Escuela de Ciencias Aplicadas
Maestría en Ciencias de los Datos y Analítica
Minería de Datos para grandes volúmenes de Información
Proyecto

Diana Lisette Arango Cañas
Juan Felipe Renza Chavarría

Noviembre 15, 2023

1 Pregunta de Investigación y Objetivos

¿Es factible inferir las relaciones entre el consumo energético de un edificio, su área y la temperatura ambiente a través de un modelo de clusterización?

Objetivo General

Analizar el consumo de energía utilizando modelos de clustering para entender la relación entre el consumo de energía de edificios, su superficie y la temperatura ambiente.

Objetivos Específicos

1. Transformar un dataset que contenga la información del consumo de energía de edificios para adecuar los datos de tal forma que habilite el uso de modelos de clustering.
2. Investigar las posibles influencias de la temperatura ambiente en el consumo energético de los edificios.
3. Investigar las posibles influencias de la superficie del edificio en su consumo energetico.
4. Concluir e interpretar los resultados de acuerdo al objetivo general propuesto.
5. Disponer en una plataforma la transformación de los datos, adicionalmente los algoritmos de aplicación de los modelos de clustering, para su posterior implementación y control de cambios.

2 Revisión de Literatura, estado del arte y bibliografía

Se realiza la búsqueda en la base de datos scopus y web of science utilizando el siguiente query: (TITLE-ABS-KEY (clustering) AND TITLE-ABS-KEY (energy AND consumption) AND TITLE-ABS-KEY (building))

En esta búsqueda se resaltan dos documentos por la similaridad que tienen con el desarrollo de este proyecto y la reciente publicación de ambos, el primero, *Cluster analysis of energy consumption mix in the Japanese residential sector*(Delage and Nakata, 2023) y el segundo *Changes in energy use profiles derived from electricity smart meter readings of residential buildings in Milan before, during and after the COVID-19 main lockdown*(Ferrando et al., 2023), a continuación se detalla el contenido de estos dos artículos:

- **Changes in energy use profiles derived from electricity smart meter readings of residential buildings in Milan before, during and after the COVID-19 main lockdown:** Este estudio tiene como objetivo entender los cambios en los hábitos de consumo de energía durante los confinamientos por el COVID-19, se basa en resultados de artículos previos donde se ha determinado que las estaciones del año tienen relevancia en el análisis de los consumos de energía, es por lo que en el set de datos se identifican periodos como invierno, verano, confinamiento y otoño. En la transformación de los datos para este estudio se realiza una separación en dos set de datos diferentes, uno con la información de los días de "trabajo" y otro con la información de los datos del fin de semana, sábados y domingos para aplicar independientemente el modelo de clustering.
  El modelo de clustering implementado es K-means con 300 iteraciones, el resultado entrega un número de clusters de 4 a 10 dependiendo de la época del año y el tipo de día, ver figura 1, lo cual concuerda con los resultados de casos similares explorados en la literatura.

Final number of clusters for each period.

Period label		2019	2020
Winter	Working days	8	5
	Weekends	5	4
Lockdown	Working days	10	10
	Weekends	10	10
Autumn	Working days	7	6
	Weekends	7	7

Figure 1: Resultados clustering Energía Milan

  Finalmente, se crean curvas de consumo de energía para para todos los edificios ubicados en un mismo cluster, donde el eje x, son las horas del día y el eje y, es el consumo de energía para encontrar un consumo de energía medio para cada cluster y así poder comparar los cambios entre los años 2019 y 2020.

- **Cluster analysis of energy consumption mix in the Japanese residential sector:** Este estudio tiene como objetivo mejorar la precisión en la división de los tipos de clientes para diseñar de forma más adecuada los futuros sistemas inteligentes de energía. La forma tradicional de catalogar los consumidores en residencial, comercial, industrial y transporte es identificándolos por su demanda de energía, sin embargo, se entiende que en el valor de dicho consumo están involucrados múltiples factores económicos, culturales, sociales, etc. Los datos consisten en información de 9505 hogares de todo japón de los cuales se tiene el consumo de energía, gas, gasolina, queroseno y diesel. Adicionalmente, se tiene datos de la ubicación geográfica, edad e ingresos de los miembros del hogar, año de construcción del edificio, superficie, número de vehículos entre otros.
  El enfoque novedoso de este estudio consiste en realizar la agrupación en clusters basados en la similaridad de valores de consumo de los diferentes tipos de energía y posteriormente utilizar la clasificación realizada por el modelo de clustering para identificar cuales de las variables socioeconómicas, geográficas y características de la vivienda son parámetros representativos de cada cluster, es decir cuales parámetros de los hogares son similares en cada cluster, esto se realiza a través de la significancia estadística del p-value. Otro aspecto importante de este estudio implica la utilización del método de clustering DBSCAN jerárquico o hierarchical DBSCAN algorithm (HDBSCAN), para llevar a cabo la agrupación este método se fundamenta en el clustering espacial basado en densidad, diseñado para manejar la presencia de ruido o datos atípicos.

Finalmente, se realiza la citación de artículos adicionales que valdría la pena explorar con mayor profundidad (Ahmad et al., 2017), (Liu et al., 2021), (Meenal et al., 2021), (Sadorsky, 2021).

3 Metodología de Investigación CRISP-DM

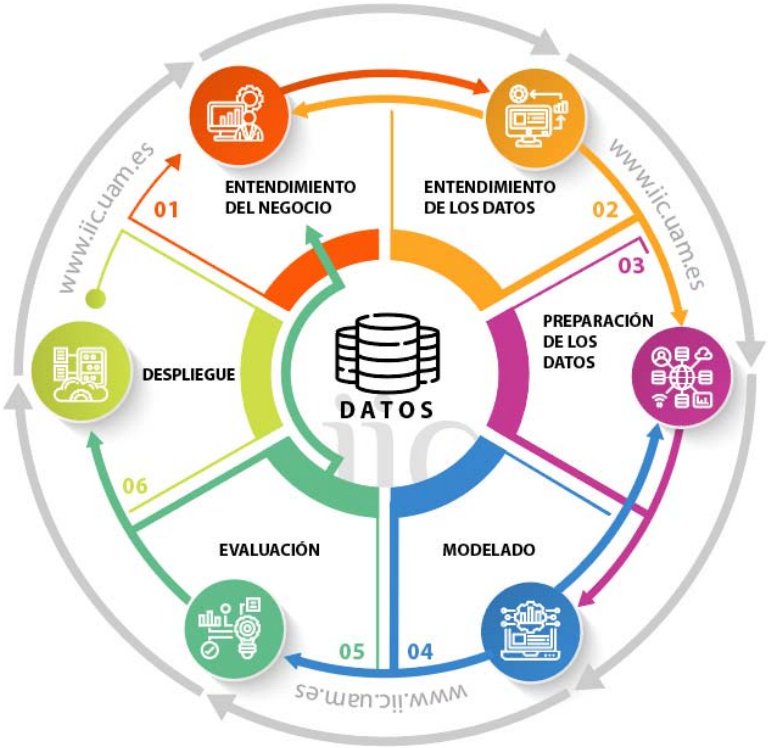

Figure 2: Metodología de Investigación

Inicialmente, se inició el proceso de comprensión del negocio. Los datos corresponden a consumos de energía de edificaciones, con diferentes tipos de granularidad. Es importante desarrollar metodologías y modelos que permitan estimar correctamente los consumos energéticos, pues el crecimiento poblacional, a nivel mundial, genera un crecimiento en la demanda; por lo tanto, el desempeño de la estimación podrá entregar a los usuarios del modelo ventajas competitivas o financieras dependiendo del uso que se le quiera dar a la predicción. (Somu et al., 2021). Una vez se inició el proceso de comprensión de los datos, fue fácilmente validar que iba a ser necesario una proceso de limpieza inicial, pues la información está separada en diferentes archivos *.csv*, adicionalmente, la frecuencia en la que se tomaron los datos, tiene una granularidad diferente a través de los set de datos; por lo tanto, fue necesario estandarizar el conjunto final de forma consistente. Así, se aplicaron diferentes pasos para estructurar los set de datos y unificarlos en un último conjunto, que posteriormente sería usado para entrenar los modelos, con las columnas a utilizar. El modelado se realizó sobre el último set de datos, se implementaron modelos no supervisados, los cuáles fueron evaluados usando métricas adecuadas. Finalmente, al ser este un proyecto académico, no hay un despliegue del modelo en un proceso productivo, pero si se discuten posibles usos en la sesión de trabajos futuros.

4 Análisis de los datos

Los datos están compuestos por 5 archivos, los cuales contienen información de la serie de tiempo de consumos de energía de los edificios, información adicional relacionada con las temperaturas cercanas a estos e información de los días y fechas en los que fueron capturados los datos. En total se tienen más de 3 millones de registros con información de consumo de energía capturadas cada 15 minutos. Estos datos tienen un peso de 661.5 MB almacenados en AWS S3 en formato csv.
Los datos se obtuvieron en el siguiente link: Kaggle.com. A continuación una descripción de cada uno de los set de datos y el tipo de dato que se puede ver en cada una de las conlumnas de los conjuntos.

Historical Consumption

1. **obs_id (int):** ID de la observación.
2. **SiteId (int):** ID del edificio. Permite localizar los datos adicionales de cada edificación a través del datasets.
3. **ForecastId (int):** ID de la serie de tiempo.
4. **Timestamp (datetime):** Marca de tiempo al momento de tomar la medida.
5. **Value (float):** Medida del consumo de energia de la edificación.

Building Metadata

1. **SiteId (int):** ID del edificio. Permite localizar los datos adicionales de cada edificación a través del datasets.
2. **Surface (float):** Superficie de la edificación en Und2
3. **Sampling(int):** Cantidad de minutos entre cada observación.
4. **BaseTemperature (float):** Temperatura base de la edificación.
5. **IsDayOff (boolean):** True si DAY_OF_WEEK no es un día hábil.

Historical Weather Data

1. **SiteId (int):** ID del edificio. Permite localizar los datos adicionales de cada edificación a través del datasets.
2. **Timestamp (datetime):** Marca de tiempo al momento de tomar la medida.
3. **Temperature (Fahrenheit):** Temperatura medida en la estación climática.
4. **Distance (km):** Distancia entre la estación meteorológica y el edificio.

Public Holidays

1. **SiteId (int):** ID del edificio. Permite localizar los datos adicionales de cada edificación a través del datasets.
2. **Date (date):** Fecha del día feriado.
3. **Holiday (string):** Nombre del día feriado.

Es importante destacar que para efectos de este proyecto no se hizo uso de todos los set de datos disponibles. Únicamente usamos los conjuntos $test - data.csv$, $training - data.csv$, $metada.csv$ y $weather.csv$

4.1 Transformaciones de los datos:

El dataset de los consumos de energía de cada edificio tiene observaciones cada 15 minutos, es decir que para una misma hora de un determinado día se pueden tener hasta 3 datos, ver figura 3, en contraste los datos de la variable temperatura solo tiene un dato cada hora. Para lograr concordancia entre estos dos set de datos, se debe convertir la información de consumos de energía a un único dato para cada hora y esto se realiza a través de la mediana para evitar la influencia de los datos atípicos dentro de la muestra.

	obs_id	SiteId	Timestamp	ForecastId	Value
0	323604	235	2014-01-02T19:00:00+00:00	5004	157265.446409
1	2813181	235	2014-01-02T22:00:00+00:00	5004	155498.418922
2	4006999	235	2014-01-02T22:45:00+00:00	5004	155498.418922
3	106973	235	2014-01-03T00:30:00+00:00	5004	157265.446409
4	6793052	235	2014-01-13T14:15:00+00:00	5005	91885.429363

Figure 3: Consumos energía

Continuamos las transformaciones con el archivo de variables climáticas, en este la información de las temperaturas tomadas en las estaciones meteorológicas son observaciones que tienen estampas de tiempos irregulares, es decir el minuto donde se obtiene el dato no siempre es el mismo, ver figura 4. Para lograr concordancia entre los datos de consumos de energía y temperaturas se debe convertir la información de temperaturas a un único dato para cada hora y esto se realiza, al igual que con los consumos de energía, a través de la mediana.

	Timestamp	Temperature	Distance	SiteId
0	2017-03-03T19:00:00+00:00	10.6	27.489346	51
1	2017-03-03T19:20:00+00:00	11.0	28.663082	51
2	2017-03-03T20:00:00+00:00	6.3	28.307039	51
3	2017-03-03T21:55:00+00:00	10.0	29.797449	51
4	2017-03-03T23:00:00+00:00	5.4	28.307039	51

Figure 4: Temperaturas

Con el objetivo de mejorar la agrupación de los datos y los resultados finales basados en la información obtenida de la literatura explorada, se realizaron transformaciones adicionales a los datos. Estas consistieron en obtener valores únicos para cada día, es decir pasar de analizar los datos con frecuencia horaria a frecuencia diaria. La decisión no modifica el objetivo del proyecto y por el contrario mejora la interpretabilidad de los resultados.
La trasformación mencionada consistió en sumar los datos de consumo y el resultado dividirlo por el numero de horas del día, para obtener así un valor único valor para cada uno de los días en el set de datos.
De forma similar y para guardar la concordancia entre ambos set de datos, se agruparon los valores de temperatura en un dato único para cada día por cada uno de los edificios, para que llegar al valor nuevamente se utiliza la mediana.

El último paso de la transformación de los datos consiste en unir el conjunto de temperaturas, consumos e información del tamaño de los edificios ya que cada uno de estos se encuentran en archivos diferentes, esto se realiza a través de inner join. Con esta acción obtenemos el data set definitivo con el cuál se continuarán los análisis y depuraciones para aplicar modelo. Finalmente, se convierte en un data set de Pyspark para aprovechar las facilidades que ofrece en el manejo de grandes volúmenes de información ver figura 15,

SiteId	hourlyValue	Temperature	Surface
2	41099.85530538901	21.85	6098.278376070084
2	43842.76590953585	20.0	6098.278376070084
2	38438.75222184868	17.1	6098.278376070084
2	38593.308226343644	17.0	6098.278376070084
2	36562.415813026004	18.5	6098.278376070084
2	38627.456375728485	18.7	6098.278376070084
2	120814.59481115872	19.0	6098.278376070084
2	120990.88145617132	18.5	6098.278376070084
2	133063.28521485085	20.5	6098.278376070084
2	127743.95142526618	22.0	6098.278376070084
2	116607.74387337094	22.0	6098.278376070084
2	47835.657866690744	21.0	6098.278376070084
2	46833.26440622998	21.0	6098.278376070084
2	135316.43336072244	22.0	6098.278376070084
2	112236.11789576599	20.025	6098.278376070084
2	128526.67429291939	18.0	6098.278376070084
2	112607.53840120707	17.625	6098.278376070084
2	91462.38784571202	18.0	6098.278376070084
2	34947.73089887514	18.0	6098.278376070084
2	31343.537902560187	19.4	6098.278376070084

only showing top 20 rows

Figure 5: Datos

4.2 Descripción de los datos:

En esta sección analizaremos las características del dataset que se usará para aplicar el clustering. Las transformaciones realizadas en la etapa anterior tuvieron un efecto de mejoría en los datos debido a las agrupaciones de las variables consumo de energía y temperatura, por lo tanto el dataset no tiene valores nan ni valores nulos, ver figura 6 y 7.

```
valores_nan_por_columna.show()
```

SiteId	hourlyValue	Temperature	Surface
0	0	0	0

Figure 6: Datos

```
valores_nulos_por_columna.show()
```

SiteId	hourlyValue	Temperature	Surface
0	0	0	0

Figure 7: Datos

Finalmente, se tiene una descripción estadística de los datos con información de mínimos, máximos, desviación estándar, etc, aquí se observa que para la variable consumos se tiene un valor mínimo de cero sin embargo esto

no se considera como un valor incorrecto que se deba imputar ya que es un dato valido si se tienen en cuenta que en los primeros días cuando se inician a habitar los edificios estos consumos no están presentes, a continuación se muestran los datos mencionados, ver figura 8.

summary	SiteId	hourlyValue	Temperature	Surface
count	27043	27043	27043	27043
mean	24.026513330621604	77764.74078552173	15.071969640942221	8080.8110070801085
stddev	15.580142720736013	109143.54337082528	7.014628751213149	9561.62262334408
min	2	0.0	-11.0	14.884534437442198
25%	9	13198.277092868075	10.25	1032.735063490691
50%	22	30688.138890945993	15.0	6098.278376070084
75%	39	96966.13333420357	19.65	10985.29263411183
max	57	864531.5992355075	38.0	45941.71643756564

Figure 8: Datos

5 Uso de herramientas de Big Data

Teniendo en cuenta la necesidad de trabajar con grandes volúmenes de información, la arquitectura se diseñó de tal forma que los datos no estén almacenados en el mismo lugar del código. Existe una conexión con un **bucket S3** de AWS. Este servicio, está diseñado para almacenamiento pseudo infinito; por lo tanto, cualquier cantidad de información puede ser almacenada en el bucket y posteriormente acceder a esta a través de la conexión creada. Adicionalmente, el código tiene una mezcla entre PySpark y python, este último siendo usado únicamente para describir el modelo de forma gráfica. El procesamiento del código se realizó en un *Codespace* de GitHub, en dónde se desplegó una máquina virtual de 4 núcleos y 16GB de memoria RAM. Se tomó la decisión de utilizar esta arquitectura, pues en la plataforma Databricks, en su versión estudiantil, encontramos limitaciones para el almacenamiento de información. En el gráfico siguiente se observa la arquitectura utilizada.

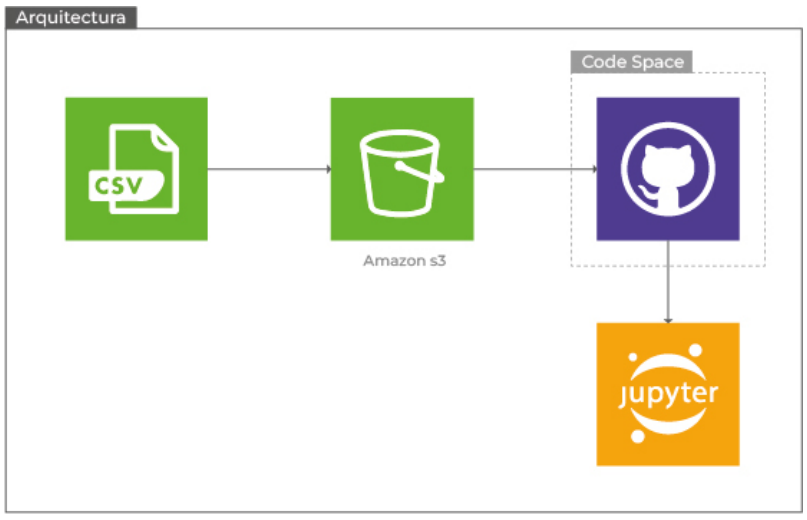

Figure 9: Arquitectura de Trabajo

6 Entregables y su descripción

El entregable está alojado en un repositorio público GitHub.com. La arquitectura se define como una mezcla entre **S3** y Jupyter Notebook. Los datos están almacenados en un bucket público de S3 llamado **mindatos-project** en el repositorio, el archivo **project.py** se encarga de hacer uso de las funciones definidas en **aws-s3** para acceder a la

información y almacenarla en un diccionario llamado **data_frames**.
A partir del diccionario, se asignan variables para cada uno de los conjuntos de datos. Después del procesamiento de datos, se realizaron tres modelos en los que se busca la agrupación teniendo en cuenta 3 variables: $Temperature, Value, Surface$

K-means: Scikit-learn

Haciendo uso del método del codo, se tomó la decisión de entrenar el modelo un $K = 4$; posteriormente, se hace uso del modelo Kmeans importado desde sklearn.cluster. Los resultados se grafican haciendo uso de la librería Seanborn, además del gráfico que representa el método del codo usado para elegir la cantidad de clusters.

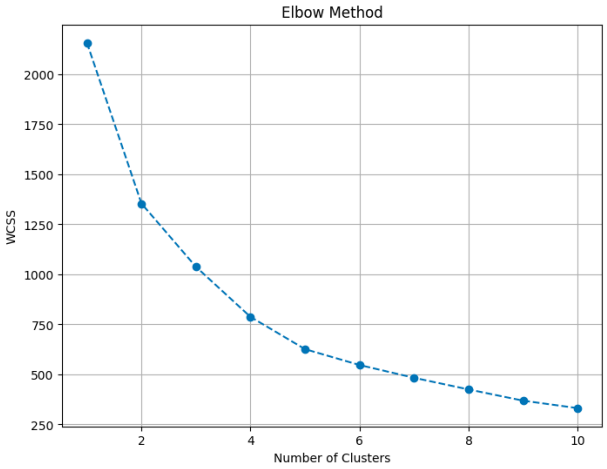

Figure 10: Método del codo

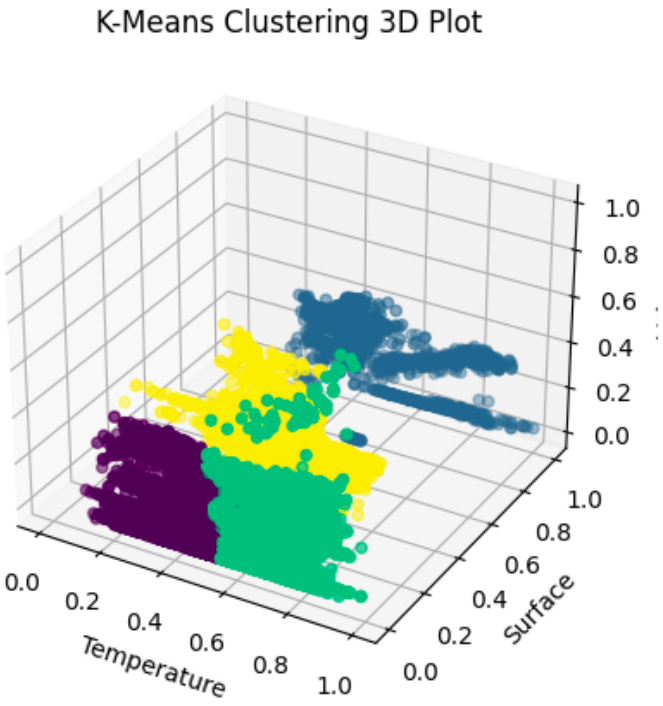

Figure 11: 3D plot: Clusters - Scikit-learn

K-means: PySpark

Haciendo uso del costo y de la silueta, se itera a través de los diferentes números de K para elegir el número óptimo. En este caso, se realiza el modelo haciendo uso de Kmeans importado desde pyspark.ml.clustering. En este caso, el número óptimo se considera $K = 6$. Nuevamente, se hace uso de la librería Seaborn para graficar los clusters y también, se muestra el gráfico del método del codo.

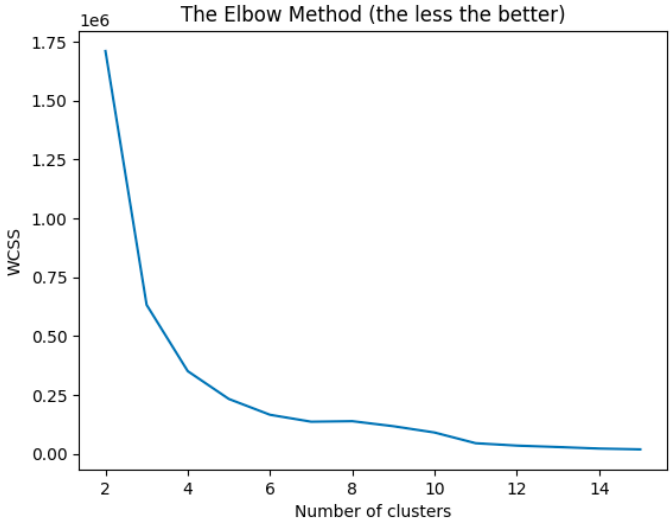

Figure 12: Método del codo

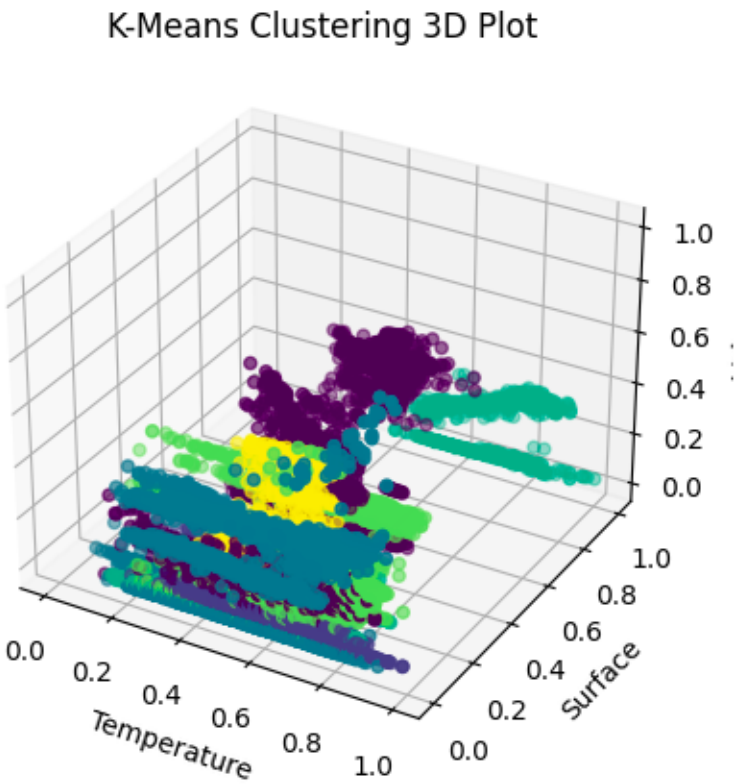

Figure 13: 3D plot: Clusters - pySpark

Por último, se calcula la silueta, la cual, obtuvo un valor $silueta = 0.7566$

Gaussian Mixture

En uno de sus paper, Li et al. (2020) hace uso de una Gaussian Mixture para incrementa el desempeño de la predicción de consumos energéticos. Adicionalmente, también se menciona en la literatura, el uso de $GaussianMixtureModels$ es común para dar solución a problemas no supervisados de agrupación (Mirzal, 2022). Es por eso, que para este proyecto, se implementa también este modelo con el objetivo de observar sus resultados. Importando desde sklearn.mixture, se usó el modelo GaussianMixture y posteriormente se graficó la cantidad de clusters contra el número BIC. Los resultados pueden ser observados a continuación:

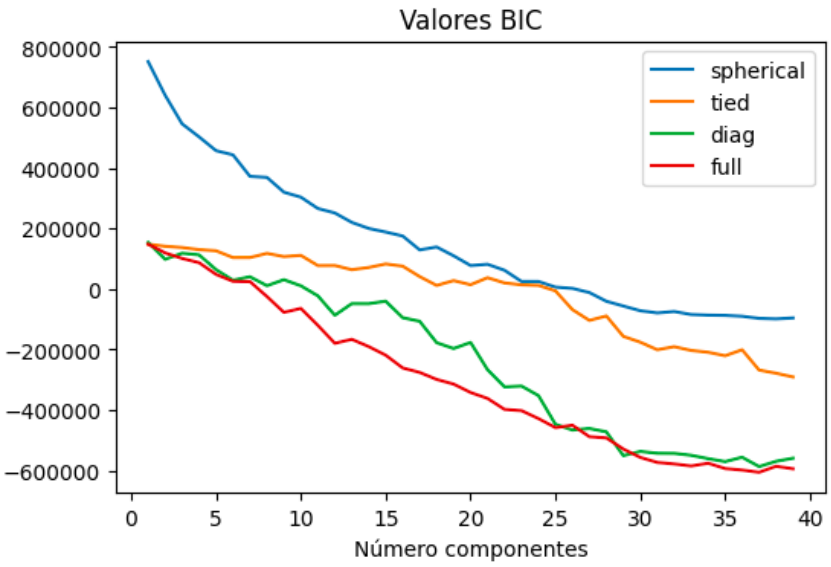

Figure 14: Gaussian Mixture - Valores BIC

De acuerdo a los resultados obtenidos, se entrega el modelo con $n\_components = 6$ con lo que se obtiene un log de probabilidad predicha.

7 Conclusiones y trabajo futuro

Conclusiones

- Los resultados del modelo de kmeans desde pyspark aplicados para generar agrupaciones de los edificios según el consumo de energía están basados en la superficie del mismo, dejando a un lado la variable temperatura, lo que indica que esta no es un factor decisivo en el modelo.
- En el modelo de kmeans de la librería skitlearn tiene relevancia la variable temperatura, particularmente genera una clasificación para los edificios de valores de superficies pequeños, para los otros 2 clusters no es relevante, sin embargo esta clasificación no se ve adecuada debido a que los valores de consumo no cambian por los cambios de temperatura.
- A partir de la agrupación de los datos se observa que la variable temperatura tiene un influencia poco significativa en el resultado, por lo tanto se puede evaluar la eficiencia y pertinencia de tener un registro tan frecuente de esta información.

- Se considera que la cantidad más adecuada de clusters es 6,esto se define a través de las métricas de costo y silueta las cuales tienen mejor desempeño con esta cantidad de agrupaciones y adicionalmente coincide con uno de los artículos evaluados en la revisión bibliográfica de este mismo tipo clustering.
- Se encontró como la mejor estrategia el uso de un bucket de S3 para alojar los datos debido a su gran volumen, ya que adicionalmente permite acceder con facilidad a los mismos desde diferentes herramientas como databricks o github. También presenta ventajas en el manejo sencillo e intuitivo de la herramienta.
- Las herramientas como aws o databricks tienen importantes restricciones para el manejo de grandes volumenes de información en la modalidad estudiantil o gratuita, para el desarrollo de este proyecto se utilizó github ya que garantizó el consumo y procesamiento de los datos.
- Por medio de la aplicación del modelo guassiano se evidencia que asumir los datos con geometría esférica y por lo tanto aplicar una medida de distancia euclidiana no es la mejor forma de agrupar los datos, por lo tanto el kmeans puede mejorarse al cambiar la medida de distancia.

Trabajos Futuros

- En un trabajo posterior se podría probar un modelo de clustering eliminando la variable temperatura e incluyendo como variables los días de la semana para entender si este es un factor determinante, adicionalmente separar los data sets como fue realizado en el articulo *Changes in energy use profiles derived from electricity smart meter readings of residential buildings in Milan before, during and after the COVID-19 main lockdown* (Ferrando et al., 2023)
- Posterior a la implementación de este modelo se pueden tomar los resultados para etiquetar el set de datos, con el objetivo de implementar un modelo supervisado que permita la predicción del consumo energético.
- Con el objetivo de obtener un mejor desempeño en la agrupación de los datos, se propone aplicar el modelo de clustering haciendo una intervención a la medida de distancia, cambiando la distancia euclidiana que supone datos esféricos por una que se ajuste mejor a la geometría de los datos.

8 Ejecución del plan

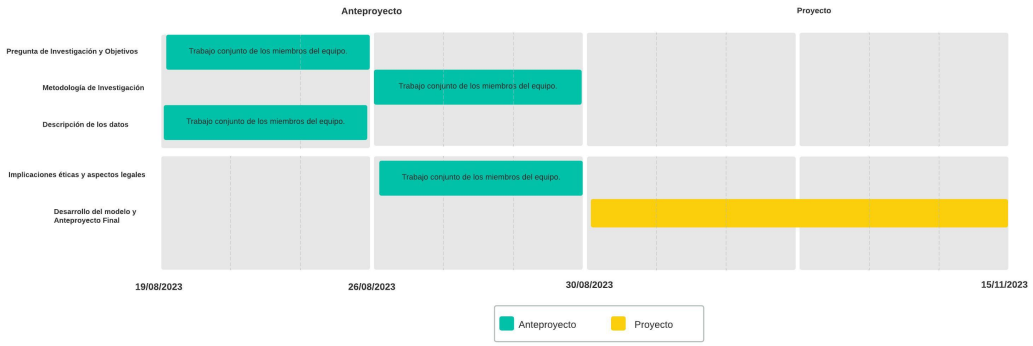

Figure 15: Hoja de ruta

En la figura anterior se muestra un diagrama con los pasos del anteproyecto y del proyecto. La ejecución real, del proyecto, representada en el diagrama con una línea amarilla, comprende un análisis más detallado de los datos y de las transformaciones necesarias para su posterior modelado. Uno de los retos más significativos del proyecto fue precisamente la limpieza de los datos, ya que los diferentes conjuntos no contenían la misma frecuencia de

medición; por lo tanto, fue necesario estandarizar los datos, en el sentido de ponerlos todos en una misa frecuencia, para poder relacionarlos. Teniendo en cuenta las aproximaciones que se habrian hecho de forma distinta, tenemos que el conjunto de datos tenía diferentes variables adicionales que en esta implementación no fueron usadas en el modelo de clustering. Algo que no se tuvo en cuenta fue la cantidad de tiempo que iba a consumir la limpieza de los datos y las limitaciones de las versiones gratuitas de las plataformas de Big Data.

9 Implicaciones éticas

En un proyecto de análisis de consumo energético en edificios, es importante considerar varias implicaciones éticas:

1. **Privacidad de Datos:** La recopilación y análisis de datos de consumo energético podría implicar información sensible de los residentes o usuarios de los edificios. Es necesario asegurarse de manejar los datos de manera anónima y cumplir con las regulaciones de privacidad correspondientes.

2. **Sesgos y Equidad:** Los modelos predictivos podrían estar influenciados por sesgos si los datos históricos reflejan desigualdades en el consumo energético entre diferentes grupos demográficos. Es esencial abordar estos sesgos y garantizar que las recomendaciones no perpetúen inequidades.

3. **Transparencia y Explicabilidad:** Los modelos utilizados para predecir el consumo energético deben ser transparentes y comprensibles para los usuarios y partes interesadas. Las decisiones basadas en estos modelos pueden tener un impacto significativo, por lo que la claridad en su funcionamiento es esencial.

4. **Consentimiento Informado:** Si los datos se recopilan directamente de los residentes o usuarios de los edificios, es fundamental obtener su consentimiento informado. Deben entender cómo se utilizarán sus datos y tener la opción de negarse sin consecuencias negativas.

5. **Impacto Ambiental y Social:** Las recomendaciones para mejorar la eficiencia energética podrían tener impactos en el entorno y en las personas. Es esencial considerar estos impactos y encontrar un equilibrio entre la eficiencia y el bienestar de las personas.

10 Aspectos legales y comerciales

Los datos utilizados y resultados obtenidos en este proyecto son considerados como privados y anónimos, por lo tanto no se deben compartir los mismos con ninguna empresa privada, evitando así el ofrecimiento de propuestas de productos o servicios que lleven a una mayor eficiencia en el consumo de energía. Así mismo, se recomienda acompañar la interpretación de los resultados de este trabajo por profesionales especializados en eficiencia energética que entreguen acciones viables y costo eficientes.

References

Ahmad, M. W., Mourshed, M., and Rezgui, Y. (2017), "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, 147.

Delage, R. and Nakata, T. (2023), "Cluster analysis of energy consumption mix in the Japanese residential sector," *Smart Energy*, 12.

Ferrando, M., Banfi, A., and Causone, F. (2023), "Changes in energy use profiles derived from electricity smart meter readings of residential buildings in Milan before, during and after the COVID-19 main lockdown," *Sustainable Cities and Society*, 99.

Li, T., Wang, Y., and Zhang, N. (2020), "Combining Probability Density Forecasts for Power Electrical Loads," *IEEE Transactions on Smart Grid*, 11.

Liu, Y., Chen, H., Zhang, L., and Feng, Z. (2021), "Enhancing building energy efficiency using a random forest model: A hybrid prediction approach," *Energy Reports*, 7.

Meenal, R., Michael, P. A., Pamela, D., and Rajasekaran, E. (2021), "Weather prediction using random forest machine learning model," *Indonesian Journal of Electrical Engineering and Computer Science*, 22.

Mirzal, A. (2022), "Statistical Analysis of Microarray Data Clustering using NMF, Spectral Clustering, Kmeans, and GMM," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19.

Sadorsky, P. (2021), "A Random Forests Approach to Predicting Clean Energy Stock Prices," *Journal of Risk and Financial Management*, 14.

Somu, N., R, G. R. M., and Ramamritham, K. (2021), "A deep learning framework for building energy consumption forecast," *Renewable and Sustainable Energy Reviews*, 137.