Examen

Nombre: Ibarra (Apellido Paterno) Ramírez (Apellido Materno) Sergio (Nombre(s)) **Fecha**: Mayo de 2023

Calificación: _____

Objetivo: *El participante identificará los objetivos y fases del modelo CRISP-DM para un caso de estudio proporcionado por el participante.*

Instrucciones:

1. Llenar cada uno de los recuadros siguientes explicando el caso de estudio a tratar, así como sus objetivos y los objetivos de minería de datos.
2. Llenar cada uno de los recuadros explicando cada una de las fases del modelo CRISP-DM para el caso de estudio elegido por el participante.

Explicar el caso de estudio: **Valor 1 punto**

Históricamente se ha dado importancia al pronóstico de demanda de hidrocarburos ya que son base de la economía, transporte y manufactura de cada país, por ello resulta de vital importancia contar con estudios y proyecciones sobre la demanda de dichos bienes. En este sentido se propone generar un sistema de pronóstico y recomendación de demanda de gas natural para el caso de México

Objetivos del caso de estudio: **Valor 1.5 puntos**

El principal objetivo es contar con varios métodos que permitan llevar a cabo un pronóstico de demanda de gas natural en los sectores eléctrico, petrolero y residencial para los siguientes 12-18 meses con el menor error posible.

Objetivos de la minería de datos: **Valor 1.5 puntos**

El principal objetivo de la minería de datos sería contar con los datos históricos de variables predictoras asociadas con demanda de gas natural (por ejemplo, datos históricos de precio de importación del hidrocarburo, del cambio peso-dólar, del número de habitantes, del número de casas, etc). La idea es contar con una base histórica limpia y normalizada de datos desde 2005 y hasta 2022 de las diferentes variables seleccionadas, tratar datos faltantes o *outliers* de la mejor manera posible y determinar las variables más significativas usando seguramente la técnica de Principal Component Analysis (PCA)

Fases del modelo CRISP_DM

1. Comprensión del negocio **Valor 1 punto**

Es importante estudiar que variables o situaciones externas suelen impactar con mayor grado a la cantidad de gas natural que se demanda en los diferentes sectores en México, Será necesario entender la dinámica de la compra-venta de gas, las condiciones en que se llevan a cabo las negociaciones y entender cualquier factor que pueda afectar a nuestra variable objetivo, por ejemplo, será interesante determinar si parámetros no cuantitativos como la política energética nacional, el tipo de gobierno (derecha o izquierda), entre otras tienen un efecto en el valor del gas demandado

2. Comprensión de los datos **Valor 1 punto**

Se llevará a cabo revisión de la literatura y trabajos pasados que aborden el tratamiento de datos similares para comprender como trataron datos faltantes, outliers, datos categóricos, variables dummy, etc, así como la dimensión de los datos

3. Preparación de los datos **Valor 1 punto**

Será necesario una ardua búsqueda y limpieza de datos históricos de las variables predictoras de la cantidad de gas demandado. Se piensa que aquellos datos económicos históricos como el PIB, el precio de importación serán los más relevantes, sin embargo, también habrá que explorar la posible inclusión de variables de tipo cualitativos (como la política energética nacional) que permitan tener una base de datos sólida para entrenar a los modelos.

Se pretende que todas las tablas estén en su forma Normal y que se trate de manera adecuada los faltantes y outliers (Media histórica u omisión del dato) así como decidir una técnica adecuada de codificación de variables categóricas (e.j. *Hot-Encoding*). Por último, será importante verificar si para alguno(s) de los modelos empleados requieren que los datos estén estandarizados, en cuyo caso se aplicará de igual manera una transformación matemática a nuestras variables predictoras del tipo: xi = xi-mu / Xmax- Xmin

4. Modelado **Valor 1 punto**

La idea principal es utilizar por una parte modelos lineales generalizados (GLM por sus siglas en ingles) y por otro lado Redes Neuronales (NN) que permitan modelar a la cantidad de gas demandado a tiempo futuro, como una función de las variables predictoras más significativas y las relaciones entre ellas.

En el caso de los modelos GLM: Se deberá determinar que las variables predictoras no tengan una relación lineal grande entre ellas y se deberá verificar posibles “variables dummy” que permitan manejar predictores categóricos, así como decidir que coeficientes formarán parte del modelo final, con base en la significancia que éstos tengan.

En el caso de las NN: Se probarán diferentes tipos y arquitecturas de Redes Neuronales como del tipo CNN y Recursivas, etc , se entenderá las ventajas y desventajas de cada tipo respecto al pronóstico de demanda de gas natural

5. Evaluación **Valor 1 punto**

Se evaluarán los modelos haciendo pronósticos de los últimos 12 -18 meses de datos históricos conocidos y evaluando diferentes tipos de errores (MAE, MAPE, RMSE) buscando elegir el modelo que menor error arroje. Además se evaluará el balance dificultad del modelo vs likelihood de probabilidad en resultado con criterios como AIC o BIC con el objetivo de calificar y ponderar a los diferentes modelos tanto de GLM como de NN en sus diferentes aspectos.

Se pretende discutir todos los resultados y evaluaciones con el equipo de expertos en Gas Natural, quienes darán guía y opinión sobre lo obtenido, así como posibles mejoras en el modelo, pronósticos, datos de entrenamiento y/o consideraciones

6. Despliegue y explotación **Valor 1 punto**

Una vez construidos los modelos, se pretende tener un ‘ambiente de producción más amigable’ y además replicable en el futuro, que permita al usuario alimentar la base de datos nuevamente, es decir, proveer a los datos de train con los valores que vayan siendo registrados en 2023, 2024, etc para ser capaces de pronosticar valores futuros solamente cambiando el conjunto de datos de entrenamiento.

Valor total 10 puntos