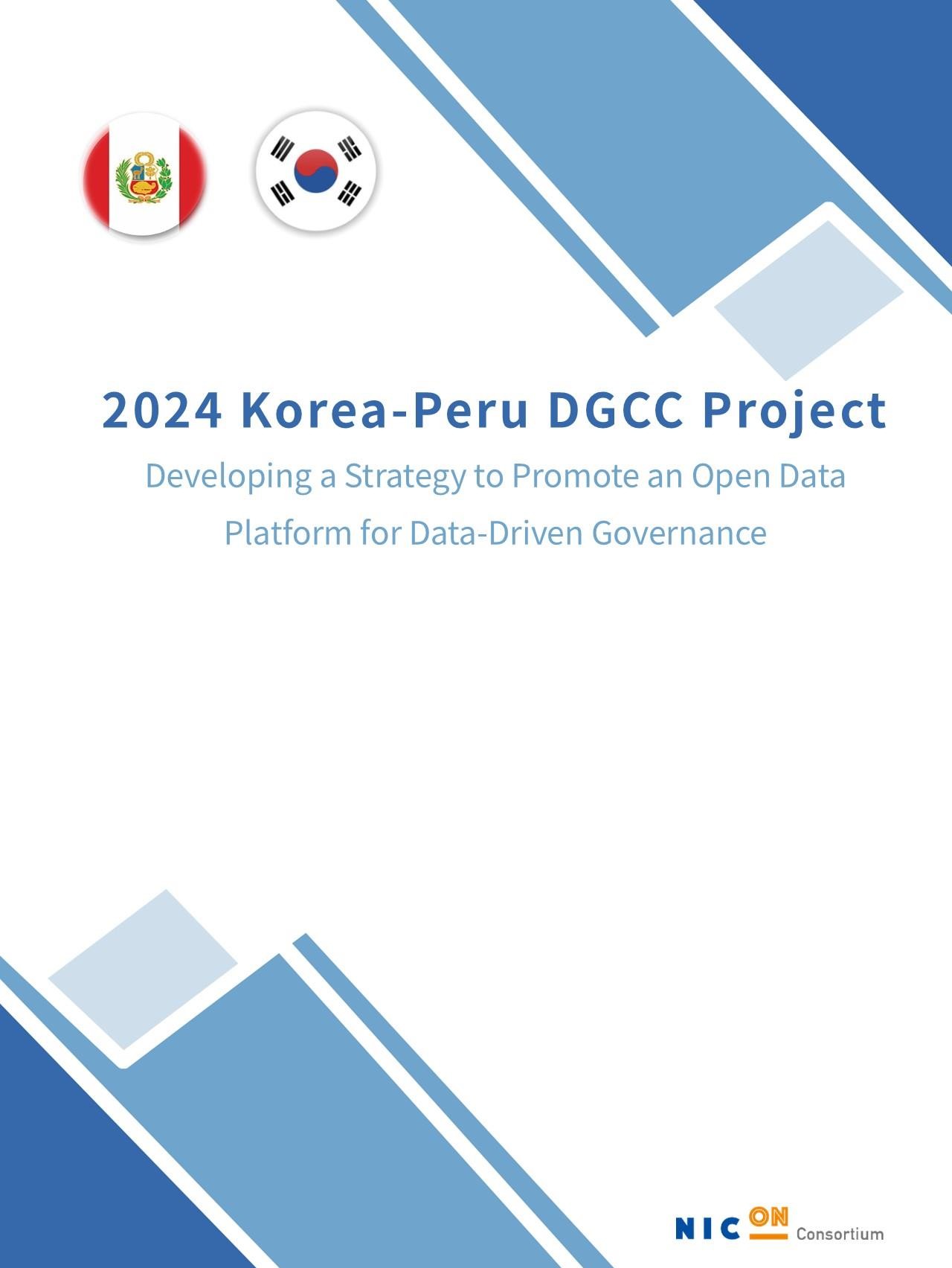
Desarrollo de una estrategia para promover la plataforma de IA en el gobierno público

**Directrices para la evaluación del desempeño de la IA en el sector público**

2024.12



**Índice**

[1. Introducción 4](#_Toc189467091)

[1.1. Necesidad y beneficios esperados de la adopción de IA 4](#_Toc189467092)

[1.2. Importancia de la evaluación del desempeño de la IA en el sector público 4](#_Toc189467093)

[1.3. Propósito y destinatarios de las directrices 5](#_Toc189467094)

[2. Proceso de adopción de IA, necesidad de evaluación del desempeño y factores de evaluación del desempeño 6](#_Toc189467095)

[2.1. Resumen del proceso de adopción de IA 6](#_Toc189467096)

[2.2. Necesidad de evaluación del desempeño en cada etapa y factores de evaluación del desempeño 7](#_Toc189467097)

[3. Indicadores y metodología de evaluación del desempeño de la IA 11](#_Toc189467098)

[3.1. Indicadores de evaluación cuantitativa 11](#_Toc189467099)

[3.2. Indicadores de evaluación cualitativa 16](#_Toc189467100)

[3.3. Herramientas y marco de evaluación del desempeño 19](#_Toc189467101)

[4. Casos de evaluación del desempeño de la IA por sector público 24](#_Toc189467102)

[4.1. Caso de evaluación del desempeño de la IA en la gestión de quejas y reclamos 24](#_Toc189467103)

[4.2. Caso de evaluación de la IA para la clasificación automática de documentos administrativos 24](#_Toc189467104)

[4.3. Caso de evaluación del desempeño de chatbots basados en IA 25](#_Toc189467105)

[4.4. Caso de evaluación de la IA para el análisis de videos e imágenes 26](#_Toc189467106)

[5. Estrategia para la implementación del proceso de evaluación del desempeño de la IA 27](#_Toc189467107)

[5.1. Establecimiento del sistema de evaluación del desempeño de la IA en instituciones públicas 27](#_Toc189467108)

[5.2. Establecimiento de los criterios y procedimientos de evaluación del desempeño de la IA 27](#_Toc189467109)

[5.3. Estrategia para la utilización de los resultados de la evaluación del desempeño 28](#_Toc189467110)

[5.4. Monitoreo continuo del desempeño de la IA y estrategias de mejora 29](#_Toc189467111)

**Índice de tablas**

[<Tabla 1 - Indicadores de evaluación: Precisión, Exactitud, Recuerdo y Puntuación F1> 11](#_Toc189746615)

[<Tabla 2 - Indicadores de evaluación: Tasa de error, Error absoluto medio, Raíz del error cuadrático medio> 13](#_Toc189746616)

<Tabla 3 - Indicadores de evaluación: Velocidad de procesamiento y tiempo de respuesta>[< 14](#_Toc189746617)

[<Tabla 4 - Indicadores de evaluación: Escalabilidad y mantenibilidad> 15](#_Toc189746618)

[<Tabla 5 - Indicadores de evaluación: Explicabilidad e interpretabilidad> 16](#_Toc189746619)

[<Tabla 6 - Indicadores de evaluación: Equidad y sesgo> 17](#_Toc189746620)

[<Tabla 7 - Indicadores de evaluación: Criterios éticos y cumplimiento legal> 19](#_Toc189746621)

[<Tabla 8 - Funcionalidades principales de Scikit-learn> 20](#_Toc189746622)

[<Tabla 9 - Ejemplo de código fuente para la evaluación del desempeño utilizando Python> 22](#_Toc189746623)

[<Tabla 10 - Funcionalidades principales de TFMA> 22](#_Toc189746624)

[<Tabla 11 - Funcionalidades principales de Fairness Indicators> 22](#_Toc189746625)

[<Tabla 12 - Funcionalidades principales de SHAP> 23](#_Toc189746626)

[<Tabla 13 - Funcionalidades principales de LIME> 23](#_Toc189746627)

[<Tabla 14 - Funcionalidades principales de Apache JMeter> 23](#_Toc189746628)

[<Tabla 15 - Comparación de herramientas de evaluación del desempeño de la IA> 24](#_Toc189746629)

[<Tabla 16 - Ejemplo de estructura del sistema de evaluación del desempeño de la IA> 27](#_Toc189746630)

[<Tabla 17 - Ejemplo de criterios de evaluación del desempeño de la IA> 28](#_Toc189746631)

[<Tabla 18 - Ejemplo de procedimiento de evaluación del desempeño de la IA en el sector público> 28](#_Toc189746632)

[< Tabla 19 - Ejemplo de estrategias para la utilización de los resultados de la evaluación del desempeño de la IA> 29](#_Toc189746633)

[<Tabla 20 - Ejemplo de monitoreo del desempeño de la IA> 29](#_Toc189746634)

1. Introducción
   1. Necesidad y beneficios esperados de la adopción de IA
      1. Necesidad de la adopción de IA en el sector público de Perú

* Mejora de la eficiencia administrativa
* El uso de IA permite analizar grandes volúmenes de datos de manera rápida y facilita la automatización de tareas administrativas, reduciendo la carga de trabajo de los funcionarios públicos.
* Mejora de los servicios para los ciudadanos
* Las tecnologías basadas en IA, como chatbots, reconocimiento de voz y análisis de imágenes, mejoran la accesibilidad a los servicios gubernamentales y permiten respuestas más rápidas y precisas.
* Aceleración de la transformación digital
* El gobierno de Perú está implementando una estrategia nacional de transformación digital, y la IA actúa como una de las tecnologías clave para su realización.
* Apoyo a la toma de decisiones basada en datos
* El uso de modelos de análisis de IA permite a los responsables de políticas desarrollar estrategias más precisas y fundamentadas en evidencia.
  + 1. Beneficios esperados
* Reducción de costos mediante la automatización administrativa
* La implementación de IA en la gestión de quejas, clasificación automática de documentos oficiales y control del tráfico reduce el tiempo de trabajo y disminuye los costos humanos y financieros.
* Apoyo en la formulación de políticas precisas
* El análisis de datos basado en IA permite desarrollar políticas más efectivas en áreas como la prevención del crimen, la predicción de la propagación de enfermedades y la optimización de la planificación urbana.
* Mayor comodidad para los ciudadanos
* Las personas con baja alfabetización digital pueden acceder más fácilmente a los servicios públicos a través de soluciones personalizadas basadas en IA.
* Fortalecimiento de la inclusión digital
* Las personas con baja alfabetización digital pueden acceder más fácilmente a los servicios públicos a través de soluciones personalizadas basadas en IA.
  1. Importancia de la evaluación del desempeño de la IA en el sector público
     1. Necesidad de la evaluación del desempeño para garantizar la fiabilidad
* Garantía de la calidad de los servicios públicos
* Si la IA en los servicios públicos comete errores, puede generar una pérdida de confianza por parte de los ciudadanos y reducir la credibilidad de los servicios administrativos. Por ello, es esencial garantizar la fiabilidad de los modelos de IA mediante evaluaciones de desempeño.
* Operación responsable de la IA
* Cuando la IA influye directamente en las políticas públicas, el gobierno debe evaluar su desempeño de manera transparente y garantizar su explicabilidad.
* Verificación de la equidad y ética de la IA
* En el sector público, si la IA discrimina a ciertos grupos o toma decisiones incorrectas, pueden surgir controversias sociales. Es fundamental evaluar si los modelos de IA cumplen con estándares justos y éticos.
  + 1. Necesidad de la evaluación del desempeño de la IA en las instituciones públicas de Perú
* IA para la automatización de la gestión de quejas y reclamos
* Si la IA clasifica y responde automáticamente a las quejas con baja precisión o respuestas sesgadas, puede aumentar la insatisfacción de los ciudadanos.
* Análisis de datos médicos basado en IA
* Dado que un error en el diagnóstico generado por la IA puede afectar directamente la salud de la población, es fundamental evaluar su precisión y fiabilidad.
* IA para la gestión del tráfico y la planificación urbana
* Si un sistema de control de semáforos basado en IA funciona de manera ineficiente, puede agravar la congestión del tráfico, por lo que su rendimiento debe ser verificado rigurosamente.
* IA para la predicción y respuesta ante desastres
* Si la IA proporciona datos inexactos en la predicción de desastres naturales o cambios climáticos, será difícil prevenir daños, lo que hace imprescindible una evaluación exhaustiva.
  1. Propósito y alcance de la guía
     1. Principales destinatarios de la guía
* Presentación de criterios y procedimientos objetivos para la evaluación del desempeño
* Establecer criterios y procedimientos que permitan evaluar objetivamente el desempeño de la IA en el sector público de Perú.
* Provisión de servicios de IA óptimos para los ciudadanos peruanos
* Verificar la fiabilidad, equidad y eficiencia de los modelos de IA antes de su implementación en los servicios públicos, garantizando así un servicio óptimo para los ciudadanos.
* Guía estandarizada para la evaluación del desempeño de la IA en el sector público
* Proporcionar una metodología estandarizada y directrices prácticas para que las instituciones públicas puedan llevar a cabo evaluaciones de desempeño de la IA de manera autónoma.
  + 1. Principales destinatarios de las directrices
* Gobierno de Perú e instituciones públicas
* Organismos que buscan implementar IA en la administración y los servicios públicos.
* Departamentos de datos públicos y tecnología de la información
* Unidades gubernamentales encargadas de la construcción y operación de sistemas de IA.
* Proveedores y desarrolladores de soluciones de IA
* Empresas e instituciones de investigación que colaboran con el gobierno en el desarrollo de servicios de IA.
* Entidades de políticas y regulación
* Organismos gubernamentales responsables de establecer y supervisar los estándares de desempeño y ética de la IA.
* Ciudadanos y ONG
* Partes interesadas que supervisan la transparencia y equidad de los sistemas de IA públicos y proporcionan retroalimentación.

1. Procedimiento de adopción de IA, necesidad de evaluación del desempeño y factores de evaluación
   1. Resumen del procedimiento de adopción de IA

La adopción de IA puede realizarse a través de dos enfoques: la implementación de soluciones existentes o el desarrollo propio. Para obtener información detallada sobre los métodos de adopción de IA, se recomienda consultar la "Guía para la Adopción y Operación de IA en el Sector Público de Perú". Para que las instituciones públicas en Perú implementen la IA de manera efectiva, deben seguir un procedimiento estructurado. La aplicación de sistemas de IA en los servicios públicos requiere garantizar su fiabilidad y equidad, lo que hace imprescindible la evaluación del desempeño en cada etapa del proceso.

A continuación, se presentan los procedimientos de adopción de IA según el enfoque basado en soluciones existentes y el desarrollo propio.

* + 1. Procedimiento de adopción de soluciones de IA

El procedimiento de adopción de soluciones de IA se lleva a cabo en las siguientes etapas: definición de requisitos, evaluación preliminar de riesgos, selección del proveedor de soluciones, implementación de la solución de IA y operación y gestión del desempeño del servicio de IA.

<Figura 1 - Procedimiento de adopción de soluciones de IA>

* **Definición de requisitos**: Etapa en la que se establecen los requisitos alineados con los objetivos y propósitos del servicio de IA en el sector público.
* **Evaluación preliminar de riesgos:** Fase en la que se identifican y minimizan los riesgos técnicos, operativos, legales y éticos antes de la adopción de la solución de IA.
* **Selección del proveedor de soluciones:** Proceso mediante el cual la institución pública elige un proveedor de soluciones de IA que cumpla con sus requisitos de servicio.
* **Implementación de la solución de IA:** Etapa en la que la solución de IA seleccionada se implementa en la institución pública.
* **Operación y gestión del desempeño del servicio de IA:** Fase en la que, tras la implementación, se monitorea el desempeño del sistema de IA en la etapa operativa.
  + 1. Procedimiento de desarrollo propio de IA

El procedimiento de desarrollo propio de IA se lleva a cabo en las siguientes etapas: planificación, diseño y construcción, preparación de datos y entrenamiento del modelo, validación e integración del modelo de IA, y operación y mejora continua.

<Figura 2 - Procedimiento de desarrollo propio de IA>

* **Fase de planificación:** Definir claramente los objetivos y propósitos de la adopción de IA, estableciendo la dirección general del proyecto y su plan de ejecución.
* **Fase de diseño y construcción**: Diseñar las funciones y la arquitectura tecnológica del servicio de IA, y construir el sistema y la infraestructura para crear un entorno operativo viable.
* **Fase de preparación de datos y entrenamiento del modelo**: Recopilar, limpiar y preprocesar los datos necesarios para el entrenamiento del modelo de IA, utilizando los algoritmos más adecuados para su aprendizaje.
* **Fase de validación e integración del modelo de IA**: Verificar el rendimiento y la estabilidad del modelo entrenado e integrarlo con los sistemas existentes para su implementación en un entorno real.
* **Fase de operación y mejora continua:** Operar el servicio de IA, monitorear su desempeño en tiempo real y realizar mejoras y actualizaciones constantes.
  1. Necesidad de evaluación del desempeño en cada etapa y factores de evaluación
     1. Enfoque de adopción de soluciones de IA
        1. Fase de definición de requisitos
* **Necesidad**
* Es fundamental definir claramente el propósito de la adopción de IA y establecer las funciones y objetivos de desempeño requeridos para el servicio público en cuestión.
* Se deben establecer los principales indicadores de desempeño del modelo de IA, considerando la naturaleza de la administración pública, como precisión (Accuracy), velocidad de procesamiento, fiabilidad y escalabilidad.
* Se debe determinar un estándar de calidad de datos y una política de protección de la privacidad, teniendo en cuenta el tipo y alcance de los datos que manejará la IA.
* **Factores de evaluación del desempeño**
* Selección de la tecnología de IA adecuada según el propósito del servicio (por ejemplo, chatbot, análisis de imágenes, clasificación de texto, etc.).
* Establecimiento del nivel de desempeño esperado (por ejemplo, precisión de respuestas superior al 90% en un sistema de IA para respuestas automáticas a quejas y reclamos).
* Evaluación de la compatibilidad técnica entre el entorno operativo de la institución pública y la solución de IA.
  + - 1. Fase de evaluación preliminar de riesgos
* **Necesidad**
* Es fundamental revisar si la solución de IA podría comprometer la fiabilidad y transparencia de los servicios públicos.
* Se debe evaluar si los datos de entrenamiento del modelo presentan sesgos (Bias) y garantizar que se cumplan los principios de equidad (Fairness) y explicabilidad (Explainability).
* Se debe verificar el cumplimiento de la solución de IA con la legislación vigente (Ley de Protección de Datos Personales, Guías de Ética en IA, etc.).
* **Factores de evaluación del desempeño**
* Análisis de si el modelo de IA genera impactos discriminatorios en determinados grupos (género, edad, región, etc.).
* Revisión de si el proceso de toma de decisiones de la IA es transparente y explicable.
* Evaluación de la implementación de medidas adecuadas para la seguridad de los datos y la protección de la privacidad.
  + - 1. Fase de selección del proveedor de soluciones
* **Necesidad**
* Se debe verificar si el modelo de IA propuesto por el proveedor cumple con los requisitos establecidos.
* Es fundamental probar el desempeño del modelo de IA en diversos entornos para evaluar su aplicabilidad en la práctica.
* Se debe considerar la viabilidad de mantenimiento y actualización del sistema, así como la estructura de soporte técnico, para garantizar su operatividad a largo plazo.
* **Factores de evaluación del desempeño**
* Comparación del desempeño del modelo de IA utilizando un conjunto de datos de prueba común (precisión, velocidad de respuesta, eficiencia de recursos, etc.).
* Prueba de estabilidad del modelo de IA en el entorno real (sistemas administrativos, infraestructura en la nube, etc.).
* Evaluación de los planes del proveedor para la mejora del rendimiento de la IA y la disponibilidad de soporte de mantenimiento.
  + - 1. Fase de implementación y aplicación de la solución de IA
* **Necesidad**
* Es fundamental analizar las diferencias de desempeño entre el entorno de desarrollo y el entorno real de operación en los servicios públicos.
* Se debe verificar si la solución de IA mantiene consistentemente el nivel de desempeño esperado.
* Es necesario optimizar el modelo de IA incorporando retroalimentación de los usuarios reales y realizar mejoras cuando sea necesario.
* **Factores de evaluación del desempeño**
* Medición de la precisión y velocidad de respuesta de la IA después de su implementación en los servicios públicos.
* Análisis de la frecuencia de errores y casos de mal funcionamiento (por ejemplo, tasa de clasificación errónea en un sistema de clasificación automática de quejas y reclamos).
* Encuestas de satisfacción a usuarios (funcionarios y ciudadanos) y recopilación de solicitudes de mejora.
  + - 1. Fase de operación y gestión del desempeño del servicio de IA
* **Necesidad**
* Se debe monitorear si la solución de IA mantiene un rendimiento óptimo con el tiempo, evitando la degradación del desempeño (Drift).
* A medida que cambian los datos procesados por la IA, es necesario evaluar si se requiere reentrenamiento y actualización del modelo con nuevos datos.
* Con base en los resultados de la evaluación del desempeño, se deben establecer estrategias de mejora y garantizar la gestión continua de la calidad de la solución de IA.
* **Factores de evaluación del desempeño**
* Revisión periódica del desempeño del modelo de IA en operación y elaboración de informes de evaluación.
* Análisis de la necesidad de reentrenamiento y actualización del modelo para mejorar el servicio.
* Supervisión continua del cumplimiento de los estándares legales y éticos en el servicio de IA.
  + 1. Enfoque de desarrollo propio de IA
       1. Fase de planificación
* **Necesidad**
* Definir claramente el propósito y los beneficios esperados de la aplicación de IA en la gestión administrativa, estableciendo criterios concretos para la evaluación del desempeño.
* Establecer de antemano los objetivos de desempeño (por ejemplo, precisión superior al 90%, tiempo de procesamiento inferior a 1 segundo) y evaluar constantemente el cumplimiento de estos durante el desarrollo.
* Considerar previamente la ética, equidad y cumplimiento legal del modelo de IA, incorporando estos aspectos en los criterios de evaluación del desempeño.
* **Factores de evaluación del desempeño**
* Definición de indicadores de evaluación alineados con los objetivos y beneficios esperados de la IA (precisión, velocidad, impacto en la reducción de costos, etc.).
* Inclusión de criterios de equidad (Fairness) y alineación con las guías éticas.
* Revisión del cumplimiento con regulaciones legales y normas de protección de datos.
  + - 1. Fase de diseño y construcción
* **Necesidad**
* Evaluar si la arquitectura del sistema de IA (basada en la nube, on-premise, etc.) está optimizada para su implementación.
* Comparar diferentes algoritmos de IA y evaluar el rendimiento inicial del modelo seleccionado antes de su implementación.
* Garantizar que la arquitectura de IA esté diseñada para ser escalable y mantenible a largo plazo, asegurando su eficiencia operativa.
* **Factores de evaluación del desempeño**
* Pruebas comparativas entre diferentes algoritmos de IA (aprendizaje profundo, aprendizaje automático, etc.).
* Evaluación de la escalabilidad y capacidad de mantenimiento del sistema.
* Prueba del rendimiento inicial del modelo (velocidad de entrenamiento, eficiencia computacional, capacidad de procesamiento de datos, etc.).
  + - 1. Fase de preparación de datos y entrenamiento del modelo
* **Necesidad**
* La calidad de los datos de entrenamiento es el factor más crítico en el rendimiento del modelo de IA, por lo que es esencial evaluarla minuciosamente.
* Se debe revisar el equilibrio, diversidad y coherencia de los datos para garantizar un entrenamiento justo y sin sesgos (Bias).
* Tras el entrenamiento del modelo, se debe analizar la presencia de sobreajuste (Overfitting) o subajuste (Underfitting) para prever su rendimiento en un entorno real.
* **Factores de evaluación del desempeño**
* Evaluación de la calidad de los datos (eliminación de ruido, detección de datos duplicados).
* Análisis de sesgos (Bias) en los datos y verificación de equidad en el entrenamiento.
* Evaluación del sobreajuste y capacidad de generalización del modelo.
  + - 1. Fase de validación e integración del modelo de IA
* **Necesidad**
* Se debe verificar si el modelo de IA desarrollado cumple con el rendimiento esperado en un entorno real de servicio público.
* Es fundamental evaluar la fiabilidad, precisión y velocidad de procesamiento del modelo utilizando diversos conjuntos de datos de prueba para seleccionar el modelo óptimo.
* Se debe analizar la compatibilidad del modelo con los sistemas administrativos existentes y comprobar que su integración no degrade el rendimiento del sistema.
* **Factores de evaluación del desempeño**
* Validación del rendimiento del modelo de IA mediante conjuntos de datos de prueba (Precisión, Recall, F1-score, etc.).
* Pruebas de compatibilidad e integración con los sistemas internos de las instituciones públicas (conexión API, flujo de datos, etc.).
* Evaluación de la velocidad de respuesta en tiempo real y la capacidad de procesamiento del modelo de IA.
  + - 1. Fase de operación y gestión del desempeño del servicio de IA
* **Necesidad**
* Se debe evaluar periódicamente si la IA mantiene un rendimiento óptimo en su uso dentro del servicio público.
* Es fundamental detectar la degradación del rendimiento del modelo debido al Concept Drift (deriva del concepto) y garantizar actualizaciones oportunas.
* A partir del análisis de los resultados de evaluación del desempeño, se deben establecer estrategias de mejora del servicio y aumentar la utilidad del sistema de IA en la formulación de políticas públicas.
* **Factores de evaluación del desempeño**
* Verificación de la precisión y mantenimiento del rendimiento del modelo de IA en un entorno real.
* Evaluación de la necesidad de actualización y reentrenamiento del modelo de IA.
* Elaboración de un plan de mejora del servicio de IA basado en la retroalimentación de los usuarios (funcionarios y ciudadanos).

1. Indicadores y metodología de evaluación del desempeño de la IA

Para evaluar objetivamente el desempeño de la IA, es necesario utilizar tanto indicadores cuantitativos como cualitativos. Los indicadores cuantitativos permiten medir el rendimiento en valores numéricos, mientras que los indicadores cualitativos se centran en evaluar qué tan efectivamente opera el sistema de IA en un entorno de servicio real.

Además, el uso de diversas herramientas de evaluación de código abierto permite validar el desempeño del modelo con mayor fiabilidad y garantizar una evaluación precisa y transparente.

* 1. Indicadores de evaluación cuantitativa
     1. Precisión (Accuracy), Precisión Positiva (Precision), Sensibilidad (Recall), Puntuación F1 (F1-score)

Indicadores clave para evaluar el rendimiento predictivo de un modelo de IA, especialmente utilizados en modelos de clasificación.

|  |  |  |
| --- | --- | --- |
| Indicadores de evaluación | Descripción  detallada | Fórmulas |
| Precisión (Accuracy) | Proporción de predicciones correctas sobre el total de predicciones realizadas. |  |
| Precisión Positiva (Precision) | Proporción de casos en los que la IA predijo la clase positiva correctamente. |  |
| Sensibilidad (Recall) | Proporción de muestras realmente positivas que la IA predijo correctamente |  |
| F1-Score | Valor de la media armónica que equilibra la precisión (Precision) y la sensibilidad (Recall). |  |

< Tabla 1 - Indicadores de evaluación: Precisión (Accuracy), Precisión Positiva (Precision), Sensibilidad (Recall) y Puntuación F1 (F1-score) >

|  |  |  |
| --- | --- | --- |
| Símbolos | Descripción detallada | Ejemplo |
| TP(True Positive) | Cuando la IA predice una clase como positiva y realmente es positiva. | En un sistema de IA para el diagnóstico de COVID-19, cuando un paciente realmente infectado es correctamente identificado como infectado. |
| TN(True Negative) | Cuando la IA predice una clase como negativa y realmente es negativa. | Cuando una persona no infectada por COVID-19 es correctamente clasificada como negativa. |
| FP(False Positive) | Cuando la IA predice una clase como positiva, pero en realidad es negativa. | Cuando la IA de diagnóstico de COVID-19 clasifica erróneamente a una persona no infectada como infectada. |
| FN(False Negative) | Cuando la IA predice una clase como negativa, pero en realidad es positiva. | Cuando la IA de diagnóstico de COVID-19 clasifica erróneamente a una persona infectada como negativa. |

* **Ejemplo de cálculo de precisión (Accuracy)**
* Supongamos que el gobierno de Perú utiliza IA para clasificar automáticamente las quejas en tres categorías: tráfico, medio ambiente y salud.

|  |  |  |
| --- | --- | --- |
| Quejas reales | Resultados predichos por la IA | Exactitud de la predicción |
| Tráfico | Tráfico | TP |
| Medio ambiente | Medio ambiente | TN |
| Salud | Salud | TN |
| Tráfico | Medio ambiente | FP |
| Medio ambiente | Salud | FN |

* Resultado: Accuracy = (TP + TN) / (TP + TN + FP + FN) = 3/5 = 60%.
* **Ejemplo de cálculo de precisión (Precision)**
* Supongamos que la IA determina la aprobación de préstamos.

|  |  |  |  |
| --- | --- | --- | --- |
| Solicitante | Elegibilidad real para el préstamo | Predicción de la IA | Relación con la precisión (Precision) |
| Cliente A | Posible | Aprobado | TP |
| Cliente B | Posible | Aprobado | TP |
| Cliente C | No posible | Aprobado | FP |
| Cliente D | No posible | Rechazado | TN |

* Resultado: Precision = TP / (TP + FP) = 2/3 = 66.7%.
* Esto significa que, de los préstamos aprobados por la IA, solo el 66.7% corresponde a clientes realmente elegibles para el préstamo.
* **Ejemplo de cálculo de sensibilidad (Recall)**
* Supongamos que la IA analiza imágenes de rayos X para diagnosticar a pacientes con cáncer de pulmón.

|  |  |  |  |
| --- | --- | --- | --- |
| Paciente | Presencia real de cáncer | Predicción de la IA | Relación con la sensibilidad (Recall) |
| Paciente A | Cáncer | Cáncer | TP |
| Paciente B | Cáncer | Cáncer | TP |
| Paciente C | Cáncer | Normal | FN |
| Paciente D | Normal | Normal | TN |

* Resultado: Recall = TP / (TP + FN) = 2/3 = 66.7%.
* Esto significa que la IA solo identificó correctamente el 66.7% de los pacientes con cáncer.
* Una baja sensibilidad implica un alto riesgo de no detectar pacientes con cáncer (Falsos Negativos - FN), lo que puede generar graves problemas.
* En aplicaciones médicas, es fundamental maximizar la sensibilidad (Recall) para reducir los Falsos Negativos (FN).
* **Ejemplo de cálculo de F1-Score**
* Supongamos que la IA analiza imágenes de CCTV para predecir las áreas con alta probabilidad de ocurrencia de robos.

|  |  |
| --- | --- |
| Indicador de evaluación | Resultado del cálculo |
| Precisión (Precision) | 80% (Proporción de áreas identificadas por la IA como de alto riesgo donde realmente ocurrió un robo.) |
| Sensibilidad (Recall) | 50% (Proporción de áreas donde realmente ocurrió un robo y que fueron correctamente detectadas por la IA.) |

* Resultado: F1-Score = 2 × (0.8 × 0.5) / (0.8 + 0.5) = 2 × (0.4 / 1.3) = 0.615
* Un F1-score de 0.615 indica que se requiere un equilibrio entre precisión y sensibilidad.
* Aumentar la precisión mejora la exactitud en la predicción de robos, pero podría reducir la detección de casos reales (disminución de Recall).
* Aumentar la sensibilidad podría llevar a predecir todas las áreas como de alto riesgo, reduciendo la precisión.
* En sistemas de predicción de crímenes, es crucial optimizar el F1-score para lograr un balance adecuado.
  + 1. Tasa de error (Error Rate) y errores (MAE, RMSE)

Indicador que mide la cantidad de errores cometidos por el modelo de IA al realizar predicciones.

|  |  |  |
| --- | --- | --- |
| Indicador de evaluación | Descripción detallada | Fórmula |
| Tasa de error (Error Rate) | Proporción de predicciones incorrectas sobre el total de predicciones |  |
| Error absoluto medio (MAE: Mean Absolute Error) | Diferencia absoluta promedio entre los valores predichos y los valores reales |  |
| Raíz del error cuadrático medio (RMSE, Root Mean Square Error) | Raíz cuadrada del promedio de los errores elevados al cuadrado |  |

<Tabla 2 - Indicadores de evaluación: Tasa de error, Error absoluto medio (MAE), Raíz del error cuadrático medio (RMSE)>

* **Ejemplo de cálculo de la tasa de error (Error Rate)**
* Supongamos que el gobierno de Perú utiliza IA para clasificar automáticamente las quejas en tres categorías: "Tráfico", "Salud" y "Medio Ambiente".

|  |  |  |  |
| --- | --- | --- | --- |
| Indicador de evaluación | Predicción de la IA | Valor real | Exactitud |
| Queja relacionada con tráfico | Clasificada como tráfico | Es tráfico | Correcto (TP) |
| Queja relacionada con salud | Clasificada como medio ambiente | Es salud | Incorrecto (FN) |
| Queja relacionada con medio ambiente | Clasificada como medio ambiente | Clasificada como medio ambiente | Correcto (TN) |
| Queja relacionada con tráfico | Clasificada como salud | Es tráfico | Incorrecto (FP) |
| Queja relacionada con salud | Clasificada como salud | Es salud | Correcto (TP) |

* Resultado: De un total de 5 datos, hay 2 errores (FN = 1, FP = 1), por lo que la Tasa de Error (Error Rate) = 2/5 = 0.4 (40%).
* **Ejemplo de cálculo del Error Absoluto Medio (MAE, Mean Absolute Error)**
* Supongamos que el gobierno de Perú utiliza IA para predecir la concentración promedio diaria de partículas finas (PM2.5) en la región de Lima.

|  |  |  |  |
| --- | --- | --- | --- |
| Fecha | Concentración real de PM2.5 | Predicción de la IA | Error () |
| 1 de enero | 50 | 55 | 5 |
| 2 de enero | 40 | 42 | 2 |
| 3 de enero | 35 | 30 | 5 |
| 4 de enero | 45 | 50 | 5 |
| 5 de enero | 60 | 55 | 5 |

* Resultado: MAE = (5 + 2 + 5 + 5 + 5) / 5 = 22 / 5 = 4.4, lo que indica que la IA tiene un error promedio de predicción de 4.4.
* **Ejemplo de cálculo de la Raíz del Error Cuadrático Medio (RMSE, Root Mean Square Error)**
* Supongamos que el gobierno de Perú utiliza IA para predecir el consumo de electricidad (MWh) en la región de Lima.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fecha | Consumo real de electricidad (MWh) | Predicción de la IA | Error () | Error cuadrático |
| 1 de enero | 1000 | 1100 | 100 | 10000 |
| 2 de enero | 1200 | 1250 | 50 | 2500 |
| 3 de enero | 1100 | 1150 | 50 | 2500 |
| 4 de enero | 1300 | 1350 | 50 | 2500 |
| 5 de enero | 1250 | 1220 | 30 | 900 |

* Resultado: RMSE = , lo que indica que la predicción de la IA difiere en promedio en 60.7 MWh del valor real.
  + 1. Velocidad de procesamiento y tiempo de respuesta (Latencia)

El tiempo que tarda un modelo de IA en recibir una entrada y generar una predicción. Es un indicador clave en los servicios públicos que requieren respuestas en tiempo real.

|  |  |  |
| --- | --- | --- |
| Indicador de evaluación | Descripción detallada | Fórmula |
| Latencia (Latency) | Tiempo promedio (ms) que tarda el modelo de IA en procesar una solicitud | Latency = Tiempo de salida – Tiempo de entrada |
| Velocidad de procesamiento (Throughput) | Cantidad de solicitudes que el modelo de IA puede procesar por unidad de tiempo | Número de solicitudes procesadas por segundo |

<Tabla 3 - Indicadores de evaluación: Velocidad de procesamiento y tiempo de respuesta>

* **Ejemplo de cálculo de la velocidad de procesamiento (Throughput)**
* Supongamos que el gobierno de Perú opera un sistema de IA para clasificar automáticamente las quejas de los ciudadanos.

|  |  |
| --- | --- |
| Versión del modelo | Número de quejas procesadas por segundo (Throughput) |
| AI Modelo A | 20 quejas/segundo |
| AI Modelo B | 50 quejas/segundo |
| AI Modelo C | 80 quejas/segundo |

* Resultado: El modelo de IA C tiene la mayor velocidad de procesamiento (80 quejas/segundo), lo que permite gestionar un gran volumen de quejas de manera más rápida.
* **Ejemplo de cálculo del tiempo de respuesta (Latency)**
* Supongamos que el gobierno de Perú implementa un chatbot de IA para la atención de quejas ciudadanas.

|  |  |
| --- | --- |
| Versión del modelo | Tiempo de respuesta (Latency) |
| AI Chatbot A | 3 segundos |
| AI Chatbot B | 1 segundo |
| AI Chatbot C | 0.5 segundos |

* Resultado: El AI Chatbot C tiene el tiempo de respuesta más rápido (0.5 segundos), permitiendo a los ciudadanos recibir respuestas de manera inmediata sin esperas.
  + 1. Escalabilidad y mantenibilidad del modelo

Indicador que evalúa si el modelo de IA puede operar sin problemas en diversos entornos.

|  |  |
| --- | --- |
| Indicador de evaluación | Descripción detallada |
| Escalabilidad (Scalability) | Evalúa si el modelo puede mantener un rendimiento adecuado cuando se agregan nuevos datos. |
| Mantenibilidad (Maintainability) | Evalúa si el modelo puede actualizarse y reentrenarse fácilmente en caso de cambios. |
| Costo computacional (Compute Cost) | Mide la eficiencia operativa de la IA en entornos de nube o en servidores locales (On-Premise) dentro de las instituciones públicas. |

<Tabla 4 - Indicadores de evaluación: Escalabilidad y mantenibilidad>

* **Ejemplo de cálculo de escalabilidad (Scalability)**
* Supongamos que el gobierno de Perú opera un sistema de IA para la clasificación automática de quejas ciudadanas.

|  |  |  |
| --- | --- | --- |
| Año | Promedio diario de quejas | Desempeño del modelo de IA |
| 2024 | 10,000 quejas | Procesamiento normal |
| 2025 | 50,000 quejas | Reducción de la velocidad de procesamiento |
| 2026 | 100,000 quejas | Fallo del sistema debido a sobrecarga del servidor |

* Resultado: En 2024, la IA operó correctamente, pero a partir de 2025, el aumento en el número de quejas provocó una reducción en la velocidad de procesamiento.
* **Ejemplo de cálculo del costo computacional (Compute Cost)**
* Supongamos que el gobierno de Perú implementa un sistema basado en IA para analizar datos de tráfico en Lima y ajustar las señales en tiempo real.

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo de IA | Recursos utilizados | Costo estimado por día | Problema |
| Modelo A | CPU 32 núcleos, RAM 128GB | $200 | Costo demasiado alto |
| Modelo B | 2 GPUs, RAM 256GB | $500 | Sobrecarga de cómputo |
| Modelo C | Modelo optimizado, RAM 64GB | $50 | Optimizado |

* Modelo A: Funciona principalmente con CPU, pero el costo es relativamente alto para manejar cálculos complejos.
* Modelo B: Utiliza GPU de alto rendimiento para un procesamiento rápido, pero el costo es demasiado elevado para una operación a largo plazo.
* Modelo C: Optimiza el modelo para reducir el uso de RAM, disminuyendo significativamente los costos computacionales.
  1. Indicadores de evaluación cualitativa
     1. Explicabilidad e interpretabilidad (Explainability)

Capacidad de la IA para hacer comprensible el motivo de sus decisiones. En los servicios públicos, la transparencia es esencial, por lo que la explicabilidad es un factor clave de evaluación.

|  |  |
| --- | --- |
| Técnica | Descripción detallada |
| SHAP | SHAP (Shapley Additive Explanations) | Método que cuantifica y explica los factores que influyeron en cada predicción individual del modelo de IA. |
| LIME | LIME (Local Interpretable Model-agnostic Explanations) | Método que analiza las predicciones individuales de la IA modificando los datos cercanos a la predicción para proporcionar una explicación. |

<Tabla 5 - Indicadores de evaluación: Explicabilidad e interpretabilidad>

* **Ejemplo de SHAP**
* Supongamos que el gobierno de Perú opera un sistema de IA para determinar la elegibilidad de los ciudadanos para recibir beneficios sociales.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Solicitante | Predicción de la IA | Influencia del nivel de ingresos | Influencia del número de dependientes | Influencia de la discapacidad | Resultado del análisis SHAP |
| A | Aprobado | Bajo (+0.7) | Alto (+0.4) | Sí (+0.5) | Factores clave para la aprobación: bajos ingresos, muchos dependientes, discapacidad |
| B | Rechazado | Alto (-0.8) | Ninguno (-0.3) | No (-0.2) | Factores clave para el rechazo: ingresos altos, sin dependientes, sin discapacidad |

* El análisis SHAP permite cuantificar la influencia de cada variable (ingresos, número de dependientes, discapacidad) en la predicción.
* Facilita la comprensión de los ciudadanos sobre las decisiones de la IA, explicando por qué se aprobó o se rechazó el beneficio.
* **Ejemplo de LIME**
* Supongamos que un hospital público en Perú opera un sistema de IA que analiza imágenes de rayos X para diagnosticar neumonía.

|  |  |  |  |
| --- | --- | --- | --- |
| Paciente | Diagnóstico de la IA | Área anómala en la radiografía | Resultado del análisis LIME |
| A | Posibilidad de neumonía 85% | Parte inferior izquierda del pulmón | La IA predice neumonía basándose principalmente en la sombra en la parte inferior izquierda del pulmón. |
| B | Normal | Ninguna | La IA no detecta ninguna anomalía evidente. |

* El uso de LIME permite visualizar qué área específica analizó la IA para diagnosticar neumonía.
* Facilita la colaboración con los médicos para verificar si la IA ha aprendido características incorrectas.
* Proporciona explicaciones adicionales al personal médico para aumentar la confianza en los resultados del diagnóstico de la IA.
  + 1. Evaluación de equidad (Fairness) y sesgo (Bias)

Es necesario evaluar el sesgo del modelo de IA para evitar que genere resultados discriminatorios hacia ciertos grupos.

|  |  |  |
| --- | --- | --- |
| Técnica de detección | Descripción detallada | Fórmula |
| Demographic Parity | * El modelo de IA debe mantener la misma tasa de predicción entre diferentes grupos. * La probabilidad de una predicción (aprobación/rechazo, etc.) debe ser igual para todos los grupos, independientemente del género, raza o edad. * La tasa de resultados positivos predichos por la IA (por ejemplo, aprobación de préstamos, selección en un proceso de contratación) debe ser la misma en todos los grupos. | P(Y^=1∣A=0)=P(Y^=1∣A=1)   * **A es un atributo protegido (Protected Attribute), por ejemplo, el género (A=0: Mujer, A=1: Hombre).** |
| Equalized Odds | * El modelo de IA debe predecir con la misma precisión en todos los grupos. * La proporción de predicciones correctas (True Positive, True Negative) debe ser uniforme en todos los grupos. * No debe haber diferencias significativas en la tasa de True Positives (TP) y False Positives (FP) entre distintos grupos. | P(=1∣Y=1,A=0)=P(=1∣Y=1,A=1) P(=0∣Y=0,A=0)=P(=0∣Y=0,A=1)   * Cuando el valor real (Y) es el mismo, la predicción de la IA () debe ocurrir en la misma proporción para todos los grupos. |
| Disparate Impact | * El modelo de IA no debe afectar negativamente a ciertos grupos (género, raza, edad, etc.). * Si un grupo recibe resultados positivos (por ejemplo, aprobación de préstamos, contratación) con menos frecuencia que otro grupo, se considera una violación de Disparate Impact. * Para garantizar la equidad, un grupo debe recibir al menos el 80% de los resultados positivos en comparación con otro grupo (Regla del 80%). | * **La tasa de predicciones positivas para el grupo desfavorecido debe ser al menos el 80% de la del grupo favorecido.** |

<Tabla 6 - Indicadores de evaluación: Equidad (Fairness) y sesgo (Bias)>

* **Ejemplo de Demographic Parity**
* Supongamos que el gobierno de Perú utiliza IA para evaluar automáticamente a los candidatos en el proceso de selección de funcionarios públicos.

|  |  |  |  |
| --- | --- | --- | --- |
| Género | Número de solicitantes | Número de aprobados por la IA | Tasa de aprobación |
| Hombres | 100 | 50 | 50% |
| Mujeres | 100 | 25 | 25% |

* El 50% de los solicitantes masculinos fueron aprobados, mientras que solo el 25% de las solicitantes femeninas fueron aprobadas → Existe un sesgo (Bias) a favor de los hombres en la IA.
* Si el modelo de IA aplicara los mismos criterios, no debería haber diferencias en la tasa de aprobación por género.
* Esto constituye una violación del principio de Demographic Parity.
* **Ejemplo de Equalized Odds**
* Supongamos que un hospital público en Perú utiliza un sistema de IA para el diagnóstico de cáncer.

|  |  |  |  |
| --- | --- | --- | --- |
| Grupo de pacientes | Número real de pacientes con cáncer | Diagnósticos positivos por IA (TP) | Sensibilidad (Recall) |
| Pacientes masculinos | 100 | 90 | 90% |
| Pacientes femeninas | 100 | 70 | 70% |

* La sensibilidad (Recall) en pacientes masculinos es del 90%, mientras que en pacientes femeninas es del 70% → La IA omite más casos de cáncer en mujeres.
* Esto constituye una violación del principio de Equalized Odds.
* El modelo de IA presenta un sesgo (Bias) al predecir el cáncer con mayor precisión en un grupo en comparación con el otro.
* **Ejemplo de Disparate Impact**
* Supongamos que un banco en Perú utiliza un sistema de IA para determinar la aprobación de préstamos.

|  |  |  |  |
| --- | --- | --- | --- |
| Grupo étnico | Número de solicitantes | Número de préstamos aprobados | Tasa de aprobación |
| Grupo A | 100 | 60 | 60% |
| Grupo B | 100 | 30 | 30% |

* La tasa de aprobación en el Grupo A es del 60%, mientras que en el Grupo B es del 30%.
* Disparate Impact = 30% / 60% = 0.5 (50%) → Viola la Regla del 80%.
* El modelo de IA desfavorece al Grupo B en la aprobación de préstamos.
  + 1. Cumplimiento de estándares éticos y normativas legales

Indicador que evalúa si el modelo de IA cumple con la Ley de Protección de Datos Personales, las guías de ética en IA y otras regulaciones aplicables.

|  |  |  |
| --- | --- | --- |
| Categoría | Indicador | Descripción detallada |
| Estándares éticos | Transparencia | El sistema de IA debe poder explicar claramente qué datos utiliza y con qué criterios toma decisiones. |
| Equidad | La IA debe operar de manera justa, sin discriminación por género, raza, edad o región. |
| Responsabilidad | Debe haber una entidad responsable claramente definida en caso de errores o decisiones incorrectas del sistema de IA. |
| Protección de la privacidad | La IA debe proteger la información sensible de las personas y no recopilar ni almacenar datos sin autorización. |
| Normativas legales | Ley de Protección de Datos Personales | No debe violar la Ley de Protección de Datos Personales de Perú. |
| Estándares de Ética en IA | No debe infringir los estándares de ética en IA de Perú. |

<Tabla 7 - Indicadores de evaluación: Cumplimiento de estándares éticos y normativas legales>

* **Ejemplo de problema ético**
* Supongamos que la policía de Perú utiliza un sistema de reconocimiento facial basado en IA en espacios públicos.

|  |  |
| --- | --- |
| Función del sistema | Problema ético |
| Búsqueda automática de sospechosos | Posible discriminación por raza/edad |
| Análisis de patrones de movimiento de ciudadanos | Riesgo de violación de la privacidad |
| Envío automático de alertas por IA | Posible generación de falsas alarmas (False Alarm) |

* Existe la posibilidad de que la IA identifique injustamente a ciertas razas o géneros como sospechosos con mayor frecuencia.
* La recopilación de datos faciales sin el consentimiento de los individuos puede generar preocupaciones sobre la violación de la privacidad.
* **Ejemplo de problema legal**
* Supongamos que el gobierno de Perú utiliza un chatbot de IA para automatizar la atención de quejas ciudadanas.

|  |  |
| --- | --- |
| Función de la IA | Problema legal |
| Almacenamiento del nombre y número de teléfono del ciudadano | Posible violación de la Ley de Protección de Datos Personales de Perú |
| Uso del historial de consultas como datos de entrenamiento para la IA | Posible uso de datos sin el consentimiento del usuario |
| Almacenamiento de datos en servidores en la nube externos | Posible riesgo de transferencia de datos gubernamentales al extranjero |

* Existe la posibilidad de que la IA almacene de manera no autorizada la información personal de los ciudadanos (nombre, número de teléfono, etc.).
* Si los datos de consulta se utilizan para el entrenamiento de la IA sin el consentimiento del usuario, podría generar problemas legales.
* Si los datos recopilados por la IA se almacenan en servidores en la nube en el extranjero, podría haber una violación de la ley de protección de datos gubernamentales.
  1. Herramientas y marcos de evaluación del desempeño

Para evaluar el rendimiento de un modelo de IA, es fundamental utilizar diversas herramientas y marcos de evaluación. En particular, las instituciones públicas pueden aprovechar herramientas de código abierto que permitan evaluar la transparencia, equidad, fiabilidad y escalabilidad del modelo.

* + 1. Introducción a herramientas de evaluación de desempeño basadas en código abierto

A continuación, se presentan algunas de las principales herramientas de código abierto utilizadas para la evaluación cuantitativa y cualitativa del rendimiento de los modelos de IA.

* + - 1. Scikit-learn (Evaluación básica de modelos de IA)

Scikit-learn es una de las bibliotecas de código abierto más utilizadas para la evaluación de modelos de aprendizaje automático.

Ofrece diversos indicadores de evaluación, como precisión (Accuracy), precisión positiva (Precision), sensibilidad (Recall), F1-score y ROC-AUC. Además, es compatible con todos los principales algoritmos de aprendizaje automático, como XGBoost, Random Forest, SVM y regresión logística.

|  |  |
| --- | --- |
| Función principal | Descripción detallada |
| classification\_report() | Permite generar un informe con precisión (Precision), sensibilidad (Recall) y F1-score. |
| confusion\_matrix() | Permite visualizar los valores de TP, FP, TN y FN. |
| roc\_auc\_score() | Evalúa el rendimiento del modelo mediante la puntuación ROC-AUC. |

<Tabla 8 - Funcionalidades principales de Scikit-learn>

|  |
| --- |
| import numpy as np  import pandas as pd  import matplotlib.pyplot as plt  from sklearn.model\_selection import train\_test\_split  from sklearn.ensemble import RandomForestClassifier  from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix, classification\_report, roc\_auc\_score, roc\_curve  # 1. Creación de datos (Datos de ejemplo)  np.random.seed(42)  data\_size = 1000  X = np.random.rand(data\_size, 5) # 5 características aleatorias (ejemplo: embeddings de texto de quejas)  y = np.random.choice([0, 1], size=data\_size, p=[0.6, 0.4]) # 0: Medio ambiente, 1: Tráfico (Clasificación binaria)  # 2. División en datos de entrenamiento y prueba  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  # 3. Entrenamiento del modelo de IA (Clasificador Random Forest)  model = RandomForestClassifier(n\_estimators=100, random\_state=42)  model.fit(X\_train, y\_train)  # 4. Predicción del modelo  y\_pred = model.predict(X\_test)  y\_pred\_proba = model.predict\_proba(X\_test)[:, 1] # Probabilidades para evaluación ROC-AUC  # 5. Cálculo de indicadores de evaluación del desempeño  accuracy = accuracy\_score(y\_test, y\_pred)  precision = precision\_score(y\_test, y\_pred)  recall = recall\_score(y\_test, y\_pred)  f1 = f1\_score(y\_test, y\_pred)  roc\_auc = roc\_auc\_score(y\_test, y\_pred\_proba)  # 6. Impresión de resultados  print("Resultados de evaluación del desempeño del modelo de IA:")  print(f" Precisión (Accuracy): {accuracy:.4f}")  print(f" Precisión Positiva (Precision): {precision:.4f}")  print(f" Sensibilidad (Recall): {recall:.4f}")  print(f" Puntuación F1 (F1-score): {f1:.4f}")  print(f" ROC-AUC: {roc\_auc:.4f}")  # 7. Impresión de la matriz de confusión  print("\n Matriz de confusión (Confusion Matrix):")  print(confusion\_matrix(y\_test, y\_pred))  # 8. Impresión del informe de evaluación detallado  print("\n Informe de evaluación detallado (Classification Report):")  print(classification\_report(y\_test, y\_pred))  # 9. Visualización de la curva ROC  fpr, tpr, \_ = roc\_curve(y\_test, y\_pred\_proba)  plt.figure(figsize=(6, 6))  plt.plot(fpr, tpr, label=f'ROC curve (area = {roc\_auc:.2f})')  plt.plot([0, 1], [0, 1], linestyle='--', color='gray') # Línea diagonal de referencia  plt.xlabel('False Positive Rate')  plt.ylabel('True Positive Rate')  plt.title('ROC-AUC Curve')  plt.legend(loc='lower right')  plt.show() |

<Tabla 9 - Ejemplo de código fuente para la evaluación del desempeño utilizando Python>

* + - 1. Tensorflow Model Analysis (TFMA)

TFMA es una herramienta desarrollada por Google para la evaluación del desempeño y equidad de modelos de IA. Permite analizar errores de predicción, sesgos (Bias) y el equilibrio en los datos, además de realizar evaluaciones a gran escala con grandes conjuntos de datos.

Sus principales funciones son las siguientes:

|  |  |
| --- | --- |
| Función principal | Descripción detallada |
| Fairness Indicators | Permite detectar sesgos relacionados con género, edad, raza, entre otros. |
| Slice-based Analysis | Permite comparar el desempeño del modelo en grupos específicos (por ejemplo, mayores de 60 años). |
| Large-scale Evaluation | Permite el análisis del rendimiento del modelo de IA utilizando grandes volúmenes de datos (Big Data). |

<Tabla 10 - Funcionalidades principales de TFMA>

* + - 1. Fairness Indicators (Evaluación de equidad en IA)

Fairness Indicators es una herramienta de código abierto desarrollada por Google para evaluar la equidad en modelos de IA. Permite medir indicadores como Demographic Parity, Equalized Odds y Disparate Impact. Además, detecta posibles casos de trato desigual por parte del modelo de IA hacia grupos específicos según género, edad o raza.

Las principales funciones son las siguientes:

|  |  |
| --- | --- |
| Función principal | Descripción detallada |
| Evaluación de equidad | Analiza si el modelo de IA discrimina a determinados grupos (género, edad, raza). |
| Detección de sesgo (Bias Detection) | Evalúa si el modelo de IA ha aprendido con datos sesgados. |
| Visualización de equidad (Fairness Visualization) | Analiza visualmente la distribución de datos y las predicciones del modelo de IA. |

<Tabla 11 - Funcionalidades principales de Fairness Indicators>

* + - 1. SHAP (Sharpley Additive Explanations)

SHAP es una herramienta que ayuda a explicar por qué un modelo de IA toma una determinada decisión.

Utilizando los valores SHAP, permite cuantificar los factores que influyen en cada predicción individual, siendo una herramienta útil para garantizar la transparencia en los servicios públicos basados en IA.

Las principales funciones son las siguientes:

|  |  |
| --- | --- |
| Función principal | Descripción detallada |
| Feature Importance | Analiza cómo el modelo de IA utiliza las variables más importantes (ingresos, edad, etc.). |
| Local Explanations | Examina los factores que influyeron en cada predicción individual (aprobación/rechazo de un préstamo). |
| Global Model Analysis | Evalúa los criterios generales con los que opera el modelo de IA. |

<Tabla 12 - Funcionalidades principales de SHAP>

* + - 1. LIME (Local Interpretable Model-agnostic Explanations)

LIME es una herramienta que explica cómo un modelo de IA toma decisiones en predicciones individuales. Analiza cómo reacciona el modelo al realizar pequeños cambios en una muestra específica. Además, permite explicar las razones detrás de decisiones como diagnósticos médicos, aprobaciones de préstamos, entre otras.

Las principales funciones son las siguientes:

|  |  |
| --- | --- |
| Función principal | Descripción detallada |
| Local Explanation | Analiza la base de la toma de decisiones de la IA para una muestra de datos específica. |
| Feature Contribution | Examina qué variables (palabras clave, características de imágenes, etc.) utilizó la IA para hacer una predicción. |
| Model Agnostic | Es compatible con todos los modelos de aprendizaje automático y aprendizaje profundo. |

<Tabla 13 - Funcionalidades principales de LIME>

* + - 1. Apache JMeter (Pruebas de velocidad de procesamiento y tiempo de respuesta de IA)

Apache JMeter es una herramienta de código abierto para evaluar el rendimiento de los servicios de IA, incluyendo velocidad de procesamiento y tiempo de respuesta. Permite medir la rapidez con la que un modelo de IA procesa datos en tiempo real. Además, ofrece funciones para optimizar el rendimiento de los servicios de IA en entornos en la nube y on-premise.

Las principales funciones son las siguientes:

|  |  |
| --- | --- |
| Función principal | Descripción detallada |
| Load Testing | Prueba la estabilidad del sistema de IA bajo alto tráfico. |
| Response Time Analysis | Mide la velocidad de predicción y el tiempo de respuesta de la IA. |
| Scalability Testing | Evalúa si el modelo de IA mantiene un rendimiento estable a medida que aumenta la cantidad de datos. |

<Tabla 14 - Funcionalidades principales de Apache JMeter>

* + - 1. Comparación de herramientas de evaluación del desempeño de IA

A continuación, se presenta una comparación de las principales herramientas de evaluación del desempeño de IA.

|  |  |  |
| --- | --- | --- |
| Herramienta principal | Funcionalidades clave | Ámbito de aplicación |
| Scikit-learn | Evaluación básica del desempeño del modelo de IA | Pruebas de desempeño en modelos de clasificación y regresión |
| TFMA | Análisis de equidad y desempeño del modelo de IA | Evaluación de sistemas de IA a gran escala |
| Fairness Indicators | Evaluación de equidad en IA | Detección de discriminación y sesgo |
| SHAP | Análisis de explicabilidad del modelo de IA | Interpretación de IA en préstamos y asistencia social |
| LIME | Explicación de predicciones individuales | Interpretación de IA en diagnóstico médico y clasificación de quejas |
| Apache JMeter | Pruebas de velocidad y escalabilidad de los servicios de IA | Evaluación de IA en chatbots y predicción del tráfico |

<Tabla 15 - Comparación de herramientas de evaluación del desempeño de IA>

1. Evaluación del desempeño de la IA por sector público
   1. Caso de evaluación del desempeño de la IA en la gestión de quejas y reclamos
      1. Descripción general

El gobierno de Perú ha implementado un sistema basado en IA para la recepción y gestión automatizada de quejas ciudadanas. La IA analiza las quejas recibidas y las clasifica automáticamente en los departamentos correspondientes, optimizando la eficiencia operativa.

* + 1. Principales indicadores de evaluación del desempeño

|  |  |  |
| --- | --- | --- |
| Indicador de evaluación | Significado | Objetivo |
| Precisión (Accuracy) | Proporción de quejas correctamente clasificadas por la IA | ≥ 85% |
| Precisión Positiva (Precision) | Proporción de quejas clasificadas en una categoría específica que realmente pertenecen a ella | ≥ 80% |
| |  | | --- | | Sensibilidad (Recall) |  |  | | --- | |  | | Proporción de quejas de una categoría específica que la IA identificó correctamente | ≥ 75% |
| |  | | --- | | Tiempo de respuesta (Latency) |  |  | | --- | |  | | Tiempo promedio que tarda la IA en analizar una queja | ≤ 2 segundos |

* + 1. Método de evaluación del desempeño
       1. Uso de datos de prueba
* Se utilizaron 10,000 registros de quejas ciudadanas recopiladas en Lima como conjunto de datos de entrenamiento.
* Los datos incluyen cinco categorías: tráfico, medio ambiente, salud, problemas de infraestructura pública, entre otros.
  + - 1. Análisis de los resultados de la evaluación del desempeño
* Precisión (Accuracy): 83.5%
* Precisión Positiva (Precision): 81.2%
* Sensibilidad (Recall): 76.8%
* Tiempo de respuesta (Latency): Promedio de 1.8 segundos
  + 1. Dirección de mejora
* Analizar los datos mal clasificados y realizar un entrenamiento adicional con datos corregidos.
* Mejorar el procesamiento de nuevos tipos de quejas mediante técnicas de aumento de datos.
  1. Caso de evaluación del desempeño de la IA para la clasificación automática de documentos administrativos
     1. Descripción general

Las instituciones públicas de Perú han implementado un sistema basado en IA para la clasificación y almacenamiento automático de documentos administrativos. Este sistema mejora la eficiencia operativa al clasificar automáticamente documentos legales, informes y materiales de políticas públicas.

* + 1. Principales indicadores de evaluación del desempeño

|  |  |  |
| --- | --- | --- |
| Indicador de evaluación | Significado | Objetivo |
| Precisión (Accuracy) | Proporción de documentos correctamente clasificados por la IA | ≥ 90% |
| Tasa de error (Error Rate) | Proporción de documentos clasificados incorrectamente | ≤ 5% |
| Tiempo de procesamiento (Processing Time) | Tiempo necesario para clasificar un documento | ≤ 1 segundo |

* + 1. Método de evaluación del desempeño
       1. Uso de datos de prueba
* Se utilizó un conjunto de datos de 50,000 documentos administrativos recopilados de instituciones públicas.
* Los documentos se clasificaron en tres categorías: documentos legales (30%), informes administrativos (40%) y documentos relacionados con servicios ciudadanos (30%).
  + - 1. Resultados de la evaluación del desempeño de la IA
* Precisión (Accuracy): 92.1%
* Tasa de error (Error Rate): 4.2%
* Tiempo de procesamiento (Processing Time): Promedio de 0.85 segundos
  + 1. Dirección de mejora
* Los errores ocurren cuando la distinción entre documentos legales y documentos de servicios ciudadanos no es clara → Es necesario mejorar el etiquetado de datos.
* Mejorar la tecnología OCR para aumentar la precisión en el reconocimiento de documentos escaneados.
  1. Caso de evaluación del desempeño de chatbots basados en IA
     1. Descripción general

El gobierno de Perú ha implementado un chatbot basado en IA para automatizar la atención de quejas de servicios públicos y consultas legales. El chatbot es capaz de comprender las preguntas de los ciudadanos y proporcionar respuestas adecuadas.

* + 1. Principales indicadores de evaluación del desempeño

|  |  |  |
| --- | --- | --- |
| Indicador de evaluación | Significado | Objetivo |
| Precisión de respuesta (Answer Accuracy) | Proporción de respuestas correctas proporcionadas por la IA | ≥ 85% |
| Tiempo de respuesta (Response Time) | Tiempo promedio que tarda la IA en responder a una pregunta | ≤ 1 segundo |
| Satisfacción del usuario (User Satisfaction) | Evaluación de la satisfacción de los usuarios del chatbot | ≥ 80% |

* + 1. Método de evaluación del desempeño
       1. Uso de datos de prueba
* Se realizó la prueba con 10,000 registros de consultas reales de instituciones públicas.
* Incluye consultas generales (40%), asesoramiento legal (30%) y consultas sobre asistencia social (30%).
  + - 1. Resultados de la evaluación del desempeño de la IA
* Precisión de respuesta (Answer Accuracy): 82.7%
* Tiempo de respuesta (Response Time): Promedio de 0.9 segundos
* Satisfacción del usuario (User Satisfaction): 78%
  + 1. Dirección de mejora
* Rendimiento reducido en consultas complejas como asesoramiento legal → Se requiere una actualización a un modelo basado en GPT.
* Incorporar retroalimentación de los usuarios y realizar un entrenamiento adicional con datos mejorados para optimizar las respuestas del chatbot.
  1. Caso de evaluación del desempeño de la IA para el análisis de videos e imágenes
     1. Descripción general

La Policía de Perú y las agencias de seguridad pública han implementado un sistema de análisis de videos mediante IA. Este sistema detecta comportamientos inusuales en espacios públicos y ayuda en la prevención del crimen.

* + 1. Principales indicadores de evaluación del desempeño

|  |  |  |
| --- | --- | --- |
| Indicador de evaluación | Significado | Objetivo |
| Precisión en la detección de objetos (Object Detection Accuracy) | Proporción de personas, vehículos u objetos correctamente detectados por la IA | ≥ 90% |
| Tasa de falsos positivos (False Positive Rate) | Proporción de detecciones incorrectas donde la IA clasifica comportamientos normales como anómalos | ≤ 10% |
| Tiempo de inferencia (Inference Time) | Tiempo requerido para analizar un solo fotograma de video | ≤ 0.5 segundos |

* + 1. Método de evaluación del desempeño
       1. Uso de datos de prueba
* Se utilizaron 100,000 grabaciones de CCTV capturadas en la ciudad de Lima, Perú, como datos de entrenamiento y evaluación para la IA.
* Incluye monitoreo en tiempo real del tráfico y detección de comportamientos inusuales en espacios públicos.
  + - 1. Resultados de la evaluación del desempeño de la IA
* Precisión en la detección de objetos (Object Detection Accuracy): 88.5%
* Tasa de falsos positivos (False Positive Rate): 12.3%
* Tiempo de inferencia (Inference Time): Promedio de 0.48 segundos
  + 1. Dirección de mejora
* Análisis de las causas de la alta tasa de falsos positivos → Se requiere entrenamiento adicional considerando factores como el clima (lluvia, baja iluminación).
* Optimización del modelo para detección en tiempo real mediante técnicas de reducción de peso del modelo y la implementación de Edge AI.

1. Estrategia para la implementación del proceso de evaluación del desempeño de la IA

Para que el gobierno de Perú implemente la IA de manera efectiva en los servicios públicos y garantice su fiabilidad, es necesario establecer un proceso sistemático de evaluación del desempeño de la IA. Se debe desarrollar un marco integral que evalúe precisión, equidad, velocidad de procesamiento, escalabilidad y cumplimiento de estándares éticos.

* 1. Establecimiento del sistema de evaluación del desempeño de la IA en instituciones públicas
     1. Descripción general

El gobierno de Perú debe establecer un sistema integral de evaluación del desempeño de la IA en los principales sectores donde se implementa la tecnología, como gestión de quejas, automatización de documentos, chatbots y análisis de videos.

* + 1. Estrategia para la estructuración del sistema de evaluación del desempeño de la IA

|  |  |  |
| --- | --- | --- |
| Fase | Descripción detallada | Institución responsable |
| Elaboración de directrices para la evaluación del desempeño de la IA | Definir los criterios de evaluación (precisión, velocidad de procesamiento, equidad, etc.) | PCM (Presidencia del Consejo de Ministros) |
| Establecimiento de procedimientos de evaluación por sector | Implementar criterios de evaluación adaptados a las características de cada institución | Ministerios correspondientes (Transporte, Salud, etc.) |
| Creación de un sistema de auditoría de IA (AI Audit) | Supervisar que la IA funcione de manera ética y transparente | Comité de Ética en IA |
| Desarrollo de conjuntos de datos y benchmarks para IA pública | Gestionar datasets para la evaluación del desempeño de la IA en el sector público | Instituto Nacional de Estadística e Informática (INEI), PCM |

<Tabla 16 - Ejemplo de estructura del sistema de evaluación del desempeño de la IA>

* 1. Establecimiento de los criterios y procedimientos de evaluación del desempeño de la IA
     1. Descripción general

Se deben establecer criterios y procedimientos comunes de evaluación del desempeño de la IA en las instituciones públicas. Además, es fundamental definir métodos de evaluación claros para garantizar que los sistemas de IA mantengan su fiabilidad de manera continua.

* + 1. Criterios de evaluación del desempeño de la IA

|  |  |  |
| --- | --- | --- |
| Categoría de evaluación | Criterios de evaluación | Casos de aplicación |
| Precisión (Accuracy) | El modelo de IA debe mantener una precisión mínima del 85% | Clasificación automática de quejas basada en IA |
| Tiempo de respuesta (Latency) | El tiempo medio de respuesta de la IA debe ser ≤ 1 segundo | Chatbots basados en IA |
| Equidad (Fairness) | La IA debe operar sin discriminación por género, edad o región | Evaluación de crédito basada en IA |
| Explicabilidad (Explainability) | La IA debe ser capaz de justificar sus decisiones | Selección de beneficiarios de asistencia social basada en IA |
| Estabilidad y escalabilidad (Scalability) | La IA debe operar de manera estable incluso bajo alto tráfico | Sistema de predicción de tráfico basado en IA |

<Tabla 17 - Ejemplo de criterios de evaluación del desempeño de la IA>

* + 1. Procedimiento de evaluación del desempeño de la IA

|  |  |  |
| --- | --- | --- |
| Fase | Descripción detallada | Herramientas de evaluación principales |
| Revisión de datos (Data Review) | Evaluación de la calidad de los datos utilizados en el entrenamiento del modelo de IA | Scikit-learn, Pandas |
| Evaluación del desempeño del modelo (Model Performance Evaluation) | Medición del rendimiento cuantitativo (precisión, velocidad) | Scikit-learn, TensorFlow Model Analysis (TFMA) |
| Análisis de equidad y ética (Fairness & Bias Analysis) | Evaluación de sesgos en el modelo de IA | Fairness Indicators, SHAP |
| Prueba en entorno real (Real-World Testing) | Validación del rendimiento de la IA en un entorno de servicio público | Apache JMeter (prueba de velocidad de procesamiento) |
| Optimización y aplicación de mejoras (Iteration & Improvement) | Implementación de mejoras para optimizar el rendimiento | LIME, Bias Mitigation Tools |

<Tabla 18 - Ejemplo de procedimiento de evaluación del desempeño de la IA en el sector público>

* 1. Estrategia para la utilización de los resultados de la evaluación del desempeño
     1. Descripción general

Los resultados de la evaluación del desempeño de la IA deben utilizarse activamente para la toma de decisiones políticas, la mejora de los servicios públicos y el fortalecimiento de la confianza ciudadana.

* + 1. Estrategia para la utilización de los resultados de la evaluación del desempeño de la IA

La evaluación del desempeño de la IA puede optimizar los servicios públicos dirigidos a los ciudadanos peruanos y respaldar la toma de decisiones en políticas públicas.

|  |  |
| --- | --- |
| Ámbito de aplicación | Estrategia de utilización |
| Optimización de servicios públicos | Mejorar continuamente los sistemas de IA basándose en los resultados de la evaluación del desempeño. |
| Apoyo a la toma de decisiones | Utilizar los datos de evaluación de IA para formular políticas nacionales de IA (ejemplo: mejora de políticas de predicción del tráfico basada en IA). |
| Fortalecimiento de la confianza ciudadana | Publicar los resultados de la evaluación de IA en instituciones públicas para aumentar la transparencia. |
| Reentrenamiento y mejora del modelo de IA | Aplicar los resultados de la evaluación para eliminar sesgos y optimizar el rendimiento del modelo de IA. |

<Tabla 19 - Ejemplo de estrategias para la utilización de los resultados de la evaluación del desempeño de la IA>

* 1. Estrategia para el monitoreo continuo del desempeño de la IA y su mejora
     1. Descripción general

Se debe establecer un proceso continuo de monitoreo y mejora del desempeño de los sistemas de IA para evitar la degradación de su rendimiento con el tiempo. Además, es fundamental implementar un sistema de detección de Concept Drift y actualización automática del modelo para garantizar su precisión y relevancia en diferentes condiciones.

* + 1. Procedimiento de monitoreo continuo del desempeño de la IA

El desempeño de los sistemas de IA debe ser monitoreado de manera continua durante su operación. Dado que los modelos de IA aprenden y procesan nuevos datos constantemente, su rendimiento puede variar con el tiempo, lo que hace necesario un seguimiento sistemático para detectar y corregir posibles degradaciones.

|  |  |  |
| --- | --- | --- |
| Fase | Descripción detallada | Herramientas de evaluación principales |
| Monitoreo en tiempo real (Real-time Monitoring) | Supervisar en tiempo real los datos procesados por la IA en los servicios públicos | Prometheus, TensorFlow Model Analysis |
| Detección de cambios en el rendimiento (Concept Drift Detection) | Analizar si los cambios en los patrones de datos afectan el desempeño de la IA | Evidently AI, DataDrift Detection |
| Reentrenamiento y optimización del modelo (Re-training & Optimization) | Actualizar el modelo de IA incorporando los datos más recientes | AutoML, MLOps Pipeline |
| Incorporación de retroalimentación (Feedback Integration) | Mejorar el desempeño de la IA con el feedback de ciudadanos y funcionarios públicos | Encuestas a usuarios |

<Tabla 20 - Ejemplo de monitoreo del desempeño de la IA>

**Referencias**

Solaiman, I. (2023). The gradient of generative AI release: Methods and considerations. Hugging Face. https://arxiv.org/abs/2302.04844

Luccioni, S., & Mitchell, M. (2022, October 24). Evaluating language model bias with Evaluate. Hugging Face. https://huggingface.co/blog/evaluating-llm-bias

Kim, D. (2018). BLEU Score. https://donghwa-kim.github.io/BLEU.html

Google Developer Program. ML Concepts: Classification: ROC and AUC. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>

Evaluating language competence of Llama 2-based models: The BLEU Score. Medium. https://medium.com/@geronimo7/evaluating-language-competence-of-llama-2-based-models-the-bleu-score-d44c651a5e58

Park, Sangho. BLEU del conjunto de datos de chatbots en coreano. https://github.com/sxngho/ABChat

Kim, Junseok (2017). Rendimiento y latencia de Microsoft Azure. Naver Papago.

Scikit-learn developers. Plot precision-recall. https://scikit-learn.org/1.5/auto\_examples/model\_selection/plot\_precision\_recall.html

Hiroki Nakagami et al. Rendimiento y tiempo de latencia de Microsoft Azure. <https://learn.microsoft.com/ko-kr/azure/ai-services/openai/how-to/latency>

Asociación de Tecnología de la Información y Comunicación de Corea (TTA). Guía para el desarrollo de inteligencia artificial confiable. Asociación de Tecnología de la Información y las Comunicaciones de Corea