**Optimización de las inversiones y planes de mantenimiento sobre los circuitos eléctricos de media tensión en Enel Colombia.**

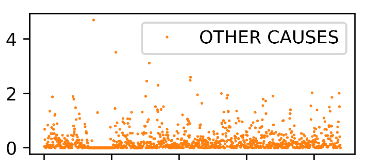
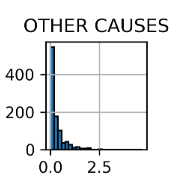
1. Estadísticas descriptivas de los datos “limpios”, acompañado de una descripción de las mismas y del proceso de limpieza de datos. Recuerden describir el tipo de variables con las que cuentan, sus estadísticos descriptivos, qué hicieron con los datos faltantes, qué hicieron con las variables categóricas, etc. Incluyan en su descripción una comparación detallada entre la base de datos original y la base de datos final.

Se analizó un dataset con 1.097 registros, en los cuales el tratamiento que se realizó para los datos faltantes fue el rellenado con ceros de acuerdo a criterio de experto, dado que se ubicaban en variables comerciales y de indicadores.

El dataset cuenta con 122 variables de las cuales 8 son categóricas y 114 de tipo numérica. Las variables categóricas no se trasformaron ya que se trataban de variables técnicas de etiquetado y ubicación geográfica que no aportarían al modelo.



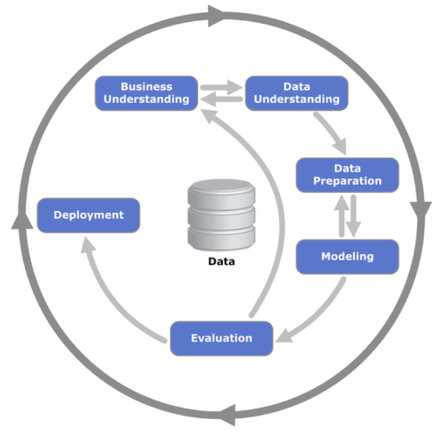
Se analizaron las distribuciones de las variables numericas, encontrado variedad y necesidad de escalar los datos ya que se veian diferentes magnitudes, aquí algunos ejemplos:

Para más detalle del analisis descriptivo de los datos abrir el siguiente link:

<https://bit.ly/3U230Mv>

1. Basándose en la retroalimentación que recibieron de su propuesta, describan en detalle el algoritmo que plantearon utilizar.

  
Fig. X Fases del modelo de referencia CRISP-DM[2]

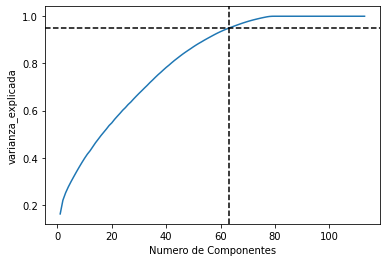
El ciclo de vida CRISP-DM consiste en 6 áreas de conocimiento (Identificación y entendimiento del problema de negocios, entendimiento de los datos, preparación de los datos, modelamiento, evaluación y despliegue). El presente trabajo se enfocará en las cinco (5) primeras áreas que permitirán determinar las acciones de EnEl Colombia como parte del descubrimiento de patrones de falla en el sistema de distribución eléctrica.

Para el modelamiento, la metodología consiste en:

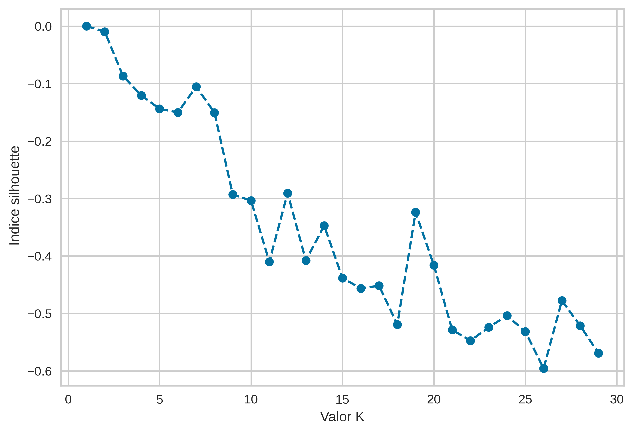
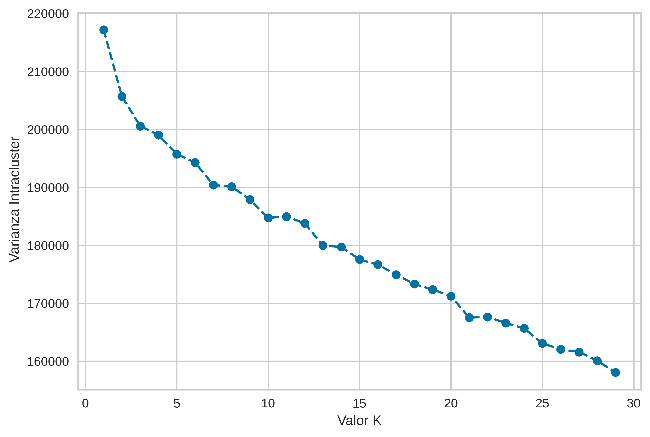
* Estandarización de los datos
* Aplicación de análisis de Componentes principales
* Determinación de numero de componentes principales
* Análisis de clusterizacion a través de K -Means
* Selección de clústeres
* Evaluación de desempeño
* Selección operacional de escenarios basados en los siguientes criterios: Patrón de falla normal con alto impacto en el punto final de uso; patrón de falla normal con bajo impacto en el punto final de uso; Patrón de falla alto con bajo impacto en el punto final de uso y Patrón de falla alto con alto impacto en el punto final de uso.

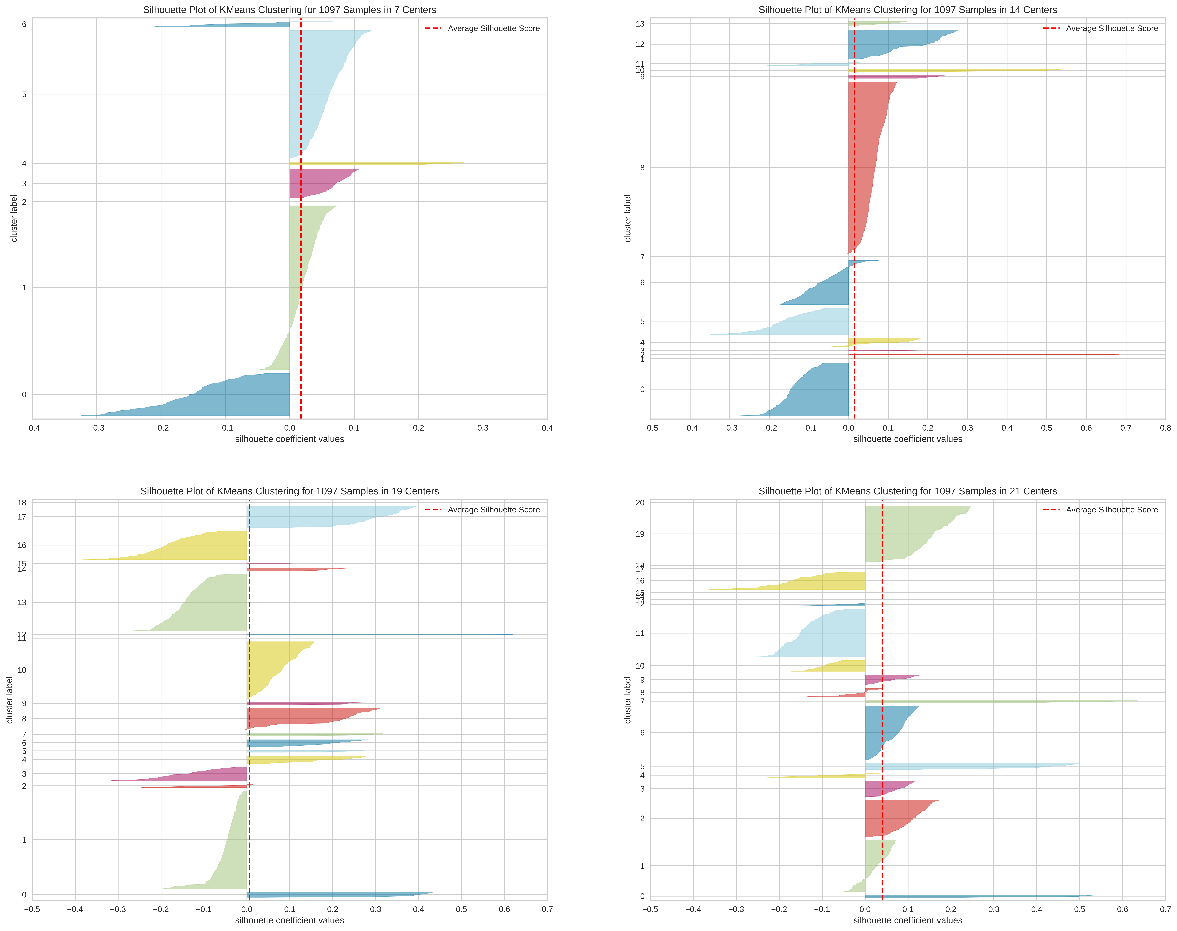
1. Haciendo uso de los datos limpios, implementen el algoritmo propuesto y presente los resultados preliminares utilizando tablas y/o visualizaciones. Describan los criterios de decisión que tomaron para obtener y elegir los resultados expuestos. Por ejemplo: si utilizaron PCA, deben describir qué criterio(s) utilizaron para escoger el número de componentes principales.

Luego de escalar los datos para lograr la estandarización se aplica el análisis de componentes principales PCA.



Encontrando que el número de componentes que explican el 95% de la varianza explicada es 64, se disminuyen en un 56% la cantidad de variables.





1. Bibliografía