginas) bajo el marco de **70 preguntas e indicadores** metodológicos. Se requería que los resultados fueran comparables, auditables y con trazabilidad absoluta.

- **La Solución:** Creación de un *pipeline* de **Recuperación Aumentada Generativa (RAG)**. Este dividió los textos en fragmentos, generó representaciones semánticas (*embeddings*) y aplicó una política de respuesta estrictamente basada en el contexto para evitar que el modelo opinara libremente. Si no había evidencia documental, el sistema debía declararlo de forma explícita.

- **Resultado:** Procesamiento sistemático de las sentencias en **dos días** (tras dos semanas previas de calibración humana), obteniendo valoraciones cualitativas estructuradas y trazabilidad exacta hasta la página y fragmento de origen.

Bloque 1: Ventajas y Desafíos Identificados

A partir de las votaciones en vivo, el público identificó las **implicaciones éticas** y la **calidad de los datos** como los desafíos más apremiantes. Los expertos desglosaron los aprendizajes del ecosistema real en las siguientes dimensiones:

Ventajas Tangibles

1. **Capacidad de escala:** Viabilidad de procesar de manera masiva e integral el 100% del universo documental (informes, actas, anexos), mitigando el riesgo de omitir hallazgos por limitaciones de tiempo.

2. **Homogeneidad en el análisis:** Reducción de la variabilidad operativa producida por el cansancio humano o las discrepancias subjetivas de criterio en equipos evaluadores grandes.

3. **Trazabilidad:** Posibilidad de almacenar y auditar la cadena metodológica completa (documento original\fragmento\vector\prompt\puntaje).

4. **Eficiencia económica y temporal:** Reducción drástica de costos operativos y tiempos de lectura preliminar, permitiendo reasignar el valor de los expertos a tareas críticas.
5. **Eficiencia**: Se requieren menos recursos tanto de tiempo, como presupuestales y humanos para realizar tareas complejas como el análisis de grandes volúmenes de texto.

6. **Efectividad de la interacción experto-máquina**: El human in the loop en todo el ciclo de vida del proyecto, desde el diseño hasta la validación de resultados es clave para garantizar la calidad y precisión del uso de los modelos LLM.
7. **Iteraciones y pilotos**: La posibilidad de realizar pilotos en corto tiempo para verificar resultados es otro elemento clave que el uso de estos modelos permite a pesar de las restricciones de tiempo
8. **Exhaustividad analítica**: Es posible garantizar y tener trazabilidad de la revisión del 100% de documentos disponibles para el análisis.

Desafíos Complejos

- **Calidad de los datos:** Documentos mal escaneados o carentes de metadatos degradan inmediatamente el rendimiento algorítmico.

- **Diseño de preguntas y prompts:** Las dimensiones complejas requieren traducirse en instrucciones precisas. Los modelos del pasado se saturaban con directrices extensas; por ejemplo, Econometría tuvo que segmentar su estrategia utilizando dos modelos simultáneos (uno para criterios mínimos y otro para calidad alta) para optimizar la atención de la API de OpenAI.

- **Calibración y estabilidad:** Al no ser los modelos de lenguaje (LLM) sistemas deterministas, se deben implementar controles estadísticos iterativos para asegurar la reproducibilidad de los resultados ante los mismos insumos.

- **Riesgo de falsa confianza:** Un informe con redacción impecable y elocuente no es sinónimo de un resultado metodológicamente correcto o veraz.

Bloque 2: El Estado Actual de la IA en la Evaluación

El consenso del panel y de la audiencia determinó que el uso de la IA en América Latina se encuentra en un **nivel exploratorio y de transición**. Los LLM no operan de forma autónoma ni reemplazan el juicio de valor, pero configuran lo que se denomina **"Evaluación Aumentada"** o *Inteligencia Ampliada*.

Basándose en las directrices de redes como *BetterEvaluation*, la ONU (UNEG) y el informe de la OCDE (2025) *"Gobernar con la inteligencia artificial"*, el uso actual se concentra en:

- **Apoyo en el diseño:** Síntesis acelerada de evidencia empírica y literatura previa mediante herramientas específicas como *Covidence, DistillerSR* o *EPPI-Reviewer*.

- **Apoyo analítico:** Empleo de *machine learning* clásico para modelos predictivos (*ex-ante*) y estimación de impactos (*ex-post*). En el plano cualitativo, herramientas como *CoLoop* o

los módulos de *Atlas.ti* auxilian en la codificación deductiva mediante taxonomías preestablecidas, agilizando el tratamiento de grupos focales y entrevistas abiertas.

- **Participación ciudadana:** Procesamiento y mapeo conceptual de las opiniones de las partes interesadas y comunidades recolectadas a través de plataformas digitales.

Bloque 3: Perspectiva Futura, Ética y Estándares

El cierre de la discusión proyectó cinco grandes tendencias técnicas para el sector de la evaluación social:

1. **Automatización documental masiva:** Como un estándar obligatorio de inicio en los proyectos.

2. **Evaluación continua:** Tránsito de las evaluaciones tradicionales estáticas (fotografías temporales tomadas cada cierta cantidad de años) hacia sistemas dinámicos de retroalimentación en tiempo real.

3. **Integración metodológica:** Fusión de técnicas estadísticas cuantitativas (*clustering*) con las capacidades narrativas y de síntesis de la IA generativa.

4. **Trazabilidad técnica obligatoria** para evitar el uso de "cajas negras" algorítmicas en la sustentación de políticas públicas.

5. **Profesionalización del evaluador:** Desarrollo de habilidades críticas para formular preguntas, auditar la evidencia recolectada y parametrizar métricas de desempeño.

Alertas Críticas y Ética

Adriana Cárdenas compartió un experimento personal donde evidenció que al interrogar colecciones documentales cerradas en plataformas avanzadas , se identificaron hasta 6 párrafos de afirmaciones falsas o "alucinaciones". En paralelo, advirtieron sobre la alarmante tendencia global en la consultoría social de reducir los tiempos de las evaluaciones de 4 meses a solo 1 mes lo que puede estar motivado por la eficiencia técnica de la IA, lo cual puede poner en serio riesgo el rigor de los resultados.

Asimismo, se expuso el **Proyecto APE (Evaluación Autónoma de Políticas)** de la Universidad de Zúrich. Un experimento netamente académico que utiliza un ecosistema agéntico para redactar evaluaciones empíricas de principio a fin sin intervención humana. En torneos automatizados de comparación, las evaluaciones humanas alcanzan un puntaje promedio de 1.817 puntos, mientras que las de la IA promedian 1.264 puntos. Aunque la brecha es moderada, los creadores del proyecto aclaran explícitamente que los textos

automatizados contienen alucinaciones y no deben usarse bajo ninguna circunstancia para la toma de decisiones reales.

Marcos de Gobernanza de Referencia

Para mitigar los sesgos algorítmicos, la delegación indebida de decisiones y la pérdida de habilidades cognitivas en los equipos, el panel instó a revisar e incorporar los siguientes estándares internacionales recomendados:

- **Principios de la OCDE (Actualizados en 2024):** Centrados en la transparencia, explicabilidad, robustez, seguridad y rendición de cuentas (*accountability*).
- **UNESCO (2021):** Recomendación global sobre la ética de la inteligencia artificial.
- **IEEE 7000:** Ingeniería orientada a la integración de valores éticos desde el diseño del software.
- **ISO 42001 (2023):** Primer estándar internacional certificable para Sistemas de Gestión de Inteligencia Artificial.
- **Marco de Gestión de Riesgos del NIST (EE. UU., 2023):** Estructurado en las funciones de Gobernar, Mapear, Medir y Gestionar.
- **Ley de IA de la Unión Europea (2024):** Primer marco legal vinculante con enfoque basado en niveles de riesgo y sanciones económicas explícitas.

Conclusiones Finales

El panel concluyó con un rechazo a la idea de delegar el 100% de una evaluación social a un entorno automatizado, o al menos en la actualidad. El desafío imperante no es de naturaleza tecnológica, sino de responsabilidad ética y metodológica. El verdadero valor radica en un esquema de **supervisión humana por defecto (*Human-in-the-loop*)**, donde la máquina asuma la carga del procesamiento masivo y el ser humano mantenga de forma exclusiva la responsabilidad ética, la empatía social y la interpretación final de los resultados.

La IA debe encargarse de procesar, organizar, recuperar y sintetizar grandes volúmenes de información, mientras que los profesionales humanos retienen de forma exclusiva e irremplazable las facultades de interpretar el contexto, deliberar, aplicar el juicio de valor y responder éticamente por las recomendaciones entregadas.