CIENCIA DE DATOS EN EL INE

Conferencias Ciudadanas

Unidad de Gobierno de Datos

Julio 2025

Presentación

Speaker

Ignacio Agloni

Jefe Unidad de Gobierno y Ciencia de Datos - Instituto Nacional de Estadísticas

Magister (c) en Tecnologías de la Información - Universidad de Chile

Sociólogo - Universidad de Chile

ifaglonij@ine.gob.cl

https://github.com/ignacioagloni

Presentación

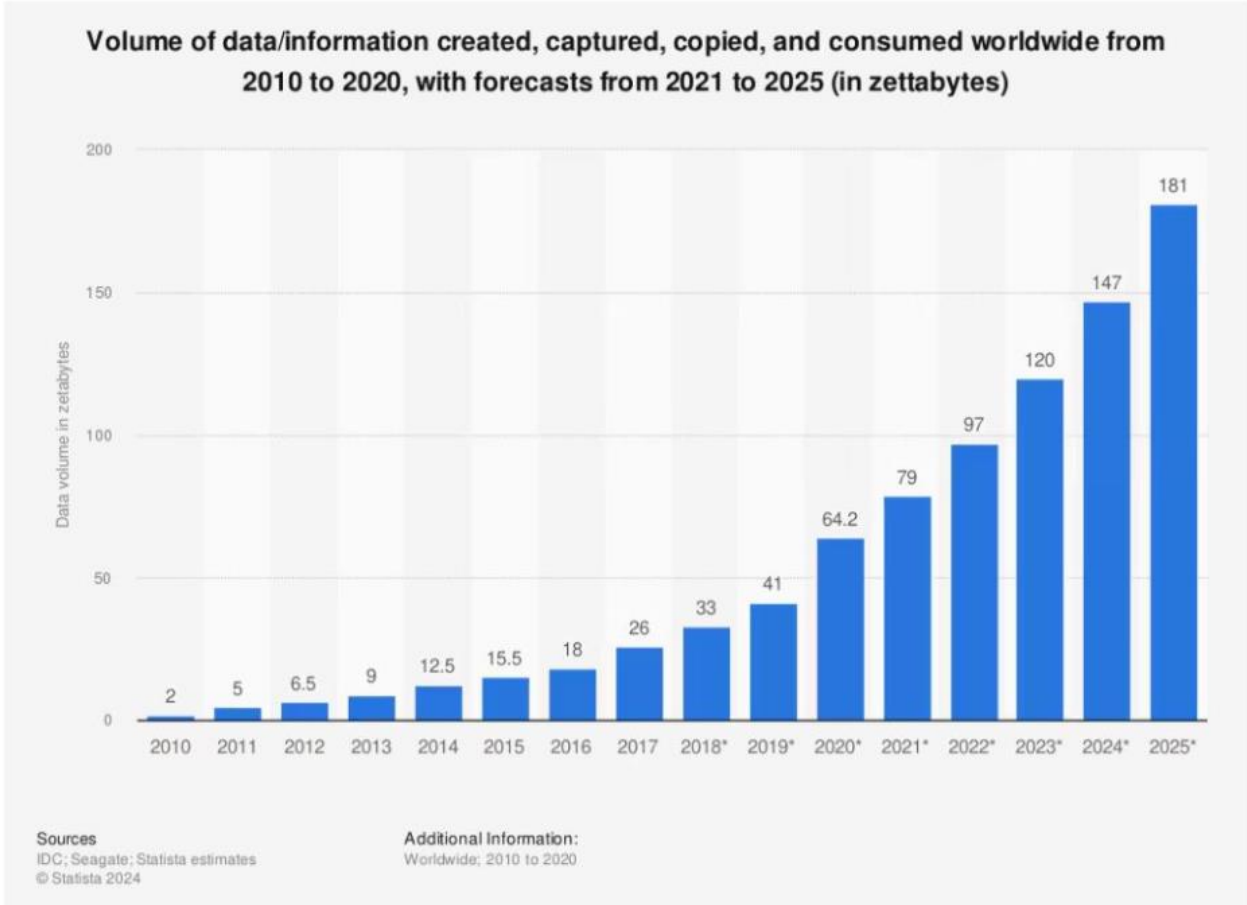
Link a presentación en formato HTML en Github

Temas a revisar

- *Contexto*
- *Ciencia de datos en el INE*
- *Organización del equipo*
- *Procesamiento de lenguaje natural*
- *Aprovechamiento de imágenes satelitales*
- *Visión computacional*
- *Otros desarrollos*
- *AI-Readiness en las ONE*
- *Reflexiones finales*

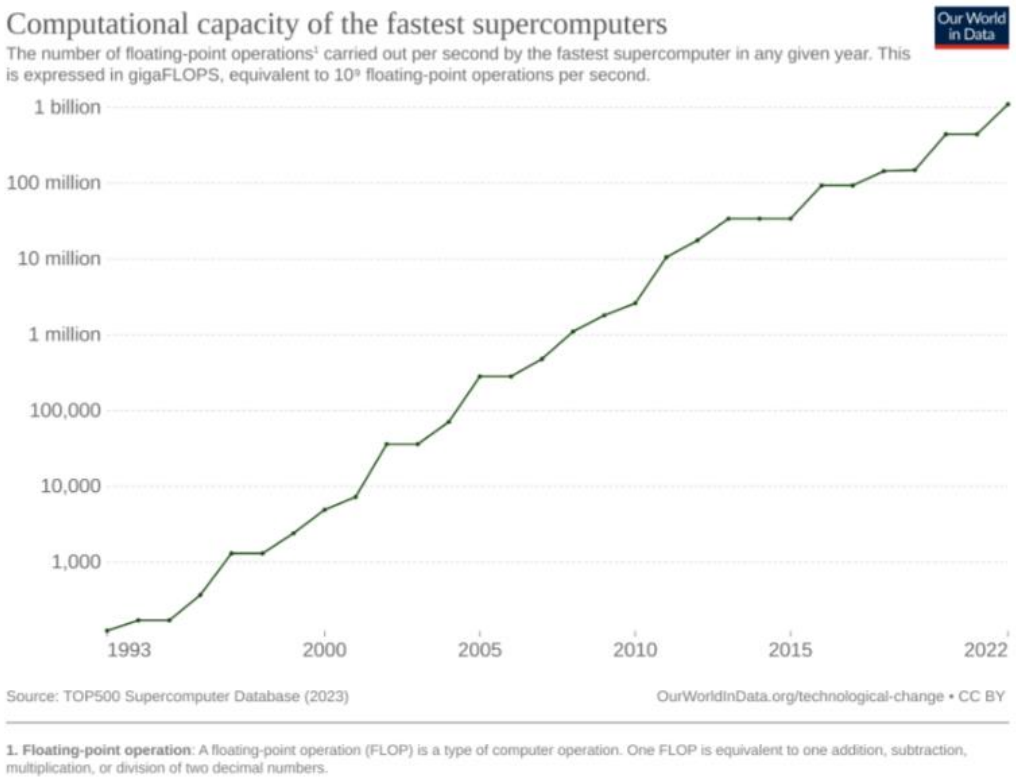
Contexto

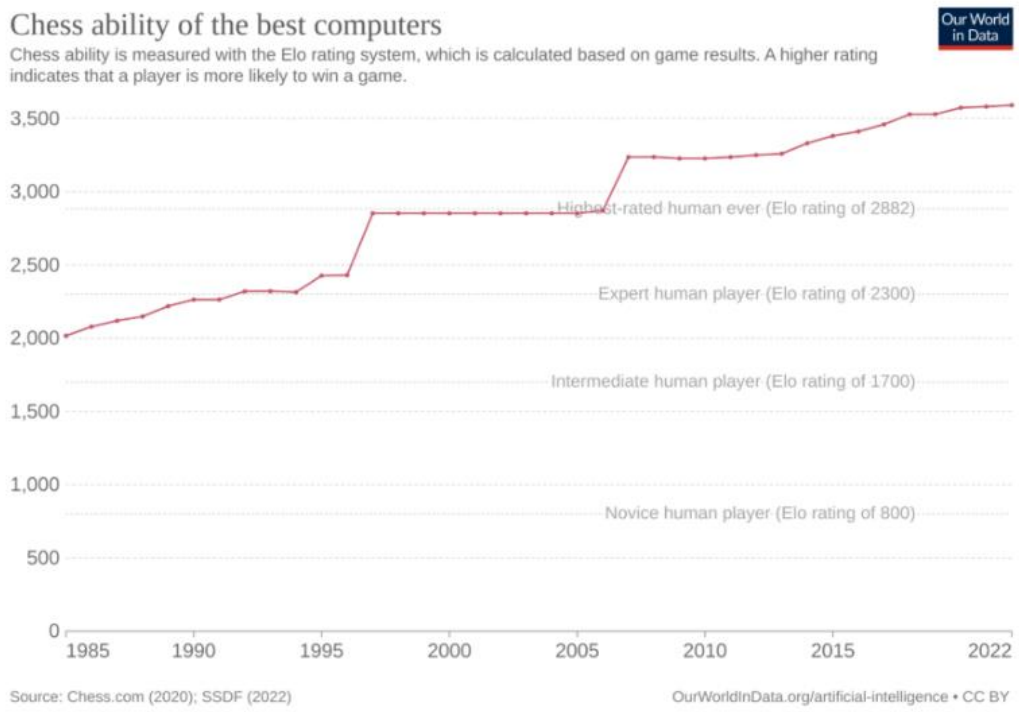
Cantidad de datos creados al año



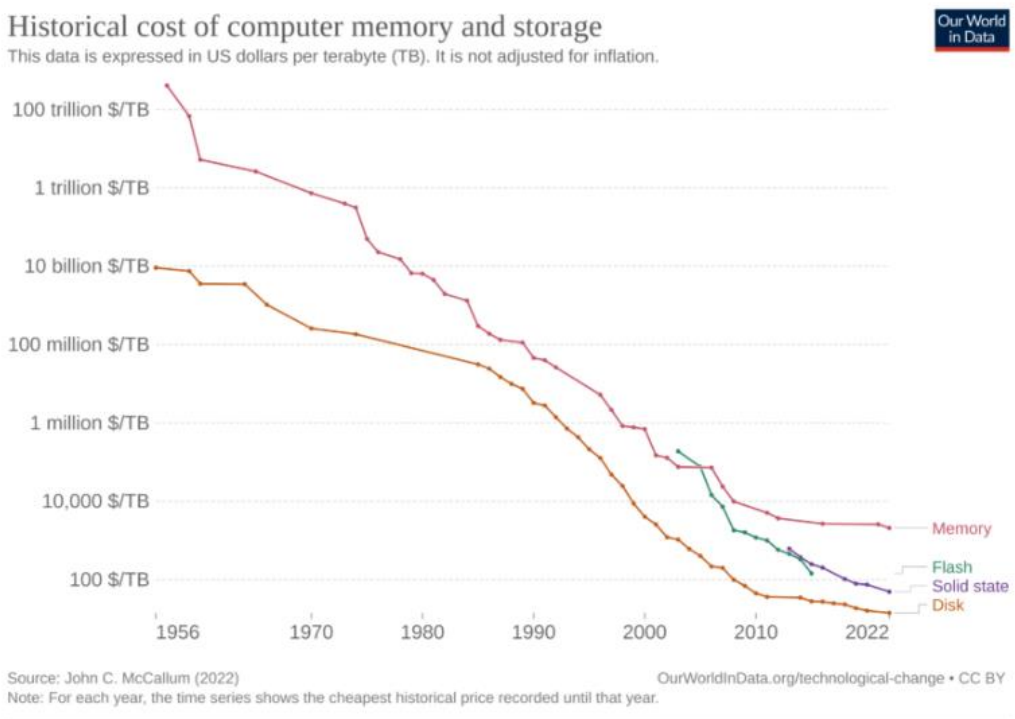
$$zettabyte = 10^{12} gb \approx 88.888.888.888 \text{ horas de video en 4k}$$

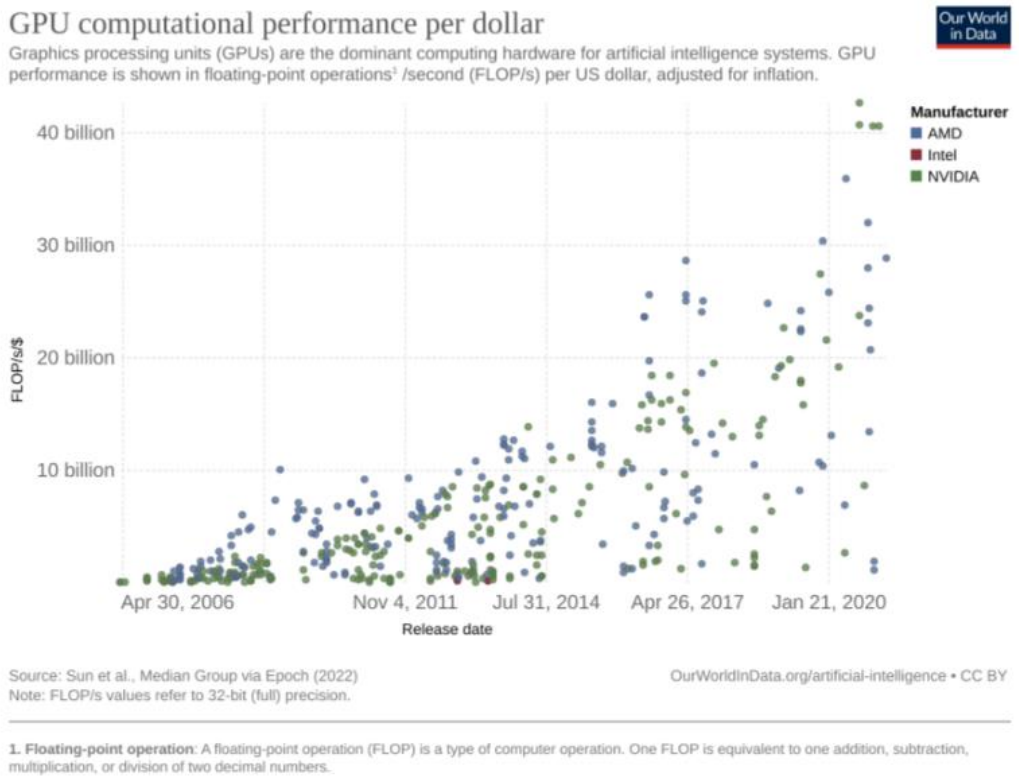
Capacidad de cómputo





Abaratamiento de la tecnología





Escenario de…

- Digitalización de la comunicación
- Aumento en la capacidad de almacenamiento y procesammiento
- Abaratamiento del *hardware*

Se crean, procesan y almacenan datos en todo momento

Empresas que manejan muchos datos

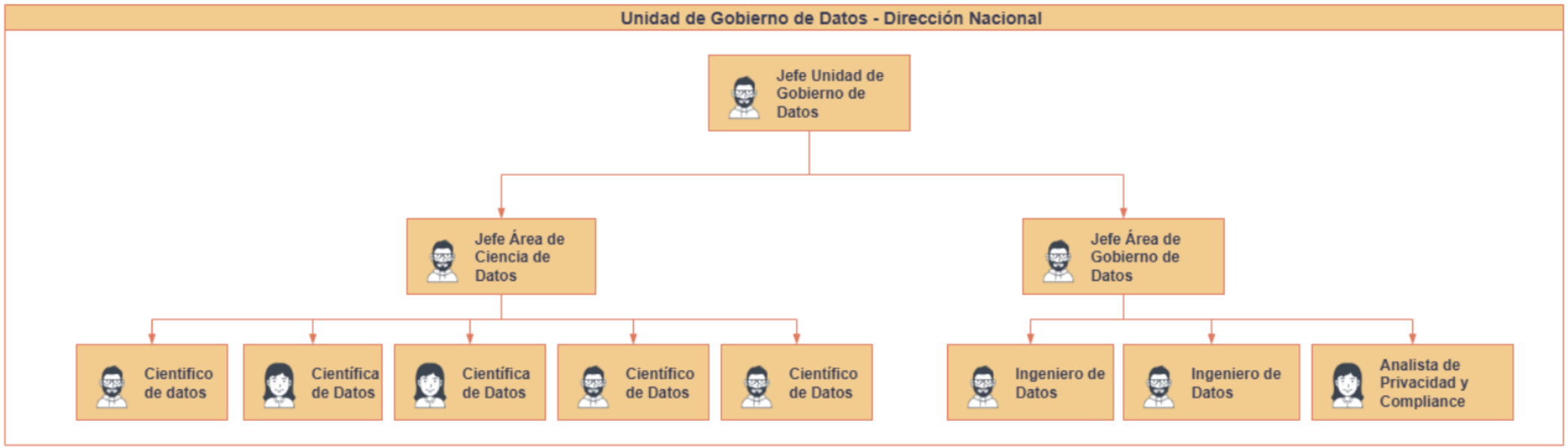
Comisión de Estadística de Naciones Unidas

Llamado de la Comisión de Estadística en su 56° periodo de sesiones

- ✔ Mantener la **colaboración internacional** en IA y ciencia de datos
- ✔ Fortalecer el uso de **fuentes alternativas de datos**
- ✔ Consolidar la **preparación para la IA** en ámbitos como **interoperabilidad**, **metadatos** y uso de **principios FAIR** (datos fáciles de encontrar, accesibles, interoperables y reutilizables)
- ✔ Solicitó el **establecimiento urgente de un equipo de tareas** sobre el uso de **grandes modelos lingüísticos**

Ciencia de datos e IA en el INE

Organización del equipo de ciencia de datos



Principales proyectos y aplicaciones

Procesamiento de lenguaje natural

Cod automática | CIUO y CAENES | Clasificación delitos | Data drift

b1_1 ¿Cuál es el oficio, labor u ocupación que [nombre de persona] realizó la semana pasada en su actividad principal?

☐ 1 Analista socioeconómico

☐ 88 No sabe
☐ 99 No responde

b1_2 ¿Qué tareas realizó en esta ocupación?

☐ 1 Procesar y analizar datos de encuestas

☐ 88 No sabe
☐ 99 No responde

Procesamiento de lenguaje natural

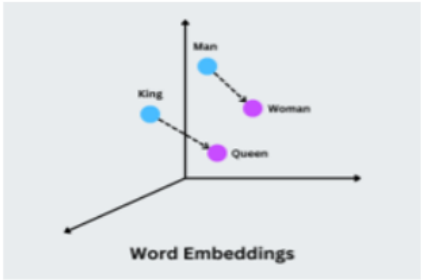
Instituto Nacional de Estadísticas · Chile

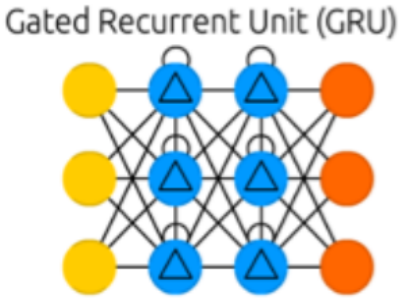
Cod automática | CIUO y CAENES | Clasificación delitos | Data drift

CIUO-08: es el Clasificador Internacional Uniforme de Ocupaciones
CAENES: es el Clasificador de Actividades Económicas Nacional para Encuestas Sociodemográficas

Modelo desarrollado para la **codificación de ocupación y actividad económica** en diversas operaciones estadísticas basadas en encuestas.

Capa de embeddings + **GRU** (Gated Recurrent Unit)





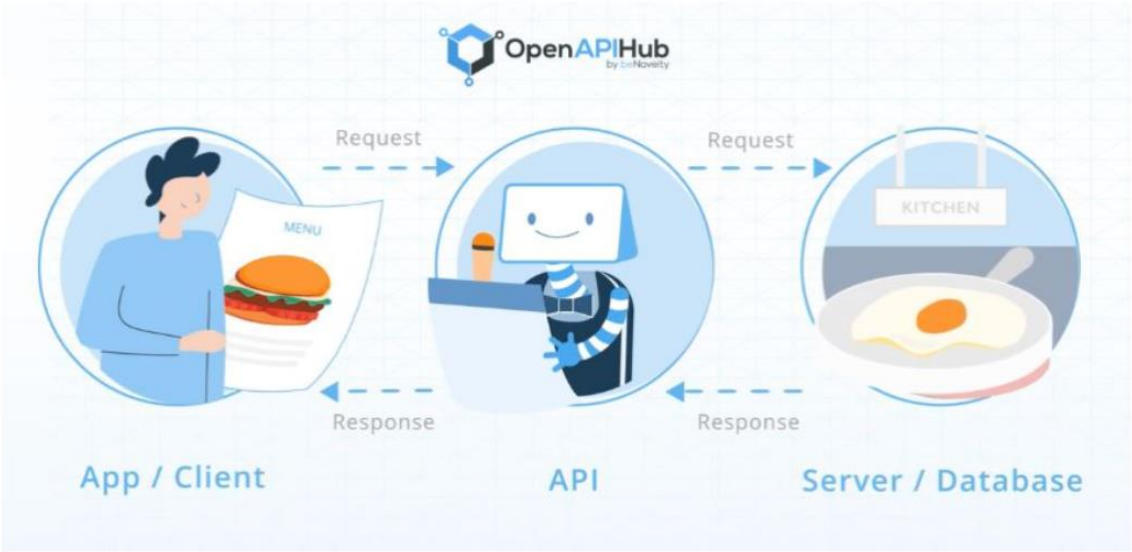
Instituto Nacional de Estadísticas · Chile

Procesamiento de lenguaje natural

Instituto Nacional de Estadísticas · Chile

Cod automática | CIUO y CAENES | Clasificación delitos | Data drift

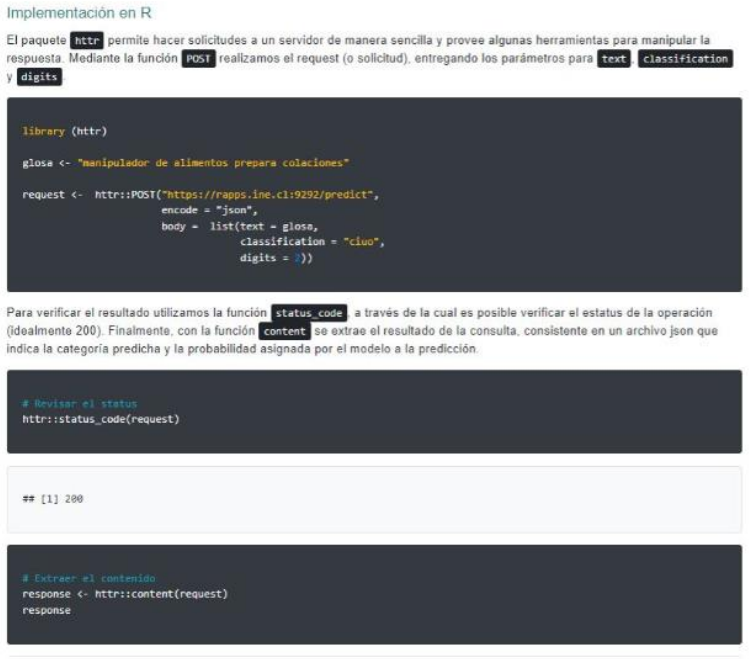
¿Cómo podemos ofrecer un **servicio** de codificación automática?



Puedes acceder al tutorial de uso de esta API a través de este enlace

Procesamiento de lenguaje natural

Cod automática | CIUO y CAENES | Clasificación delitos | Data drift



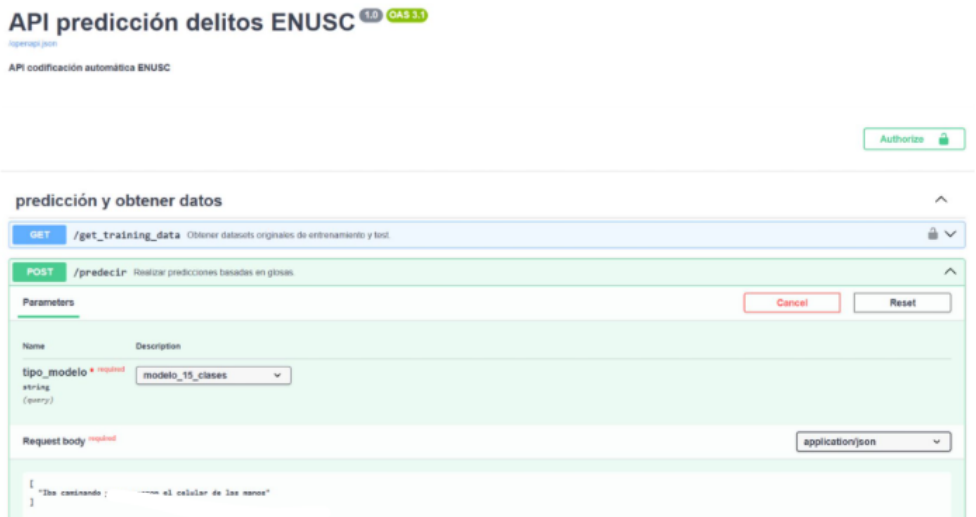
[Puedes acceder al tutorial de uso de esta API a través de este enlace]

Procesamiento de lenguaje natural

Cod automática | CIUO y CAENES | Clasificación delitos | Data drift

Modelo desarrollado para la **revisión de calidad de la encuesta de victimización (ENUSC)**

Capa de embeddings + **LSTM** (Long Short-Term Memory)



Procesamiento de lenguaje natural

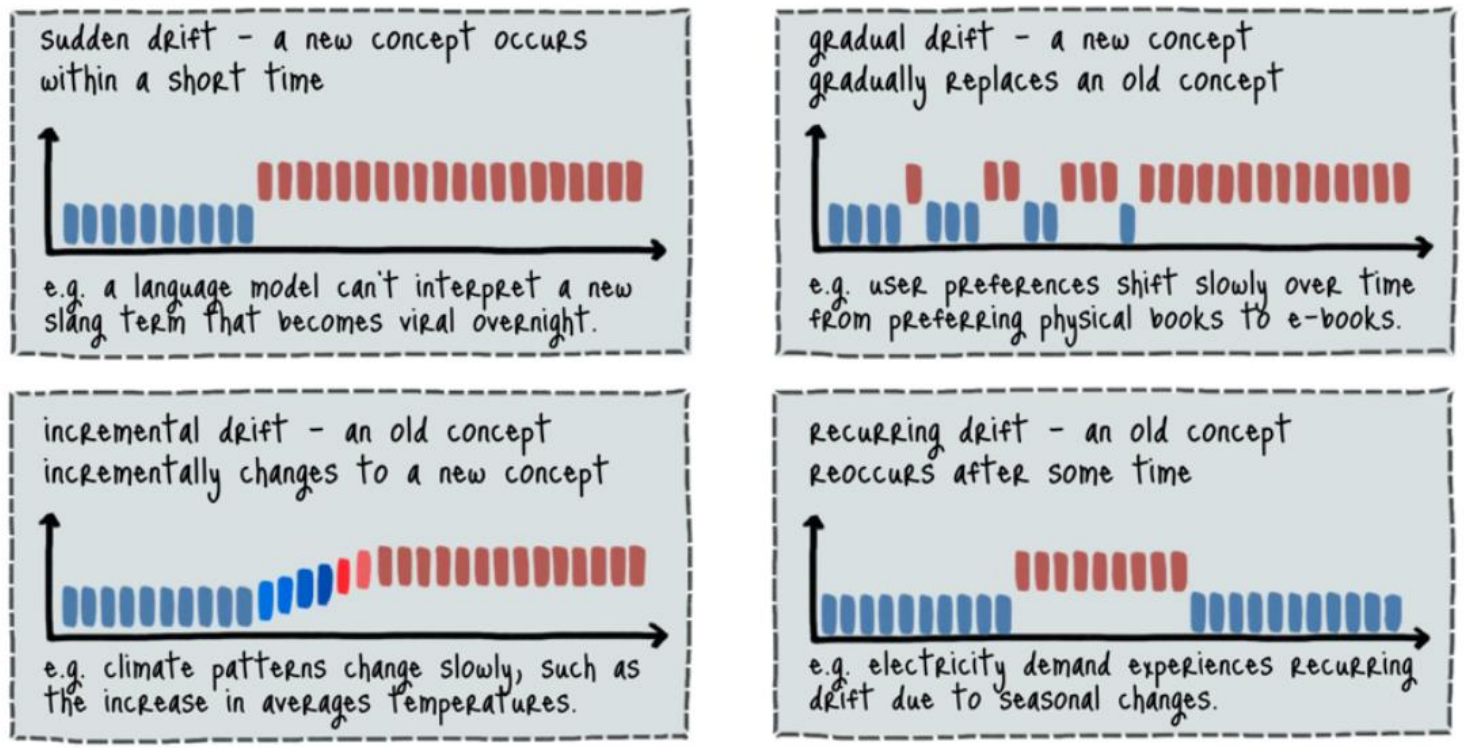
Cod automática | CIUO y CAENES | Clasificación delitos | Data drift


```
{
  "probabilidades": [
    {
      "AMENAZA": "9.063535e-07",
      "CIBER_ACOSO": "5.93864e-07",
      "CIBER_DESTRUC": "1.0372355e-06",
      "CIBER_HACKEO": "9.057918e-07",
      "ESTAFA": "1.2200211e-05",
      "FRAUDE": "9.422963e-06",
      "HURTO": "0.016638247",
      "LESIONES": "1.5515674e-05",
      "ROBO_DESDE_VEHIC": "2.830686e-05",
      "ROBO_SORPRESA": "0.97862625",
      "ROBO_VEHIC": "2.522674e-06",
      "ROBO_VIOLENCIA": "0.0046238205",
      "ROBO_VIVIENDA": "3.844919e-05",
      "VANDAL_VEHIC": "1.3687213e-06",
      "VANDAL_VIV": "2.541593e-07"
    }
  ],
  "predicciones": [
    "ROBO_SORPRESA"
  ]
}
```

Procesamiento de lenguaje natural

Instituto Nacional de Estadísticas · Chile

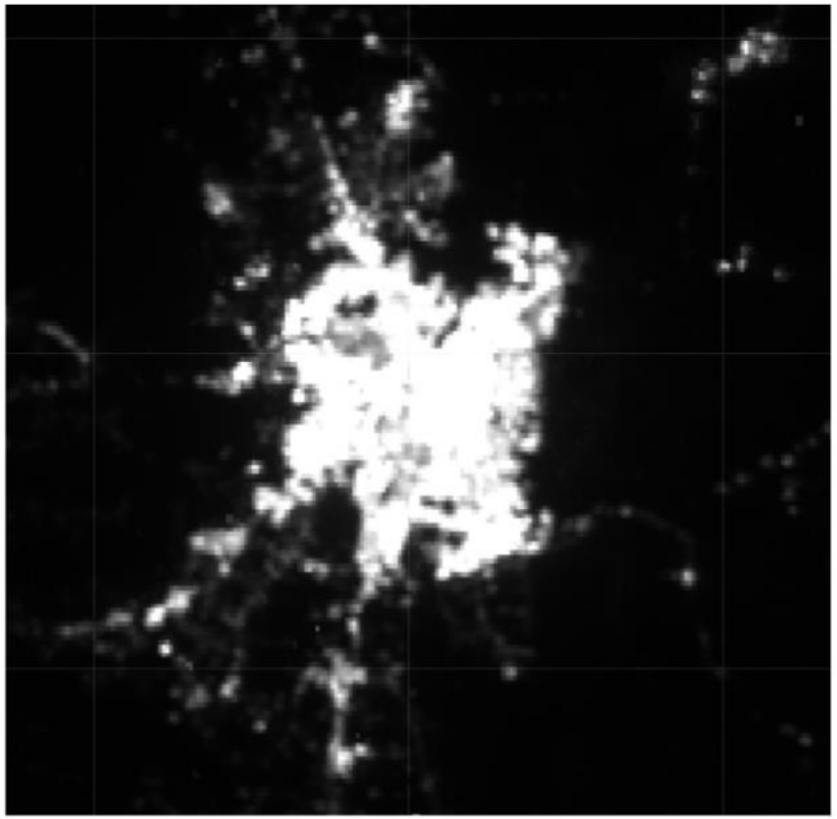
Cod automática | CIUO y CAENES | Clasificación delitos | Data drift



Instituto Nacional de Estadísticas · Chile

Aprovechamiento de imágenes satelitales

Luces Nocturnas | Estratificación



Google Earth Engine

Aprovechamiento de imágenes satelitales

Luces Nocturnas | Estratificación

Visión computacional

Innominación de documentos con DL

Requerimiento | Modelo | App

Requerimiento de Transparencia al Sector público de **públicar informes de actividades de personal a honorarios**

No es posible divulgar nombre junto con rut y firmas de personas

INFORME DE CUMPLIMIENTO
CONVENIO A HONORARIOS A SUMA ALZADA
CON PERSONAS NATURALES

N° DEL INFORME 1
FECHA INFORME 28-mar-24

IDENTIFICACIÓN DEL PRESTADOR DE SERVICIO

NOMBRE: Javiera Eduarda
RUT : 12.345.678-9
PERÍODO DE DESEMPEÑO: DEL 01-mar-24 AL: 31-02-2024
REGIÓN: Metropolitana

MOTIVO DE LA CONTRATACIÓN

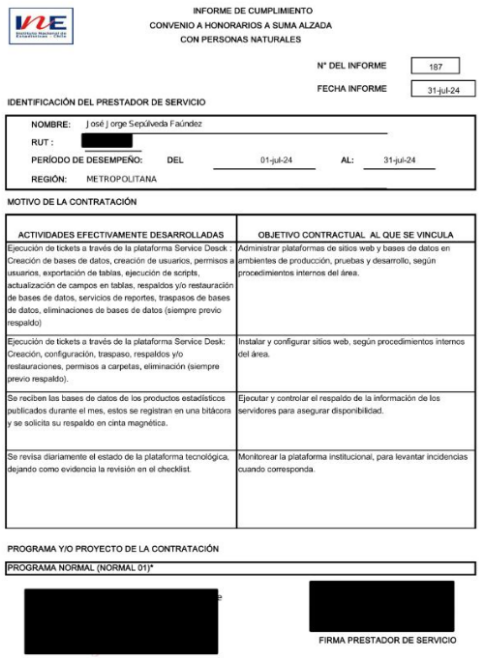
ACTIVIDADES EFECTIVAMENTE DESARROLLADAS	OBJETIVO CONTRACTUAL AL QUE SE VINCULA
1. Entrenamiento de modelo de deep learning para innominación de información en los informes de cumplimiento del personal honorarios a suma alzada de la institución. 2. Elaborar presentación sobre la innominación de informes para visita de INEGI.	INVESTIGAR E IMPLEMENTAR METODOLOGÍAS DE MACHINE LEARNING Y DEEP LEARNING RELATIVAS PRINCIPALMENTE AL RECONOCIMIENTO VISUAL CON DEEP LEARNING, PARA GENERAR MODELOS Y ALGORITMOS PARA SU APLICACIÓN EN LA ALIANZA CHILE-MEXICO

Encuesta de Seguridad Ciudadana

V° B° COORDINADOR DEL CONVENIO
FIRMA PRESTADOR DE SERVICIO

SUBDIRECCIÓN ADMINISTRATIVA

Debe publicarse así →



INFORME DE CUMPLIMIENTO
CONVENIO A HONORARIOS A SUMA ALZADA
CON PERSONAS NATURALES

N° DEL INFORME 187
FECHA INFORME 31-jul-24

IDENTIFICACIÓN DEL PRESTADOR DE SERVICIO

NOMBRE: José Jorge Sepúlveda Fuéndez
RUT : [redacted]
PERÍODO DE DESEMPEÑO: DEL 01-jul-24 AL: 31-jul-24
REGIÓN: METROPOLITANA

MOTIVO DE LA CONTRATACIÓN

ACTIVIDADES EFECTIVAMENTE DESARROLLADAS	OBJETIVO CONTRACTUAL AL QUE SE VINCULA
Ejecución de tickets a través de la plataforma Service Desck : Creación de bases de datos, creación de usuarios, permisos a usuarios, exportación de tablas, ejecución de scripts, actualización de campos en tablas, respaldos y/o restauración de bases de datos, servicios de reportes, traspasos de bases de datos, eliminaciones de bases de datos (siempre previo respaldo)	Administrar plataformas de sitios web y bases de datos en ambientes de producción, pruebas y desarrollo, según procedimientos internos del área.
Ejecución de tickets a través de la plataforma Service Desk: Creación, configuración, traspaso, respaldos y/o restauraciones, permisos a carpetas, eliminación (siempre previo respaldo).	Instalar y configurar sitios web, según procedimientos internos del área.
Se reciben las bases de datos de los productos estadísticos publicados durante el mes, estos se registran en una bitácora y se solicita su respaldo en cinta magnética.	Ejecutar y controlar el respaldo de la información de los servidores para asegurar disponibilidad.
Se revisa diariamente el estado de la plataforma tecnológica, dejando como evidencia la revisión en el checklist.	Monitorear la plataforma institucional, para levantar incidencias cuando corresponda.

PROGRAMA Y/O PROYECTO DE LA CONTRATACIÓN

PROGRAMA NORMAL (NORMAL 01)*

[redacted]
[redacted]
FIRMA PRESTADOR DE SERVICIO

	N documentos	Tiempo
Enero 2024	256	~8.8 hrs.
Junio 2024	1519	~50.6 hrs.

Innominación de documentos con DL

Requerimiento | Modelo | App

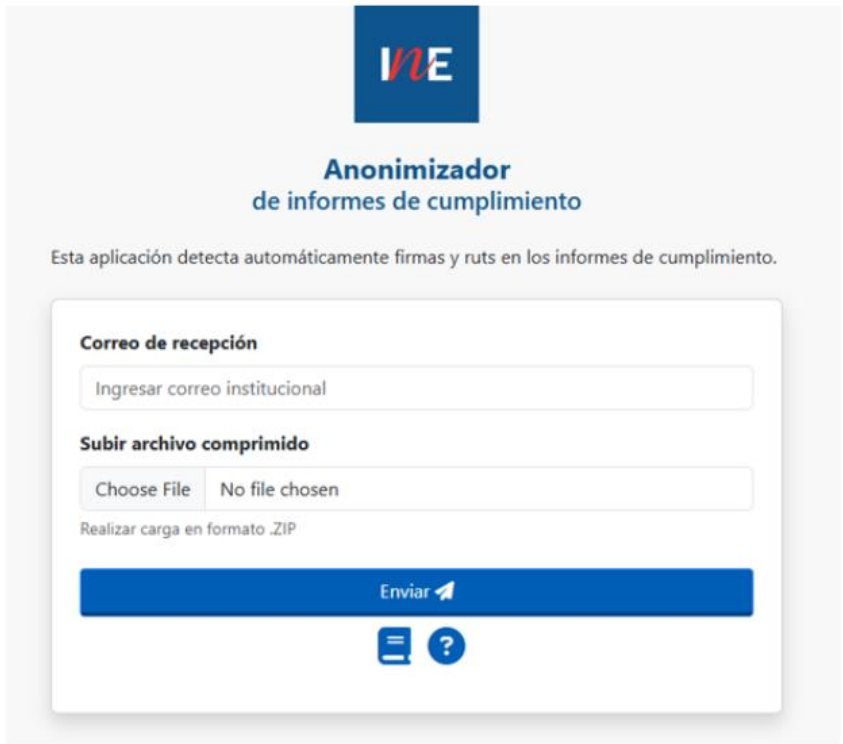
Etiquetamos internamente ~1000 imágenes

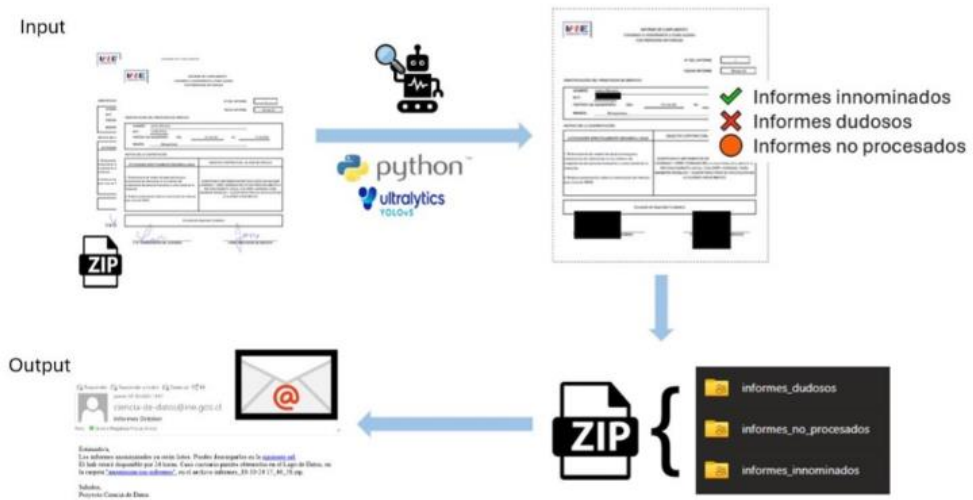
El modelo YOLOv *You Only Look Once*, es un modelo de detección de objetos proveniente del paquete Ultralytics, diseñado con *deep learning*

El modelo cuenta con 24 capas convolucionales y 2 capas de conexión completa

Innominación de documentos con DL

Requerimiento | Modelo | App



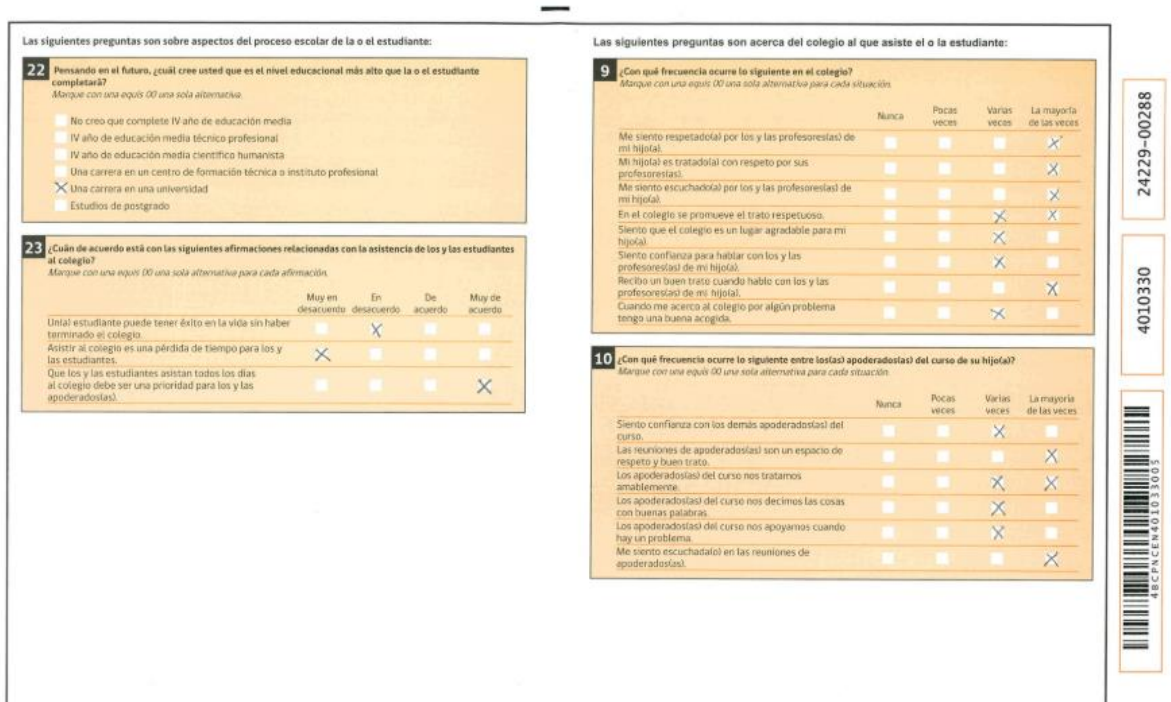


Detección de dobles marcas en SIMCE

SIMCE | OCR | Solución

Esta es una colaboración del INE a la **Agencia de Calidad de la Educación** y el **Laboratorio de Gobierno** (finales del 2024)

Objetivo: generar un modelo que detecte las **dobles marcas** en cuestionarios SIMCE



Las siguientes preguntas son sobre aspectos del proceso escolar de la o el estudiante:

22 Pensando en el futuro, ¿cuál cree usted que es el nivel educacional más alto que la o el estudiante completará?
Marque con una equis (X) una sola alternativa.

- No creo que complete IV año de educación media
- IV año de educación media técnico profesional
- IV año de educación media científico humanista
- Una carrera en un centro de formación técnica o instituto profesional
- X Una carrera en una universidad
- Estudios de postgrado

23 ¿Cuán de acuerdo está con las siguientes afirmaciones relacionadas con la asistencia de los y las estudiantes al colegio?
Marque con una equis (X) una sola alternativa para cada afirmación.

	Muy en desacuerdo	En desacuerdo	De acuerdo	Muy de acuerdo
Un(a) estudiante puede tener éxito en la vida sin haber terminado el colegio.		X		
Asistir al colegio es una pérdida de tiempo para los y las estudiantes.	X			
Que los y las estudiantes asistan todos los días al colegio debe ser una prioridad para los y las apoderados(as).				X

Las siguientes preguntas son acerca del colegio al que asiste el o la estudiante:

9 ¿Con qué frecuencia ocurre lo siguiente en el colegio?
Marque con una equis (X) una sola alternativa para cada situación.

	Nunca	Pocas veces	Varias veces	La mayoría de las veces
Me siento respetado(a) por los y las profesores(as) de mi hijo(a).				X
Mi hijo(a) es tratado(a) con respeto por sus profesores(as).				X
Me siento escuchado(a) por los y las profesores(as) de mi hijo(a).				X
En el colegio se promueve el trato respetuoso.			X	X
Siento que el colegio es un lugar agradable para mi hijo(a).			X	
Siento confianza para hablar con los y las profesores(as) de mi hijo(a).			X	
Recibo un buen trato cuando hablo con los y las profesores(as) de mi hijo(a).				X
Cuando me acerco al colegio por algún problema tengo una buena acogida.			X	

10 ¿Con qué frecuencia ocurre lo siguiente entre los(as) apoderados(as) del curso de su hijo(a)?
Marque con una equis (X) una sola alternativa para cada situación.

	Nunca	Pocas veces	Varias veces	La mayoría de las veces
Siento confianza con los demás apoderados(as) del curso.			X	
Las reuniones de apoderados(as) son un espacio de respeto y buen trato.				X
Los apoderados(as) del curso nos tratamos amablemente.			X	X
Los apoderados(as) del curso nos decimos las cosas con buenas palabras.			X	
Los apoderados(as) del curso nos apoyamos cuando hay un problema.			X	
Me siento escuchado(a) en las reuniones de apoderados(as).				X

24229-00288

4010330

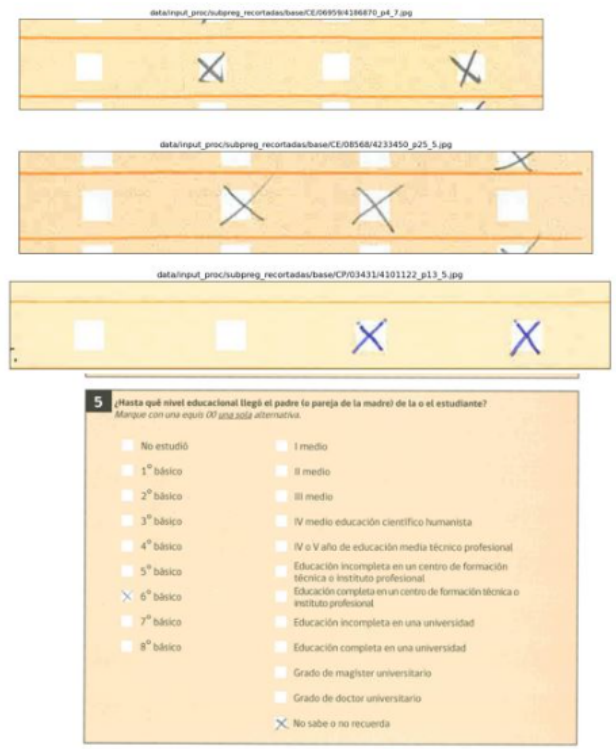
48CPACEN4010330005

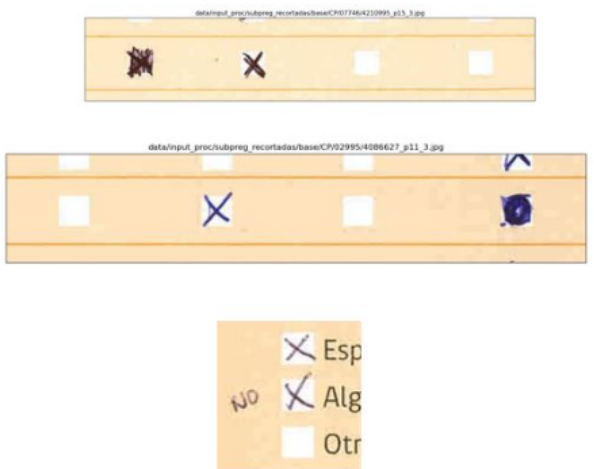
Detección de dobles marcas en SIMCE

SIMCE | OCR | Solución

El **SIMCE se revisa por completo por personas**, que corroboran lo que registra el OCR

El OCR reacciona ante estímulos identificando **doble marca**





Detección de dobles marcas en SIMCE

SIMCE | OCR | Solución

El **modelo fue implementado en el SIMCE 2024** en 4° básico, 2° medio y 6° básico, con muy buenos resultados

Otros desarrollos

Seguimiento operativo con paradatos

Contexto | Dashboard

¿Qué son los paradatos?

Registran **todos los eventos** que se realizan en un dispositivo movil de captura durante una entrevista

Son datos grandes: entre 60 y 200 millones de registros en una encuesta normal

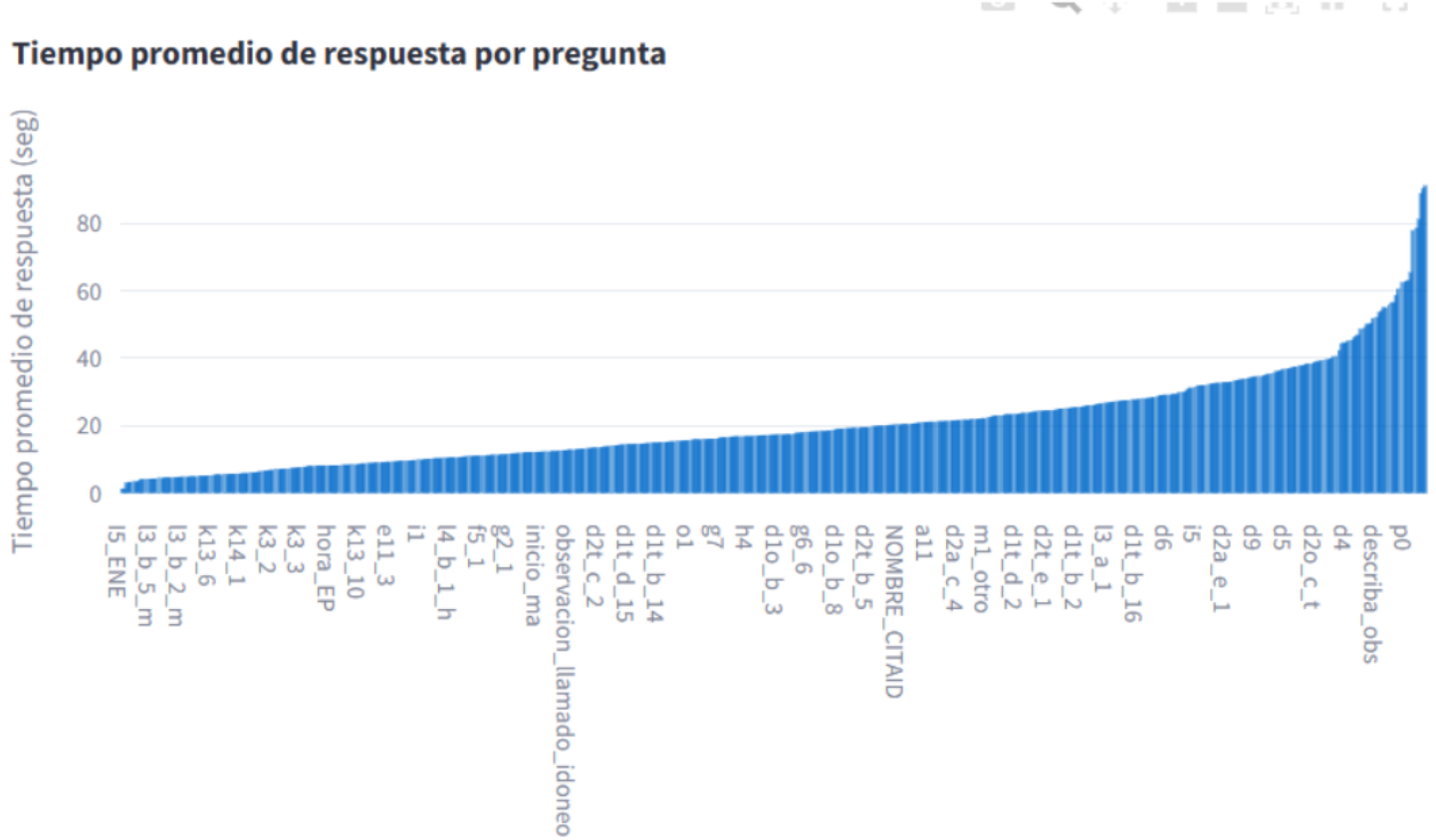
Permiten **detectar desviaciones** en la correcta aplicación de una encuesta

Esta aplicación vive en un **Lago de Datos**

trino

Seguimiento operativo con paradatos

Contexto | Dashboard

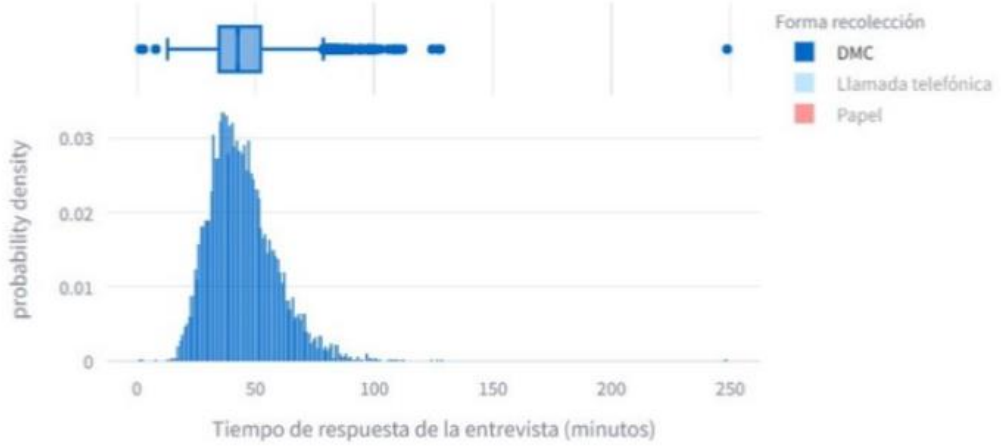


Seguimiento operativo con paradatos

Contexto | Dashboard

Cantidad de respuestas por entrevista según recolector

Tiempo de respuesta de entrevistas según recolector



Paquete calidad en R

*Es el **primer paquete en R desarrollado en el INE** y con **colaboración de CEPAL***

Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares

Tiene **Más de 13k descargas totales** hasta ayer

Tiene **414 descargas** el último mes

Tiene **76 descargas** la última semana

Tuvo **19 descargas** ayer

Paquete calidad en R

UPDATE: Este año también incluimos el estándar de calidad para **encuestas económicas**

https://github.com/inesscc/calidad

En este link pueden encontrar el tutorial del paquete calidad en la web institucional

Tutorial paquete calidad

Para facilitar el proceso de aplicación de la evaluación de estándares de calidad, el Instituto Nacional de Estadísticas ha desarrollado el paquete *calidad*, una herramienta en R que permite aplicar de forma sencilla estándares oficiales para la evaluación de calidad en encuestas de hogares y económicas. Este paquete busca apoyar tanto a equipos técnicos como a personas interesadas en mejorar la comprensión y uso de las estadísticas oficiales.

El paquete *calidad* está construido sobre la base de dos grandes pasos:

1. Estimación y generación de insumos
2. Evaluación de calidad

En este tutorial se presentan casos de uso para estándares de encuestas de hogares y de encuestas económicas. Para ello, utilizaremos dos conjuntos de datos incluidos en el paquete:

- Encuesta Nacional de Empleo 2020 (ENE)
- VII Encuesta Longitudinal de Empresas (ELE).

Procesamiento de datos:

Ambos conjuntos de datos se cargan automáticamente al iniciar la sesión con el paquete. En el caso de la ENE, se realizará un procesamiento adicional para permitir la estimación en subpoblaciones:

```
# cargando paquetes
if (!require("calidad")) install.packages("calidad")
if (!require("dplyr")) install.packages("dplyr")
if (!require("survey")) install.packages("survey")

library(calidad)
library(dplyr)
library(survey)

# edicion base de datos ENE
ene <- ene %>%
       mutate(fdt = if_else(cae_especifico >= 1 & cae_especifico <= 9, 1, 0), # fuerza de trabajo
              ocupado = if_else(cae_especifico >= 1 & cae_especifico <= 7, 1, 0),
              desocupado = if_else(cae_especifico >= 8 & cae_especifico <= 9, 1, 0),
              hombre = if_else(sexo == 1, 1, 0),
              mujer = if_else(sexo == 2, 1, 0))
```

AI-readiness

Datos listos para la IA en las ONE

LLMs y Motores de búsqueda

Quiero el producto interno de Chile para el año 2020

En 2020, el Producto Interno Bruto (PIB) de Chile se contrajo un **5,8%**. Esta fue la mayor caída desde 1982, y estuvo fuertemente influenciada por el impacto de la pandemia de COVID-19.

dame el valor en dólares

El valor del Producto Interno Bruto (PIB) de Chile en 2020 fue de aproximadamente **277.674 millones de dólares estadounidenses.**

Quiero saber cuál era el PIB per cápita de Chile en 2020

El PIB per cápita de Chile en 2020 fue de **$13,163 dólares estadounidenses.**

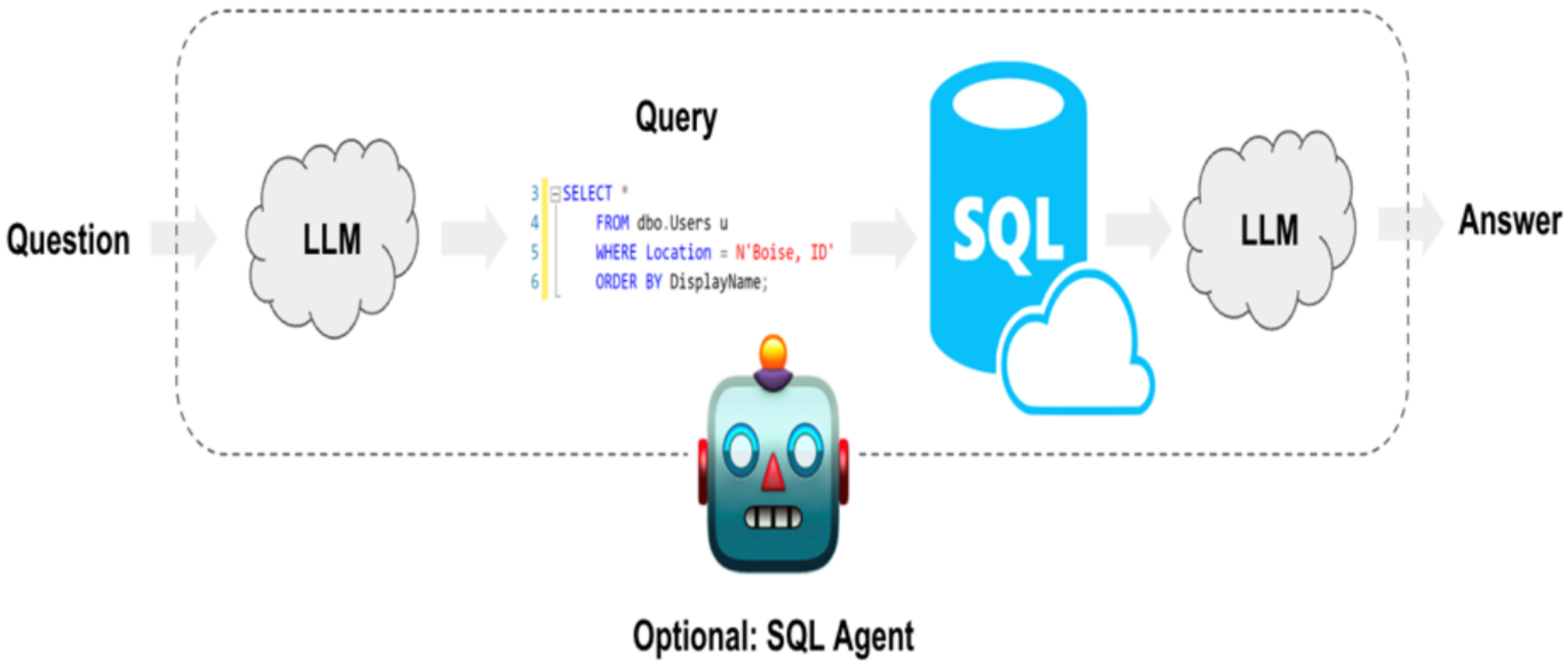
per cápita

En 2020, el **Producto Interno Bruto (PIB) per cápita de Chile** fue de aproximadamente **13.088 dólares estadounidenses.**

Worklflow

Datos listos para la IA en las ONE

LLMs y Motores de búsqueda | Worklflow



Publicación en revista Stats

Paper en revista Open Access de MDPI, generado en **colaboración entre el INE e INEGI, pronto a publicarse.**

stats

MDPI

Article

Automated Classification of Crime Narratives Using Machine Learning and Language Models in Official Statistics

Klaus Lehmann [1,*], Elio Villaseñor [2,*], Alejandro Pimentel [2], Javiera Preuss [1], Nicolás Berhó [1], Oswaldo Diaz [2] and Ignacio Agloni [1]

[1] Instituto Nacional de Estadísticas (INE), Morandé 801, Santiago, Chile, 8340148; kilehmannm@ine.gob.cl (K.L.); jmpreussa@ine.gob.cl (J.P.); nberhom@ine.gob.cl (N.B.); ifaglonij@ine.gob.cl (I.A.)

[2] Instituto Nacional de Estadística y Geografía (INEGI), Heroe de Nacozari 2301, Aguascalientes, Mexico, 20276; elio.villasenor@inegi.org.mx (E.V.); alejandro.pimentel@inegi.org.mx (A.P.); oswaldo.diaz@inegi.org.mx (O.D.)

* Correspondence: elio.villasenor@inegi.org.mx (E.V.); kilehmannm@ine.gob.cl (K.L.)

Reflexiones finales

- No ha sido fácil llegar a donde estamos (tensión entre lo **urgente** y lo **estratégico**)

- La ciencia de datos para **mejorar** y **ampliar la oferta estadística**

- Se abre la oportunidad de **aprovechar fuentes no tradicionales** de información (imágenes satelitales, datos de telefonía celular, etc.)

- Con el avance tecnológico, de la ciencia de datos y la inteligencia artificial, se ha abierto un **importante campo interdisciplinario entre los equipos de negocio y de tecnología**

- Se alza un importante desafío en cuanto a **no quedarse fuera del avance de la IA**

GRACIAS

CIENCIA DE DATOS EN EL INE

Conferencias Ciudadanas

Unidad de Gobierno de Datos

Julio 2025