

# **Relatório Projeto Final — IAA 2025/2026**

Consumo Energético — Clustering e Previsão

110942 - Luís Lourenço | 101018 - Miguel Bento

## **Estrutura**

1. Business Understanding
2. Data Understanding
3. Data Preparation e Feature Engineering
4. Modeling (Clustering)
5. Evaluation (Previsão)
6. Conclusions

## **1 Business Understanding**

O Município da Maia disponibilizou dados de consumo energético de edifícios institucionais, recolhidos a cada 15 minutos. O objetivo deste projeto é analisar esses consumos de forma a apoiar a gestão energética municipal, recorrendo a técnicas de clustering e previsão.

Em particular, pretende-se:

- identificar grupos de edifícios com padrões de consumo semelhantes;
- caracterizar perfis energéticos típicos e detectar comportamentos atípicos;
- avaliar a previsibilidade do consumo futuro em diferentes tipos de edifícios;
- comparar modelos de previsão simples e avançados com um baseline semanal.

O sucesso do projeto é avaliado pela capacidade de obter clusters interpretáveis e previsões superiores ao baseline, com potencial utilidade para planeamento energético e optimização de custos.

## **2 Data Understanding**

O dataset contém medições de potência activa e reactiva para múltiplos CPEs, ao longo de vários anos, com periodicidade de 15 minutos.

A análise exploratória revelou:

- distribuições fortemente assimétricas, com muitos valores baixos e poucos consumos muito elevados;
- a presença de edifícios com comportamento claramente distinto, incluindo um grande outlier energético;
- padrões diurnos e semanais bem definidos na maioria dos CPEs, com redução do consumo durante a noite e fins de semana;

- correlação moderada entre potência activa e potências reactivas, sugerindo influência do tipo de equipamentos instalados.

A variável **Consumo** encontra-se sem valores no dataset fornecido, pelo que todas as análises e modelos utilizaram exclusivamente a **Potência Ativa** como variável de interesse.

Estes resultados confirmam que os dados possuem estrutura temporal e variabilidade suficiente para clustering e previsão.

Importa salientar que a análise exploratória apresentada neste relatório corresponde a uma **síntese dos principais resultados e padrões relevantes** para os objectivos do trabalho.

O notebook Jupyter associado ao projeto contém uma análise exploratória substancialmente mais extensa, incluindo visualizações adicionais, estatísticas detalhadas por CPE e análises intermédias que fundamentaram as decisões metodológicas adoptadas ao longo do processo.

Por razões de clareza, concisão e limite de extensão, o presente relatório foca-se apenas nos aspetos essenciais para a compreensão dos dados, dos modelos e dos resultados obtidos.



Figura 1: Distribuição da Potência Ativa.

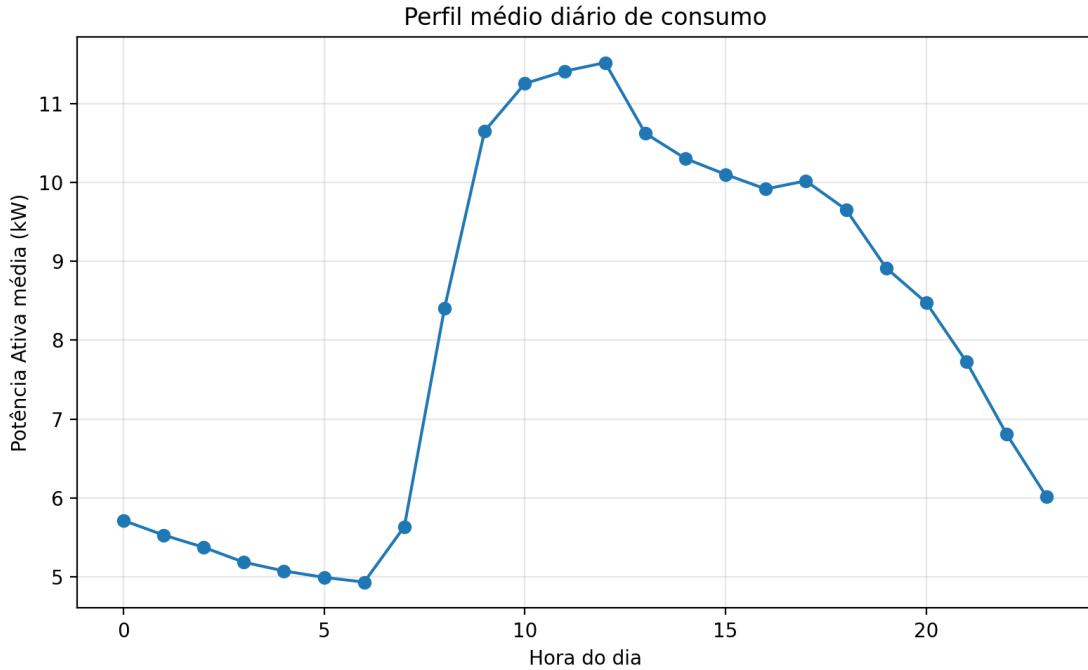


Figura 2: Perfil médio diário de consumo.

### 3 Data Preparation e Feature Engineering

As séries temporais foram preparadas para garantir consistência e comparabilidade entre CPEs.

Os principais passos incluíram:

- reindexação de cada série para uma grelha regular de 15 minutos;
- interpolação linear apenas para falhas curtas (até 1 hora);
- remoção de CPEs com séries demasiado incompletas ou curtas;
- normalização das features para evitar enviesamento por escala.

Após filtragem, permaneceram **84 CPEs válidos**.

Para o clustering e os modelos de previsão foram extraídas features que capturaram:

- perfis horários e semanais médios;
- nível médio de consumo e variabilidade;
- autocorrelação diária (lag de 24h);
- indicadores estruturais como base load, picos diários e comportamento reativo.

Este conjunto de features fornece uma representação compacta e informativa do comportamento energético de cada edifício, adequada tanto para segmentação como para modelação preditiva.

## 4 Clustering

### 4.1 K-Means — Segmentação Global dos Consumidores

O algoritmo K-Means foi aplicado sobre o conjunto de features normalizadas com o objectivo de identificar grupos globais de edifícios com comportamentos energéticos semelhantes.

A escolha do número de clusters baseou-se na análise conjunta dos métodos **Elbow** e **Silhouette**, tendo-se observado que  $k = 3$  oferece o melhor compromisso entre qualidade de separação e interpretabilidade.

A visualização dos clusters através de **PCA** revelou uma estrutura clara:

- um ponto isolado correspondente a um grande consumidor energético;
- dois grupos principais, diferenciados sobretudo pelo nível médio de consumo e pela variabilidade diária.

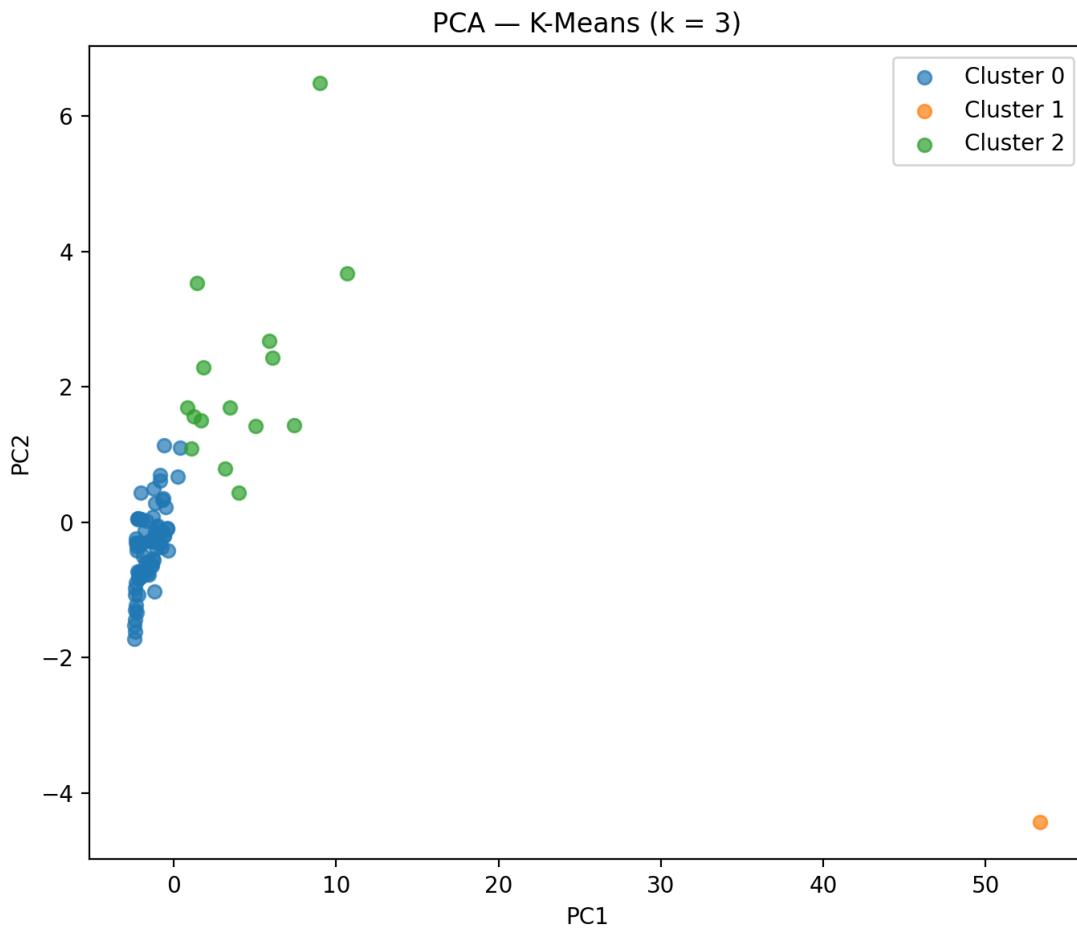


Figura 3: Projecção PCA dos clusters obtidos com K-Means ( $k=3$ ). Destaca um grande outlier energético e dois grupos principais.

#### 4.1.1 Interpretação dos clusters ( $k = 3$ )

A caracterização estatística e a análise dos perfis horários e semanais permitem interpretar os clusters da seguinte forma:

- **Cluster 0 — Edifícios de baixo consumo e elevada regularidade** Inclui a maioria dos CPEs, com potência média reduzida, baixa variabilidade e elevada autocorrelação diária. Representa edifícios pequenos ou instalações com funcionamento previsível, sendo os mais fáceis de prever.
- **Cluster 1 — Outlier energético (1 CPE)** Contém um único edifício com consumo médio e máximo muito superiores aos restantes, elevada variabilidade e comportamento complexo. A validação com dados reais confirma a correspondência com os Paços do Concelho / Torre do Lidor.
- **Cluster 2 — Edifícios intermédios com padrões irregulares** Agrupa instalações com consumo moderado a elevado e maior variabilidade diária, como escolas, pavilhões, piscinas e infraestruturas desportivas.

#### 4.1.2 Associação entre CPEs, edifícios reais e clusters

De forma a validar a interpretabilidade dos clusters obtidos, foi realizada uma associação entre os identificadores técnicos (CPE), os edifícios reais correspondentes e o cluster atribuído pelo modelo.

Tabela 1: Associação entre edifícios reais, CPE e cluster atribuído

Edifício	CPE	Cluster
Torre do Lidor / Paços do Concelho – Servidores	PT0002000068856781NM	2
Torre do Lidor / Paços do Concelho	PT0002000078441876HB	1
Complexo de Educação Ambiental da Quinta da Gruta	PT0002000082549706RH	0
Quinta dos Cónegos	PT0002000068856906VS	0
Complexo Municipal de Piscinas de Folgosa	PT0002000100113293JT	2
Pavilhão Municipal de Gueifães II	PT0002000110090564GD	0
Aeródromo de Vilar de Luz	PT0002000077394934QY	2
Parque Central da Maia	PT0002000081997398TD	2

A Tabela 1 apresenta alguns exemplos representativos, permitindo relacionar diretamente os perfis energéticos identificados com tipologias reais de edifícios municipais.

Observa-se que edifícios com perfis de funcionamento intensivo e elevada variabilidade, como complexos desportivos, infraestruturas técnicas ou edifícios administrativos de grande dimensão, tendem a agrupar-se nos clusters de maior consumo e irregularidade. Por outro lado, edifícios de menor escala ou com utilização mais previsível concentram-se no cluster de baixo consumo e elevada regularidade.

Esta correspondência reforça a validade semântica dos clusters obtidos e confirma que a segmentação reflete padrões energéticos reais e interpretáveis.

A correspondência entre os clusters obtidos e a tipologia real dos edifícios confirma a validade da segmentação produzida pelo K-Means.

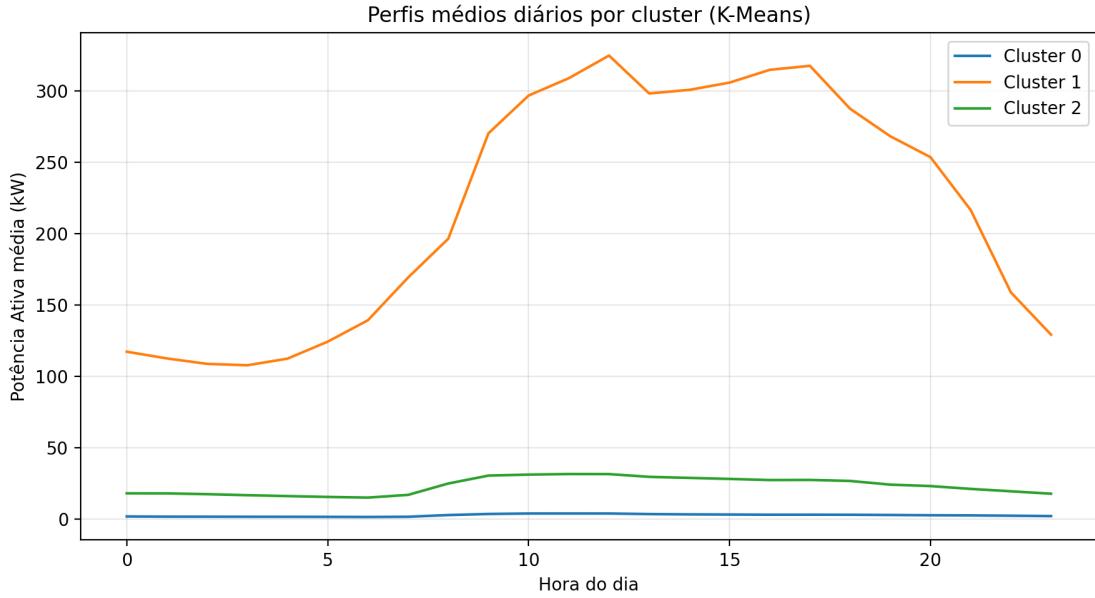


Figura 4: Perfis médios diários por cluster (K-Means). Evidenciam diferenças claras de regularidade e intensidade de consumo.

## 4.2 DBSCAN — Detecção de Clusters Densos e Outliers

Para complementar a análise, foi aplicado o algoritmo **DBSCAN**, que agrupa pontos com base em densidade local e permite identificar explicitamente outliers.

Após redução de dimensionalidade com PCA, o parâmetro **eps = 0.60** foi seleccionado como configuração final, por produzir uma segmentação estável e interpretável.

O DBSCAN identificou:

- **3 clusters densos**, correspondentes a perfis energéticos bem definidos;
- **59 outliers**, representando edifícios cujo comportamento não forma grupos suficientemente compactos.

Este resultado evidencia a elevada heterogeneidade dos dados e reforça a existência de instalações com comportamento energético atípico.

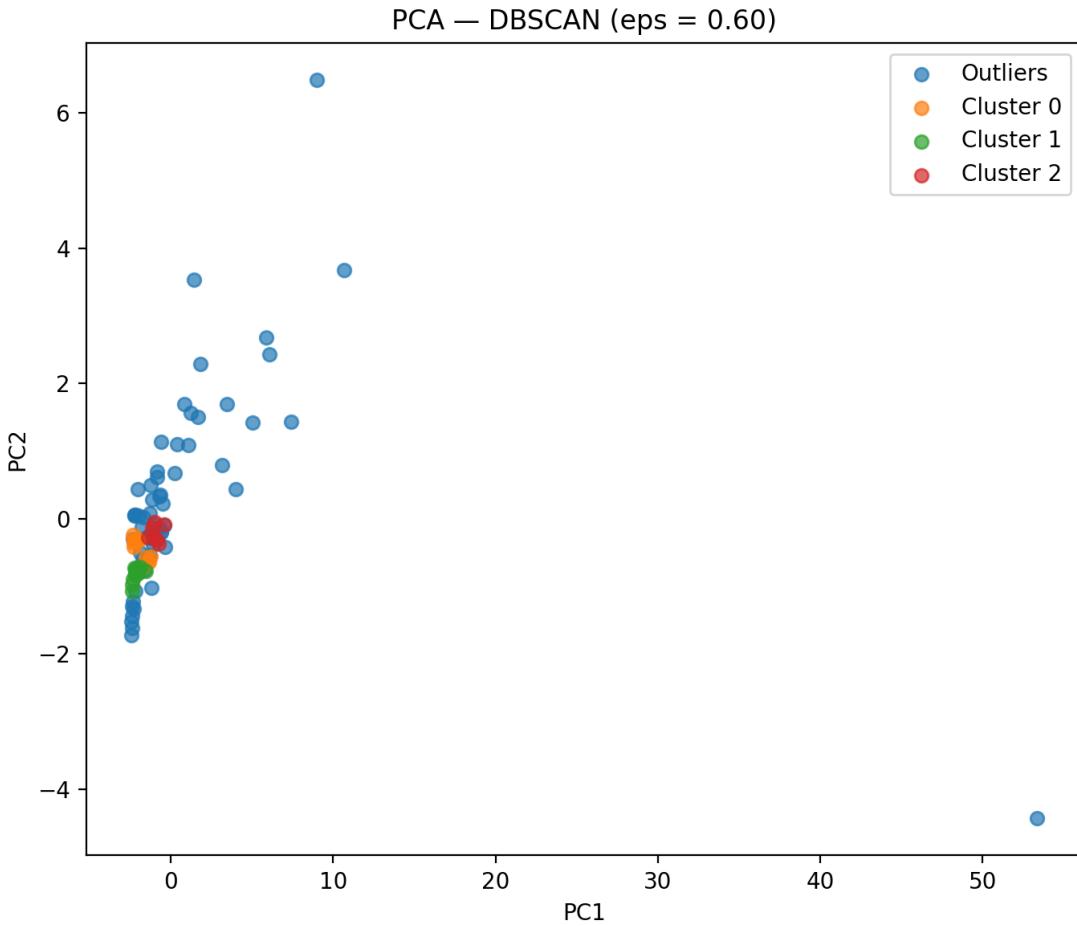


Figura 5: Clusters densos e outliers identificados pelo DBSCAN após redução de dimensionalidade com PCA.

### 4.3 Comparação entre K-Means e DBSCAN

A comparação entre os dois algoritmos revela diferenças fundamentais:

- O **K-Means** força a atribuição de todos os CPEs a um cluster, sendo adequado para uma segmentação global do parque edifício.
- O **DBSCAN** é mais conservador, identificando apenas grupos densos e classificando uma parte significativa dos edifícios como outliers.

A matriz de contingência entre os rótulos dos dois métodos mostra baixa correspondência directa, o que é esperado dado que:

- o K-Means assume clusters esféricos e homogéneos;
- o DBSCAN depende da densidade local dos dados.

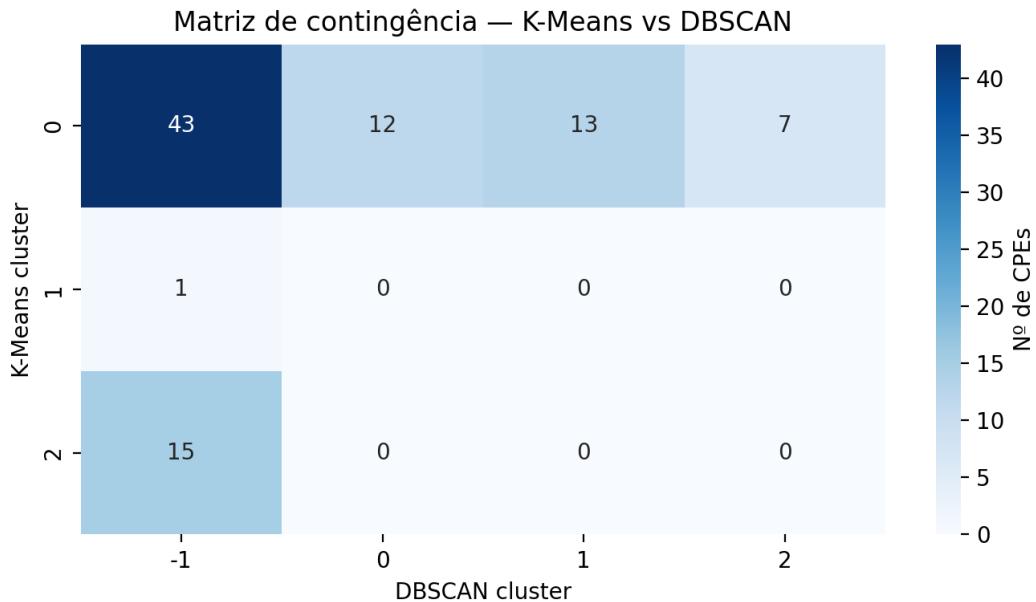


Figura 6: Matriz de contingência entre os clusters obtidos pelo K-Means e pelo DBSCAN.

Apesar disso, ambos os métodos concordam na identificação do grande outlier energético e na existência de grupos com comportamento regular e previsível.

Assim, os algoritmos devem ser vistos como **complementares**:

- o K-Means fornece uma segmentação global e interpretável;
- o DBSCAN destaca padrões densos inequívocos e edifícios atípicos.

Enquanto o K-Means impõe uma segmentação global, o que força a todos os edifícios a pertencer a um grupo, o DBSCAN adopta uma abordagem conservadora baseada em densidade local. Num dataset energético altamente heterogéneo, esta diferença conceptual conduz a resultados distintos: o K-Means revela a estrutura global do parque edifício, enquanto o DBSCAN identifica apenas padrões inequívocos e destaca explicitamente a maioria dos edifícios como comportamentos atípicos.

#### 4.4 Conclusão do Clustering

O clustering revelou que os edifícios municipais apresentam perfis energéticos claramente distintos, com diferenças significativas ao nível do consumo médio, variabilidade e regularidade temporal.

A segmentação obtida permite:

- caracterizar diferentes tipos de consumidores;
- identificar edifícios estáveis e edifícios problemáticos;
- apoiar a análise da previsibilidade do consumo energético.

Estes resultados são utilizados directamente na secção seguinte para interpretar o desempenho dos modelos de previsão por cluster.

## 5 Previsão do Consumo Energético

O objectivo desta secção é avaliar a capacidade de previsão do consumo energético dos edifícios municipais, comparando diferentes abordagens: um baseline simples, um modelo clássico de séries temporais e modelos baseados em *feature sets*.

A avaliação foi realizada com uma separação **estritamente temporal**, utilizando os **70% iniciais de cada série para treino** e os **30% finais para teste**, garantindo ausência de *data leakage*.

Para a análise visual dos resultados de previsão, foi definido um critério de erro aceitável baseado no erro absoluto relativo. Em particular, considerou-se que uma previsão é aceitável quando o erro absoluto se mantém abaixo de **15%** do valor real observado.

Este limiar foi usado nos gráficos que seguem como referência visual, o que permite distinguir previsões adequadas de desvios relevantes, sobretudo em edifícios com padrões de consumo mais regulares.

Este valor foi escolhido por representar um compromisso entre tolerância operacional e exigência preditiva em contexto de gestão energética municipal, sendo particularmente adequado para comparação qualitativa entre modelos.

Importa salientar que as visualizações apresentadas ao longo desta secção correspondem sempre a um **CPE individual representativo de cada cluster** (concretamente, o elemento mais central no espaço de features) e não à média agregada do cluster.

Consequentemente, estas visualizações podem, em alguns casos, induzir a percepção de um desempenho inferior ao real desempenho médio do modelo no cluster, especialmente em edifícios com elevada variabilidade temporal.

Adicionalmente, a **escala do erro absoluto varia entre modelos e entre CPEs**, o que limita a comparação direta exclusivamente visual entre diferentes abordagens de previsão. Por este motivo, a comparação quantitativa global entre modelos deve ser feita com base na Tabela 2, que apresenta métricas agregadas e comparáveis (MAE, RMSE, SMAPE e  $R^2$ ), constituindo uma visão mais robusta do desempenho relativo dos modelos.

## 5.1 Baseline Semanal

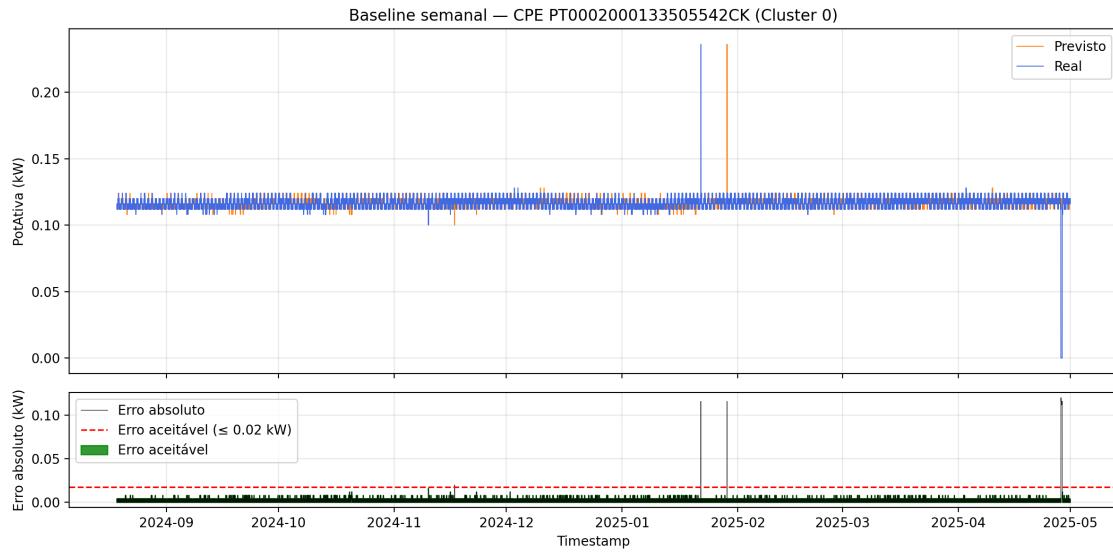


Figura 7: Comparação entre valores reais e o baseline semanal para o CPE mais central do Cluster 0.

O baseline utilizado assume que o consumo num determinado instante é igual ao consumo observado no mesmo instante uma semana antes.

Os resultados mostram que:

- o baseline apresenta desempenho quase perfeito em edifícios com padrões altamente regulares;
- os erros aumentam significativamente em instalações com maior variabilidade e picos de consumo;
- o erro varia fortemente entre clusters, sendo mínimo no Cluster 0 e máximo no outlier energético (Cluster 1).

Este comportamento confirma que a repetição semanal é uma forte referência, mas insuficiente para perfis mais complexos.

## 5.2 ARIMA — Modelo Clássico de Séries Temporais

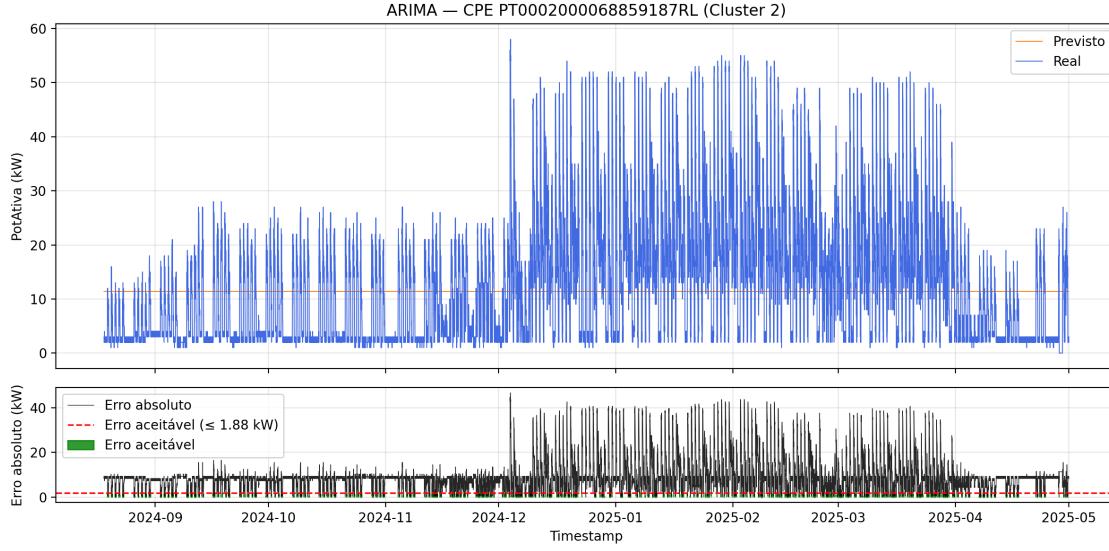


Figura 8: Comparação entre valores reais e previsões ARIMA para o CPE mais central do Cluster 2. Observa-se forte suavização das previsões.

Foi aplicado um modelo **ARIMA univariado**, treinado individualmente por CPE, utilizando apenas a série de potência activa.

O desempenho do ARIMA foi consistentemente **inferior ao baseline semanal**, tanto em MAE como em RMSE. As previsões apresentaram forte suavização e convergência para valores médios, refletindo a incapacidade do modelo em capturar padrões diários e semanais dominantes.

Este resultado era esperado, dado que:

- o modelo é estritamente univariado;
- não inclui componentes sazonais;
- não utiliza variáveis exógenas.

Apesar do desempenho fraco, o ARIMA cumpre o seu papel como **baseline estatístico clássico**, evidenciando as limitações de abordagens tradicionais neste contexto energético.

O ARIMA serve assim como um controlo experimental: demonstra que, mesmo em séries com elevada autocorrelação, modelos que ignoram estrutura sazonal explícita e informação contextual são inadequados para previsão energética realista.

## 5.3 Modelos Baseados em Features

Dado que a Random Forest e o XGBoost pertencem à mesma classe de modelos baseados em features e exibem comportamentos qualitativos semelhantes nas visualizações temporais, optou-se por apresentar apenas um exemplo visual representativo (XGBoost), complementando a análise da Random Forest com métricas quantitativas globais.

### 5.3.1 Random Forest

A Random Forest apresentou uma melhoria muito significativa face ao baseline, reduzindo o erro absoluto médio (MAE) em cerca de **60%**. O modelo capturou eficazmente padrões de curto prazo e dinâmicas intra-diárias.

No entanto:

- o RMSE manteve-se elevado em edifícios com picos abruptos;
- nenhum dos modelos conseguiu antecipar totalmente eventos extremos, devido à ausência de variáveis exógenas.

A análise por cluster mostrou que:

- o Cluster 0 é facilmente previsível;
- o Cluster 2 apresenta dificuldade intermédia;
- o Cluster 1 continua a ser extremamente difícil de prever, embora a RF reduza substancialmente o erro face ao baseline.

### 5.3.2 XGBoost

O modelo XGBoost foi progressivamente optimizado através de um conjunto de experiências exploratórias, testando diferentes combinações de features temporais, estatísticas e estruturais, bem como ajustes nos principais hiperparâmetros do modelo.

Em particular, foram avaliadas variações ao nível:

- do conjunto de *lags* temporais incluídos;
- da incorporação de features agregadas (médias e desvios móveis);
- da inclusão de variáveis estruturais por CPE;
- da complexidade do modelo (profundidade das árvores e regularização).

A versão designada por **XGBoost BEST** corresponde à configuração que apresentou o melhor compromisso entre erro médio absoluto (MAE), capacidade de generalização e estabilidade dos resultados entre diferentes clusters energéticos. Esta seleção baseou-se na avaliação quantitativa global e na análise qualitativa do comportamento temporal das previsões.

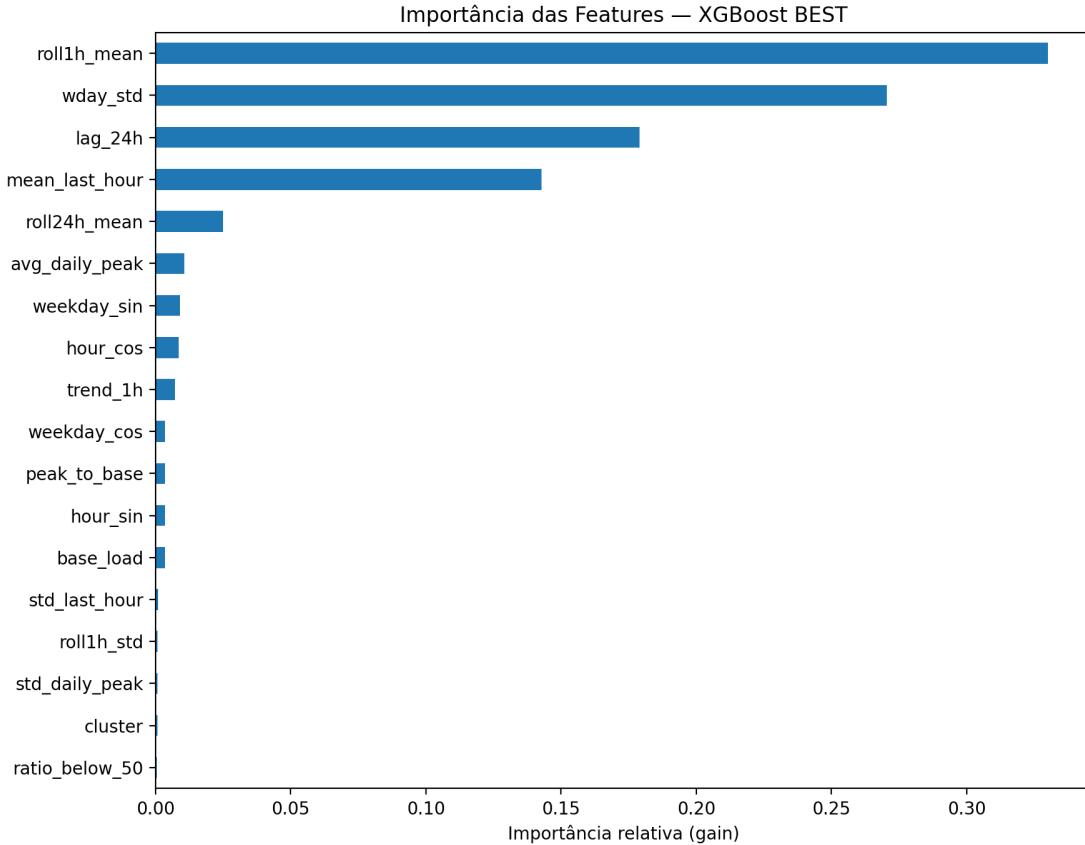


Figura 9: Importância das features no modelo XGBoost BEST. Observa-se a predominância de variáveis temporais de curto prazo e indicadores estruturais do consumo.

O **XGBoost BEST** apresentou:

- o menor **MAE global** entre todos os modelos testados;
- desempenho particularmente forte em edifícios com variabilidade moderada;
- maior sensibilidade a picos extremos, refletindo-se num RMSE ligeiramente superior ao da Random Forest.

#### 5.4 Comparação Global dos Modelos

Tabela 2: Comparação global do desempenho dos modelos de previsão, avaliada através das métricas MAE, RMSE, SMAPE e  $R^2$ .

Modelo	MAE	RMSE	SMAPE (%)	$R^2$
Baseline (t7)	2.187	7.683	39.62	0.880
ARIMA	5.177	6.648	87.52	-0.278
Random Forest	0.800	2.487	43.15	0.988
XGBoost (BEST)	0.624	3.050	42.59	0.982

A comparação entre os modelos revela que:

- o **baseline semanal** é adequado apenas para edifícios altamente regulares;
- o **ARIMA** é claramente superado por todas as restantes abordagens;
- a **Random Forest** apresenta maior robustez face a picos, resultando no melhor RMSE global;
- o **XGBoost BEST** oferece a melhor precisão média (MAE), sendo o modelo mais eficaz na maioria dos edifícios.

O SMAPE penaliza fortemente erros relativos em períodos de baixo consumo, o que explica valores mais elevados nos modelos baseados em features, apesar do melhor desempenho em MAE e RMSE.

Os resultados mostram que não existe um modelo universalmente superior. A Random Forest apresenta maior robustez a valores extremos, enquanto o XGBoost optimizado oferece maior precisão média. Esta diferença está directamente ligada à natureza dos perfis energéticos identificados no clustering.

## 5.5 Análise Visual — Real vs Previsto

Os CPEs seleccionados correspondem aos elementos mais centrais de cada cluster no espaço de features, garantindo representatividade estatística dos perfis energéticos analisados.

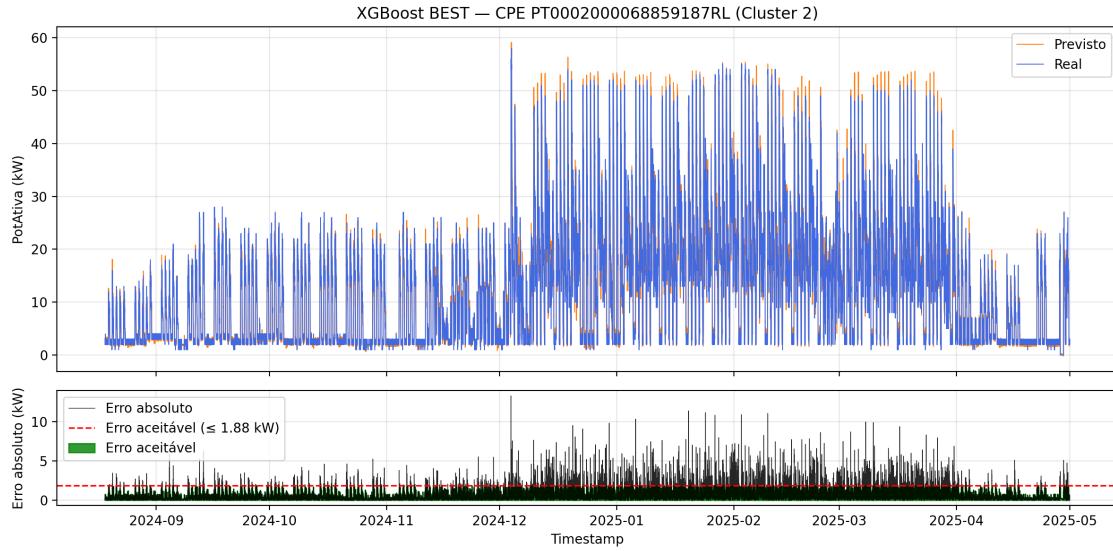


Figura 10: Comparação entre valores reais e previstos pelo XGBoost BEST para o CPE mais central do Cluster 2.

As visualizações *Real vs Previsto*, confirmam de forma clara as conclusões quantitativas:

- em edifícios estáveis, todos os modelos produzem previsões próximas da realidade;
- em perfis intermédios, os modelos baseados em features capturam melhor a dinâmica do consumo;
- no outlier energético, todos os modelos enfrentam dificuldades, embora RF e XGBoost superem claramente o baseline e o ARIMA.

Estas visualizações evidenciam que a **dificuldade de previsão está fortemente ligada ao cluster energético** identificado na fase de clustering.

## 5.6 Síntese da Previsão

Em síntese:

- a previsibilidade do consumo varia significativamente entre edifícios;
- modelos baseados em features oferecem ganhos substanciais face a abordagens clássicas;
- a ausência de variáveis exógenas limita a previsão de picos abruptos;
- a segmentação por clusters é fundamental para interpretar correctamente os erros de previsão.

# 6 Conclusões

## 6.1 Síntese do trabalho realizado

Neste trabalho foi aplicado o processo CRISP-DM ao conjunto de dados de consumo energético da Câmara Municipal da Maia, com dois objectivos principais: - Caracterizar os diferentes consumidores através de técnicas de clustering - Avaliar a capacidade de previsão do consumo energético utilizando diferentes abordagens de modelação.

O estudo combinou análise exploratória, extração de features, aprendizagem não supervisionada e modelos de previsão, permitindo obter uma visão integrada dos padrões de consumo e da sua previsibilidade.

## 6.2 Conclusões do Clustering

A aplicação de técnicas de clustering permitiu identificar perfis energéticos distintos entre os CPEs analisados.

- O K-Means revelou-se eficaz na separação global dos consumidores, produzindo clusters interpretáveis e coerentes com diferentes níveis de consumo, variabilidade e regularidade.
- O DBSCAN, por sua vez, identificou apenas padrões densos e inequívocos, classificando muitos CPEs como outliers, o que evidencia a heterogeneidade dos comportamentos energéticos.

A comparação entre ambos mostrou que os métodos são complementares: o K-Means é adequado para uma segmentação global, enquanto o DBSCAN é útil para detecção de perfis muito específicos ou atípicos.

A validação com edifícios reais confirmou que os clusters obtidos correspondem a tipologias plausíveis de consumo (edifícios estáveis, perfis intermédios e grandes consumidores/outliers).

## 6.3 Conclusões da Previsão

A avaliação dos modelos de previsão evidenciou diferenças claras entre abordagens:

- O baseline semanal mostrou-se surpreendentemente forte para edifícios com comportamento estável, confirmando que grande parte do consumo apresenta elevada repetibilidade semanal.
- O ARIMA univariado, sem componente sazonal, apresentou previsões praticamente constantes, demonstrando limitações claras na captura de padrões diários e semanais característicos do consumo energético.
- Os modelos baseados em feature sets (Random Forest e XGBoost) superaram consistentemente o baseline e o ARIMA, capturando melhor a dinâmica temporal e a variabilidade dos dados.

- O XGBoost, após afinação e selecção de features, apresentou o melhor desempenho global, especialmente em edifícios com maior variabilidade e picos de consumo.

As visualizações de Real vs Previsto mostraram de forma clara que a dificuldade de previsão varia significativamente entre clusters, sendo maior em perfis mais irregulares e com eventos extremos.

#### **6.4 Considerações metodológicas**

A separação treino–teste foi realizada de forma estritamente temporal e individual por CPE, utilizando os 70% iniciais de cada série para treino e os 30% finais para teste. Esta abordagem garante ausência de data leakage e avalia correctamente a capacidade de previsão do consumo futuro de edifícios já observados.

As features foram construídas apenas com informação passada ou com características estruturais dos consumidores, assegurando consistência metodológica.

#### **6.5 Limitações e trabalho futuro**

Apesar dos bons resultados obtidos, existem limitações relevantes:

- A ausência de variáveis exógenas (meteorologia, ocupação, eventos) limita a capacidade de previsão de picos abruptos.
- Os modelos avaliados não foram optimizados para previsão multi-horizonte de longo prazo.
- O ARIMA poderia ser melhorado com componentes sazonais explícitas, embora tal não fosse o foco deste trabalho.

Como trabalho futuro, seria interessante integrar dados externos, explorar modelos híbridos e avaliar estratégias de previsão específicas por cluster.

#### **6.6 Conclusão final**

Em síntese, este trabalho demonstra que:

- Os dados de consumo energético apresentam estrutura suficiente para segmentação significativa;
- A previsibilidade depende fortemente do perfil energético do consumidor;
- Modelos baseados em features oferecem ganhos substanciais face a abordagens clássicas univariadas.

Estes resultados evidenciam o potencial da combinação entre segmentação de consumidores e modelos preditivos para apoiar decisões de gestão energética em contexto municipal real.

### **Reprodutibilidade e Código**

O notebook Jupyter utilizado neste trabalho, bem como todos os ficheiros do projeto (código, dados processados e figuras), encontram-se disponíveis no repositório GitHub:

<https://github.com/lcmlo/IAA-Consumo-Energetico-Clustering-Previsao>

## Repositório e organização do projeto

O notebook e os ficheiros associados a este projeto encontram-se alojados num repositório GitHub.

A estrutura do repositório é a seguinte:

- `IAA_notebook_final.ipynb` — notebook principal do projeto, contendo todo o código, análise exploratória, clustering e modelos de previsão;
- `figures/` — figuras utilizadas no relatório e geradas pelo notebook;
- ficheiros `.pkl` — caches intermédias usadas para acelerar a execução dos modelos;
- `README.md` — descrição geral do projeto, objectivos e dependências;
- `IAA_relatorio_latex.pdf` — relatório final em PDF, sem código.

Todo o código, resultados intermédios e figuras são reproduzíveis através do notebook disponibilizado no repositório GitHub.