

# **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 ativa e reativa 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.

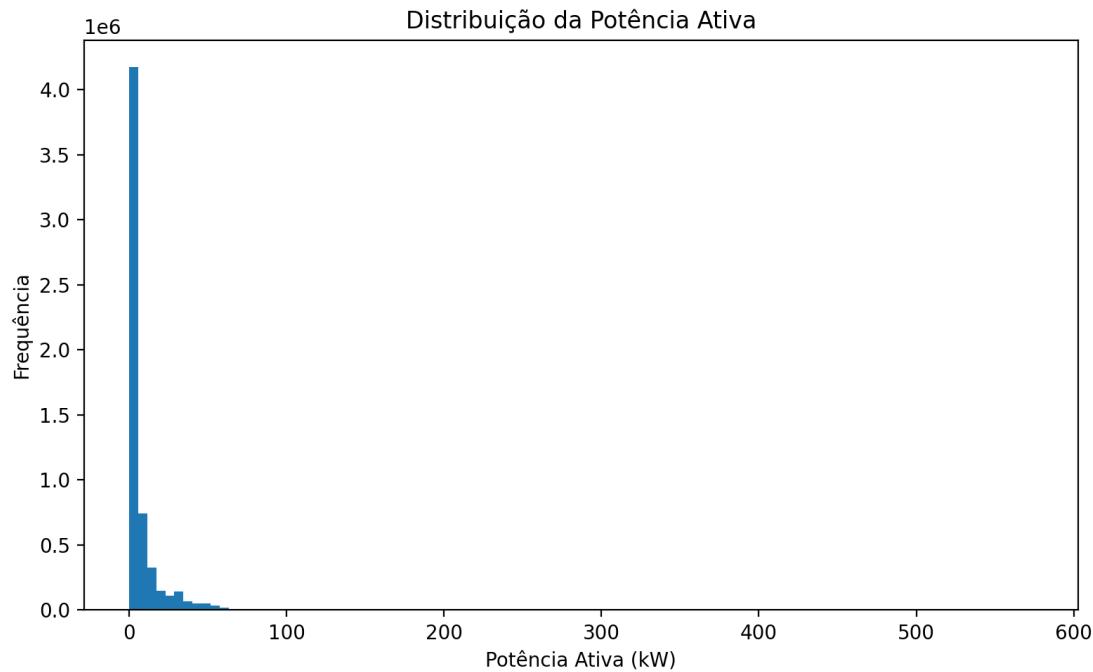


Figura 1: Distribuição da Potência Ativa.

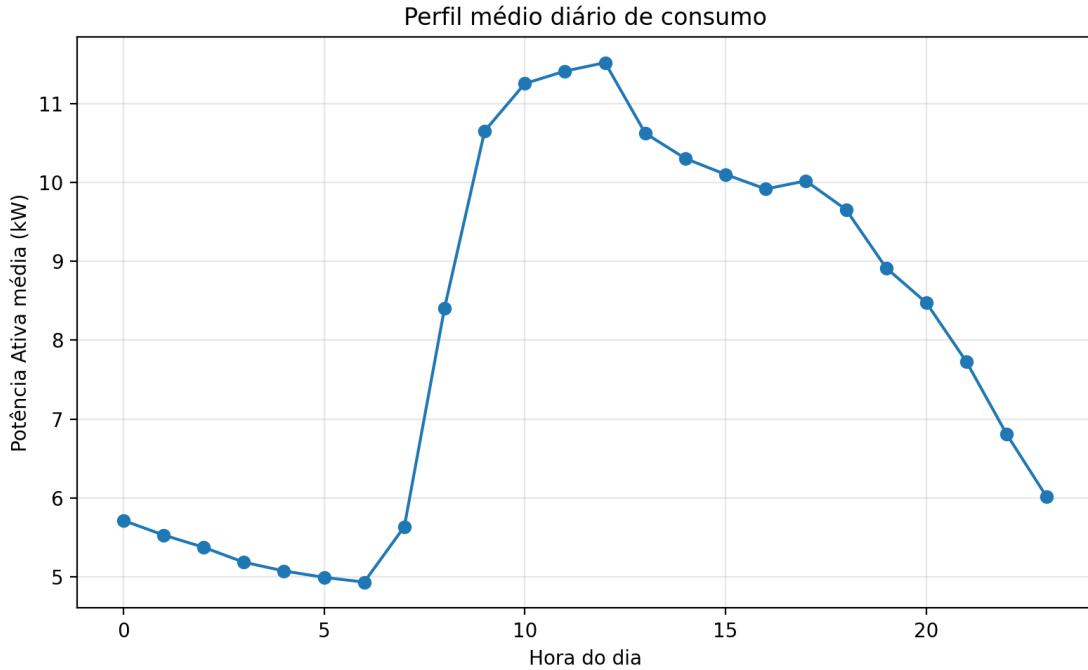


Figura 2: Perfil médio diário de consumo.

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 substancialmente mais extensa, incluindo visualizações adicionais, estatísticas detalhadas por CPE e análises intermédias que fundamentaram as decisões metodológicas adoptadas ao longo do processo.

### 3 Data Preparation e Feature Engineering

As séries temporais foram preparadas para garantir consistência e comparabilidade entre CPEs, foram reindexadas para uma grelha regular de 15 minutos, garantindo consistência temporal entre CPEs. Foram aplicadas interpolações lineares apenas em falhas curtas (até 1 hora), enquanto CPEs com séries demasiado incompletas ou curtas foram removidos, (edifícios com menos de um ano de dados disponíveis e/ou mais de 30 por cento de valores em falta), para garantir séries temporais suficientemente longas e estáveis para o clustering e para a avaliação dos modelos de previsão. As features foram posteriormente normalizadas, de forma a evitar enviesamento por escala e assegurar comparabilidade entre edifícios.

Após filtragem, permaneceram **84 CPEs válidos** de um total inicial de 91.

Para o clustering e os modelos de previsão foram extraídas features que capturam informação complementar sobre o comportamento energético dos edifícios, incluindo perfis horários e semanais médios, nível médio de consumo, variabilidade temporal, autocorrelação diária (lag de 24h) e 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, caracterizada pela presença de um ponto isolado correspondente a um grande consumidor energético e por dois grupos principais, diferenciados sobretudo pelo nível médio de consumo e pela variabilidade diária.

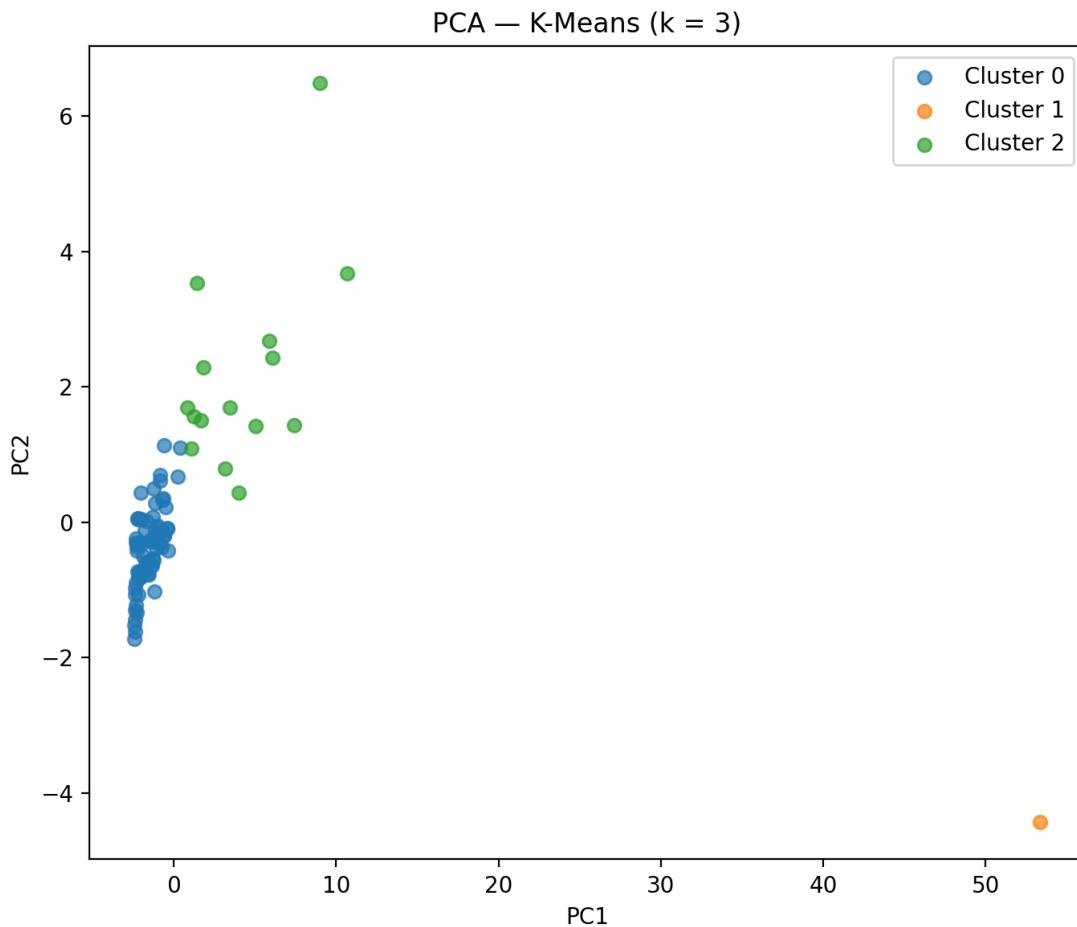


Figura 3: Projecção PCA dos clusters obtidos com K-Means ( $k=3$ ). Destaca um grande outlier energético e dois grupos principais.

#### 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.

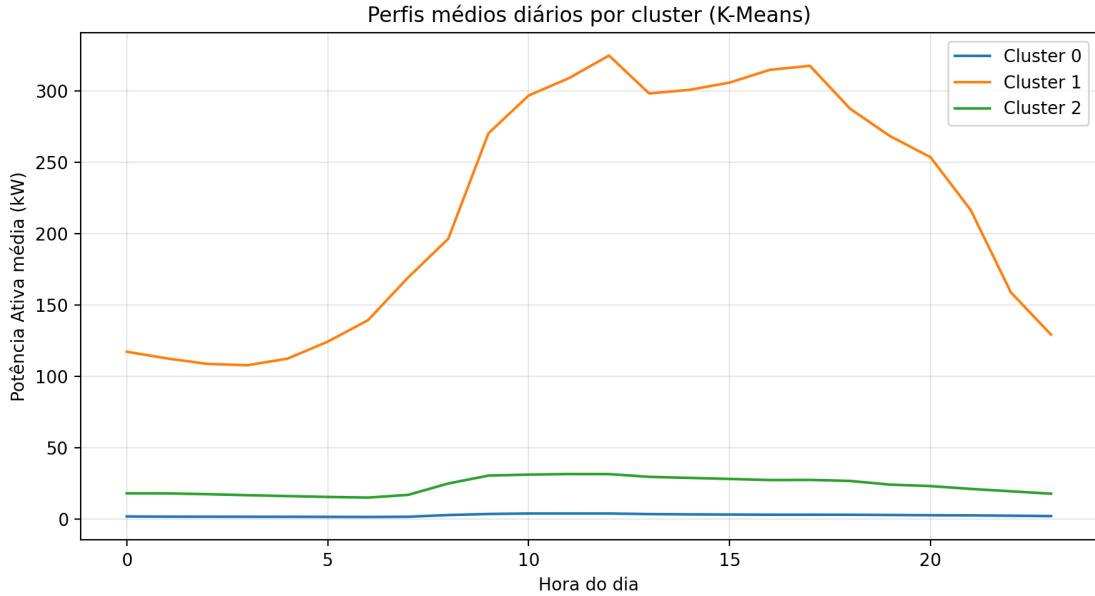


Figura 4: Perfis médios diários por cluster (K-Means). Evidenciam diferenças claras de regularidade e intensidade de consumo.

## 4.2 DBSCAN — Detecção de Clusters Densos e Outliers

Para complementar a análise, foi aplicado o algoritmo **DBSCAN**, com o objectivo de identificar grupos densos de edifícios e detectar explicitamente comportamentos atípicos.

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 três clusters densos, correspondentes a perfis energéticos bem definidos, e um conjunto significativo de outliers (59 CPEs), representando edifícios cujo comportamento não forma grupos suficientemente compactos no espaço de features.

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

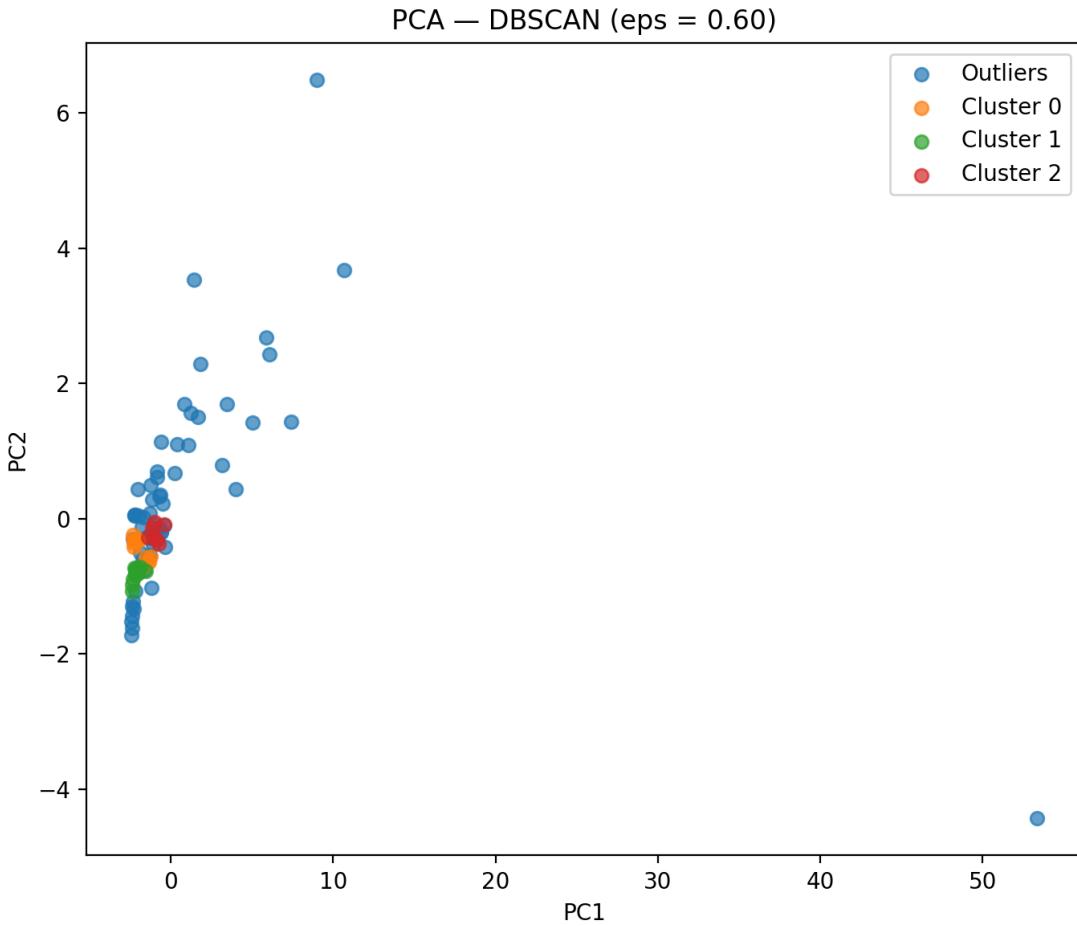


Figura 5: Clusters densos e outliers identificados pelo DBSCAN após redução de dimensionalidade com PCA.

### 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 direta, o que é esperado, uma vez que o K-Means assume clusters globais, esféricos e homogêneos, enquanto o DBSCAN depende fortemente da densidade local dos dados.

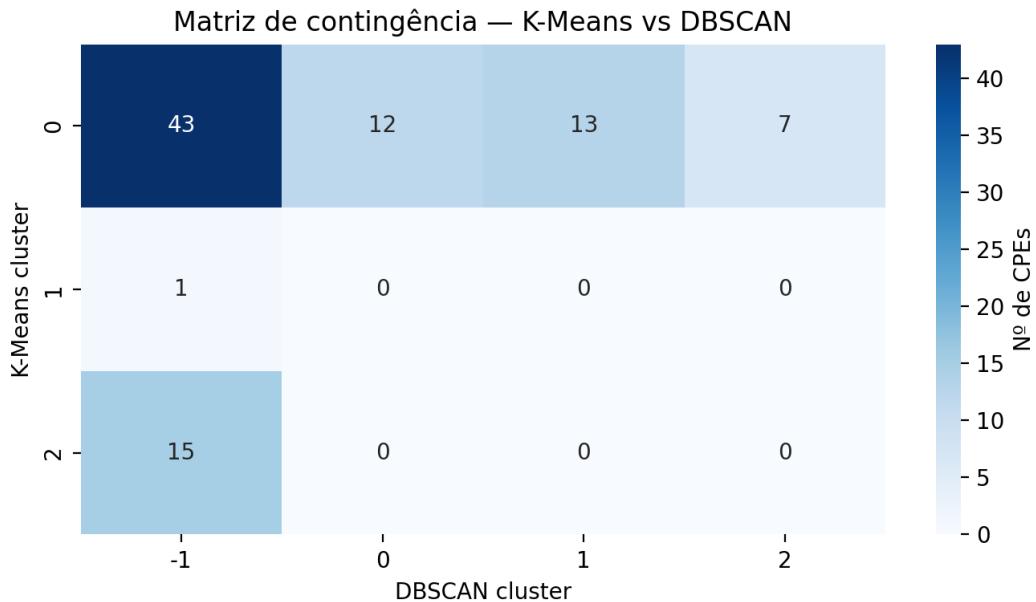


Figura 6: Matriz de contingência entre os clusters obtidos pelo K-Means e pelo DBSCAN.

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 edificado, 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 edifício com menor distância ao centróide do cluster) 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

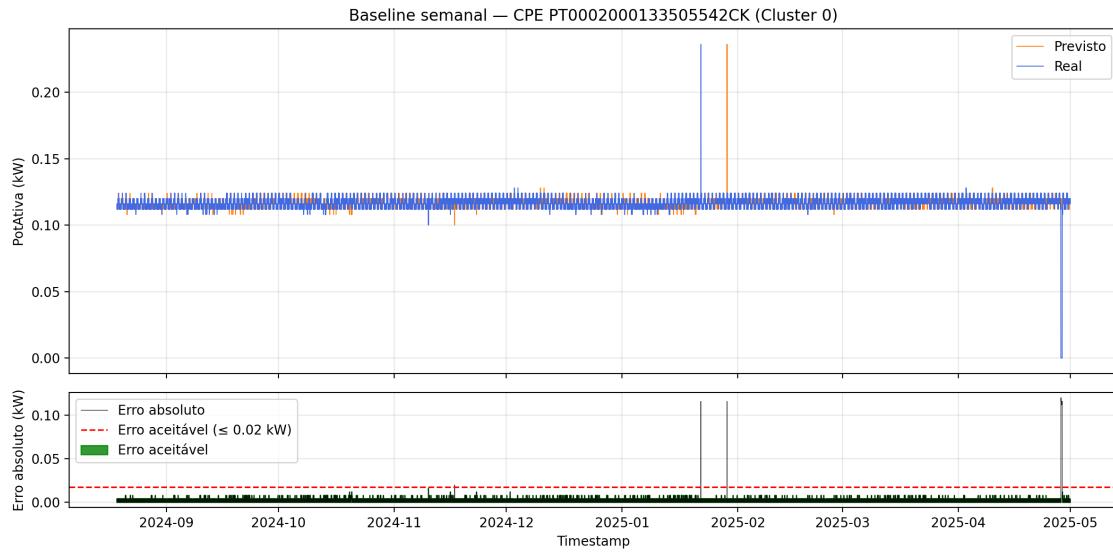


Figura 7: Comparaçāo entre valores reais e o baseline semanal para o CPE mais central do Cluster 0.

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, enquanto os erros aumentam significativamente em instalações com maior variabilidade e picos de consumo. Observa-se ainda uma forte dependência do erro em relação ao cluster energético, 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

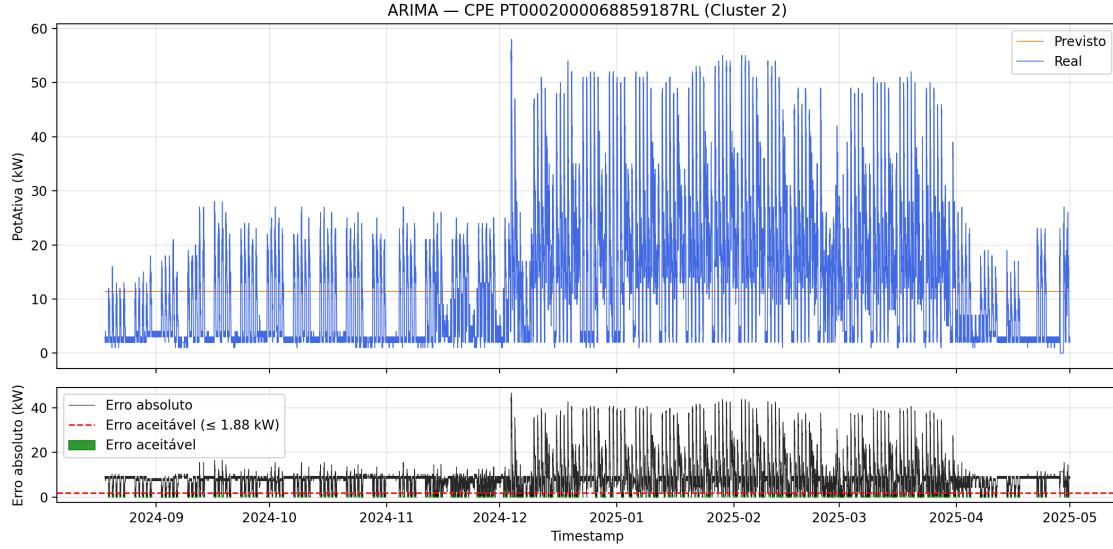


Figura 8: Comparação entre valores reais e previsões ARIMA para o CPE mais central do Cluster 2. Observa-se forte suavização das previsões.

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.

Este resultado era esperado, uma vez que o modelo é estritamente univariado, não inclui componentes sazonais explícitas e não utiliza variáveis exógenas, o que limita a sua capacidade de capturar padrões diários e semanais dominantes.

Apesar do desempenho fraco, o ARIMA cumpre o seu papel como **baseline estatístico clássico**, o que evidencia 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

A Random Forest e o XGBoost pertencem à mesma classe de modelos baseados em features e exibem comportamentos qualitativos semelhantes nas visualizações temporais. Por este motivo, são apresentados exemplos visuais representativos de ambos os modelos para um perfil intermédio, complementando a análise qualitativa com uma comparação quantitativa global baseada em métricas agregadas.

### 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.

A análise por cluster mostrou que o Cluster 0 é facilmente previsível, enquanto que o Cluster 2 apresenta dificuldade intermédia. O Cluster 1, correspondente ao outlier energético, continua a ser extremamente difícil de prever, embora a Random Forest reduza substancialmente o erro face ao baseline.

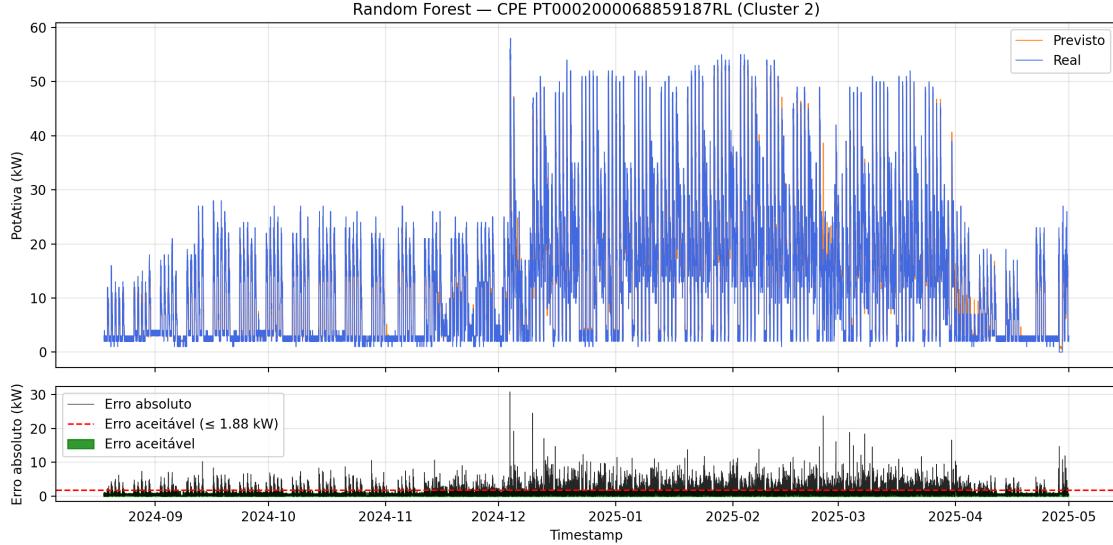


Figura 9: Comparaçāo entre valores reais e previstos pela Random Forest para o CPE mais central do Cluster 2.

Apesar do bom desempenho médio, o RMSE mantém-se elevado em edifícios com picos abruptos, o que reflete a dificuldade dos modelos em antecipar eventos extremos na ausência de variáveis exógenas.

### 5.3.2 XGBoost

O modelo XGBoost foi progressivamente otimizado 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 e da complexidade do modelo (profundidade das árvores e regularização).

A versão designada por **XGBoost BEST** no notebook 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.

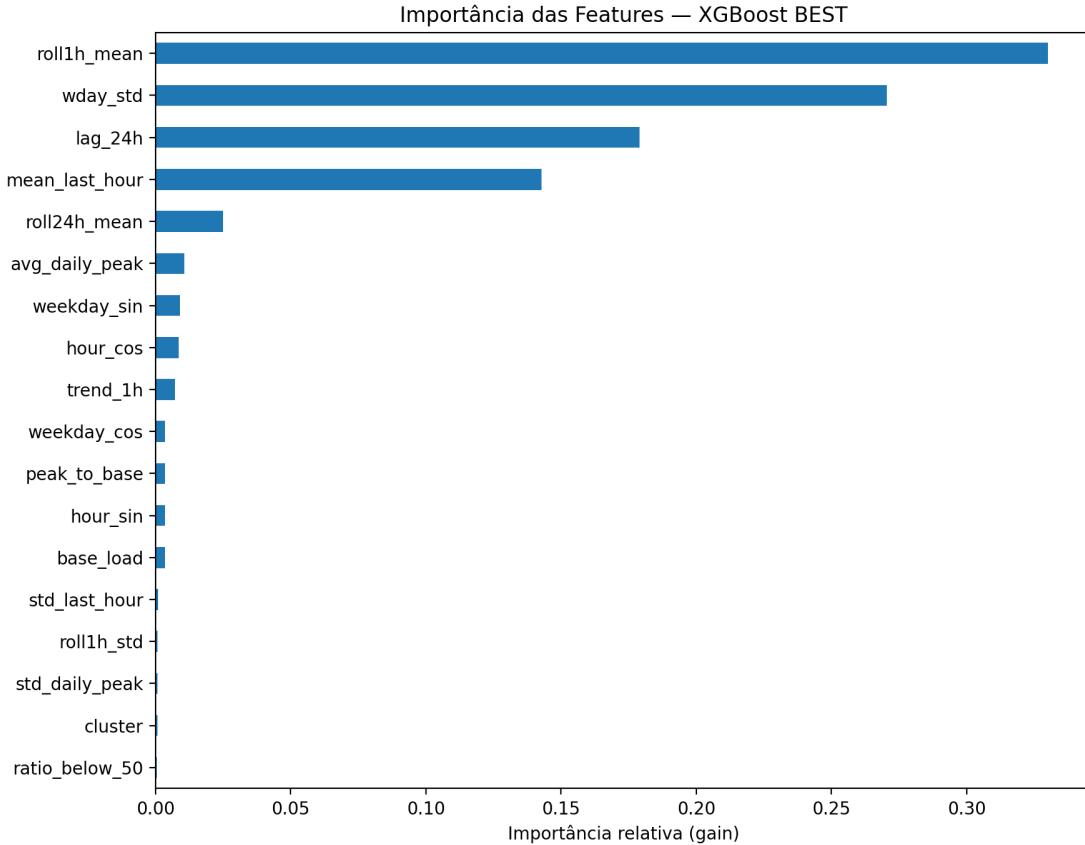


Figura 10: Importância das features no modelo XGBoost BEST. Observa-se a predominância de variáveis temporais de curto prazo e indicadores estruturais do consumo.

A análise da importância das features evidencia o papel dominante da informação temporal recente e de indicadores estruturais do consumo, confirmando que a capacidade preditiva do modelo resulta sobretudo da captura de dependências de curto prazo e padrões recorrentes.

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

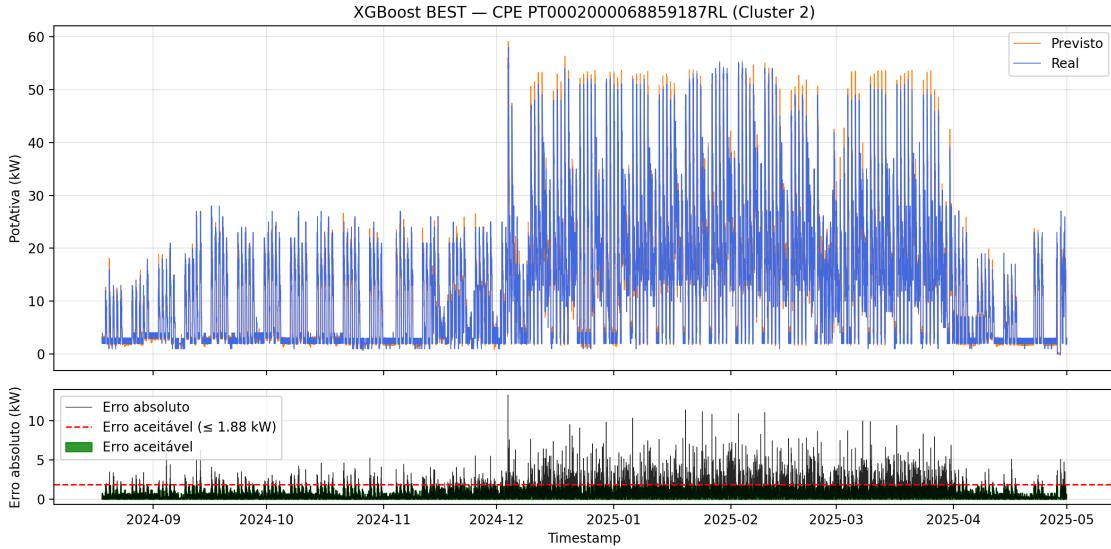


Figura 11: Comparação entre valores reais e previstos pelo XGBoost BEST para um edifício com perfil intermédio.

A Figura 11 ilustra o comportamento temporal típico das previsões produzidas pelo XGBoost em perfis intermédios, mostrando uma boa capacidade de acompanhamento da dinâmica do consumo, com desvios mais pronunciados associados a variações abruptas e picos extremos.

Esta análise visual complementa a avaliação quantitativa apresentada na secção seguinte, permitindo interpretar os erros observados à luz dos perfis energéticos identificados na fase de clustering.

#### 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 (t-7)	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 para todos os perfis energéticos analisados. 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.

#### 5.4.1 Desempenho dos modelos por cluster

De forma a aprofundar a análise global apresentada na Tabela 2, esta secção apresenta a comparação do desempenho dos modelos de previsão separadamente por cluster energético. Esta análise permite avaliar de que forma o desempenho relativo de cada modelo varia em função do perfil energético dos edifícios.

Tabela 3: Desempenho dos modelos de previsão no Cluster 0 (elevada regularidade).

Modelo	MAE	RMSE	SMAPE (%)	$R^2$
Baseline (t-7)	0.954	1.765	47.052	0.728
ARIMA	2.271	2.954	93.698	-0.313
Random Forest	0.411	0.695	50.495	0.876
XGBoost (BEST)	0.256	0.389	50.993	0.417

Tabela 4: Desempenho dos modelos de previsão no Cluster 1 (outlier energético extremo).

Modelo	MAE	RMSE	SMAPE (%)	$R^2$
Baseline (t-7)	45.547	63.956	23.045	0.614
ARIMA	91.319	104.722	45.741	-0.036
Random Forest	13.863	19.959	7.486	0.962
XGBoost (BEST)	23.022	31.993	11.563	0.903

Tabela 5: Desempenho dos modelos de previsão no Cluster 2 (variabilidade temporal intermédia).

Modelo	MAE	RMSE	SMAPE (%)	$R^2$
Baseline (t-7)	5.322	9.855	25.246	0.685
ARIMA	12.609	16.855	62.277	-0.135
Random Forest	1.778	3.034	10.719	0.959
XGBoost (BEST)	1.287	2.362	7.833	0.978

A análise por cluster evidencia que o desempenho relativo dos modelos varia significativamente em função do perfil energético dos edifícios. No Cluster 0, caracterizado por elevada regularidade e baixa variância do consumo, o baseline semanal apresenta desempenho competitivo, enquanto os modelos baseados em features oferecem ganhos marginais em termos de erro absoluto. Nestes casos, métricas como o coeficiente de determinação ( $R^2$ ) tornam-se menos informativas, uma vez que a reduzida variância penaliza modelos mais flexíveis, apesar do seu bom desempenho em MAE e RMSE.

No Cluster 2, associado a maior variabilidade temporal, a Random Forest e o XGBoost superam claramente as abordagens clássicas, refletindo a vantagem de modelos capazes de capturar padrões não lineares e dinâmicas intra-diárias mais complexas. No outlier energético (Cluster 1), todos os modelos apresentam degradação de desempenho, embora os modelos baseados em features reduzam substancialmente o erro face ao baseline semanal.

Estes resultados demonstram que não existe um modelo universalmente ótimo, sendo metodologicamente mais adequado adoptar estratégias de previsão diferenciadas por cluster energético, ajustadas às características específicas de cada perfil de consumo.

## 5.5 Síntese da Previsão

Em síntese, a previsibilidade do consumo energético varia significativamente entre edifícios e está fortemente associada ao perfil energético identificado na fase de clustering. Os modelos baseados em features oferecem ganhos substanciais face a abordagens clássicas, mas a ausência de variáveis exógenas continua a limitar a previsão de picos abruptos e eventos extremos.

# 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. O primeiro consiste na caracterização dos diferentes consumidores através de técnicas de clustering. O segundo visa 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. Em particular,

os resultados sugerem que estratégias de previsão e gestão energética em contexto municipal devem ser diferenciadas por tipologia de edifício, em vez de adoptar abordagens uniformes para todo o parque edificado.

## 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.