



CIENCIA DE DATOS EN EL INE
Conferencias Ciudadanas
Unidad de Gobierno de Datos
Julio 2025

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

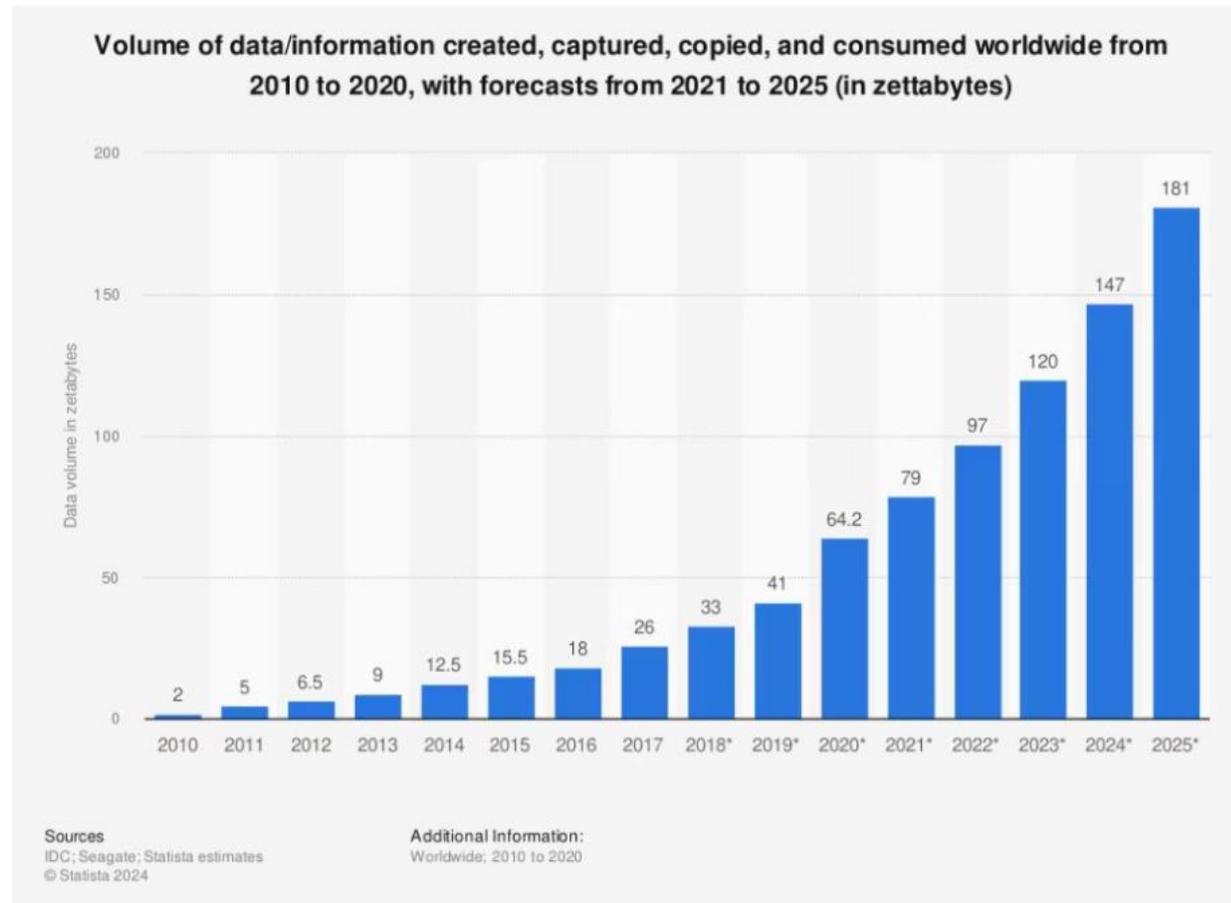
 ifaglonij@ine.gob.cl

 <https://github.com/ignacioagloni>

[Link a presentación en formato HTML en Github](#)

- *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

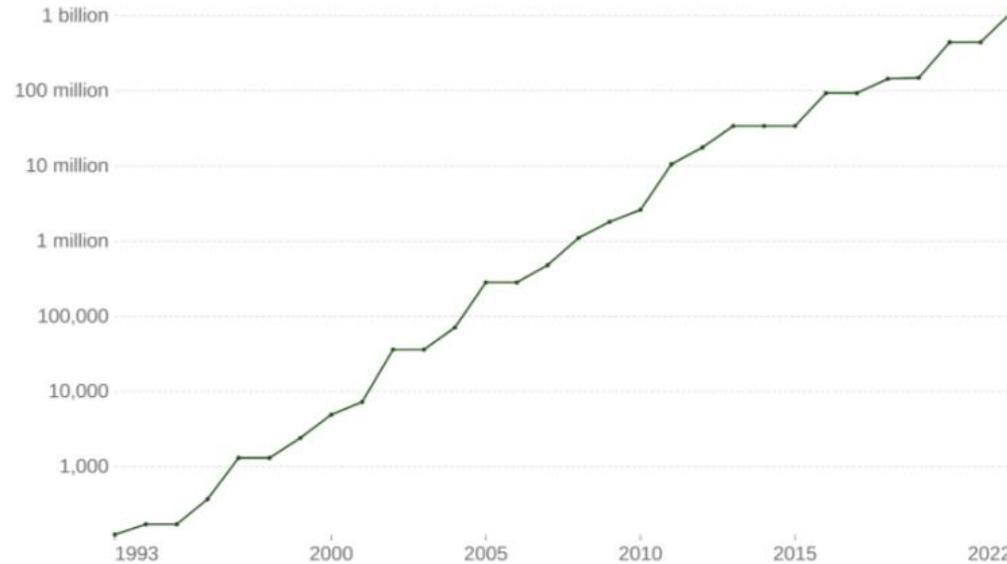


$zettabyte = 10^{12} gb \approx 88.888.888.888$ horas de video en 4k

Computational capacity of the fastest supercomputers

The number of floating-point operations¹ carried out per second by the fastest supercomputer in any given year. This is expressed in gigaFLOPS, equivalent to 10⁹ floating-point operations per second.

Our World in Data



Source: TOP500 Supercomputer Database (2023)

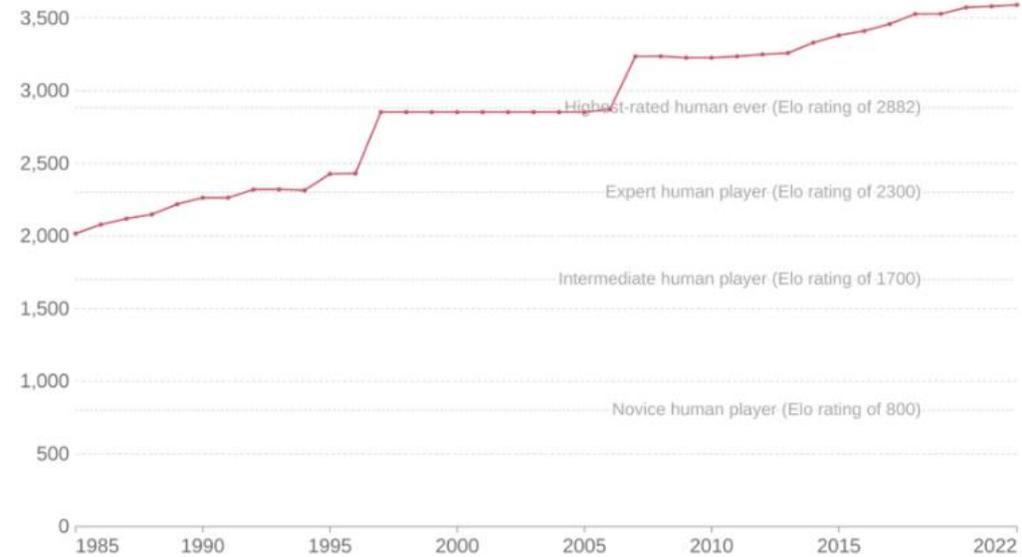
OurWorldInData.org/technological-change • CC BY

1. **Floating-point operation:** A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

Chess ability of the best computers

Chess ability is measured with the Elo rating system, which is calculated based on game results. A higher rating indicates that a player is more likely to win a game.

Our World in Data

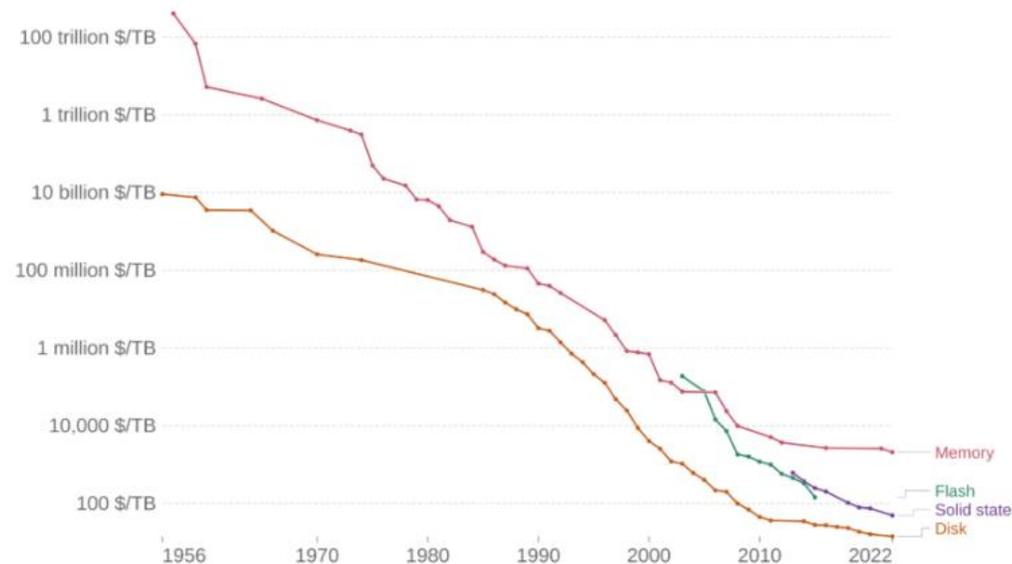


Source: Chess.com (2020); SSDF (2022)

OurWorldInData.org/artificial-intelligence • CC BY

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.



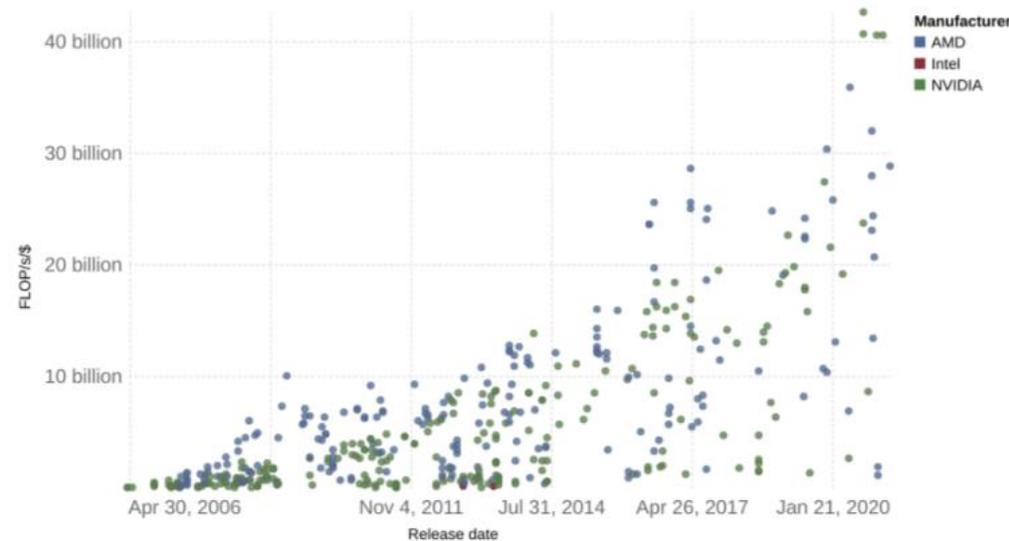
Source: John C. McCallum (2022)

OurWorldInData.org/technological-change • CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year.

GPU computational performance per dollar

Graphics processing units (GPUs) are the dominant computing hardware for artificial intelligence systems. GPU performance is shown in floating-point operations¹ /second (FLOP/s) per US dollar, adjusted for inflation.



Source: Sun et al., Median Group via Epoch (2022)

Note: FLOP/s values refer to 32-bit (full) precision.

OurWorldInData.org/artificial-intelligence • CC BY

1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

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

Se crean, procesan y almacenan datos en todo momento

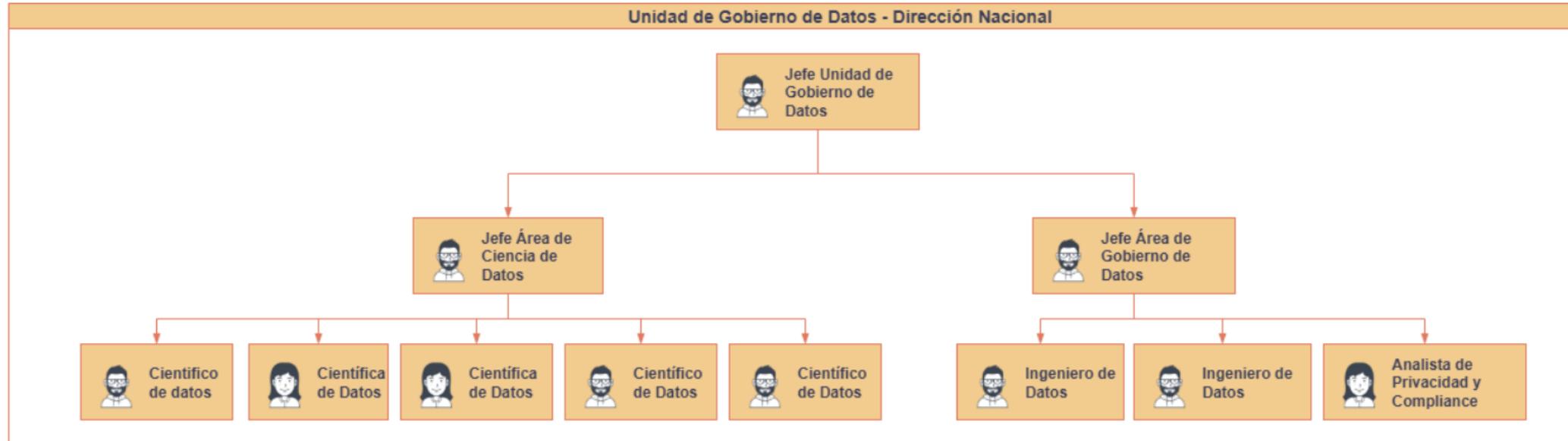
Empresas que manejan muchos datos



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



Principales proyectos y aplicaciones

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

Cod automática

CIUO y CAENES

Clasificación delitos

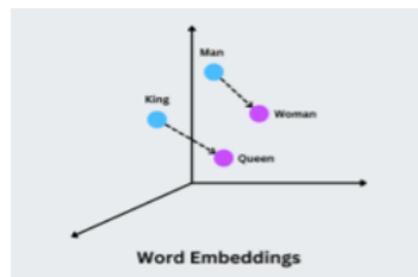
Data drift

CIUO-08: es el Clasificador Internacional Uniforme de Ocupaciones

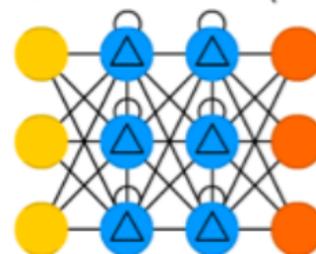
CAENES: es el Clasificador de Actividades Económicas Nacional para Encuestas Sociodemográficas

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

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



Gated Recurrent Unit (GRU)



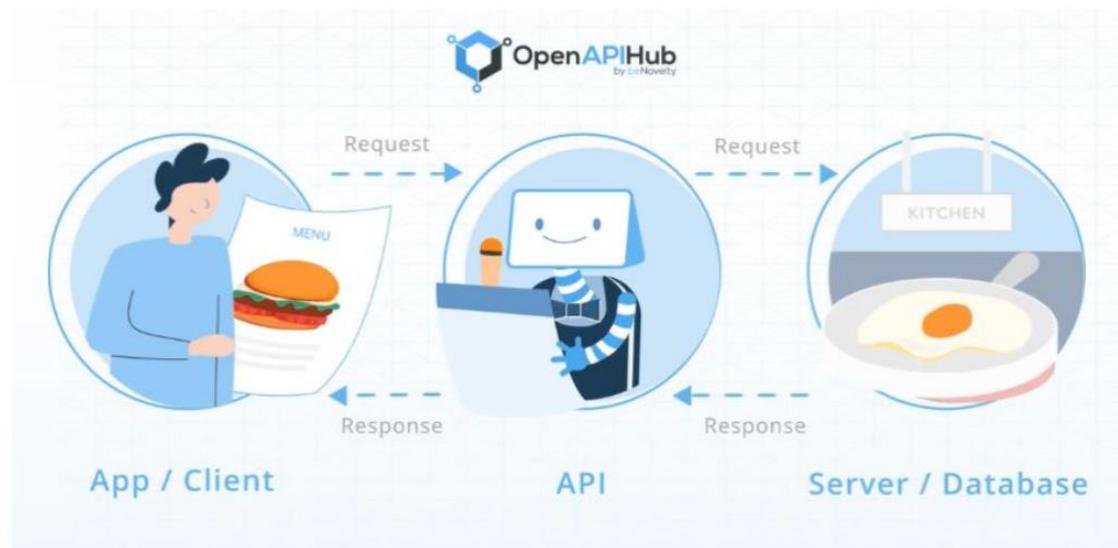
Cod automática

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](#)

Cod automática

CIUO y CAENES

Clasificación delitos

Data drift

Inicio > Calidad Estadística > Clasificaciones > API codificación

Clasificaciones

> Directrices
Metodológicas

> Código de Buenas
Prácticas

Tutorial API de codificación automática

Con el objeto de hacer más eficiente el uso de recursos y mejorar la calidad de los datos publicados por el INE, durante los últimos años la institución ha avanzado en estrategias automatizadas de codificación, principalmente basadas en técnicas de aprendizaje de máquinas (machine learning). Este trabajo se encuentra a la base de la API de codificación automática que el presente tutorial busca acercar a las personas usuarias.

Esta API para la codificación automática pone a disposición de los usuarios y usuarias modelos para clasificar rama de actividad económica (CAENES) y ocupación (CIUO-08 CL) de las personas, al nivel de desagregación de 1 y 2 dígitos, de acuerdo a como sea parametrizada. Los datos de entrenamiento provienen principalmente de la coyuntura de la Encuesta Nacional de Empleo, de modo que los modelos deberían ser utilizados sobre glosas cuya recolección tenga características similares a las implementadas en el trabajo de campo de dicha encuesta.

El etiquetado de los datos y el entrenamiento de los modelos fueron realizados en el marco del Proyecto Estratégico Servicios Compartidos para la Producción Estadística, radicado en la Subdirección Técnica del INE. Para mayor información acerca del proceso de etiquetado manual y de la arquitectura de los modelos, diríjase al documento "Codificación automática de clasificadores CIUO-08 CL y CAENES a partir de técnicas de machine learning. Creación de sets de entrenamiento y optimización de algoritmos", disponible en este mismo sitio para su consulta y descarga.

En la presente viñeta se muestra, a partir de algunos ejemplos, la forma de interactuar con la API de codificación automática mediante R y Python. Esta guía está orientada a usuarios y usuarias con un manejo intermedio de R y/o Python y con conocimientos básicos de machine learning. Para una aproximación más formal a los métodos de la API, diríjase al siguiente sitio: https://rapps.ine.cl:9292/_docs_/

Documento metodológico servicio de codificación automática:
Codificación automática de clasificadores CIUO-08 CL y CAENES a partir de técnicas de machine learning

Implementación en R

El paquete `httr` permite hacer solicitudes a un servidor de manera sencilla y provee algunas herramientas para manipular la respuesta. Mediante la función `POST` realizamos el request (o solicitud), entregando los parámetros para `text`, `classification` y `digits`.

```
library(httr)

glosa <- "manipulador de alimentos prepara colaciones"

request <- httr::POST("https://rapps.ine.cl:9292/predict",
  encode = "json",
  body = list(text = glosa,
             classification = "ciuo",
             digits = 1))
```

Para verificar el resultado utilizamos la función `status_code` a través de la cual es posible verificar el estatus de la operación (idealmente 200). Finalmente, con la función `content` se extrae el resultado de la consulta, consistente en un archivo json que indica la categoría predicha y la probabilidad asignada por el modelo a la predicción.

```
# Revisar el status
httr::status_code(request)
```

```
## [1] 200
```

```
# Extraer el contenido
response <- httr::content(request)
response
```

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

Cod automática

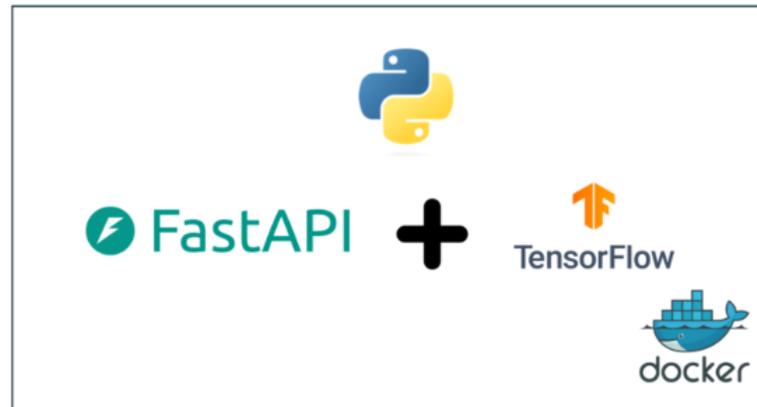
CIUO y CAENES

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)



API predicción delitos ENUSC ^{1.0} ^{QAS 3.1}

openapi.yaml

API codificación automática ENUSC

Authorize 

predicción y obtener datos

GET /get_training_data Obtener datasets originales de entrenamiento y test

POST /predecir Realizar predicciones basadas en glosas

Parameters

Name	Description
tipo_modelo ^{required}	modelo_15_clases
string (body)	

Request body ^{required}

application/json

```
{  "Des continando : el celular de las manos"}
```

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"
  ]
}
```

Cod automática

CIUO y CAENES

Clasificación delitos

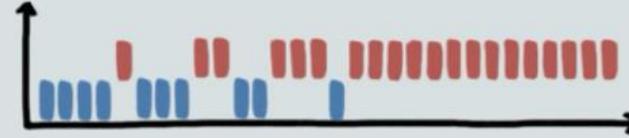
Data drift

sudden drift - a new concept occurs within a short time



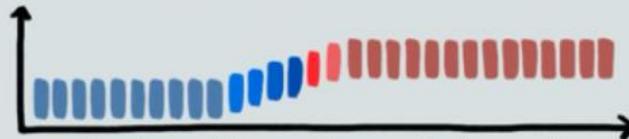
e.g. a language model can't interpret a new slang term that becomes viral overnight.

gradual drift - a new concept gradually replaces an old concept



e.g. user preferences shift slowly over time from preferring physical books to e-books.

incremental drift - an old concept incrementally changes to a new concept



e.g. climate patterns change slowly, such as the increase in average temperatures.

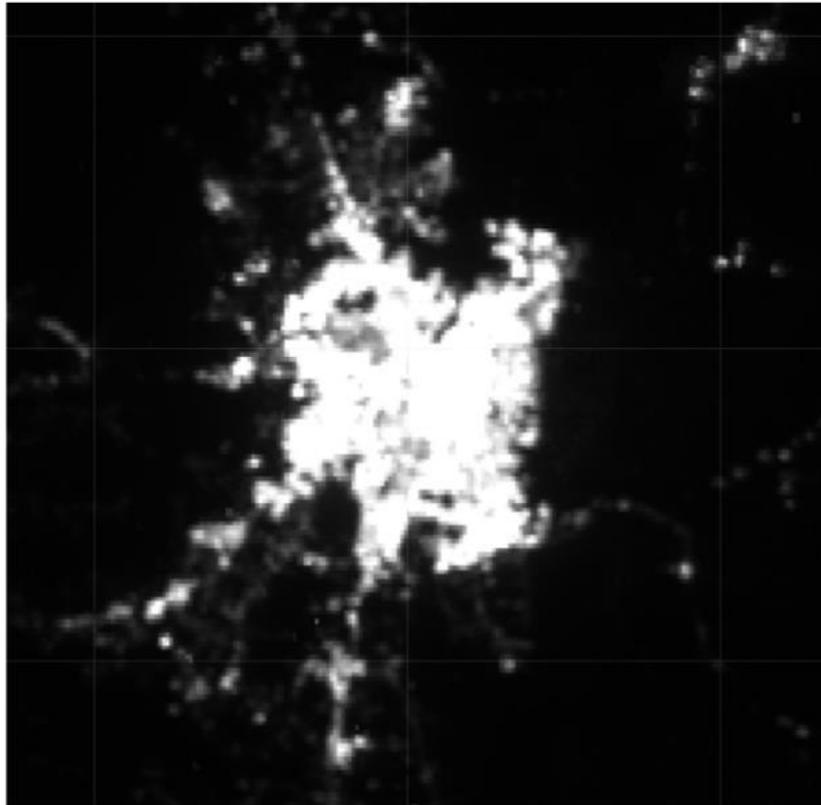
recurring drift - an old concept reoccurs after some time



e.g. electricity demand experiences recurring drift due to seasonal changes.

Luces Nocturnas

Estratificación



Luces Nocturnas

Estratificación



Visión computacional

Innominación de documentos con DL

Requerimiento

Modelo

App

Requerimiento de Transparencia al Sector público de publicar 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]
FIRMA PRESTADOR DE SERVICIO: [Firma]

Debe publicarse así →



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

N° DEL INFORME: 187
FECHA INFORME: 31-ju-24

IDENTIFICACIÓN DEL PRESTADOR DE SERVICIO

NOMBRE: José Jorge Sepúlveda Faúndez
RUT: [Redacted]
PERÍODO DE DESEMPEÑO: DEL 01-ju-24 AL 31-ju-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 Desk. 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.	Epouatar 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]
FIRMA PRESTADOR DE SERVICIO

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

Requerimiento

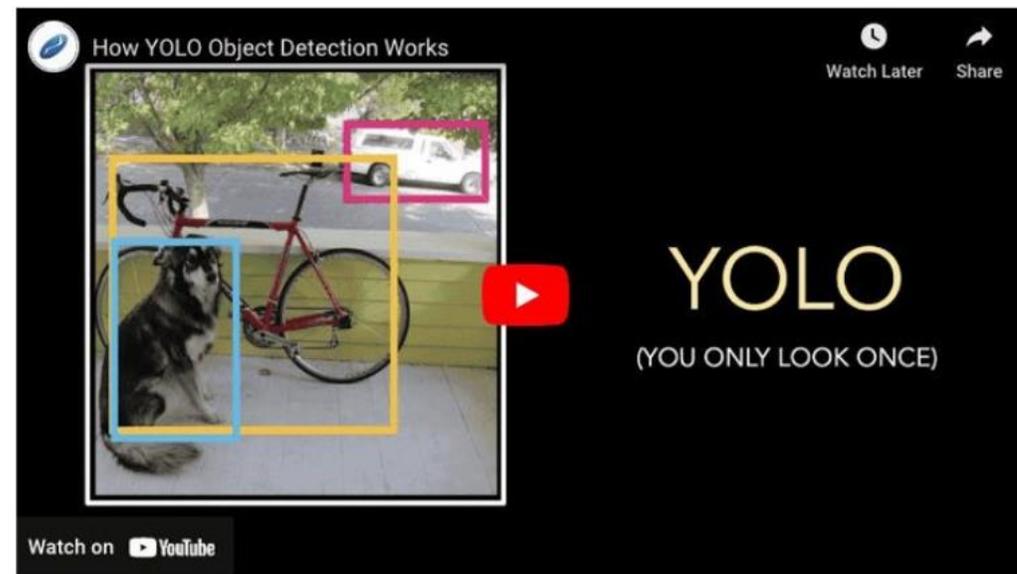
Modelo

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*

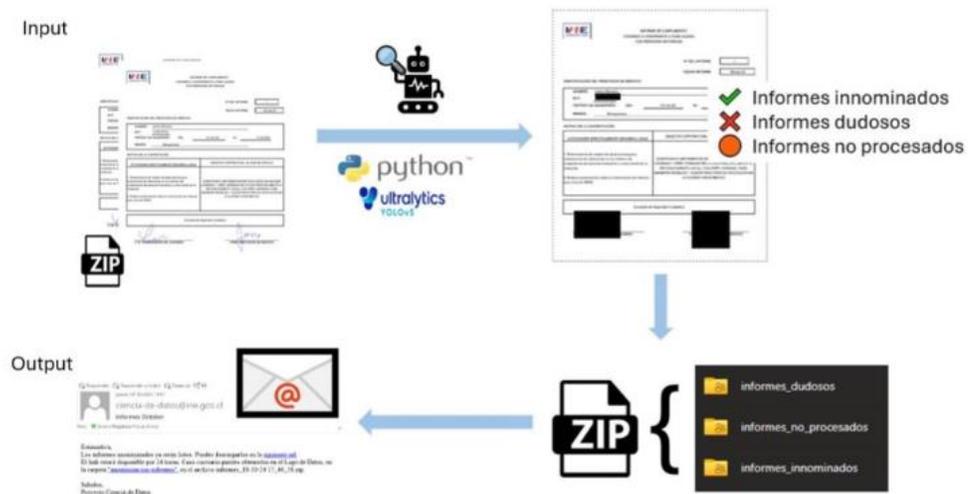
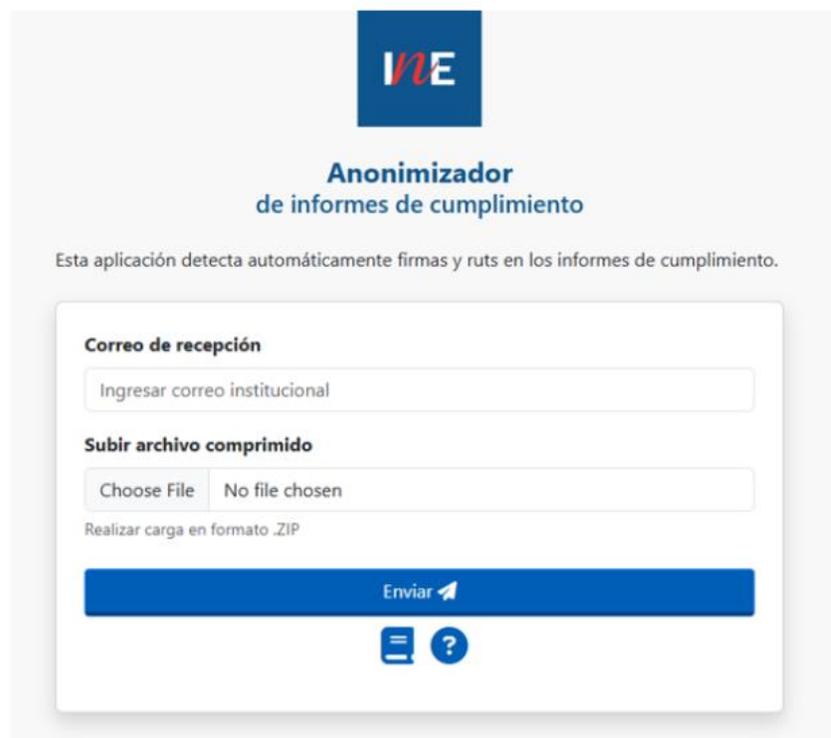
El modelo cuenta con 24 capas convolucionales y 2 capas de conexión completa



Requerimiento

Modelo

App



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.

<input type="checkbox"/>	No creo que complete IV año de educación media
<input type="checkbox"/>	IV año de educación media técnico profesional
<input type="checkbox"/>	IV año de educación media científico humanista
<input type="checkbox"/>	Una carrera en un centro de formación técnica o instituto profesional
<input checked="" type="checkbox"/>	Una carrera en una universidad
<input type="checkbox"/>	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.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Asistir al colegio es una pérdida de tiempo para los y las estudiantes.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Que los y las estudiantes asistan todos los días al colegio debe ser una prioridad para los y las apoderados(as).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

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 profesoras(as) de mi hijo(a).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Mi hijo(a) es tratado(a) con respeto por sus profesoras(as).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Me siento escuchado(a) por los y las profesoras(as) de mi hijo(a).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
En el colegio se promueve el trato respetuoso.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Siento que el colegio es un lugar agradable para mi hijo(a).	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Siento confianza para hablar con los y las profesoras(as) de mi hijo(a).	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Recibo un buen trato cuando hablo con los y las profesoras(as) de mi hijo(a).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Cuando me acerco al colegio por algún problema tengo una buena acogida.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

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.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Las reuniones de apoderados(as) son un espacio de respeto y buen trato.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Los apoderados(as) del curso nos tratamos amablemente.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Los apoderados(as) del curso nos decimos las cosas con buenas palabras.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Los apoderados(as) del curso nos apoyamos cuando hay un problema.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Me siento escuchado(a) en las reuniones de apoderados(as).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

4010330

24229-00288

4010330

4010330



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**

datainput_pro/subpreg_recordadas/base/CE08958438870_p4_7.jpg

datainput_pro/subpreg_recordadas/base/CE085684233450_p25_5.jpg

datainput_pro/subpreg_recordadas/base/CP034314101122_p13_5.jpg

datainput_pro/subpreg_recordadas/base/CP073964210995_p15_3.jpg

datainput_pro/subpreg_recordadas/base/CP02995408627_p11_3.jpg

5 ¿Hasta qué nivel educacional llegó el padre (o pareja de la madre) de la o el estudiante?
Marque con una equis (X) una o más alternativas.

- No estudió
- 1° básico
- 2° básico
- 3° básico
- 4° básico
- 5° básico
- 6° básico
- 7° básico
- 8° básico
- I medio
- II medio
- III medio
- IV medio educación científico humanista
- IV o V año de educación media técnico profesional
- Educación incompleta en un centro de formación técnica o instituto profesional
- Educación completa en un centro de formación técnica o instituto profesional
- Educación incompleta en una universidad
- Educación completa en una universidad
- Grado de magister universitario
- Grado de doctor universitario
- No sabe o no recuerda

NO Esp
 Alg
 Otr

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

Contexto

Dashboard

¿Qué son los paradata?

Registran **todos los eventos** que se realizan en un dispositivo móvil 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**

MINIO

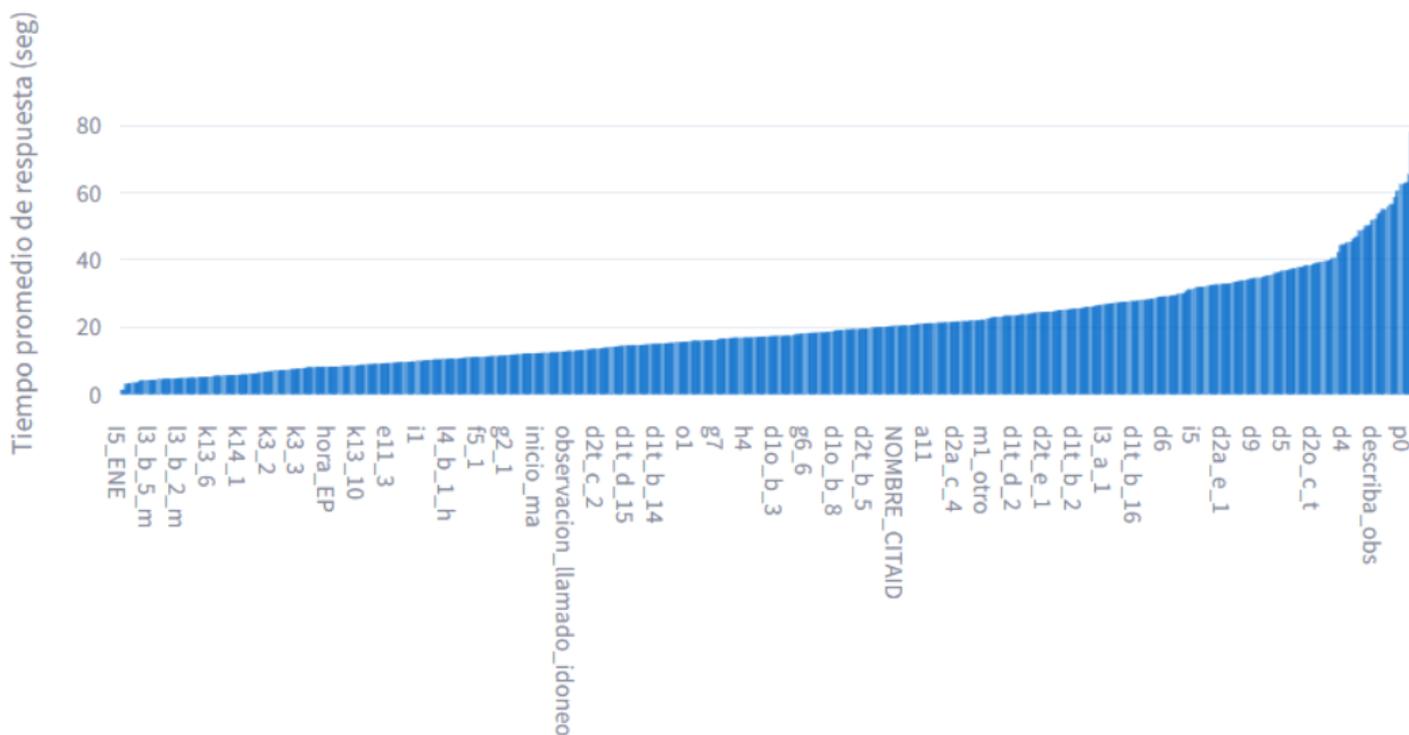


trino

Contexto

Dashboard

Tiempo promedio de respuesta por pregunta



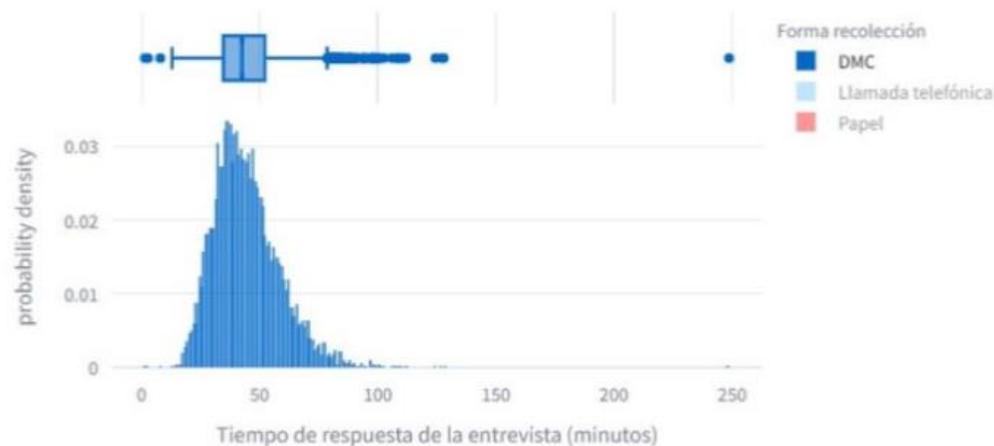
Contexto

Dashboard

Cantidad de respuestas por entrevista según recolector



Tiempo de respuesta de entrevistas según recolector



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



Tiene Más de 13k descargas totales hasta ayer

Tiene 414 descargas el último mes

Tiene 76 descargas la última semana

Tuvo 19 descargas ayer

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)

# edición 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

LLMs y Motores de búsqueda

Workflow

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.

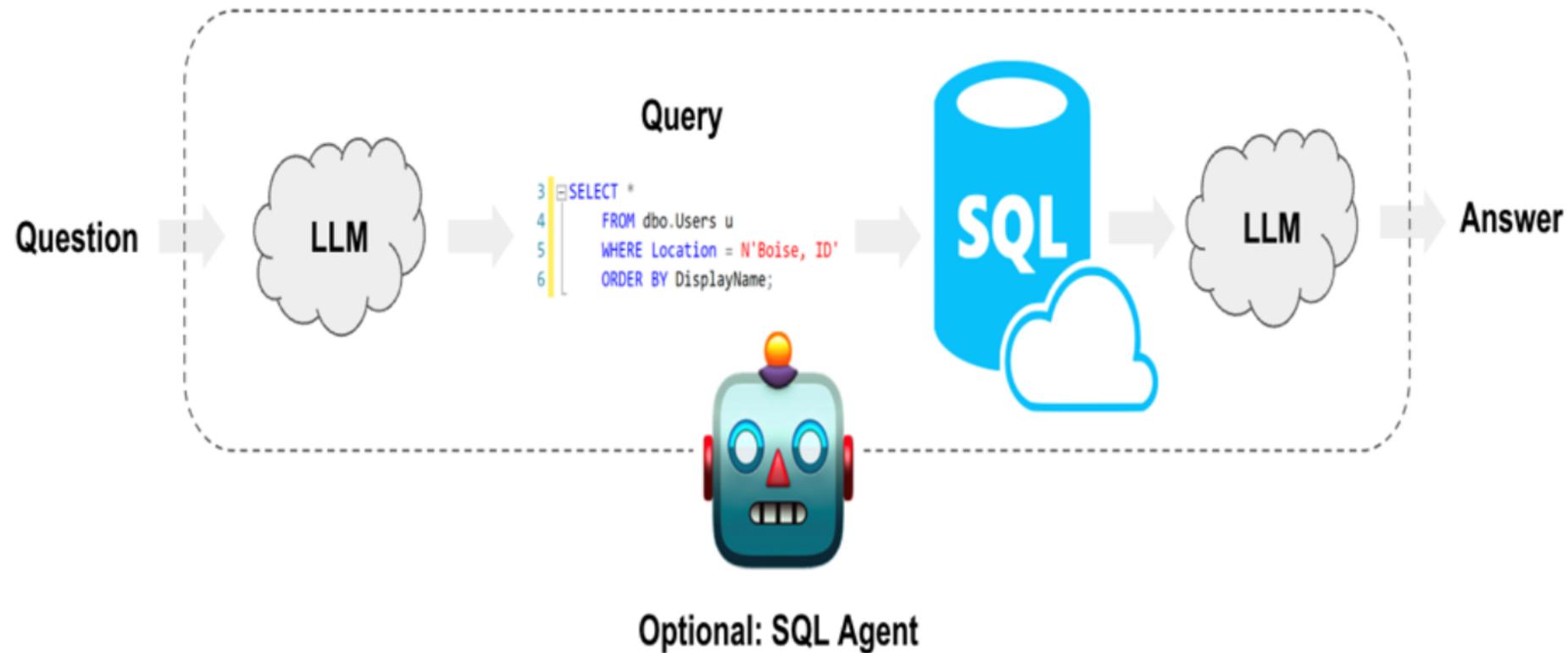
per cápita

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

LLMs y Motores de búsqueda

Workflow



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



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.)

- 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