OFICINA DE SISTEMAS

DOCUMENTO TÉCNICO DE LA SOLUCIÓN PROPUESTA PARA EL RETO DE ACELERACIÓN DEL MINISTERIO DE TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIONES DE COLOMBIA "PUBLICACIÓN Y USO DE DATOS ESTRATÉGICOS CON IA"

PROTOTIPO DE CHATBOT AFINADO CON LOS DATOS ABIERTOS DEL SIPSA EN SU COMPONENTE DE ABASTECIMIENTO

NOVIEMBRE DEL 2024

CONTENIDO

Lista de tablas

Lista de figuras

1. Introducción

El concurso de "Máxima velocidad - potencia de vida" organizado por el Ministerio de Tecnologías de la Información y las Comunicaciones (MINTIC) busca incentivar la creación de soluciones innovadoras que aborden problemas relevantes en el ámbito nacional mediante el uso de tecnologías emergentes y herramientas digitales avanzadas. Con desafíos diseñados para distintos niveles de competencia y especialización, este concurso invita a los participantes a proponer soluciones a los retos de: Máxima PotencIA, turbo Boosters y aceleración. Respecto a este último, se tienen tres (3) retos: marco de referencia de arquitectura empresarial - v3, seguridad de la información y publicación y uso de datos estratégicos con IA.

Puntualmente, el reto de "Publicación y uso de datos estratégicos con IA" tiene como objetivo lograr que las entidades públicas del orden nacional y territorial realicen la publicación y aprovechamiento de los datos abiertos estratégicos del país, haciendo uso de herramientas de inteligencia artificial que faciliten su entendimiento y análisis, y vinculando a los actores del ecosistema de datos abiertos. Lo anterior, basado en tres (3) pilares: transparencia y acceso a la información, toma de decisiones basadas en datos y cumplimiento de la normatividad vigente. Para ello, los participantes deben: i) seleccionar un conjunto de datos estratégico de la Hoja de Ruta de Datos Abiertos Estratégicos para el Estado Colombiano u otro conjunto de datos estratégico de la entidad, con mínimo 1000 registros y justificar su sentido estratégico, ii) publicar en el Portal de Datos Abiertos del Estado Colombiano (https://datos.gov.co) el conjunto de datos estratégico identificado en el punto anterior y iii) elaborar un uso del conjunto de datos abiertos estratégico publicado en el punto anterior, que evidencie análisis descriptivo o predictivo.

Frente a este reto, el Departamento Administrativo Nacional de Estadística (DANE) realizo la actualización del conjunto de datos estratégico del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario Componente Abastecimiento de Alimentos - SIPSA - A - 2018 – 2024 con las vigencias de los tres (3) primeros cuatrimestres del 2024 en la plataforma de Datos Abiertos Colombia, puntualmente en el enlace https://www.datos.gov.co/Estad-sticas-Nacionales/Sistema-de-Informaci-n-de-Precios-y-Abastecimiento/ugru-ez98/about_data cuya última actualización obedece al 23 de agosto del presente año. Además, respecto al segundo apartado la Entidad realizo un desarrollo in house como ejercicio experimental para demostrar uno de los posibles usos que se le pueden dar a este conjunto de datos implementando tecnologías emergentes de Inteligencia Artificial.

El ejercicio experimental obedeció al desarrolló un prototipo de chatbot en Python, diseñado para responder consultas relacionadas con el abastecimiento de los productos de la canasta de alimentos. Además, aunque no fue diseñado específicamente para ello, el chatbot fue creado con la capacidad de responder a **consultas sobre temas predictivos de los datos de insumo (forecasting), no obstante, el DANE ha habilitado esta funcionalidad exclusivamente para los fines del concurso, aclarando que esta opción no forma parte de su misión institucional**. Para lograr esto, se utilizó el modelo GPT 3.5 (General Pretrained Transformer, en su versión 3.5), un modelo de Inteligencia Artificial desarrollado por OpenAI que puede ser accedido vía API utilizando la librería openai. El modelo fue ajustado específicamente para el dominio del conjunto de datos del SIPSA en el Componente Abastecimiento de Alimentos 2024 con el fin de optimizar las respuestas.

El presente documento se encuentra organizada en cinco (5) capítulos: Introducción, marco teórico, materiales y métodos, resultados y conclusiones. En el primer capítulo se introduce el contexto del concurso de MINTIC y la propuesta del DANE en respuesta al reto de Publicación y uso de datos estratégicos con IA. En el segundo capítulo se presenta el marco conceptual principal de las tecnologías implementadas para el desarrollo de la solución de Inteligencia Artificial. En el tercer capítulo se presentan los materiales y métodos utilizados para el cumplimiento del reto. En el cuarto capítulo se presentan los resultados de la solución desarrollada. Finalmente, en el quinto capitulo se presentan las consideraciones finales en función de los objetivos del reto planteado.

2. Marco teórico

2.1 Sistema de Información de Precios y Abastecimiento del Sector Agropecuario Componente Abastecimiento de Alimentos - SIPSA – A

El Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA) es el encargado de informar los precios mayoristas de los productos agroalimentarios que se comercializan en el país, así como la información de insumos y factores asociados a la producción agrícola y el nivel de abastecimiento de alimentos en las ciudades (DANE, 2024). El SIPSA se encuentra integrado por tres (3) componentes: Precios mayoristas, insumos y abastecimientos.

- **Precios mayoristas:** Informa sobre las cotizaciones mayoristas de los principales productos agrícolas que componen la canasta de los alimentos.
- **Insumos:** Reporta los precios minoristas de los insumos y factores asociados a la producción agrícola y pecuaria.
- **Abastecimiento:** Reporta el volumen de abastecimiento de productos en las principales ciudades del país.

Puntualmente, la operación estadística del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario **Componente de Abastecimiento de Alimentos** tiene como objetivo generar información sobre las cantidades y procedencias de los alimentos frescos y procesados que ingresan a los mercados mayoristas del país, para la toma de decisiones (DANE, 2023).

2.2 Procesamiento de Lenguaje Natural

Es un campo interdisciplinario de la inteligencia artificial y la lingüística computacional que tiene como objetivo permitir a los computadores comprender e interpretar el lenguaje natural en forma legible. Abarca una variedad de tareas que incluyen el análisis de sentimientos, traducción automática, extracción de información, generación de lenguaje natural, entre otros (Bird et al., 2009).

2.3 Inteligencia Artificial Generativa

La Inteligencia Artificial Generativa (GAI) es un enfoque dentro del campo de la inteligencia artificial que se centra en el desarrollo de sistemas capaces de crear contenidos de manera autónoma tales como imágenes, música, texto o incluso videos, imitando el estilo y la estructura de los datos de entrada (Goodfellow et al., 2016).

2.4 Modelo Transformer – GPT

Los modelos Transformers son una arquitectura de redes neuronales diseñada para procesar secuencias de datos, introducida por (Vaswani et al., 2017), se caracteriza por el uso de un mecanismo de self-attention que permite capturar relaciones contextuales a lo largo de toda la secuencia, sin la necesidad de procesamiento secuencial. Su estructura principal incluye capas de codificador y decodificador que facilitan tareas como traducción de idiomas, análisis de texto y generación de lenguaje, y su capacidad para el entrenamiento en paralelo ha impulsado grandes avances en modelos de lenguaje.

Por su parte, GPT (Generative Pre-trained Transformer) es una serie de modelos de lenguaje desarrollados por OpenAI, basados en la arquitectura de Transformer y enfocados en la generación de texto. Estos modelos utilizan un enfoque de entrenamiento pre-entrenado y ajustado (fine-tuning) para comprender el contexto y generar texto coherente y relevante a partir de una entrada inicial.

2.5 ChatBot

Un chatbot es un programa de software diseñado para simular y procesar conversaciones humanas a través de texto o voz, permitiendo a usuarios interactuar con sistemas digitales de forma conversacional (Weizenbaum, 1966). Estos pueden utilizar reglas predefinidas, procesamiento de lenguaje natural (NLP) o inteligencia artificial avanzada para responder a preguntas, realizar tareas y entablar diálogos. Son comunes en aplicaciones de atención al cliente, asistentes virtuales y plataformas de mensajería, facilitando interacciones automáticas y eficientes.

3. Materiales y métodos

3.1 Metodología de Team Data Science Process

Para el desarrollo del prototipo de chatbot se implementó la metodología de Team Data Science Process. Esta, es una metodología ágil desarrollada por Microsoft que busca estructurar el desarrollo de proyectos de ciencia de datos en varios niveles: ciclo de vida de ciencia de datos, estructuración del proyecto, infraestructura y recursos, herramientas y utilidades. Se compone por cuatro etapas: entendimiento del negocio, adquisición y entendimiento de los datos, modelamiento y despliegue. Además, se caracteriza por ser una metodología ágil que busca estructurar el desarrollo de proyectos de ciencia de datos en varios niveles: ciclo de vida de ciencia de datos, estructuración del proyecto, infraestructura y recursos y herramientas y utilidades (Microsoft, 2016). Las etapas para el desarrollo de un proyecto de ciencia de datos se integran en la **Figura 1**:

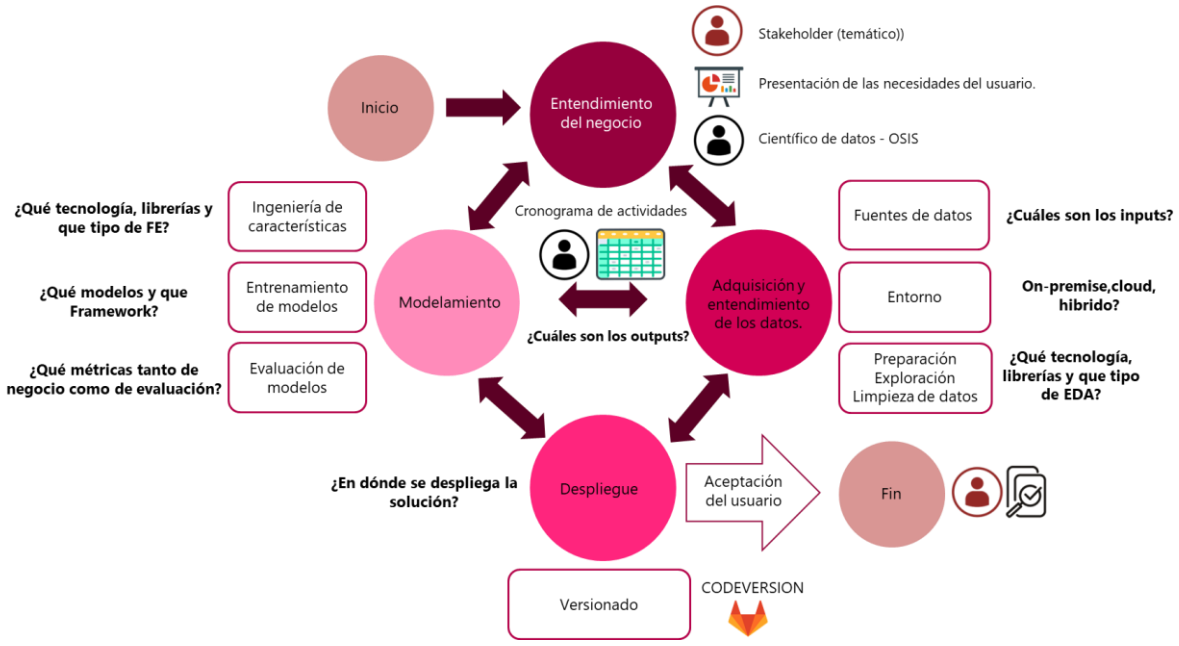

Figura 1. Etapas de un proyecto según la metodología de Team Data Science Process

3.2 Descripción del conjunto de datos

A continuación se presenta una visión general detallada del conjunto de datos, incluyendo la descripción de los atributos, su distribución y características estadísticas clave, lo que es fundamental para la fase de análisis exploratorio y la preparación de datos para el modelado.

Origen de los datos

Fuente	Método de Obtención	Ubicación	Estructura de Datos	Formato / Extensión	Tamaño	Versión
Archivo: **SIPSA_A Icuatrim2024.csv** Plataforma web de Datos Abiertos Colombia https://www.datos.gov.co/Estad-sticas-Nacionales/Sistema-de-Informaci-n-de-Precios-y-Abastecimiento/ugru-ez98/about_data	Extracción directa (Descarga)	Plataforma web de Datos Abiertos Colombia (https://www.datos.gov.co)	Estructurada	CSV	704.062 registros, 10 columnas y tamaño total de 79,5 MB	Versión 1.0
Archivo: **SIPSA_A IIcuatrim2024.csv** Plataforma web de Datos Abiertos Colombia https://www.datos.gov.co/Estad-sticas-Nacionales/Sistema-de-Informaci-n-de-Precios-y-Abastecimiento/ugru-ez98/about_data	Extracción directa (Descarga)	Plataforma web de Datos Abiertos Colombia (https://www.datos.gov.co)	Estructurada	CSV	735.964 registros, 10 columnas y tamaño total de 82 MB	Versión 1.0

Tabla 1. Origen de los datos utilizados para el reto

Diccionario de datos

A continuación, se presenta el diccionario de datos, que detalla la estructura del conjunto de datos con descripciones de cada variable y su tipo de dato. Este sirve como referencia para entender las características y el contenido de los datos utilizados en el reto.

Variable	Descripción	Tipo de dato	Fuente de datos
Fuente	Mercado mayorista en donde se captura la información	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
FechaEncuesta	Fecha en la que se realizó la encuesta	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Cod. Depto Proc.	Código de la DIVIPOLA del Departamento desde donde es remitido el alimento	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Cod. Municipio Proc.	Código de la DIVIPOLA del Municipio desde donde es remitido el alimento	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Departamento Proc.	Nombre del Departamento desde donde es remitido el alimento	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Municipio Proc.	Nombre del Municipio desde donde es remitido el alimento	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Grupo	Clasificación del alimento en la canasta de SIPSA	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Codigo CPC	Código de la Clasificación Central de Productos	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Ali	Nombre del alimento	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary
Cant Kg	Cantidad en Kilogramos	Carácter	Archivo Nacional de Datos ANDA https://microdatos.dane.gov.co/index.php/catalog/697/data-dictionary

Tabla 2. Diccionario del conjunto de datos utilizados en el reto

4. Resultados

En respuesta al reto, el equipo del DANE desarrolló un prototipo de **chatbot en Python**, diseñado para responder consultas relacionadas con el abastecimiento de los productos de la canasta. Además, aunque no fue diseñado específicamente para ello, el chatbot fue creado con la capacidad de responder a **consultas sobre temas predictivos de los datos de insumo (forecasting), no obstante, el DANE ha habilitado esta funcionalidad exclusivamente para los fines del concurso, aclarando que esta opción no forma parte de su misión institucional**. Para lograr esto, se utilizó el **modelo GPT 3.5 (General Pretrained Transformer, en su versión 3.5), un modelo de Inteligencia Artificial** desarrollado por OpenAI que puede ser accedido vía API utilizando la librería de Python openai. Adicionalmente, se incorporó el modelo de procesamiento de lenguaje natural whisper, que permite realizar transcripción de voz a texto, permitiendo que el usuario le hable a la herramienta por medio del micrófono de su computador. La interfaz gráfica de la solución se puede apreciar en la **Figura 2**.

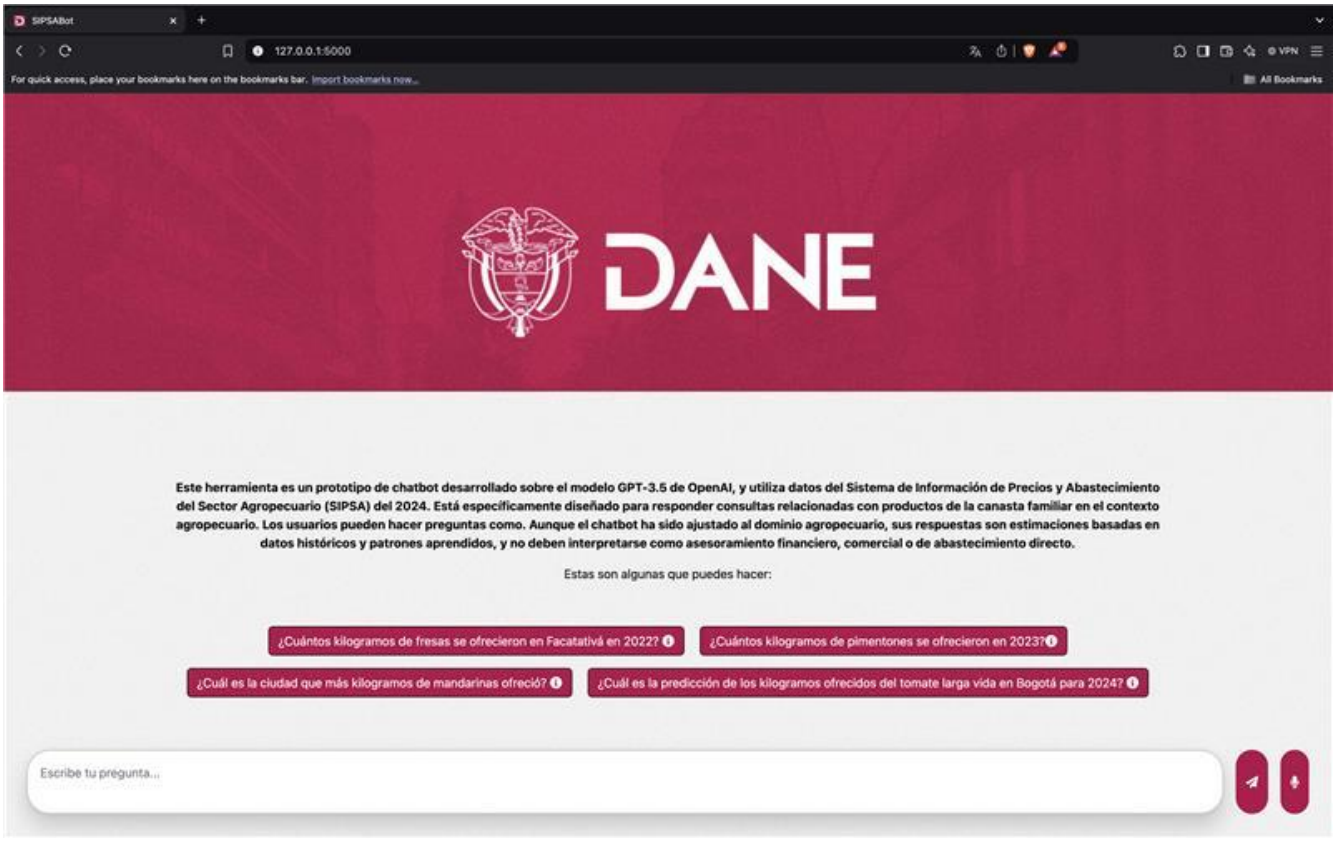

Figura 2. Interfaz gráfica del prototipo del Chatbot

El modelo GPT fue ajustado específicamente para el contexto de los datos abiertos del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario **Componente de Abastecimiento de Alimentos** (SIPSA – A) del 2024 (**Figura 3**) por medio de un ejercicio de prompt-engineering. Adicionalmente, con el objetivo de construir instrucciones de lenguaje natural (en español) en instrucciones de Python para consultar cifras de los datos de SIPSA, se utilizó la tecnología LlamaIndex, que corresponde a una tecnología, que en conjunción con modelos de lenguaje, permite consultar diferentes sistemas de almacenamiento de datos en varios idiomas. **Es importante indicar que las respuestas del chatbot son estimaciones basadas en datos históricos y patrones aprendidos del conjunto de datos señalado previamente**.

Figura 3. Conjunto de datos abiertos del SIPSA – A, redirigido desde https://www.datos.gov.co/Estad-sticas-Nacionales/Sistema-de-Informaci-n-de-Precios-y-Abastecimiento/ugru-ez98/about_data

En la **Figura 4** se presentan dos ejemplos del tipo de preguntas que se le pueden realizar al chatbot. Por ejemplo: ¿Cuál es la ciudad que más kilogramos de mandarinas ofreció? La interfaz muestra el mensaje de la respuesta justo debajo de la pregunta, en un formato de conversación. Este ejemplo demuestra cómo el chatbot proporciona datos específicos en respuesta a preguntas sobre el abastecimiento de alimentos, cumpliendo su función de consulta rápida para los datos abiertos del SIPSA en el Componente de Abastecimiento del 2024.

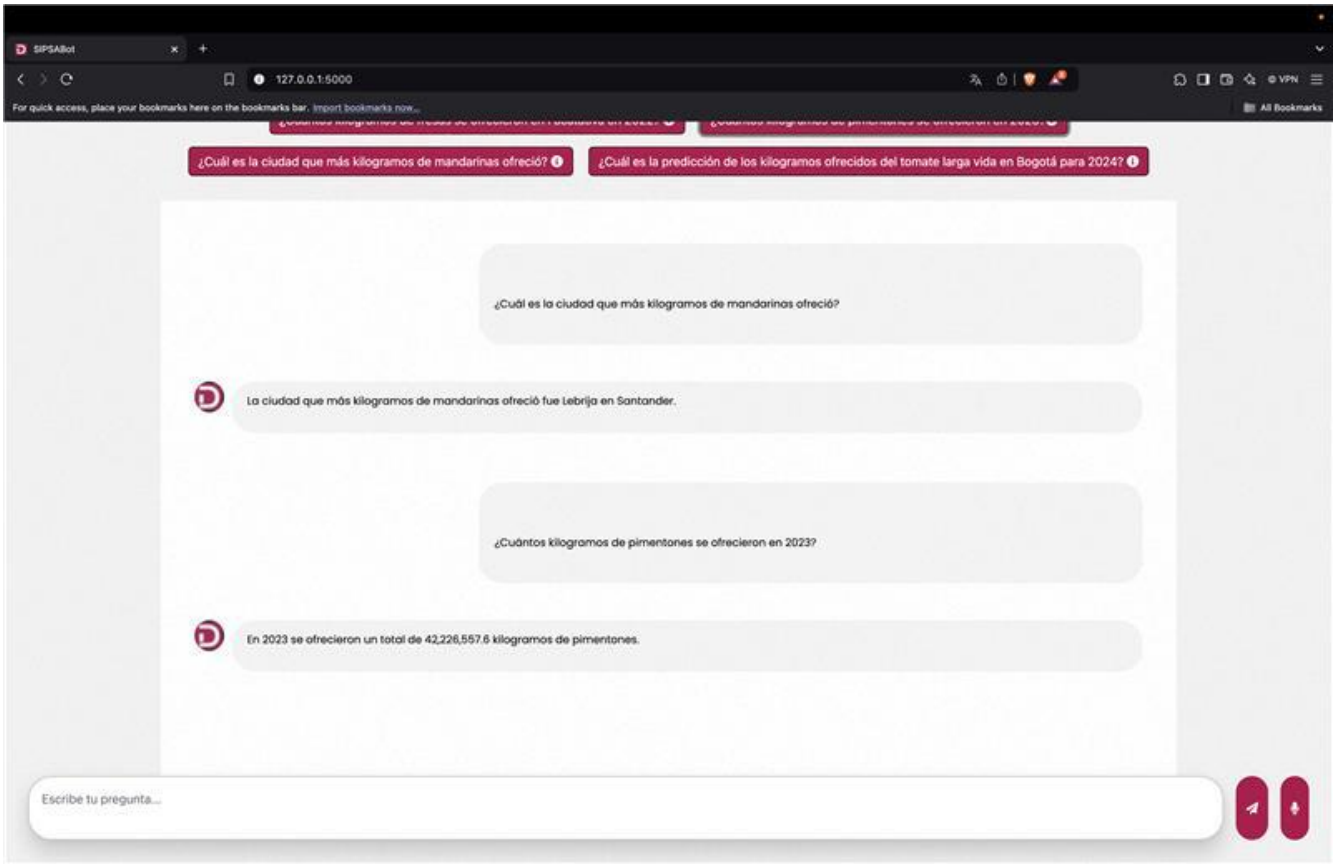

Figura 4. Ejemplo de consultas a realizar al prototipo de chatbot

Es importante hacer un par de aclaraciones sobre el prototipo de chatbot; fue diseñado para responder preguntas de manera independiente y no tiene memoria, lo que significa que no retiene información de interacciones previas ni establece un contexto entre preguntas consecutivas. Cada respuesta que proporciona se basa únicamente en la consulta actual. Para responder a las preguntas, el chatbot realiza consultas al conjunto de datos de SIPSA alojada en el portal de datos abiertos del Gobierno de Colombia, datos.gov.co, utilizando información actualizada del sistema de precios y abastecimiento del sector agropecuario, según lo publicado en este portal.

4.1 Evidencia de posible uso de análisis predictivo

Para responder al desafío de incorporar capacidades de análisis predictivo en la solución presentada, el equipo del DANE habilitó en el prototipo de chatbot una funcionalidad que permite responder a consultas sobre predicciones (forecasting) sobre los datos insumo señalados en las secciones anteriores. Sin embargo, esta funcionalidad ha sido activada exclusivamente para los fines del concurso y el DANE aclara que esta opción no corresponde a su misión institucional. Para efectos del concurso, se presenta en la **Figura 5** un ejemplo ilustrativo en el que, a partir de una consulta en lenguaje natural, el chatbot proporciona respuestas mediante análisis predictivo.

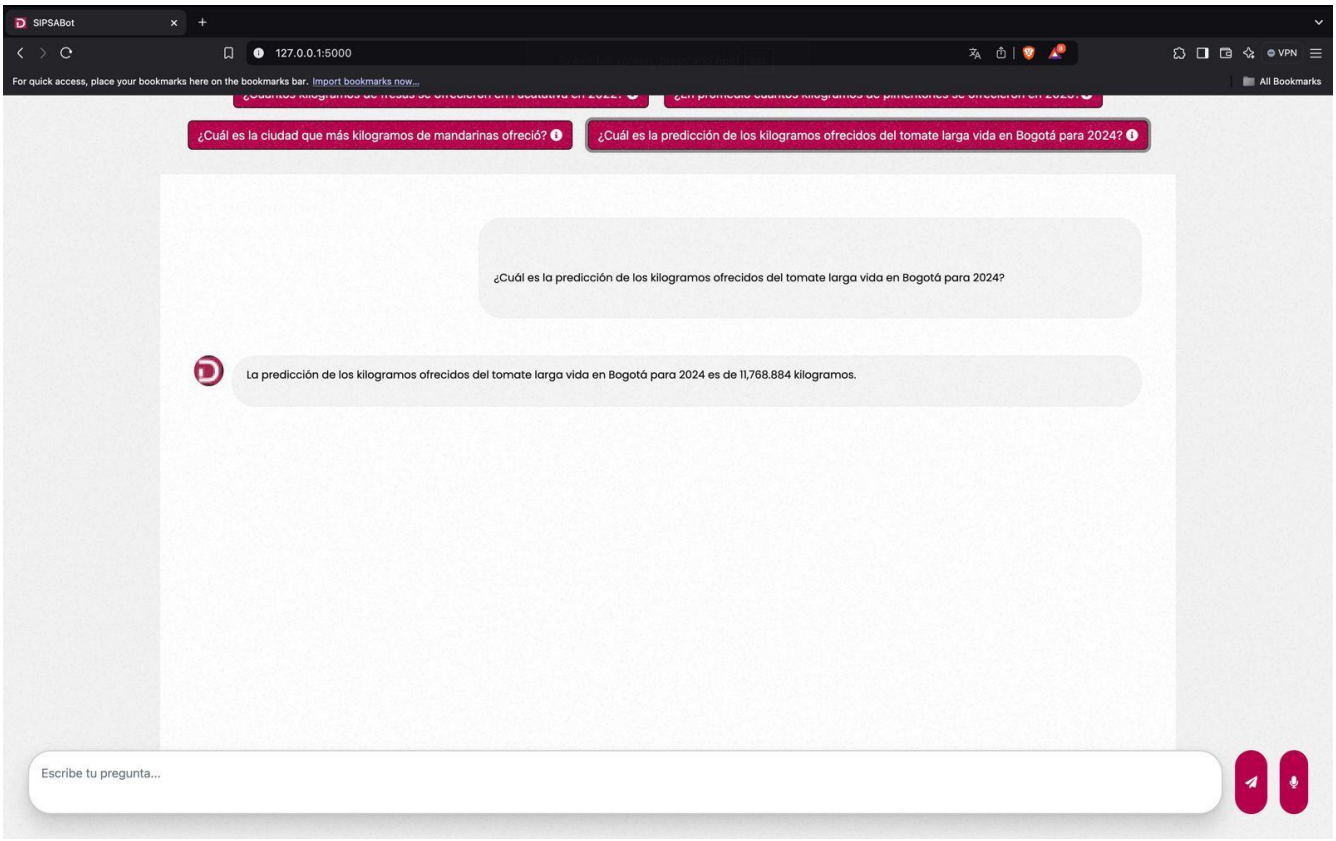

Figura 5. Ejemplo de análisis predictivo a partir de una consulta

5. Conclusiones

- El desarrollo de un prototipo de chatbot con la capacidad de realizar consultas a una fuente de datos mediante lenguaje natural representa una herramienta eficiente para el acceso rápido y amigable a la información, eliminando la necesidad de conocimientos técnicos avanzados.

- El prototipo de chatbot representa una tecnología que permite a usuarios de diferentes perfiles interactuar con datos complejos de manera intuitiva, formulando preguntas en lenguaje natural y recibiendo respuestas precisas en tiempo real.

- El prototipo de chatbot demuestra cómo pueden desarrollarse herramientas de inteligencia artificial basadas en datos públicos nacionales, facilitando el acceso a la información y reduciendo barreras técnicas.

Bibliografía

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (O'Reilly Media (ed.)).

DANE. (2023). *Ficha metodológica Sistema de Información de Precios y Abastecimiento del Sector Agropecuario Componente de Abastecimiento de Alimentos SIPSA – A*. https://www.dane.gov.co/files/operaciones/SIPSA/fmet-SIPSAbastecimiento.pdf

DANE. (2024). *Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA)*. https://www.dane.gov.co/index.php/estadisticas-por-tema/agropecuario/sistema-de-informacion-de-precios-sipsa

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Microsoft. (2016). *Team Data Science Process (TDSP)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Weizenbaum, J. (1966). ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 36–45.