DIPLOMADO Ciencia de Datos

Examen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nombre**: | Ibarra | Ramírez | Sergio | **Fecha**: | Mayo de 2023 |
|  | Apellido Paterno | Apellido Materno | Nombre(s) |  |  |
|  | **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 y es necesario para los países, contar con estudios y proyecciones sobre la demanda de dichos bienes. En este sentido se propone generar un sistema de pronóstico y recomendación de compra de gas natural para el caso de México | |
| **Objetivos del caso de estudio:** | **Valor 1.5 puntos** |
| El principal objetivo es contar con uno o 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, 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 | |
| 1. **Comprensión de los datos** | **Valor 1 punto** |
|  | |
| 1. **Preparación de los datos** | **Valor 1 punto** |
| Será necesario una ardua búsqueda y limpieza | |
| 1. **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) | |
| 1. **Evaluación** | **Valor 1 punto** |
| Se evaluarán los modelos haciendo pronósticos de los últimos 12 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 | |
| 1. **Despliegue y explotación** | **Valor 1 punto** |
|  | |

**Valor total 10 puntos**