NLP INEGI

Plan inicial basado en la metodología CRISP-DM

AUTOR:

Equipo 2

María del Carmen Vargas V.	**A0082850**
Alejandro Díaz Carrillo	**A01650603**
Rodrigo René Henríquez Paguaga	**A00827198**
Antonio Patjane Ceballos	**A01657978**
Ingrid Giselle Paz Ramírez	**A00826973**
Gerardo del Valle Cuellar	**A01284200**

VERSION: 0.0.0

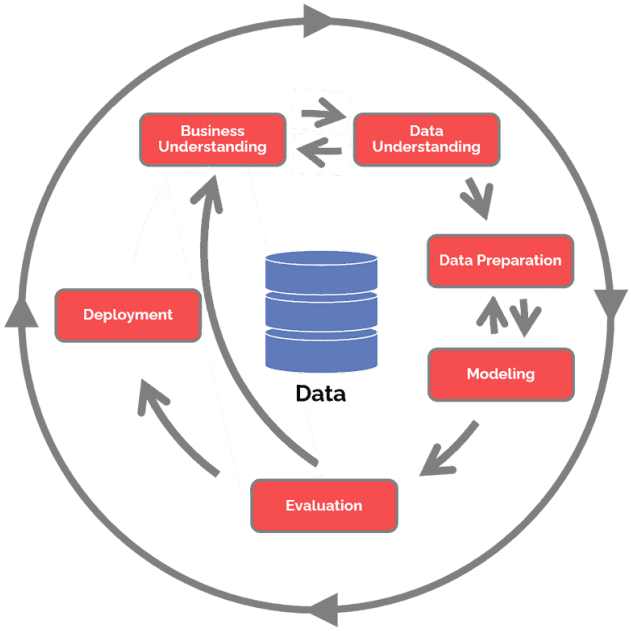
05/10/2022

VERSION HISTORY				
VERSION	**APPROVED BY**	**REVISION DATE**	**DESCRIPTION OF CHANGE**	**AUTHOR**
V.0		05/10/2022	Creación del documento	Equipo 2

Tabla de contenidos

1. Introducción

Para la solución a la problemática se optó por seguir la metodología CRISP-DM (Cross-industry standard process for data mining). Está metodología sirve como base para los procesos de ciencia de datos y cuenta con 6 fases secuenciales mostradas a continuación:



Trabajar bajo CRISP-DM nos permite tener muchas ventajas a lo largo de nuestra planeación e implementación de la solución:

- Flexibilidad: Facilidad de alterar decisiones a medida que las necesidades de la empresa cambian o en caso de cometer errores inicialmente.
- Estrategia a largo plazo: La metodología permite iterar sobre el proyecto para implementar mejoras y cambios de la estrategia a medida que el ciclo continúa.
- Entendimiento: Al centrarse en el entendimiento del negocio, tanto en la etapa inicial como la final, se evalúa si realmente la implementación se enfoca a las necesidades del negocio.

2. Business Understanding

Business Understanding

A partir de las sesiones con el socio formador (INEGI), se definió la problemática y se comprendieron las necesidades del negocio. Con los datos del INEGI es necesario buscar una solución que garantice que los ciudadanos puedan acceder y hacer uso de la información con la que se cuenta, debido a la alta desinformación sobre las capacidades que ofrece esta institución y sus herramientas.

3. Data Understanding

Data Understanding

Se cuentan con 32 bases de datos que corresponden a cada uno de los estados de la república mexicana. Cada una de las bases contiene 618 variables a nivel manzana, las cuales contienen categorías con información estadística y geográfica de interés nacional. Para el proyecto, se seleccionará una sola base de datos, y se reducirá el alcance hacía una sola categoría de información, la cuál queda por definir. Sin embargo, ya tenemos pensado darle un enfoque en el sector económico.

4. Data Preparation

Data Preparation

Los datos ya se encuentran listos para el análisis, ya que el INEGI realizó la labor de limpieza. Para poder utilizarlos son necesarias algunas herramientas:

- Spark
- SQL
- QGis

Estas nos sirven para poder hacer la selección de los datos que van a ser utilizados para el modelo.

5. Modeling

Modeling

La idea es implementar un asistente virtual dentro de un sistema de mensajería (chatbox). Para esto es necesario que el software de este asistente virtual reconozca y haga sentido de los mensajes que el usuario envíe. Por esta razón, gran parte del proyecto se dedicará a utilizar la técnica de NLP en español ya implementada en una API ofrecida por los servicios de Neuraan, una empresa desarrolladora de software, y conectarla con la base de datos del INEGI para que pueda permitirle al asistente virtual realizar la búsqueda de datos que pida el usuario.

6. Evaluation

Evaluation

Los resultados de la implementación del modelo se evalúan generando escenarios de prueba. Se verificará que para cada escenario, el modelo identifique la pregunta, realice la búsqueda dentro de la base de datos y

muestre al usuario la respuesta dentro del periodo de tiempo determinado. A partir de los casos de prueba se calcularán métricas para evaluar el desempeño del chatbot.

7. Deployment

Deployment
Al haber recorrido el ciclo de la metodología CRISP-DM varias veces, se planea publicar la implementación con el objetivo de ser utilizado por la empresa socio formadora. Este se verá implementado en la página inicial del INEGI como un chatbot en donde los usuarios podrán realizar búsquedas de información de acuerdo a sus necesidades.