





Documentación del proceso de aprendizaje automático

<div>Tarea de Aprendizaje<div>?</div></div> <div><ul style="list-style-type: none">Tipo de aprendizaje:<ul style="list-style-type: none">Aprendizaje supervisado.¿Qué se predice?<ul style="list-style-type: none">Se predice si una noticia es falsa o real.Posibles resultados:<ul style="list-style-type: none">Clase 1: Noticia real.Clase 0: Noticia falsa.¿Cuándo se observan los resultados?<ul style="list-style-type: none">Se obtienen instantáneamente tras el procesamiento del texto (segundos o milisegundos).</div>	<div>Decisiones<div>↔</div></div> <div><ul style="list-style-type: none">Uso de los resultados:<ul style="list-style-type: none">Alertar a los usuarios o verificadores de noticias sobre contenido potencialment e falso.Filtrar automáticamente contenido falso en plataformas de noticias.Ayudar a periodistas a verificar información antes de su publicación.Acciones procesables:<ul style="list-style-type: none">Etiquetado de noticias con advertencias sobre confiabilidad.Generación de reportes sobre tendencias de noticias falsas.Bloqueo de noticias falsas en plataformas automatizadas.</div>	<div>Propuesta de Valor<div>🎁</div></div> <div><ul style="list-style-type: none">Beneficiario final:<ul style="list-style-type: none">Plataformas de noticias, verificadores de información, usuarios de redes sociales, entidades gubernamental es.Empresa / Institución:<ul style="list-style-type: none">Proyecto académico en detección de noticias falsas.Problemas abordados:<ul style="list-style-type: none">Desinformación política.Manipulación de la opinión pública.Impacto en elecciones y políticas públicas.Riesgos para el beneficiario:<ul style="list-style-type: none">Falsos positivos: Noticias reales clasificadas como falsas pueden afectar la credibilidad de medios legítimos.</div>	<div>Recolección de datos<div>⬇</div></div> <div></div>	<div>Fuentes de datos<div>🗄</div></div> <div><ul style="list-style-type: none">Fuente principal:<ul style="list-style-type: none">Noticias de periódicos en línea como <i>Público</i>, <i>La Marea</i>, <i>El Común</i>.Fuente secundaria:<ul style="list-style-type: none">Bases de datos con noticias etiquetadas como reales o falsas.Método de obtención:<ul style="list-style-type: none">Descarga de datos estructurados en formato CSV.Web Scraping (en caso de futuras expansiones).¿Se pueden usar estos datos para el análisis?<ul style="list-style-type: none">Sí, ya que están etiquetados y listos para entrenamiento.</div>
---	---	--	---	--

		<ul style="list-style-type: none"> ○ Falsos negativos: Noticias falsas no detectadas pueden seguir desinformando. 		
<p>Simulación de Impacto </p> <p>Costo/beneficio de decisiones correctas e incorrectas:</p> <ul style="list-style-type: none"> • Decisión correcta (noticia real detectada como real / falsa detectada como falsa): <ul style="list-style-type: none"> ○ Evita la propagación de desinformación. ○ Mejora la confianza en los medios de comunicación. • Decisión incorrecta (falsos positivos / falsos negativos): <ul style="list-style-type: none"> ○ Puede generar censura indebida. ○ Puede permitir la difusión de noticias falsas. • Criterios de éxito: <ul style="list-style-type: none"> ○ Precisión y F1-score superiores al 85%. ○ Reducción de la desinformación en 	<p>Aprendizaje (Uso del modelo) </p> <ul style="list-style-type: none"> • ¿Es por lotes o en tiempo real? <ul style="list-style-type: none"> ○ Puede usarse en tiempo real para detectar noticias falsas al momento de su publicación. ○ También puede ejecutarse en lotes para analizar grandes volúmenes de noticias. • Frecuencia de uso: <ul style="list-style-type: none"> ○ Continuo (cada vez que se publica una noticia). 		<p>Construcción de modelos </p> <p>¿Cuántos modelos se necesitan en producción?</p> <ul style="list-style-type: none"> • Dos modelos activos (Random Forest + Desicion Tree o Regresión Logística) y uno en pruebas continuas. <p>¿Cuándo deben actualizarse?</p> <ul style="list-style-type: none"> • Cada 3-6 meses o cuando el desempeño baje significativamen te. <p>¿Cuánto tiempo está disponible para esto?</p>	<p>Ingeniería de Características </p> <p>¿Qué representaciones se usan para las entidades en tiempo de predicción?</p> <ul style="list-style-type: none"> • TF-IDF, Word Embeddings (Word2Vec, FastText, BERT) y N-gramas. <p>¿Qué transformaciones se aplican a los datos?</p> <ul style="list-style-type: none"> • Limpieza de texto, normalización, vectorización y extracción de características lingüísticas.

<p>plataformas de noticias.</p> <ul style="list-style-type: none"> • Restricciones de equidad: <ul style="list-style-type: none"> ○ Evitar sesgos políticos o ideológicos en el modelo. ○ Transparencia en los criterios de clasificación. 			<ul style="list-style-type: none"> • 4-6 horas por ciclo (featurización, entrenamiento y validación). <p>¿Qué recursos computacionales se usan?</p> <ul style="list-style-type: none"> • Infraestructura en la nube (Google Cloud, AWS) o local con GPU y 16GB+ de RAM. 	
	<p>Monitoreo</p> 