Universidade do Rio de Janeiro
Pós-Graduação em Modelagem Computacional
Aprendizagem de Máquina
Trabalho 1

Nome do Autor: Guilherme Rodrigues Cler Gabril

Data: 11/10/2025

1 Descrição dos Dados

A base de dados utilizada neste trabalho, é a **Steel Industry Energy Consumption**, que representa o consumo energético de uma indústria da Coreia do Sul, no qual o objetivo é classificar o tipo de carga elétrica (*Load Type*) em três categorias: *Light Load*, *Medium Load* e *Maximum Load*.

Os dados foram coletados ao longo do tempo e contêm medições de energia, potência, fator de potência, emissões de CO_2 e informações temporais (dia da semana e status do dia).

1.1 Variáveis de Entrada (Features)

- **date**: Data da medição (não utilizada na modelagem).
- **Usage_kWh**: Consumo de energia da indústria (kWh).
- **Lagging_Current_Reactive.Power_kVarh**: Potência reativa atrasada (kVarh).
- **Leading_Current_Reactive_Power_kVarh**: Potência reativa adiantada (kVarh).
- **CO2(tCO2)**: Emissão de dióxido de carbono associada ao consumo (ppm).
- **Lagging_Current_Power_Factor**: Fator de potência na condição de corrente defasada (%).
- **Leading_Current_Power_Factor**: Fator de potência na condição de corrente adiantada (%).
- **NSM**: Número de segundos desde a meia-noite (s).
- **WeekStatus**: Indica se o dia é de semana (1) ou fim de semana (0).
- **Day_of_week**: Dia da semana (Sunday, Monday, ..., Saturday).

1.2 Variável Alvo (Classe)

- **Load_Type**: Tipo de carga elétrica observada no momento da medição, podendo assumir três classes:
  - *Light Load*
  - *Medium Load*
  - *Maximum Load*

1.3 Resumo Teórico

O conjunto de dados caracteriza um problema de **classificação**, onde as variáveis contínuas e categóricas são utilizadas para prever o tipo de carga elétrica registrada (**Load_Type**).

1.4 Análise Técnica dos Dados

Durante o processo de preparação dos dados, algumas etapas de verificação e limpeza foram realizadas antes da aplicação dos algoritmos de aprendizado de máquina.

1.4.1 Seleção de Variáveis

As variáveis **date**, **WeekStatus** e **Day_of_week** foram removidas do conjunto de dados, pois estão relacionadas apenas a informações de tempo e dia da semana, não apresentando influência direta sobre o tipo de carga elétrica (**Load_Type**).

1.4.2 Balanceamento das Classes

O conjunto de dados é composto pelas seguintes quantidades de amostras para cada classe:

Classe	Quantidade de Amostras
Light Load	18072
Medium Load	9696
Maximum Load	7272

Tabela 1: Distribuição das classes no conjunto de dados.

Com base nesses valores, observa-se que o conjunto de dados é desbalanceada.

1.4.3 Valores Nulos e Duplicados

Foi realizada uma verificação de valores ausentes e registros duplicados no conjunto de dados. Nenhum valor nulo ou duplicado foi encontrado, garantindo a integridade dos dados para a análise.

1.4.4 Detecção de Outliers

A identificação de valores atípicos (*outliers*) foi realizada por meio de gráficos do tipo *boxplot*, aplicados às variáveis numéricas do conjunto de dados.

Inicialmente, foi gerado um boxplot considerando todas as variáveis, conforme mostrado na Figura 1. Nesse gráfico, observa-se que a variável **NSM** apresenta uma escala muito superior às demais, o que dificulta a visualização adequada dos possíveis outliers nas outras variáveis.

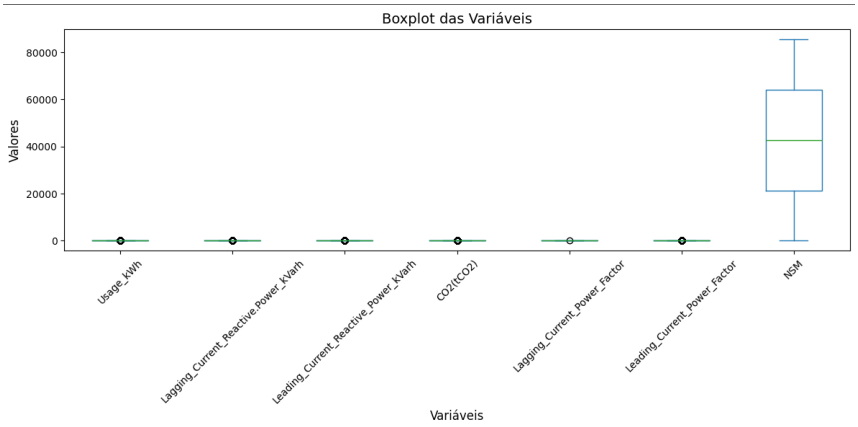

Figura 1: Boxplot com todas as variáveis numéricas, incluindo a variável **NSM**.

Para uma análise mais detalhada, foi gerado um segundo boxplot excluindo a variável **NSM**, apresentado na Figura 2. Essa visualização permitiu observar com mais clareza a presença de valores atípicos em algumas variáveis do conjunto de dados.

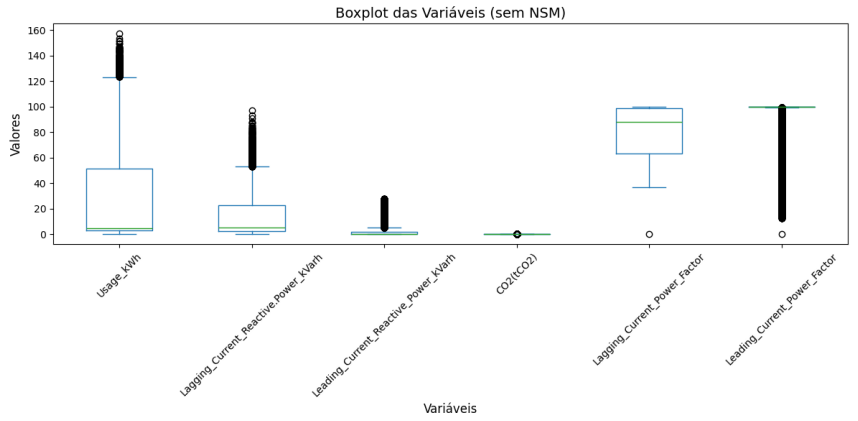

Figura 2: Boxplot das variáveis numéricas sem a variável **NSM**, evidenciando a presença de outliers.

Com base nas Figuras 1 e 2, verifica-se a existência de **outliers** em algumas variáveis. Esses valores, embora destoem do restante dos dados, foram mantidos para análise.

2 Análise dos Dados

2.1 Estatísticas Descritivas

A Tabela 2 apresenta a média, o desvio padrão e a variância das variáveis numéricas do conjunto de dados.

Atributo	Média	Desvio Padrão	Variância
Usage_kWh	27.3869	33.4444	1118.527
Lagging_Current_Reactive.Power_kVarh	13.0354	16.3060	265.886
Leading_Current_Reactive_Power_kVarh	3.8709	7.4245	55.1227
CO2(tCO2)	0.0115	0.0162	0.000261
Lagging_Current_Power_Factor	80.5781	18.9213	358.016
Leading_Current_Power_Factor	84.3679	30.4565	927.600
NSM	42750.0000	24940.5343	622030300

Tabela 2: Média, desvio padrão e variância das variáveis numéricas.

2.2 Matriz de Correlação

As Tabelas 3 e 4 apresentam as correlações entre as variáveis do conjunto de dados. Essa análise permite identificar relações lineares entre os atributos, destacando quais estão mais fortemente associados e quais são praticamente independentes.

Parte 1

	Usage_kWh	Lagging_Q	Leading_Q	CO2
Usage_kWh	1.0000	0.8962	-0.3249	0.9882
Lagging_Current_Reactive.Power_kVarh	0.8962	1.0000	-0.4051	0.8869
Leading_Current_Reactive_Power_kVarh	-0.3249	-0.4051	1.0000	-0.3328
CO2(tCO2)	0.9882	0.8869	-0.3328	1.0000
Lagging_Current_Power_Factor	0.3860	0.1445	0.5268	0.3796
Leading_Current_Power_Factor	0.3536	0.4077	-0.9440	0.3600
NSM	0.2346	0.0827	0.3716	0.2317

Tabela 3: Matriz de correlação (colunas de Usage_kWh até CO2).

Parte 2

	Lagging_PF	Leading_PF	NSM
Usage_kWh	0.3860	0.3536	0.2346
Lagging_Current_Reactive.Power_kVarh	0.1445	0.4077	0.0827
Leading_Current_Reactive_Power_kVarh	0.5268	-0.9440	0.3716
CO2(tCO2)	0.3796	0.3600	0.2317
Lagging_Current_Power_Factor	1.0000	-0.5200	0.5653
Leading_Current_Power_Factor	-0.5200	1.0000	-0.3606
NSM	0.5653	-0.3606	1.0000

Tabela 4: Matriz de correlação (colunas de Lagging_Current_Power_Factor até NSM).

Análise

Observa-se uma **forte correlação positiva** entre **Usage_kWh**, **Lagging_Current_Reactive.Power_kVarh** e **CO2(tCO2)**, indicando que o aumento do consumo de energia está diretamente relacionado ao aumento das emissões de CO e da potência reativa em atraso. Também há uma **forte correlação negativa** entre **Leading_Current_Reactive_Power_kVarh** e **Leading_Current_Power_Factor**, refletindo a relação inversa entre a potência reativa e o fator de potência. Por outro lado, variáveis

como **NSM** apresentam correlações baixas com **Usage_kWh**, **Leading_Current_Reactive_Power_kVarh**, **CO2**.

2.3 Valores Mínimos e Máximos

A Tabela 5 apresenta os valores mínimo e máximo das variáveis numéricas do conjunto de dados.

Atributo	Valor Mínimo	Valor Máximo
Usage_kWh	0.0	157.18
Lagging_Current_Reactive.Power_kVarh	0.0	96.91
Leading_Current_Reactive_Power_kVarh	0.0	27.76
CO2(tCO2)	0.0	0.07
Lagging_Current_Power_Factor	0.0	100.00
Leading_Current_Power_Factor	0.0	100.00
NSM	0.0	85500.00

Tabela 5: Valores mínimo e máximo das variáveis numéricas.

3 Metodologia

3.1 Seleção de Características

A seleção de características é um passo fundamental em aprendizado de máquina, pois permite identificar os atributos mais relevantes para a tarefa de predição, eliminando dados redundantes ou irrelevantes que podem aumentar o ruído e prejudicar o desempenho do modelo.

Para este trabalho, utilizou-se o método **SelectKBest**, que pertence à classe de métodos de seleção univariada. Este método avalia cada variável individualmente em relação à variável alvo (**Load_Type**) utilizando uma função de pontuação estatística. No contexto de classificação, a função de pontuação escolhida foi o *ANOVA F-value* (`f_classif`), que mede a relação linear entre cada atributo e a classe. Quanto maior o valor de F, mais relevante é o atributo para a discriminação entre classes.

O SelectKBest permite definir o número de características k a serem selecionadas. No presente trabalho, foram escolhidas **3 características** ($k = 3$), equilibrando a redução de dimensionalidade com a preservação da informação relevante para os clusters.

Variação de parâmetros testados:

- Número de atributos selecionados (k): 3.

3.2 Método de Agrupamento

Para agrupar as observações de acordo com padrões similares, foi utilizado o método de **Agglomerative Clustering**, uma técnica de clusterização hierárquica aglomerativa. Ao contrário de métodos de clusterização não hierárquicos (como K-Means), o Agglomerative Clustering constrói uma árvore hierárquica (*dendrograma*) representando a fusão dos clusters. Inicialmente, cada ponto de dado é considerado um cluster individual. Em cada iteração, os dois clusters mais próximos são unidos, repetindo o processo até que todos os pontos estejam agrupados em um único cluster ou até atingir o número desejado de clusters.

A proximidade entre clusters pode ser medida por diferentes estratégias de ligação (*linkage*):

- **Single linkage:** distância mínima entre qualquer par de pontos de dois clusters.
- **Complete linkage:** distância máxima entre qualquer par de pontos de dois clusters.

- **Average linkage:** média das distâncias entre todos os pares de pontos de dois clusters.
- **Ward linkage:** reduz a soma das variâncias dentro de cada cluster ao combinar clusters.

Para determinar os parâmetros que resultam no melhor agrupamento, foi utilizado um **ParameterGrid** com os seguintes valores:

- Número de clusters (*n_clusters*): 2 a 7.
- Tipo de ligação (*linkage*): *ward*, *complete*, *average*, *single*.

3.3 Critério de Validação

A avaliação de agrupamentos não supervisionados requer métricas que verifiquem a qualidade dos clusters obtidos. Neste trabalho, foi utilizado o **índice de Calinski-Harabasz**, também chamado de *Calinski-Harabasz Score*.

Este índice mede a razão entre a dispersão entre clusters (B_k) e a dispersão dentro de clusters (W_k):

$$CH = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)}$$

onde:

- $Tr(B_k)$ é a soma das distâncias quadráticas entre os centros dos clusters e o centro global dos dados.
- $Tr(W_k)$ é a soma das distâncias quadráticas dentro de cada cluster.
- k é o número de clusters.
- n é o número total de amostras.

Um valor mais alto de CH indica que os clusters são mais distintos e bem separados, enquanto um valor baixo sugere sobreposição entre clusters. Esta métrica é amplamente utilizada devido à sua simplicidade, interpretabilidade e sensibilidade à separação e coesão dos clusters.

4 Experimentos Computacionais

4.1 Análise das Componentes Principais

Para avaliar a influência de cada variável nas componentes principais, foi aplicada a técnica de **PCA (Principal Component Analysis)** aos dados padronizados. As três primeiras componentes explicam praticamente toda a variância do conjunto de dados, conforme os valores abaixo:

- Componente 1: 48,54% da variância
- Componente 2: 36,90% da variância
- Componente 3: 8,21% da variância

As duas primeiras componentes acumulam aproximadamente 85,44% da variância total, sendo suficientes para representar graficamente os principais padrões dos dados.

A Tabela 6 apresenta a contribuição (peso) de cada variável em relação às três componentes principais obtidas.

Variável	Componente 1	Componente 2	Componente 3
Usage_kWh	0.499	0.224	-0.116
Lagging_Current_Reactive.Power_kVarh	0.495	0.108	-0.213
Leading_Current_Reactive_Power_kVarh	-0.353	0.426	-0.293
CO2(tCO2)	0.499	0.219	-0.108
Lagging_Current_Power_Factor	0.036	0.563	-0.141
Leading_Current_Power_Factor	0.362	-0.418	0.296
NSM	0.014	0.467	0.858

Tabela 6: Contribuição das variáveis para cada componente principal.

Conforme apresentado na Tabela 6, observa-se que as variáveis Usage_kWh, Lagging_Current_Reactive.Power_kVarh e CO2(tCO2) têm forte contribuição na Componente 1. A Componente 2 é fortemente influenciada por Lagging_Current_Power_Factor, NSM e Leading_Current_Reactive_Power_kVarh, enquanto a Componente 3 é dominada principalmente pela variável NSM. Esses resultados indicam que o consumo energético e as potências reativas estão fortemente associados aos principais padrões de variação dos dados.

4.2 Parâmetros do Agglomerative Clustering

O método de agrupamento utilizado foi o **Agglomerative Clustering**, que combina iterativamente clusters mais próximos para formar agrupamentos hierárquicos.

A busca pelos melhores parâmetros foi realizada utilizando **ParameterGrid** com as seguintes variações:

- Número de clusters (*n_clusters*): [2, 3, 4, 5, 6, 7]
- Tipo de ligação (*linkage*): *ward*, *complete*, *average*, *single*

O melhor conjunto de parâmetros encontrado foi:

linkage = ward, n_clusters = 5

4.3 Validação dos Clusters

A qualidade dos clusters foi avaliada pelo **índice de Calinski-Harabasz**, que mede a razão entre a dispersão entre clusters e a dispersão dentro dos clusters. Quanto maior o valor, melhor a separação entre os clusters.

O valor obtido para os melhores parâmetros foi:

$$\text{Calinski-Harabasz Score} = 46570.53$$

4.4 Visualização com PCA

Para a visualização, os dados foram projetados nas duas primeiras componentes principais do PCA, permitindo observar a separação entre os 5 clusters encontrados.

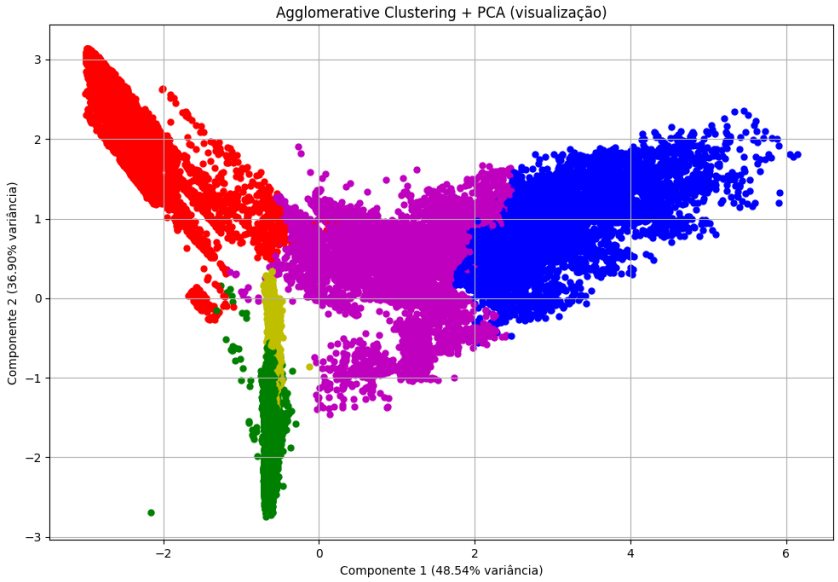

Figura 3: Visualização dos clusters obtidos pelo Agglomerative Clustering em duas componentes principais do PCA.

A Figura 3 mostra que os clusters apresentam boa separação visual, confirmando a validade do agrupamento obtido.

4.5 Segundo Agrupamento: Seleção de Características com SelectKBest

Neste segundo experimento, foi aplicada a técnica de seleção de características **SelectKBest** para reduzir o conjunto de dados às variáveis mais relevantes para a previsão do tipo de carga elétrica (**Load_Type**).

A função de pontuação utilizada foi o *ANOVA F-value* (`f_classif`), que avalia a relação estatística de cada atributo com a variável alvo. Foram selecionadas **3 características** mais relevantes:

- **Usage_kWh** — consumo energético da indústria
- **CO2(tCO2)** — emissão de dióxido de carbono
- **NSM** — número de segundos desde a meia-noite

4.5.1 Agglomerative Clustering com K-Best

O método de agrupamento permaneceu o **Agglomerative Clustering**, e foi realizado novamente um ParameterGrid para encontrar os melhores parâmetros:

- Número de clusters (*n_clusters*): [2, 3, 4, 5, 6, 7]
- Tipo de ligação (*linkage*): *ward*, *complete*, *average*, *single*

Os melhores parâmetros encontrados foram:

linkage = ward, n_clusters = 7

4.5.2 Validação dos Clusters

A avaliação da qualidade dos clusters foi realizada novamente pelo **índice de Calinski-Harabasz**. O valor obtido após a seleção de características foi:

$$\text{Calinski-Harabasz Score} = 64350.085$$

Observa-se que a pontuação aumentou em relação ao primeiro agrupamento (46570.53), indicando que a seleção de características contribuiu para formar clusters mais coesos e melhor separados.

4.5.3 Análise do PCA

Para entender melhor a estrutura dos dados e facilitar a visualização dos clusters, foi aplicada a técnica de **Análise de Componentes Principais (PCA)** às três variáveis selecionadas.

A Tabela 7 mostra a variância explicada por cada uma das três primeiras componentes principais:

Componente	**Variância Explicada (%)**
Componente 1	69.60
Componente 2	30.00
Componente 3	0.39

Tabela 7: Variância explicada pelas três primeiras componentes principais do PCA.

A soma das três componentes explica aproximadamente **100% da variância total**, indicando que o espaço reduzido mantém praticamente toda a informação original.

A Tabela 8 apresenta as contribuições das variáveis originais em cada componente principal:

Variável	**Componente 1**	**Componente 2**	**Componente 3**
Usage_kWh	0.677	-0.204	0.707
CO2(tCO2)	0.677	-0.207	-0.707
NSM	0.290	0.957	-0.002

Tabela 8: Contribuição das variáveis selecionadas para cada componente principal.

Conforme apresentado na Tabela 8, as variáveis `Usage_kWh` e `CO2(tCO2)` exercem influência predominante nas Componentes 1 e 3. Por outro lado, a variável `NSM` destaca-se como principal contribuinte na Componente 2.

4.5.4 Visualização com PCA

Para visualização, os dados selecionados foram projetados nas duas primeiras componentes principais do PCA. As duas componentes explicam uma proporção significativa da variância, permitindo observar claramente os 7 clusters obtidos.

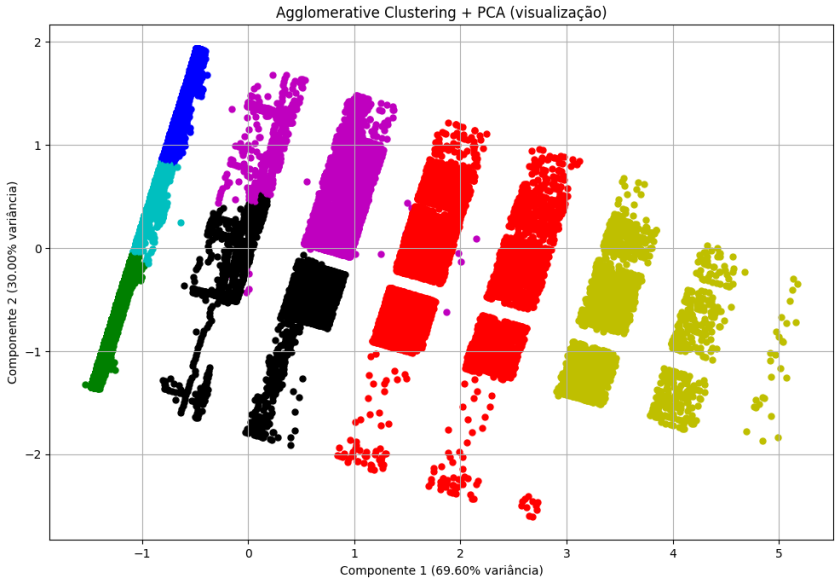

Figura 4: Visualização dos clusters após aplicação do SelectKBest, projetados nas duas primeiras componentes principais do PCA.

A Figura 4 evidencia que os clusters ficaram mais coesos e melhor separados, reforçando a importância da seleção de características na melhoria da clusterização.

5 Conclusão

A metodologia adotada, composta pela seleção de características (**SelectKBest**), agrupamento hierárquico (**Agglomerative Clustering**) e avaliação da qualidade dos clusters (**índice de Calinski-Harabasz**), mostrou-se eficaz na identificação de padrões estruturais nos dados.

Os experimentos revelaram que:

- Sem a seleção de características, o algoritmo encontrou 5 clusters com um Calinski-Harabasz Score de 46570, utilizando todas as 7 variáveis do conjunto de dados.
- Após aplicar o SelectKBest, selecionando as 3 características mais relevantes (*Usage_kWh*, *CO2(tCO2)* e *NSM*), o mesmo algoritmo encontrou 7 clusters, com uma melhoria significativa no Calinski-Harabasz Score (64350), indicando que os clusters ficaram ainda mais coesos e melhor separados.

Observa-se que, embora o conjunto de dados original possua 3 classes (*Light Load, Medium Load, Maximum Load*), o algoritmo não supervisionado identificou um número maior de clusters, sugerindo a existência de subdivisões ou variações internas nos dados que não correspondem exatamente às classes pré-definidas. Isso é esperado em tarefas de clusterização, onde os agrupamentos refletem similaridades intrínsecas nas variáveis analisadas, independentemente das classes originais.

Além disso, a análise de PCA mostrou que a maior parte da variância do conjunto de dados pôde ser representada pelas duas primeiras componentes principais, permitindo uma visualização clara dos clusters. A aplicação do SelectKBest contribuiu para reduzir a dimensionalidade e aumentar a separação entre os agrupamentos, evidenciando a importância da seleção de características em problemas de aprendizado não supervisionado.

Portanto, a metodologia adotada conseguiu identificar grupos bem definidos, ainda que em número superior ao das classes originais, fornecendo informações úteis sobre a estrutura dos dados e possíveis subcategorias de carga elétrica.

Referências

- **Base de Dados: Steel Industry Energy Consumption** — https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption
- **Agglomerative Clustering (Clusterização Hierárquica)** — https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
- **Calinski-Harabasz Score (Validação de Clusters)** — https://towardsdatascience-com.translate.goog/calinski-harabasz-index-for-k-means-clustering-evaluation-using-pytho?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=tc
- **SelectKBest (Seleção de Características)** — https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
- **PCA (Análise de Componentes Principais)** — https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- **ParameterGrid (Busca de Parâmetros)** — https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ParameterGrid.html