Instituto Tecnológico y de Estudios Superiores de Monterrey

Reporte Final

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 501)

Equipo:

1

Integrantes:

Suemy Aquino Zumaya A00828585

Javier de Golferichs García A01139500

Franco Quintanilla Fuentes A00826953

Camila P. Cusicanqui Padilla A00571258

Jesús David Núñez Rodríguez A01634928

Facundo Vecchi A01283666

Ricardo Andres Arriaga Quezada A01570553

Emilia Victoria Jácome Iñiguez A00828347

Campus Monterrey 2 de diciembre 2022

Resumen

Las aplicaciones de la Inteligencia Artificial en la industria es una amplia gama de posibilidades. EL alto impacto que puede tener esta rama de la ciencia de datos en las problemáticas socio-económicas ha llevado a la necesidad de que cada vez más Instituciones Gubernamentales decidan incorporar herramientas de Inteligencia Artificial en sus sistemas con la finalidad de procesas la gran cantidad de datos demográficos de su población que se almacena en sus bases de datos. Esta tarea puede ser muy extensa en términos de procesamiento, análisis, interpretación y subsecuente-mente en la toma de decisiones a nivel municipal, estatal o inclusive nacional.

Por esta razón, es imprescindible para las instituciones gubernamentales de poder dar un acceso fácil, rápido y eficaz a la población a sus datos demográficos con la finalidad de poder ser transparentes con su población inclusive a aquellas personas que no tienen una formación técnica o especializada en manejo de datos e información. Con eso en mente, hemos propuesto una solución que responde directamente a dicha necesidad en el que se pretende incorporar herramientas de vanguardia para proveer un producto mínimo viable y funcional al consumidor que permitirá obtener información de manera rápida, segura y directa a cualquier tipo de usuario. En colaboración con el ***INEGI*** se utilizará la base de datos del SCINCE el cual describe los datos del Censo de Población y Vivienda de 2020 a nivel Nacional.

La propuesta consiste en el agrupamiento y visualización de la base de datos antes mencionada, con ayuda de procesamiento de lenguaje natural (o sus siglas en inglés - NLP), la cual realiza un query en base al modelo de detección, donde posteriormente el interfaz mostrará una visualización de la respuesta encontrada por el bot. Esto con la finalidad de facilitar el proceso de consulta de información al usuario final que utilice la página web del INEGI y que desee obtener cifras sobre aspectos demográficos tales como: cantidad de personas en un determinado municipio/estado con un desglose por género y grupo de edad. Dicha visualización se presenta en formato de gráficas sobre la información demográfica de la población mexicana por Entidad Federativa y Municipio.

Índice

1. Metodología

1.1. Propósito

El propósito de este documento es definir todo requerimiento funcional y no funcional del sistema y además plasmar sus restricciones. De esta manera, se puede visualizar con claridad el sistema antes de la etapa de desarrollo.

Se propone la implementación de un ChatBot en la página del INEGI para que a través de la escritura o la voz con el procesamiento de lenguaje natural, se puedan mostrar datos y gráficas concretas de lo que se pide para facilitar el acceso y el manejo del conjunto de datos.

1.2. Alcance

El alcance para este documento es mostrar una guía clara para empezar a desarrollar el proyecto. Se espera que todas las funcionalidades y restricciones del sistema queden explícitas, al igual que la forma en la que pueden interactuar entre ellas. Este documento proporciona un bosquejo general de lo que se espera para el proyecto. Sin embargo, no define el diseño final del sistema.

1.3. Definiciones, acrónimos y abreviaciones

Concepto	Definición
API	Es una interfaz de programación de aplicaciones, las cuales son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos. [1]
ChatBot	Es un asistente que se comunica con los usuarios a través de mensajes de texto, o voz. [2]
NLP	El Procesamiento del Lenguaje Natural, es el área de estudio centrada en cómo los ordenadores entienden el lenguaje humano, lo interpretan y procesan. [3]

Tabla 1: Tabla de definiciones, acrónimos y abreviaciones.

1.4. Referencias

El documento está basado en el formato SRS 830 de la IEEE con la versión 1 del punto 3.

1.5. Metodología de Proyecto Aplicada

La metodología de trabajo que se aplicó a este proyecto corresponde a CRISP.

1.6. Herramientas

Las herramientas que se estarán utilizando en este proyecto corresponden a Apache Spark dado que la INEGI almacena sus datos en este formato, para la parte del proceso del lenguaje natural se utilizará tanto Neuraan como Amazon Lex para aprovechar al máximo lo mejor de ambas APIs, finalmente para la realización del front end, se realizará en React.

2. Documentación

2.1. Vista General

La solución de nuestro se divide en tres roles: frontend, backend y nuestra base datos de la INEGI.

La fig. 1 muestra el funcionamiento básico de nuestra aplicación. Primero, el usuario genera un request por medio del frontend, este mensaje se manda al backend donde con el API de Neuraan podemos determinar la intención del usuario y determinar los queries requeridos para la base de datos. La información pedida se regresa al frontend para responder y graficar la respuesta de la pregunta del usuario.

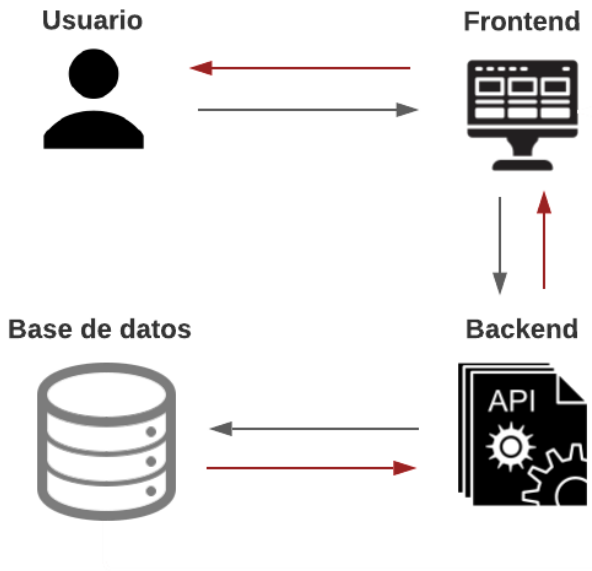

Figura 1: Diagrama de flujo de la solución del reto

El propósito de nuestra solución es facilitar el acceso y manipulación de la base de la INEGI. Por está razón, creamos un chatbot para que el usuario pueda solicitar esta información.

Hay cuatro componentes básicos del chatbot:

- Chatbot espacio del usuario: recuadro donde se ingresa el input
- Botón de chatbot: al momento de que el usuario ubica en el botón aparece el recuadro completo del chatbot.
- Barra lateral:
- Espacio para gráficas: recibe la información necesaria de la base de datos para graficar la respuesta para el usuario.

A continuación se muestran algunas capturas de pantalla que muestra el Frontend que se ha desarrollado para la implementación del Chatbot denominado como INEGIBOT dentro de la página web

de la INEGI con una interfaz amigable para los usuarios que deseen realizar consultas a las bases de datos.

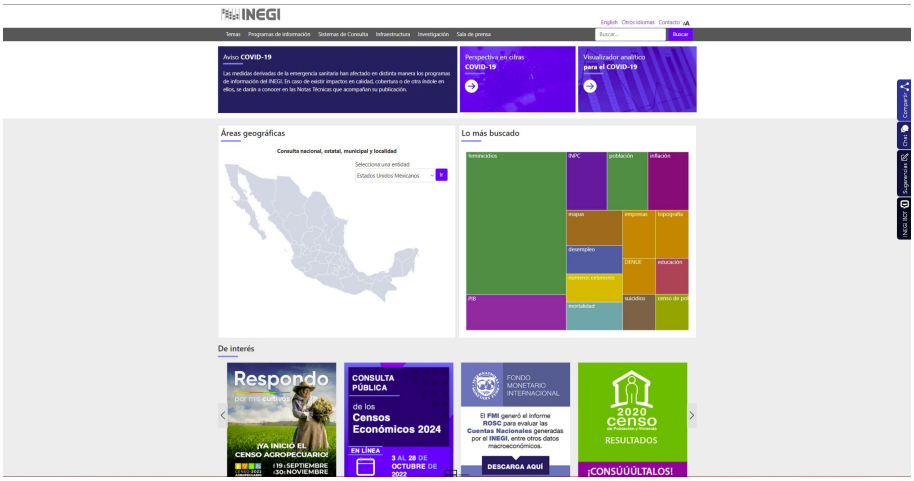

Figura 2: Vista general página de INEGI con INEGIBOT comprimido

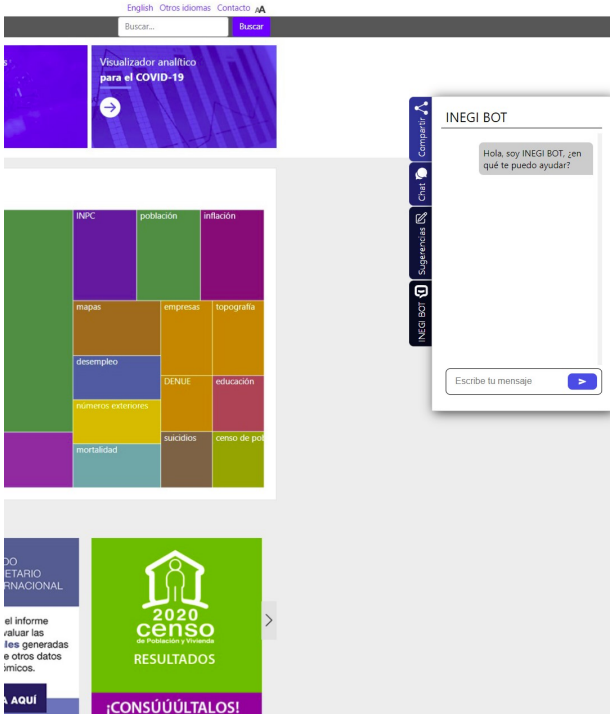

Figura 3: Ubicación de INEGIBOT dentro de la página de INEGI

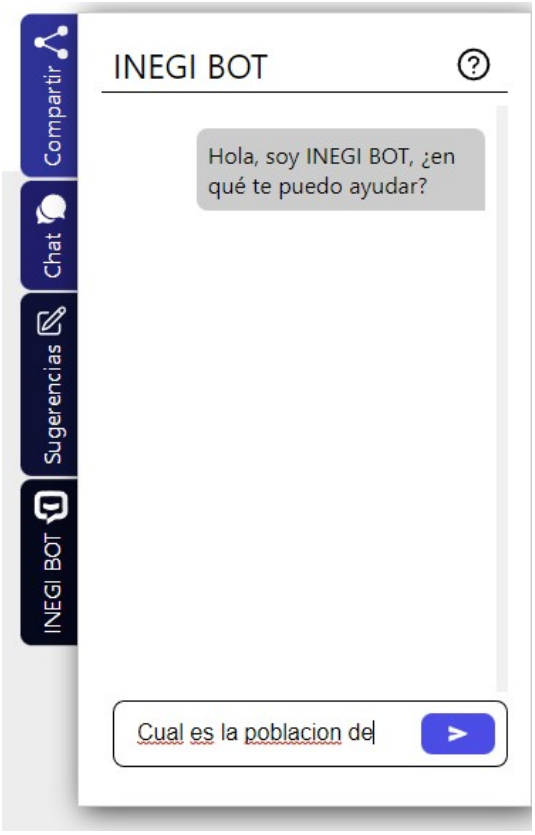

Figura 4: Saludo inicial de INEGIBOT y usuario escribiendo su consulta

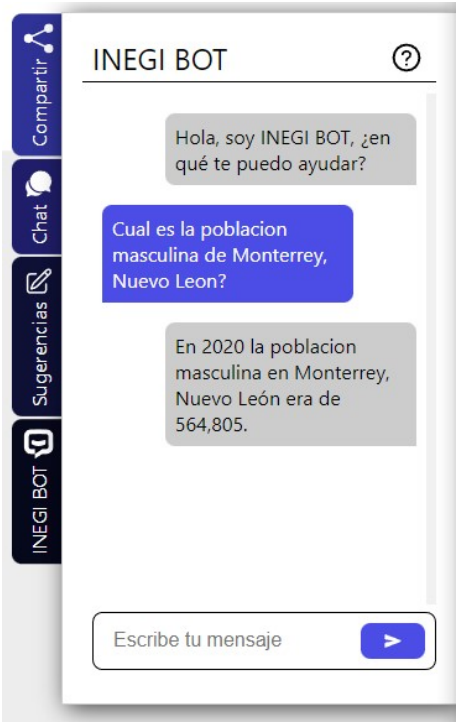

Figura 5: INEGIBOT conversación simple INEGIBOT

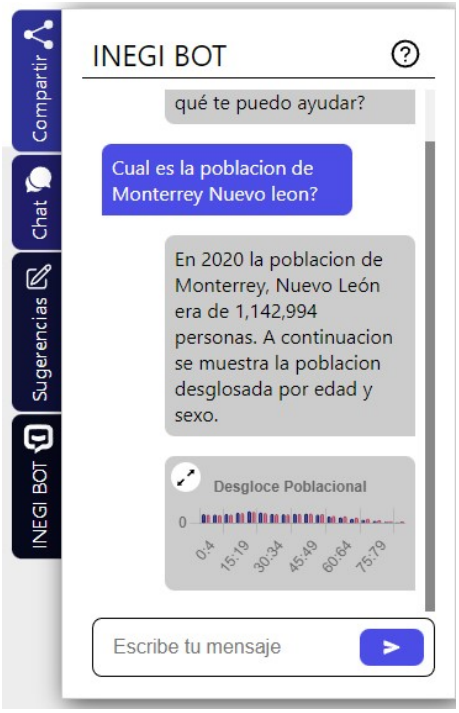

Figura 6: Respuesta a consulta en forma de gráfica por INEGIBOT

Cabe recalcar que el desarrollo actual únicamente muestra un prototipo de la forma de respuesta que se pretende dar al usuario. La información mostrada en estas gráficas únicamente tiene propósitos ilustrativos para dar visibilidad del tipo de respuestas que proporcionará el chatbot por lo que no reflejan los datos reales de la base de datos.

3. Datos

3.1. Big Data

Juzgando por la cantidad de datos con los que se tiene, se podría considerar que si se deberán aplicar técnicas de Big Data, así como de visualización y análisis de los aplicados a los datos del INEGI para poder obtener resultados relevantes a un propósito en específico. Para lo cuál se recomienda altamente utilizar la herramienta de Tableau para manejar las visualizaciones de la data.

Por otro lado, para el tema de manipulación de los datos y análisis de la información, se recomienda altamente utilizar la herramienta de Python con las librerías de Pandas, Numpy y Matplot para poder aplicar modelos de ML que permitan procesar la información. Si se busca una solución más centralizada donde se puedan realizar ambas cosas desde una sola interfaz, la recomendación es utilizar la herramienta de Dataiku que tiene más aplicaciones integradas en una misma interfaz.

3.2. Alcance

Sin embargo, para propósitos del desarrollo de este proyecto, no se aplicarán dichas técnicas ya que el alcance se limita principalmente a colocar una extensión a la página actual del INEGI con la opción de un chatbot integrado para que el usuario pueda obtener más rápidamente la información

que necesita sin tener que recurrir a realizar una consulta en lenguaje técnico. Subsecuente-mente, se manejaran los datos sin analizarlos para un contexto específico a manera de ejemplo.

3.3. NLP

El uso de NLP (Natural Language Processing) no se puede considerar como Big Data ya que es el área encargada de la interacción entre humano y computadora, la cual funcionara para obtener la solicitud del usuario y filtrar los datos proporcionados por el INEGI.

3.4. Deep Learning

Durante este proyecto no se incluyó la aplicación ni el uso de técnicas de Deep Learning dado que lo único que se realizó fue una implementación de un chatbot pre-entrenado con algoritmos internos de propiedad de Neuran. Dicho chatbot fue construido a través de la API de la empresa Neuran por lo que no se tuvo manera de modificar el algoritmo base que usa este para responder ante los intents del usuario. De modo que, para fines de este entregable en específico, no fue necesario programar un modelo de Deep Learning desde cero.

4. Modelo

Para el desarrollo del modelo, se utilizo la API de **Neuraan**, en la cual por motivos de protección de datos y confidencialidad de la información de su empresa, no nos pudieron dar más detalle del funcionamiento de su algoritmo, lo cual hace perfecto sentido por la protección de sus datos e inteligencia.

5. Evaluación

5.1. Métricas de Desempeño

Para poder hacer mediciones sobre las distintas métricas de desempeño del modelo, se realizan las siguientes acciones:

- Tiempo de respuesta del bot.
- Tiempo de cómputo.
- Respuesta del backend para realizar la query y desplegar la información del bot.
- Score del accuracy de los intends (nivel de seguridad del bot al momento de encontrar el intend a un match preestablecido).

5.2. Evaluación del Modelo

- El modelo actualmente tiene una precisión de 89 %

El modelo funciona con la API de **Neurann**, conectada al Backend de la implementación, la cual toma una entrada de tipo JSON, que con uso de procesamiento de lenguaje natural toma la entidad *(municipio y estado de la requisición)*, así como el filtro *(sexo y grupo de edad)*, tiene como salida para

el usuario se obtiene un query en el que despliega la respuesta de población y su gráfica complementaria con el desglose por edad y sexo.

5.3. Desempeño

- Se realizó una comparación con las respuestas obtenidas a los queries al solicitarlas en el buscador de Google.
- Se debe tomar en consideración que el tiempo de respuesta del bot, también considera la inicialización del backend, la cual una vez ya esta inicializada, se reduce considerablemente su tiempo de respuesta.
- Se encontró que cuando se hace una requisición el modelo tarda aproximadamente de 12 a 15 segundos en dar la salida. Mientras que cuando se hace una segunda o más con el backend ya corrido esta tarda entre 1 y 2 segundos de dar respuesta.

6. Data Ethic

Para el caso de la ética de los datos, en este proyecto nos enfocamos en 3 distintos factores para crear un chatbot con una buena estrategia de ética de datos, como lo son:

- Practicas justas
- Compilación de datos
- Confianza

Esto beneficia al chatbot y a los usuarios, ya que a la hora de generar un chatbot con una estructura transparente con los usuarios generar beneficios a los consumidores a la hora tomar decisiones, ya que nuestros datos son de uso publico y de interés global.

7. Áreas de oportunidad

Por ultimo se resaltaran algunas observación y recomendaciones de mejora que se detectaron en el proyecto:

1. A pesar de que la aplicación cumple con su función, existen algunas mejoras en el flujo de uso y la entrada de datos.
2. Actualmente el bot no cuenta con reconocimiento de flujos por lo que el usuario debe de ingresar una instrucción completa y correcta para que el bot pueda reconocer la instrucción ingresada.
3. Así mismo la entrada de datos tiene lineamientos establecidos para generar una consulta. Estos lineamientos a su vez tienen una limitante que no permita tener mas de un filtro en la búsqueda limitando cierto tipo de búsquedas.
4. Otra observación tiene que ver con la interacción de las gráficas ya que creemos que que se podrían implementar distintos tipos de gráficas como gráficas de barras agregadas.

Referencias

[1] *Cómo crear una API pública con AWS*. URL: https://aws.amazon.com/es/what-is/api/.

[2] Ramón Peris. *Chatbot: ¿Qué es, para qué sirve y cómo funcionan?* Feb. de 2021. URL: https://bloo.media/blog/por-que-implementar-chatbot-en-tu-estrategia-de-marketing/.

[3] Vive Unir. *¿Qué es el NLP y para qué sirve?* Oct. de 2021. URL: https://www.unir.net/marketing-comunicacion/revista/nlp-procesamiento-lenguage-natural/.

Anexos

A. Tutoriales seguidos

Para el desarrollo del proyecto fue fundamental el seguimiento de los tutoriales otorgados por el INEGI para su correcta descarga y uso:

Parte 1:
https://abxda.medium.com/geo-big-data-desde-cero-parte-1-e48d7ac2c7eb

Parte 2:
https://abxda.medium.com/geo-big-data-desde-cero-parte-2-85121f5d654d

Parte 3:
https://abxda.medium.com/geo-big-data-desde-cero-parte-3-a5180a86133d

Ejemplo:
https://abxda.medium.com/auto-machine-learning-geo-f540406c525a

Neuraan API:
https://neuraan.com

B. Repositorio de GitHub

Liga al repositorio de GitHub: https://github.com/facund015/TC3007C-501-Equipo1-Reto