INSTITUTO UNIVERSITÁRIO DE LISBOA

Introdução à Aprendizagem Automática

Projeto Final

Josué da Glória, Nº 134449
Ricardo Ourelo, Nº 120141

1. Plano de Implementação

O projeto foi desenvolvido de forma faseada, seguindo uma abordagem inspirada na metodologia CRISP-DM. As tarefas foram distribuídas entre os elementos do grupo, garantindo uma validação conjunta dos resultados obtidos. Esta organização permitiu uma melhor gestão do tempo durante a realização deste projeto.

Fase	Tarefa	Tempo Estimado	Responsável
1	Business Understanding e análise do enunciado	3 horas	Ambos
2	Data Understanding e EDA	5 horas	Ambos
3	Data Preparation e feature engineering	7 horas	Ambos
4	Clustering (K-Means + PCA)	6 horas	Ambos
5	Modelos de previsão (ARIMA)	6 horas	Ambos
6	Modelos supervisionados (Random Forest)	5 horas	Ambos
7	Avaliação e comparação de modelos	3 horas	Ambos
8	Escrita do relatório	8 horas	Ambos

2. Business Understanding

A gestão eficiente do consumo energético é um desafio relevante para as entidades públicas, tanto em termos de custos operacionais como de sustentabilidade ambiental. Os edifícios municipais apresentam padrões de consumo distintos, dependentes da sua função, dimensão e horários de utilização.
Neste contexto, a Câmara Municipal da Maia disponibilizou dados de consumo energético para vários pontos de consumo (CPEs). Estes dados permitem aplicar técnicas de Machine Learning com o objetivo de compreender padrões de consumo e desenvolver modelos preditivos.

O presente projeto tem como objetivos principais:

- Identificar grupos de edifícios com comportamentos energéticos semelhantes através de técnicas de clustering;
- Prever o consumo energético futuro recorrendo a abordagens de séries temporais e a modelos supervisionados;
- Comparar diferentes modelos de previsão com um baseline simples, avaliando o seu desempenho.

Os resultados pretendem apoiar a análise e a gestão energética dos edifícios municipais, bem como demonstrar a aplicabilidade prática de técnicas de aprendizagem automática a dados reais.

3. Data Understanding

3.1. Descrição Geral do Dataset

O conjunto de dados utilizado neste projeto foi fornecido pelo docente e como referido anteriormente contém medições de consumo energético com periodicidade de 15 minutos, relativas a vários Códigos de Ponto de Entrega (CPEs) pertencentes a edifícios municipais da Câmara Municipal da Maia.
As variáveis disponibilizadas incluem:

- Potência ativa, definida como variável-alvo para as tarefas de previsão;
- Potências reativas (indutiva e capacitiva), utilizadas como apoio à análise e criação de features adicionais;
- Informação temporal associada a cada registo.

3.2. Análise de Missing Values

Foi realizada uma análise dos valores em falta por variável, com o objetivo de compreender a sua distribuição e impacto potencial na análise.
Concluiu-se que as variáveis PotReactIndut e PotReactCapac contêm uma quantidade significativa de valores nulos. Isto é esperado porque só existem valores quando há cargas indutivas ou capacitivas em funcionamento.
O tratamento destes valores será feito na fase Data Preparation, conforme definido pelo CRISP-DM

Figura 1 - Distribuição dos valores em falta por variável

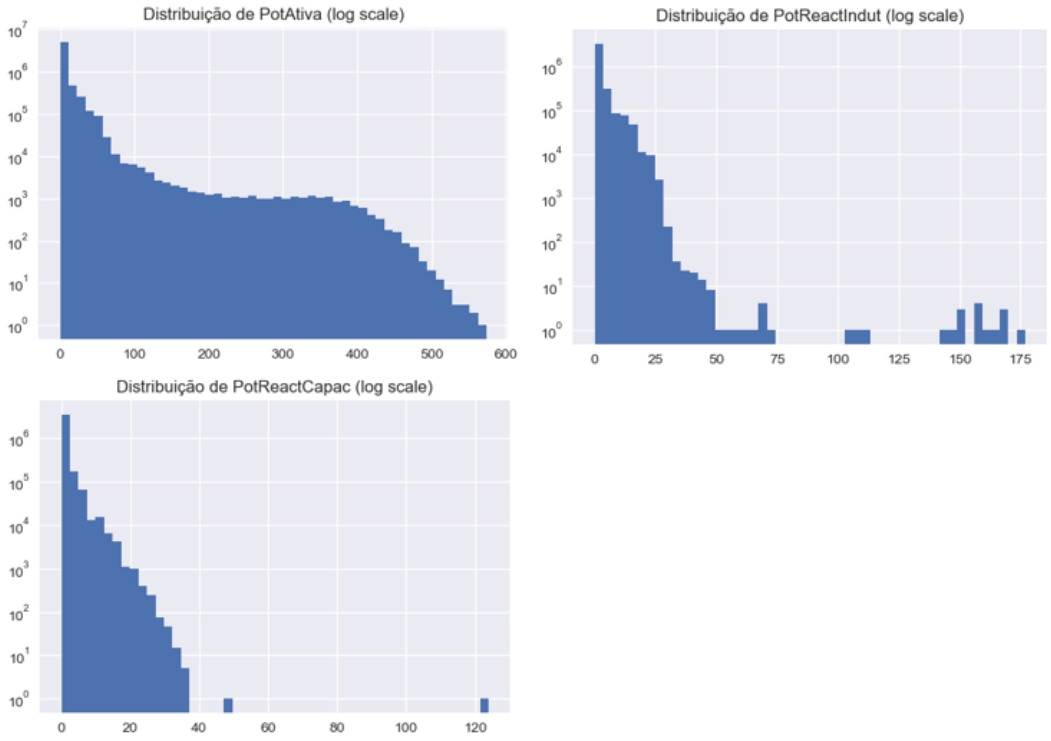
	Missing Count	Missing %
CPE	0	0.00
hora	0	0.00
DadosDeConsumo	5940182	100.00
PotAtiva	0	0.00
PotReactIndut	2271472	38.24
PotReactCapac	2271472	38.24

3.3. Análise das Variáveis Numéricas

3.3.1. Distribuições das Variáveis Numéricas

Foi efetuada a análise das distribuições das variáveis numéricas, nomeadamente da Potência Ativa e das potências reativas.
Os histogramas revelam distribuições assimétricas, com maior concentração de valores baixos e presença de picos ocasionais, que refletem edifícios com uma maior utilização energética.

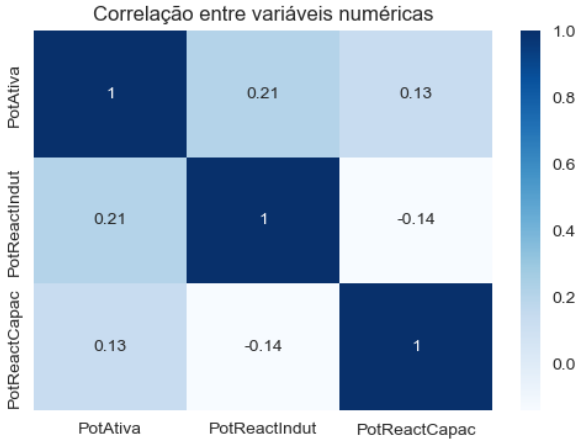
Figura 2 - Distribuições das principais variáveis numéricas



3.3.2. Correlação entre Variáveis Numéricas

Foi ainda analisada a correlação entre as variáveis numéricas através de um heatmap de correlação.
Observa-se que a potência ativa não depende fortemente das potências reativas. Quando a potência indutiva é mais alta, a capacitiva tende a ser um pouco mais baixa, e vice-versa, daí a fraca correlação. Esta análise sugere ainda que as potências reativas são úteis como qualificadores de perfis de edifícios, mas não como indicadores diretos do valor de consumo.

Figura 3 - Heatmap de correlação das variáveis numéricas

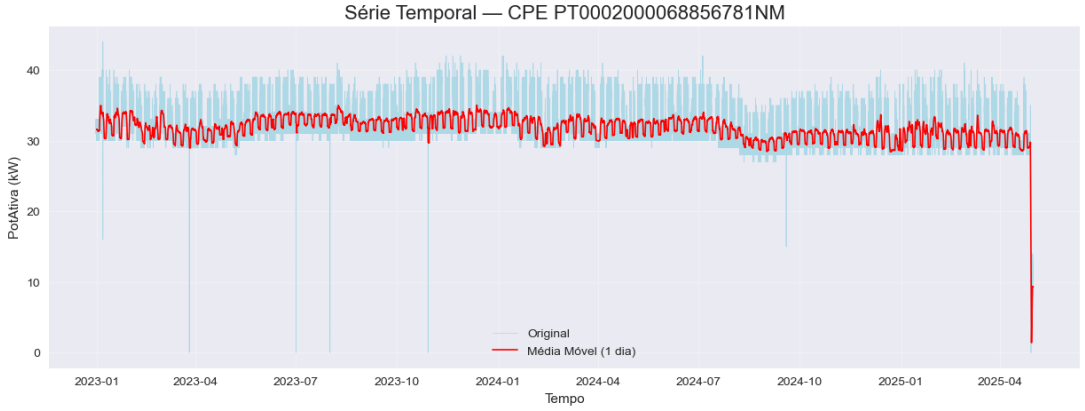


3.4. Análise Temporal do Consumo Energético

3.4.1. Evolução Temporal de um CPE Típico

Para analisar a evolução temporal da Potência Ativa foi selecionado o CPE 'PT0002000068856781NM' que representa um edifício escolar, pois era o que apresentava mais registos.
O gráfico evidencia padrões regulares de consumo, com ciclos diários e semanais bem definidos, refletindo os horários de funcionamento típicos de um edifício escolar.

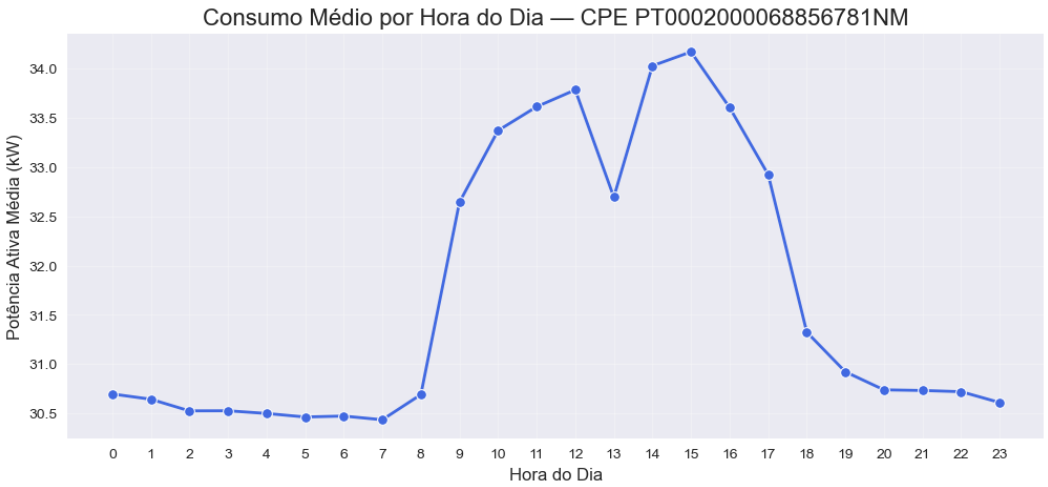
Figura 4 - Gráfico temporal do CPE com mais registos



3.4.2. Análise por Hora do Dia

A análise do consumo médio por hora do dia sobre o mesmo CPE permitiu identificar períodos de maior e menor utilização, observando-se picos de manhã entre as 8 e as 10 horas, um consumo estável durante o horário escolar das 10 às 16 horas, uma queda significativa ao final da tarde das 17 às 19 horas e um consumo mínimo durante a noite entre as 0 e as 6 horas.

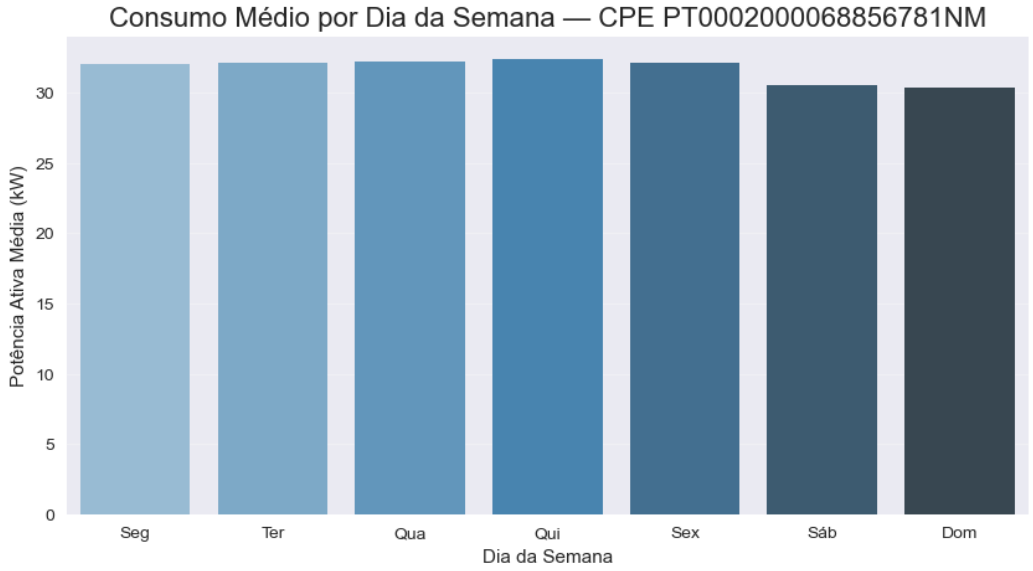
Figura 5 - Gráfico de consumo médio por hora do dia



3.4.3. Análise por Dia da Semana

A análise por dia da semana do mesmo CPE revelou diferenças entre dias úteis e fins de semana, com consumos mais elevados de segunda a sexta-feira, o que está alinhado com o funcionamento típico de edifícios escolares.

Figura 6 - Gráfico de consumo médio por dia da semana

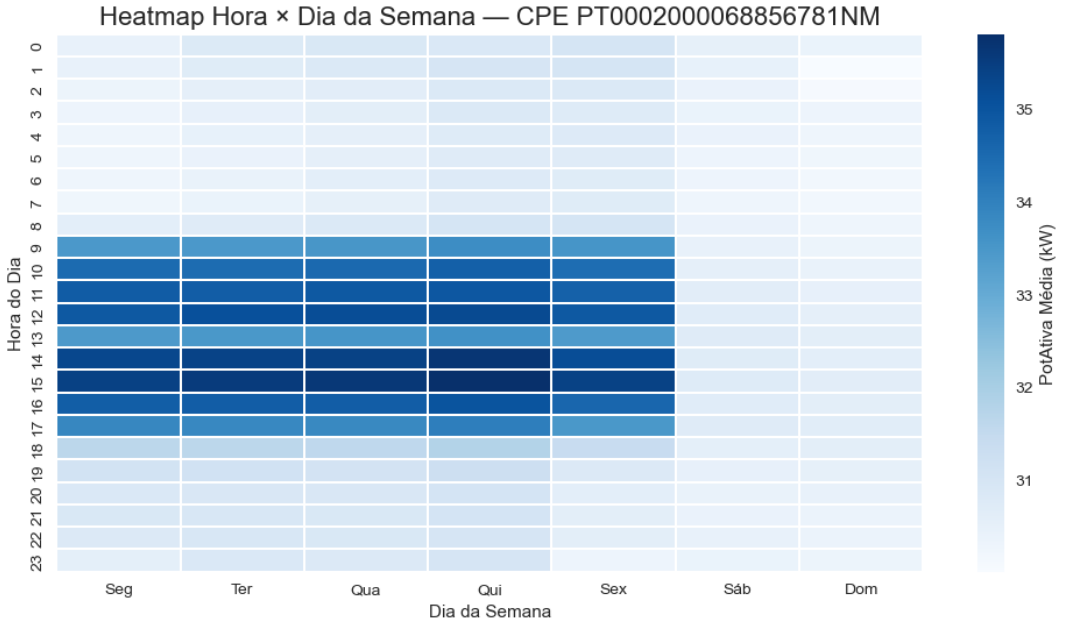


3.4.4. Análise Conjunta Hora × Dia da Semana (Heatmap)

Para aprofundar a análise temporal, foi feito um heatmap hora × dia da semana, representando a potência ativa média para o mesmo CPE que temos estado a observar.

Este tipo de visualização permite identificar padrões de consumo ao longo do dia e da semana, evidenciando de forma clara os períodos de maior e menor utilização do edifício. Observa-se um aumento significativo do consumo durante os dias úteis, concentrado sobretudo no período entre a manhã e o início da tarde, o que é consistente com horários de funcionamento típicos de edifícios escolares.

Figura 7 - Heatmap Hora × Dia da Semana



3.5. Criação de Dados Complementares por CPE

Após a análise exploratória inicial, procedeu-se à criação de um conjunto de dados agregados por CPE, com o objetivo de caracterizar globalmente o comportamento energético de cada edifício ao longo do período em análise. Esta agregação permitiu transformar a série temporal original num conjunto de features representativas, facilitando a comparação entre diferentes CPEs.
Para cada CPE foram calculadas várias métricas estatísticas da Potência Ativa, incluindo consumo medio, máximo e mínimo, desvio padrão, entre outros. Esta abordagem permitiu captar não apenas o nível médio de consumo, mas também a sua variabilidade e padrões de utilização.
Desta forma, os dados agregados por CPE funcionaram como uma ponte entre a análise temporal detalhada e os modelos de aprendizagem automática aplicados posteriormente.

4. Questão 1 – Clustering

4.1. General Dataset Preparation

A preparação geral dos dados teve como objetivo garantir a consistência, qualidade e integridade temporal do dataset original de consumo energético, criando uma base comum para todas as experiências desenvolvidas no projeto (Questões 1, 2 e 3).

4.1.1. Harmonização e seleção inicial das variáveis

Numa primeira fase, procedeu-se à renomeação das colunas, de forma a alinhar o dataset com a nomenclatura utilizada no relatório e no enunciado do projeto. Em particular, a variável temporal foi uniformizada para *hora* e a potência ativa para *PotAtiva*. A variável *DadosDeConsumo*, associada exclusivamente a fins de faturação, foi removida logo nesta fase, conforme indicado no enunciado, uma vez que não é relevante para a modelação do consumo energético e também não possuir dados.

4.1.2. Validação e tratamento da componente temporal

- A variável temporal (*hora*) foi convertida explicitamente para o formato datetime, com imposição de erros.
- Registos com timestamps inválidos ou em falta foram identificados e removidos.
- Foram igualmente eliminados registos sem identificação de CPE, uma vez que não podem ser corretamente integrados em análises por consumidor.
- Após esta limpeza, os dados foram ordenados cronologicamente por CPE e por instante temporal, garantindo a correta sequência das observações.

4.1.3. Tratamento de valores em falta

No que diz respeito ao tratamento de valores em falta, foram adotadas estratégias simples e transparentes. As variáveis de potência reativa (*PotReactIndut* e *PotReactCapac*) tiveram os seus valores em falta preenchidos com zero, permitindo manter estas variáveis disponíveis para análises exploratórias e para a extração de features, sem reduzir o número de registos. No caso da potência ativa (*PotAtiva*), os valores em falta foram imputados com a média do respetivo CPE, preservando o perfil médio de consumo de cada ponto e evitando a eliminação de observações potencialmente relevantes.
Após este tratamento, foi verificado que o dataset já não contém valores em falta nas variáveis essenciais.

4.1.4. Validação final do dataset preparado

Foi realizada uma verificação final garantindo que todos os registos apresentam:

- instante temporal válido (*hora*);
- identificação de CPE;
- valor válido de potência ativa.
- O intervalo temporal global do dataset foi identificado e documentado, assegurando transparência quanto ao período analisado.
- O dataset resultante, designado por df_prep, mantém a resolução temporal original de 15 minutos.

4.1.5. Exportação de datasets intermédios

De acordo com os requisitos do projeto, foram exportados datasets intermédios, incluindo:

- Conjuntos de features agregadas por CPE;
- Amostras reduzidas de séries temporais para execução rápida do notebook.

Estes datasets asseguram a reprodutibilidade dos resultados e o cumprimento das regras de submissão.
O dataset df_prep constitui uma base limpa, consistente e validada, adequada para:

- criação de variáveis temporais auxiliares;
- extração de features agregadas por CPE;
- construção de séries temporais por CPE;
- aplicação dos diferentes algoritmos de clustering e previsão.

Esta preparação geral é comum a todas as experiências realizadas no projeto e serve de ponto de partida para as fases de Data Preparation específica, Modelling e Evaluation em cada questão.

4.1.6. Validação dos dados

Figura 8 - Gráfico da comparação da distribuição de PotAtiva

A Figura acima apresenta a distribuição da variável *PotAtiva* antes e após a fase de preparação dos dados. Observa-se que a forma global da distribuição se mantém praticamente inalterada, indicando que os procedimentos de limpeza e preenchimento não introduziram distorções relevantes nos valores de consumo. Em particular, os valores extremos e a assimetria da distribuição foram preservados, o que é consistente com o comportamento real de consumos energéticos em edifícios com perfis heterogéneos. Esta validação é relevante, pois confirma que preparação dos dados preserva o comportamento estatístico original, enquanto assegura maior consistência e continuidade temporal para as etapas de modelação.

4.2. Data Preparation

Após a fase de preparação geral do dataset (Secção 4.1), na qual foram realizadas operações transversais de limpeza, tratamento de valores em falta, enriquecimento temporal e construção de variáveis derivadas, procedeu-se a uma preparação específica dos dados com vista à aplicação de técnicas de clustering.
O objetivo desta etapa é caracterizar cada CPE como uma entidade única, através de um conjunto de features agregadas que sintetizam os seus padrões de consumo ao longo do tempo, conforme recomendado no enunciado do projeto.

4.2.1. Definição e avaliação de múltiplos conjuntos de features

Durante o desenvolvimento do trabalho foram definidas **várias features potenciais**, incluindo:

- estatísticas globais de consumo;
- indicadores temporais detalhados;
- curvas médias horárias completas;
- percentagens de consumo por períodos do dia;
- métricas associadas à potência reativa.

Estes conjuntos foram explorados em fases preliminares com o objetivo de avaliar a sua utilidade para a separação de perfis de consumo. No entanto, para a aplicação final dos algoritmos de clustering, optou-se por uma seleção ponderada de features, privilegiando:

- capacidade de discriminação entre CPEs;
- interpretação clara dos clusters resultantes;

- redução de redundância entre variáveis;
- controlo da dimensionalidade, especialmente relevante para métodos baseados em distância.

O conjunto final utilizado para o clustering é composto pelas seguintes variáveis:

- **Estatísticas globais de consumo**
  mean_consumption, max_consumption, std_consumption
- **Características temporais do perfil de consumo**
  avg_daily_peak_time, avg_afternoon_peak_value
- **Indicadores de potência reativa**
  mean_PotReactIndut, mean_PotReactCapac, reactive_ratio
- **Distribuição do consumo por faixas horárias**
  pct_dawn, pct_morning, pct_afternoon, pct_evening
- **Comportamento semanal**
  pct_weekend

Este conjunto permite representar cada CPE de forma compacta e informativa, capturando simultaneamente:

- a intensidade e variação do consumo;
- os padrões horários dominantes;
- o comportamento diferenciado entre dias úteis e fins de semana;
- indícios do tipo de cargas elétricas através da potência reativa.

A seleção final privilegiou variáveis com significado físico claro e elevada capacidade discriminatória, reduzindo redundâncias e garantindo condições adequadas para a aplicação de algoritmos de clustering baseados em distância.

Com base neste conjunto final de features, foi construído um dataset agregado por CPE, que serviu de entrada para os algoritmos de clustering (K-Means, DBSCAN e GMM), sendo posteriormente avaliado com e sem normalização, conforme descrito nas secções seguintes de **Modelling** e **Evaluation**.

4.3. Modelling

Nesta fase foram aplicadas técnicas de aprendizagem não supervisionada com o objetivo de identificar grupos de CPEs com padrões de consumo energético semelhantes, com base no conjunto de features definido na Secção 4.2.

De acordo com o enunciado do projeto, foram implementados os algoritmos K-Means e DBSCAN. No entanto, dado que estes modelos não permitiram obter uma segmentação suficientemente estável e interpretável para todos os CPEs, foi adicionalmente introduzido o Gaussian Mixture Model (GMM) como abordagem complementar, permitindo capturar estruturas de clusters mais flexíveis.

Todos os modelos foram testados com e sem normalização, de forma a avaliar o impacto da escala das variáveis nos resultados de agrupamento.

4.3.1. K-Means

O algoritmo K-Means foi utilizado como primeira abordagem de clustering para identificar grupos de edifícios (CPEs) com perfis de consumo energético semelhantes. O método foi aplicado sobre o conjunto de features agregadas por CPE (definidas na fase de Data Preparation), que sintetizam padrões de consumo, comportamento temporal e características de potência reativa.
Como o K-Means é sensível à escala das variáveis, foram conduzidas duas experiências distintas:
(i) utilizando diretamente as features originais e
(ii) aplicando normalização (StandardScaler) antes do agrupamento.
Em ambos os casos, a escolha do número de clusters foi orientada por duas métricas: o método do cotovelo (inércia) e o Silhouette Score.

No cenário sem normalização, o método do cotovelo revela uma redução acentuada da inércia entre $k = 2$e $k = 4$, com ganhos marginais para valores superiores. Em paralelo, o Silhouette Score atinge o seu máximo em $k = 2(\approx 0.91)$, indicando uma separação extremamente forte entre dois grupos bem definidos.

Já no cenário com normalização, os valores do Silhouette Score são substancialmente mais baixos (≈ 0.33 para $k = 2$), sugerindo uma sobreposição considerável entre clusters e uma estrutura menos clara.

Assim, de acordo com ambas as métricas, o valor $k = 2$ foi selecionado para análise detalhada, sendo claramente superior no caso **sem normalização**.

Figura 9 – K-Means Sem normalização

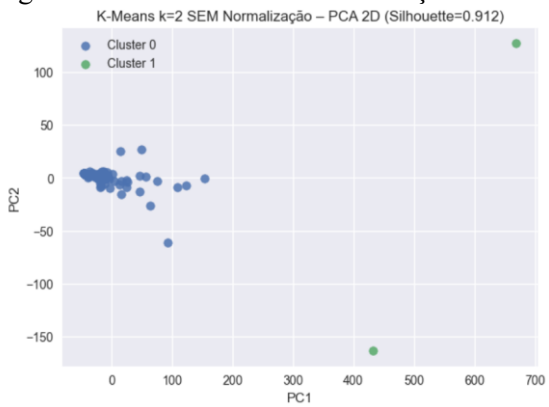
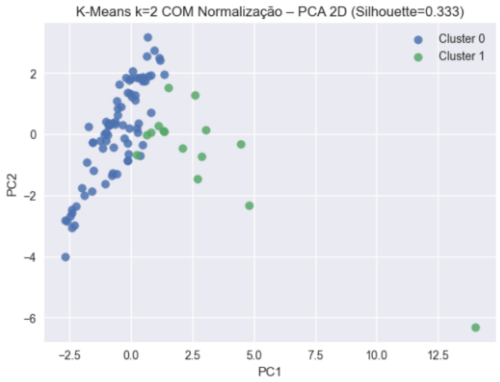


Figura 10 – K-Means Com normalização



A projeção dos dados em duas dimensões via PCA confirma os resultados quantitativos:

- Sem normalização, os dois clusters apresentam uma separação muito marcada. Um dos clusters contém a maioria dos CPEs com padrões relativamente homogéneos, enquanto o outro é composto por poucos pontos claramente afastados, correspondendo a consumidores atípicos (outliers) com valores de consumo ou potência significativamente mais elevados.
- Com normalização, a separação entre os clusters torna-se muito menos nítida. Os pontos distribuem-se de forma mais contínua no espaço PCA, o que justifica o valor baixo do Silhouette Score e indica que a normalização reduziu a influência dos CPEs extremos que dominavam a estrutura dos dados.

No cenário sem normalização, o K-Means essencialmente distingue:

- um grupo maior de **CPEs com padrões de consumo "normais"**, e
- um pequeno grupo de **consumidores de grande escala ou comportamento anómalo**, caracterizados por valores muito superiores de potência ativa e/ou reativa.

Embora esta separação seja matematicamente muito forte (Silhouette elevado), ela reflete sobretudo a presença de outliers e não uma segmentação rica de perfis de consumo (por exemplo, serviços diurnos vs. noturnos).

O K-Means revelou-se eficaz a identificar grandes discrepâncias de consumo, sobretudo na ausência de normalização, mas mostrou limitações na descoberta de subgrupos mais subtis de comportamento energético. A normalização, embora teoricamente desejável, reduziu drasticamente a separabilidade dos dados.

Por este motivo, e para capturar estruturas mais complexas e sobrepostas nos perfis de consumo dos CPEs, foi posteriormente introduzido o modelo GMM (Gaussian Mixture Model) como abordagem complementar de clustering.

4.3.2. DBSCAN

O algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) foi utilizado como alternativa ao K-Means para identificar agrupamentos de CPEs com base na densidade dos seus perfis de consumo.
O DBSCAN foi aplicado sobre o mesmo conjunto de *features* agregadas por CPE, sem normalização e com normalização (StandardScaler), utilizando $\varepsilon = 1.5$e min_samples $= 4$, conforme definido no notebook.

Figura 11 – DBSCAN Sem normalização

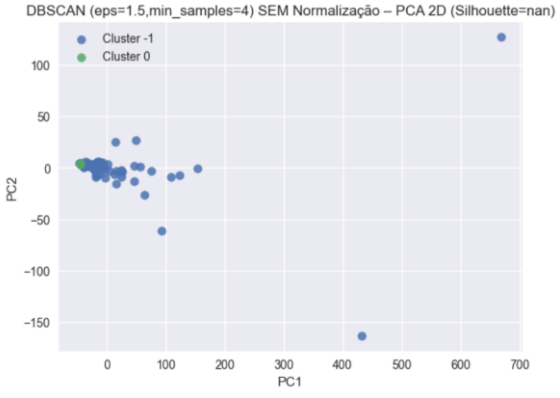
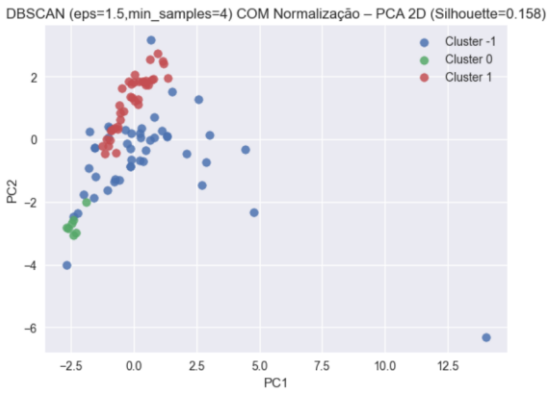


Figura 12 – DBSCAN Com normalização



Na versão sem normalização, o DBSCAN classificou a grande maioria dos CPEs como ruído (Cluster −1), identificando apenas um pequeno grupo como cluster válido. Esta situação impossibilita o cálculo do Silhouette Score (resultado *NaN*), uma vez que a métrica requer pelo menos dois clusters distintos.

A projeção em PCA 2D confirma este comportamento: os dados apresentam uma forte dispersão, com alguns CPEs extremamente afastados do grupo principal. A ausência de normalização faz com que variáveis com maior escala dominem a distância euclidiana, levando o DBSCAN a considerar a maioria dos pontos como isolados.

Assim, sem normalização, o DBSCAN funciona essencialmente como um detetor de outliers, mas não produz uma segmentação útil dos perfis de consumo.

Após normalização das *features*, o DBSCAN passa a identificar dois clusters densos (Cluster 0 e Cluster 1), mantendo simultaneamente um conjunto de pontos classificados como ruído (Cluster −1). A projeção PCA mostra agora uma estrutura mais equilibrada, em que os clusters ocupam regiões distintas do espaço latente.

O Silhouette Score obtido é de aproximadamente 0.158, indicando uma separação fraca entre os clusters. Este valor revela que, embora o DBSCAN consiga agora formar grupos, estes apresentam uma sobreposição considerável, o que limita a sua interpretabilidade.
Com normalização, o DBSCAN parece capturar pequenas diferenças locais nos perfis de consumo, separando subconjuntos de CPEs com padrões relativamente próximos, enquanto isola edifícios com comportamentos muito distintos como ruído. No entanto, a fraca qualidade do clustering sugere que os perfis dos CPEs formam um espaço relativamente contínuo, sem fronteiras densas bem definidas.

O DBSCAN mostrou-se útil para a deteção de outliers, especialmente após normalização, mas revelou limitações na identificação de grupos bem estruturados de consumidores. Em particular, o baixo Silhouette Score indica que os clusters obtidos não correspondem a perfis de consumo claramente diferenciados.

Dado que nem o K-Means nem o DBSCAN conseguiram produzir uma segmentação rica e interpretável dos CPEs, foi introduzido o Gaussian Mixture Model (GMM) como abordagem adicional, permitindo modelar distribuições sobrepostas e capturar padrões mais subtis nos dados.

4.3.3. Gaussian Mixture Models (GMM)

Após a aplicação de K-Means e DBSCAN, verificou-se que ambos apresentavam limitações relevantes na capacidade de capturar a diversidade real dos padrões de consumo dos edifícios. O K-Means mostrou forte sensibilidade a edifícios com consumos extremos, enquanto o DBSCAN revelou instabilidade e dificuldade em estruturar grupos consistentes.

Por esse motivo, foi introduzido o Gaussian Mixture Model (GMM) como abordagem complementar, por permitir:

- Modelar clusters com formas elípticas;
- Representar sobreposição entre grupos;
- Captar variações graduais entre perfis de consumo.

Foi testado k = 5, de forma consistente com os restantes métodos de clustering.

Figura 13 – GMM Sem normalização

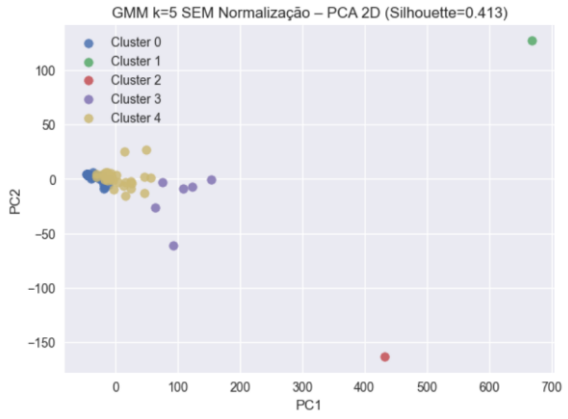
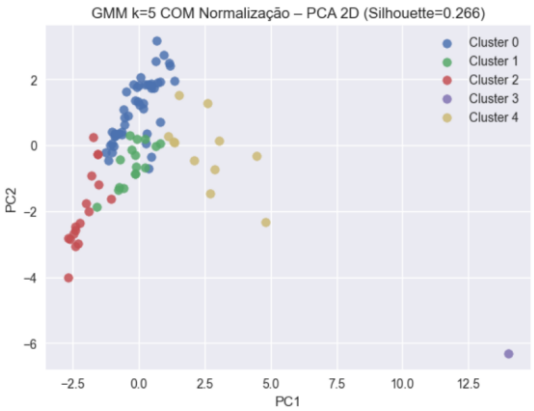


Figura 14 – GMM Com normalização



Embora o silhouette seja superior sem normalização, a análise visual e a interpretação dos clusters mostraram que a versão normalizada produz uma segmentação mais coerente do ponto de vista dos perfis energéticos.

Sem normalização a projeção PCA mostra:

- Um ou dois clusters dominados por edifícios de consumo muito elevado;
- A maioria dos edifícios comprimidos num único grupo;
- Separação fortemente influenciada pela magnitude do consumo.

Este comportamento explica o valor elevado do silhouette, mas indica que o modelo está essencialmente a separar outliers de consumo em vez de perfis de comportamento energético.

Após normalização:

- Os clusters distribuem-se ao longo de um gradiente de comportamento energético;
- Edifícios com padrões semelhantes (horários, regularidade e carga relativa) passam a agrupar-se;
- Os clusters tornam-se menos dependentes do nível absoluto de consumo.

Apesar do silhouette mais baixo, a estrutura obtida é mais consistente com o objetivo de caracterizar tipos de edifícios e perfis de utilização.

O GMM com normalização mostrou maior capacidade para:

- Identificar variações naturais entre edifícios;
- Separar padrões de utilização ao longo do dia;
- Evitar que edifícios de grande consumo dominem a segmentação.

Assim, os clusters passam a refletir diferenças de comportamento (por exemplo, edifícios com atividade diurna, edifícios com consumo noturno, edifícios com perfil misto), em vez de apenas diferenças de escala.

Apesar do silhouette score mais elevado na versão sem normalização, o GMM com normalização foi selecionado como modelo final de clustering, por demonstrar capacidade superior para

modelar variações naturais entre edifícios e produzir uma segmentação mais informativa dos perfis de consumo.

Este modelo fornece uma base mais sólida para a caracterização dos CPEs e para a interpretação dos grupos de consumidores no contexto do projeto.

4.4. Evaluation

A avaliação dos resultados do clustering teve como objetivo determinar se os grupos obtidos representam perfis de consumo energética distintos, coerentes e interpretáveis, de acordo com o objetivo do projeto de caracterizar diferentes tipos de consumidores (edifícios). Esta avaliação foi realizada com base em métricas quantitativas e na análise dos perfis obtidos a partir das variáveis explicativas.
Após a comparação entre K-Means, DBSCAN e Gaussian Mixture Models (GMM), o modelo GMM com normalização foi selecionado como solução final, por apresentar maior capacidade de modelar a variabilidade natural dos edifícios e separar perfis de consumo de forma mais realista.

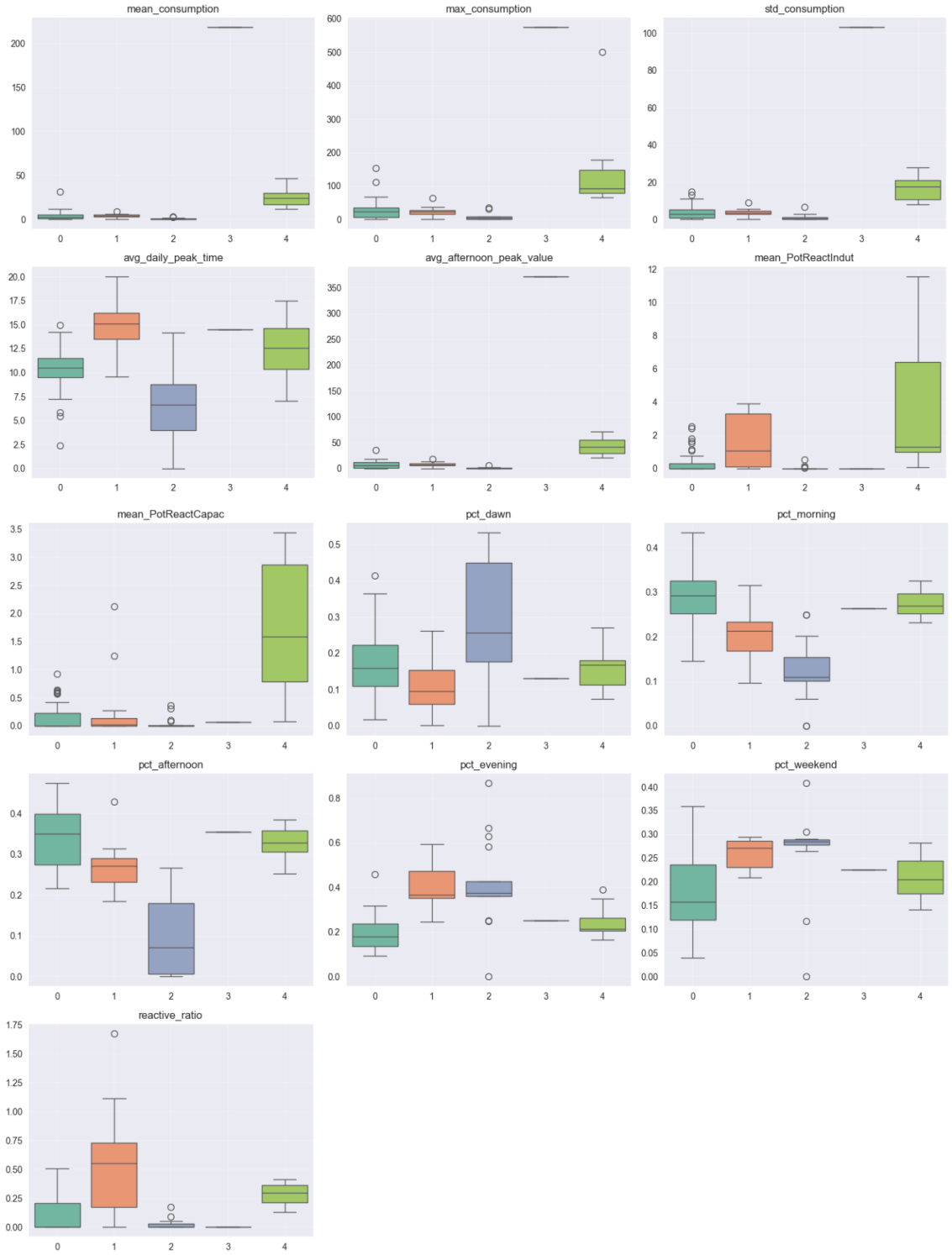
4.4.1 Qualidade dos clusters

O GMM com normalização obteve um Silhouette Score de 0,266, indicando uma separação moderada entre clusters. Embora este valor seja inferior ao observado no K-Means sem normalização, esse modelo produzia essencialmente um grande cluster dominante e alguns outliers extremos, não permitindo uma caracterização útil dos edifícios.

O GMM, por outro lado, consegue modelar sobreposição entre grupos, o que é esperado em edifícios reais, que frequentemente apresentam comportamentos híbridos (por exemplo, edifícios com atividade administrativa durante o dia e cargas técnicas à noite). Assim, mesmo com um silhouette inferior, os clusters obtidos são estruturalmente mais realistas e semanticamente mais informativos.

4.4.2 Análise dos perfis de consumo

Figura 15 – Distribuição das Features por Cluster – GMM com Normalização)



Cluster 0 – Perfil diurno estável

Apresenta consumo médio, baixa variabilidade, pico de consumo por volta do final da manhã e elevada percentagem de consumo durante o período da manhã e tarde, com baixo consumo noturno e baixa potência reativa. Este perfil é compatível com edifícios administrativos e serviços municipais.

Cluster 1 – Perfil técnico e reativo

Caracteriza-se por valores elevados de potência reativa indutiva e rácio reativo, consumo significativo à noite e ao fim de semana. Este padrão é típico de edifícios com equipamentos eletromecânicos, AVAC ou sistemas industriais.

Cluster 2 – Perfil noturno

Apresenta percentagens elevadas de consumo no período da noite, baixos valores de potência reativa e picos de consumo mais cedo no dia. Este cluster é compatível com infraestruturas automáticas, iluminação pública ou sistemas de vigilância.

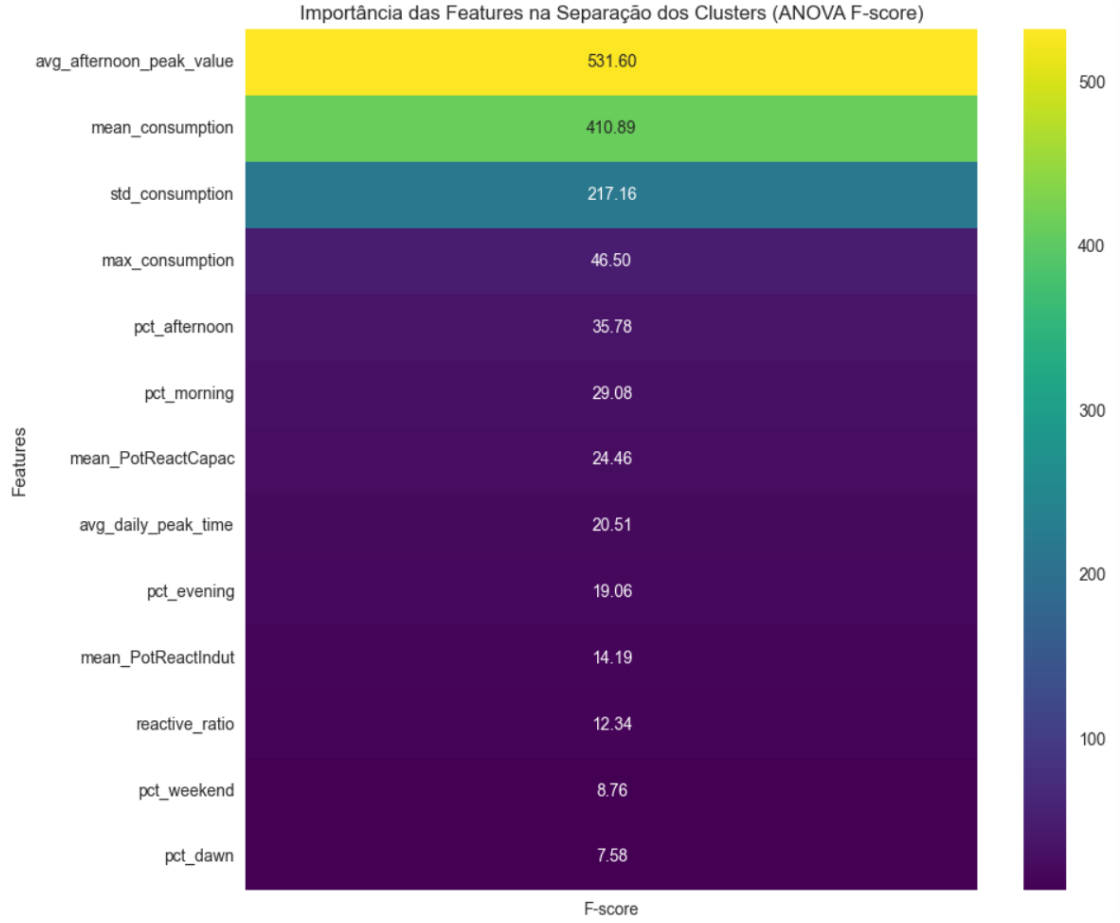
Cluster 3 – Outlier estrutural

O Cluster GMM 3 inclui apenas o CPE PT0002000078441876HB, que segundo o portal Baze (CM-Maia) corresponde à Torre do Lidador. Este edifício apresenta consumos muito elevados, picos acentuados à tarde e elevada variabilidade, o que explica a sua separação pelo GMM como um outlier estrutural, típico de uma infraestrutura de grande dimensão e carga energética intensiva.

Cluster 4 – Alto consumo e elevada variabilidade

Caracteriza edifícios com consumo médio-alto, grande desvio padrão, picos elevados à tarde e rácio reativo médio. Corresponde a edifícios com uso intensivo e comportamento operacional variável, como pavilhões, oficinas ou centros operacionais.

Figura 16 – Heatmap de "importância" das features para clusters (ANOVA F-score)



A análise dos ANOVA F-scores mostra que a separação dos clusters é dominada sobretudo por variáveis relacionadas com a intensidade do consumo. A feature mais relevante é **avg_afternoon_peak_value**, indicando que os picos de consumo no período da tarde são o principal fator de diferenciação entre os edifícios.

Seguem-se **mean_consumption** e **std_consumption**, o que confirma que tanto o nível médio de consumo como a sua variação são determinantes na formação dos grupos. As restantes features, incluindo percentagens por período do dia e indicadores de potência reativa, têm um contributo complementar, mas menos expressivo.

Em conjunto, estes resultados indicam que os clusters refletem sobretudo diferenças nos padrões de carga e nos picos de utilização, em especial durante os períodos de maior atividade.

4.4.2 Justificação da escolha do GMM

Comparativamente aos restantes métodos:

- O K-Means sem normalização agrupava apenas pela magnitude do consumo.
- O K-Means com normalização impunha fronteiras rígidas e misturava perfis distintos.

- O DBSCAN mostrou elevada sensibilidade aos parâmetros e produziu muitos pontos classificados como ruído.

O GMM com normalização revelou-se superior por:

- Modelar variâncias diferentes entre clusters
- Permitir sobreposição probabilística
- Separar edifícios híbridos e outliers de forma robusta

4.5. Conclusão

O processo de clustering permitiu identificar diferentes perfis de consumo entre os edifícios, com base em padrões temporais, intensidade de utilização e comportamento da potência reativa. A criação de features por CPE revelou-se fundamental para tornar os padrões comparáveis e interpretáveis.

Entre os métodos testados, o GMM com normalização apresentou a melhor capacidade de distinguir grupos com comportamentos distintos, acomodando variações naturais e sobreposição entre perfis, algo típico de edifícios reais. Os clusters obtidos refletem diferenças claras entre edifícios de uso diurno, infraestruturas mais intensivas e perfis com maior componente reativa.

Apesar de existirem limitações na separação perfeita dos grupos, os resultados mostram que o clustering fornece uma segmentação útil e consistente, que pode ser explorada em fases posteriores de análise e modelação preditiva.

5. Questão 2 – Time-Series Prediction (ARIMA & LSTM)

5.1. Data Preparation

Para a fase de previsão temporal, os dados foram preparados de forma a garantir coerência temporal, ausência de fuga de informação e alinhamento com o enunciado do projeto, que exige a previsão de valores futuros com base em dados históricos com pelo menos uma semana de antecedência.

O conjunto de dados limpo (df_prep) foi utilizado como base, contendo, para cada CPE, as leituras de PotAtiva com resolução de 15 minutos. A preparação seguiu os seguintes passos:

5.1.1. Construção das séries temporais por CPE

Para cada edifício (CPE), foi construída uma série temporal regular:

- Filtragem dos registos do respetivo CPE
- Conversão da coluna hora para formato datetime
- Ordenação cronológica
- Remoção de timestamps duplicados
- Reamostragem para frequência fixa de 15 minutos (15min)

- Preenchimento de falhas temporais por forward-fill (ffill)

Este processo assegura séries temporais uniformes, requisito fundamental para ARIMA e LSTM.

5.1.2. Seleção dos CPEs por cluster (GMM)

Os modelos de previsão não foram aplicados aleatoriamente a todos os CPEs. Em vez disso, foi usada a segmentação obtida pelo clustering GMM, garantindo que os CPEs utilizados representam perfis distintos de consumo.

Para cada cluster GMM:

- Foi selecionado um CPE representativo
- Apenas CPEs com histórico suficientemente longo foram usados

Este procedimento evita treinar modelos em séries demasiado curtas ou atípicas.

5.1.3. Divisão treino / teste

Para cada série temporal selecionada:

- **70% inicial** (conjunto de treino)
- **30% final** (conjunto de teste)

Esta divisão respeita a ordem temporal dos dados, evitando qualquer fuga de informação do futuro para o passado.

5.1.4. Definição do horizonte de previsão

O horizonte de interesse é uma semana, conforme o enunciado:

$$1 \text{ semana} = 7 \text{ dias} \times 96 \text{ registos/dia} = 672 \text{ observações}$$

Todos os modelos e baselines são avaliados exclusivamente sobre este horizonte.

5.1.5. Definição do baseline

O baseline exigido no enunciado foi implementado como:

"o consumo num determinado instante é igual ao consumo no mesmo instante, uma semana antes"

Isto foi operacionalizado deslocando a série temporal em 672 passos:

$$Baseline(t) = PotAtiva(t - 1 \text{ semana})$$

Apenas os instantes do conjunto de teste que tinham referência válida uma semana antes foram usados na avaliação.

Esta preparação assegura que os modelos ARIMA e LSTM são avaliados de forma rigorosa, reproduzível e totalmente alinhada com os requisitos do enunciado, permitindo uma comparação justa entre modelos e baseline para a previsão da semana seguinte.

5.2. Modelling

Para a previsão da PotAtiva a uma semana de antecedência, foram implementados dois tipos de modelos de séries temporais: ARIMA (modelo estatístico) e LSTM (modelo de deep learning). Ambos foram aplicados a séries temporais individuais de CPEs selecionados por cluster GMM, garantindo a representação de diferentes perfis de consumo.

Em ambos os casos, foi utilizado um split temporal de 70% para treino e 30% para teste, e um baseline sazonal definido como o valor da mesma hora da semana anterior.

Tanto no ARIMA como na LSTM, os modelos foram testados em duas variantes:
Sem normalização, usando diretamente os valores originais de PotAtiva;
Com normalização, aplicando StandardScaler às séries temporais antes do treino, com posterior inversão da escala para kW nas previsões.

No ARIMA, o modelo foi ajustado apenas com o período de treino e usado para prever todo o período de teste.
Na LSTM, foram usadas janelas temporais de 12 horas (48 amostras) para prever o valor de PotAtiva uma semana à frente, garantindo a coerência com o horizonte definido no enunciado.

A comparação quantitativa entre modelos e baseline, bem como o impacto da normalização, é apresentada na secção de Avaliação.

5.3. Evaluation

5.3.1. Evaluation (ARIMA & Baseline)

A avaliação do modelo ARIMA foi realizada para um CPE representativo de cada cluster GMM, de forma a testar o desempenho do modelo em perfis de consumo estruturalmente distintos. Para cada CPE, a série temporal de PotAtiva foi dividida em 70% para treino e 30% para teste, e o objetivo foi prever o consumo futuro, sendo comparado contra um baseline sazonal definido como o valor da mesma hora uma semana antes.

O desempenho foi avaliado usando MAE (Mean Absolute Error) e RMSE (Root Mean Squared Error), calculados tanto para o modelo ARIMA como para o baseline. O processo foi executado em duas variantes, sem normalização da série temporal e com normalização (StandardScaler) e posterior inversão para kW

As Tabelas abaixo mostram os resultados obtidos para ambos os cenários.

Tabela 1 – Resultados Sem normalização GMM

	Cluster_GMM	CPE	MAE_ARIMA	RMSE_ARIMA	MAE_Baseline	RMSE_Baseline
0	0	PT0002000032942455NH	0.081862	0.168831	0.070943	0.203775
1	1	PT0002000068856906VS	3.101161	4.098381	2.460469	3.290435
2	2	PT0002000032936306KX	2.668606	3.739536	0.123747	0.516918
3	3	PT0002000078441876HB	131.626721	166.532320	45.545114	63.952701
4	4	PT0002000068857897ZV	7.166093	12.430746	2.595822	5.603652

Tabela 2 – Resultados Com normalização GMM

	Cluster_GMM	CPE	MAE_ARIMA_norm	RMSE_ARIMA_norm	MAE_Baseline	RMSE_Baseline
0	0	PT0002000032942455NH	0.081860	0.168843	0.070943	0.203775
1	1	PT0002000068856906VS	3.101174	4.098392	2.460469	3.290435
2	2	PT0002000032936306KX	2.668628	3.739552	0.123747	0.516918
3	3	PT0002000078441876HB	131.626902	166.532475	45.545114	63.952701
4	4	PT0002000068857897ZV	7.166032	12.430710	2.595822	5.603652

O desempenho do modelo ARIMA foi avaliado em cada cluster GMM utilizando um CPE representativo, sendo comparado com um baseline sazonal definido como o valor da mesma hora uma semana antes. As métricas utilizadas foram o MAE (Mean Absolute Error) e o RMSE (Root Mean Squared Error), calculadas sobre o conjunto de teste (30% final da série).

Os resultados mostram que o baseline apresenta, de forma consistente, melhor desempenho médio. Em todos os clusters (0 a 4), o MAE do baseline é inferior ao do ARIMA, indicando que a regra simples baseada na semana anterior fornece previsões mais precisas em termos de erro médio. Em termos de RMSE, o baseline também supera o ARIMA em quatro dos cinco clusters, sendo a única exceção o Cluster 0, onde o ARIMA apresenta um ligeiro ganho na redução de erros extremos.
Assim, o ARIMA apenas é eficaz no Cluster GMM 0, correspondente a perfis de consumo baixos e mais regulares, enquanto falha em clusters com maior carga e variação, onde a simples repetição do padrão semanal (baseline) é mais precisa.

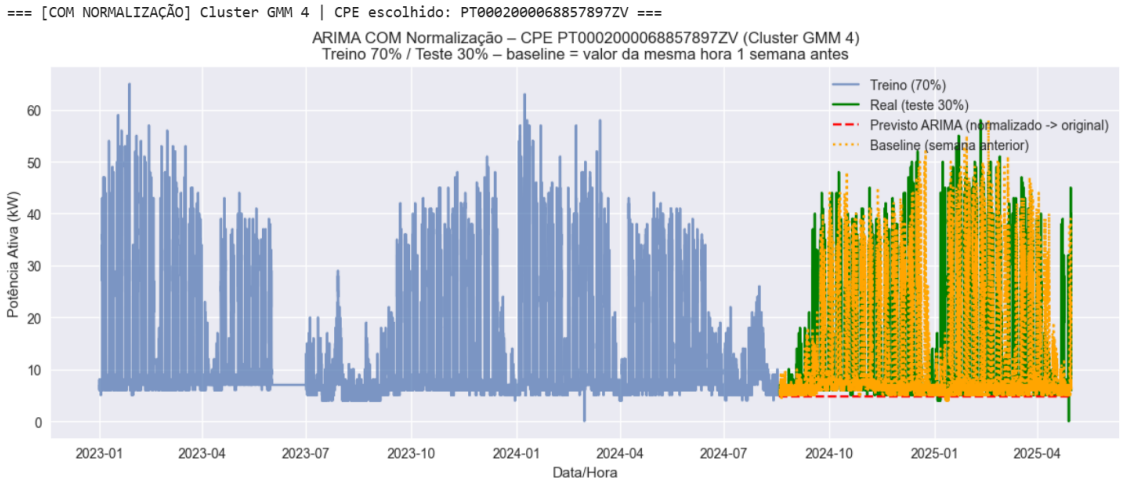
A comparação entre os resultados com e sem normalização da série revela que a normalização não teve impacto relevante no desempenho do ARIMA, uma vez que as métricas MAE e RMSE são praticamente idênticas em ambas as versões. Isto é esperado, dado que o ARIMA opera sobre diferenças e estruturas temporais, sendo invariável a transformações lineares da escala.

Em síntese, para os CPEs representativos dos clusters analisados, o ARIMA não conseguiu superar de forma consistente uma baseline sazonal simples. Este resultado sugere que, neste contexto, a forte regularidade semanal do consumo elétrico já é suficientemente capturada pela baseline, limitando o valor acrescentado do modelo ARIMA.

Para complementar a avaliação quantitativa, apresenta-se uma visualização comparativa entre a série real, a previsão do ARIMA e o baseline sazonal no conjunto de teste (30% final). A Figura

16 ilustra um exemplo representativo (Cluster GMM 4), permitindo observar a proximidade relativa do ARIMA face ao baseline

Figura 17 – ARIMA vs Baseline na previsão semanal de PotAtiva (Cluster GMM 4, CPE PT0002000068857897ZV



5.3.2. Evaluation (LSTM & Baseline)

A avaliação do modelo LSTM foi realizada por cluster GMM, comparando as previsões a 1 semana de horizonte com um baseline temporal, definido como o valor observado na mesma hora da semana anterior. O desempenho foi quantificado através das métricas MAE (Mean Absolute Error) e RMSE (Root Mean Squared Error), tanto sem normalização como com normalização da série temporal.

Tabela 3 – Resultados Sem normalização LSTM

	Cluster_GMM	CPE	MAE_LSTM	RMSE_LSTM	MAE_Baseline	RMSE_Baseline
0	0	PT0002000032942455NH	0.073268	0.166316	0.070943	0.203775
1	1	PT0002000068856906VS	2.094066	2.662999	2.460469	3.290435
2	2	PT0002000032936306KX	0.381432	0.744879	0.123747	0.516918
3	3	PT0002000078441876HB	97.154824	130.261977	45.545114	63.952701
4	4	PT0002000068857897ZV	2.816434	5.275678	2.595822	5.603652

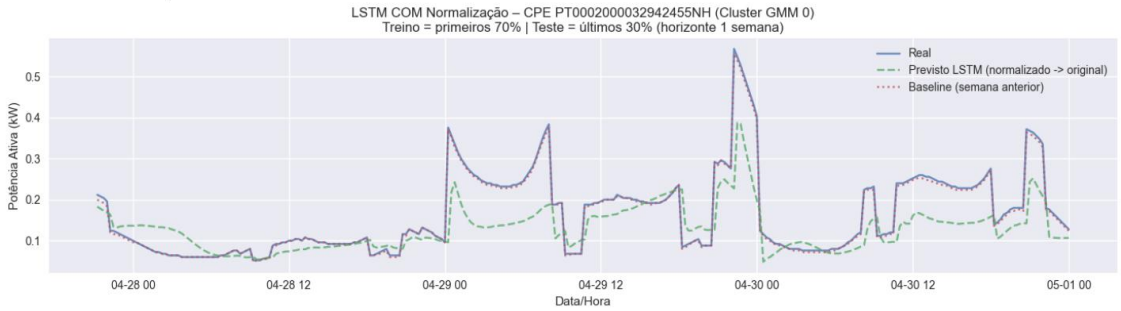
Tabela 4 – Resultados Com normalização LSTM

	Cluster_GMM	CPE	MAE_LSTM_norm	RMSE_LSTM_norm	MAE_Baseline	RMSE_Baseline
0	0	PT0002000032942455NH	0.072791	0.166161	0.070943	0.203775
1	1	PT0002000068856906VS	2.073396	2.626389	2.460469	3.290435
2	2	PT0002000032936306KX	0.408709	0.739875	0.123747	0.516918
3	3	PT0002000078441876HB	43.168720	57.025997	45.545114	63.952701
4	4	PT0002000068857897ZV	2.864192	5.261602	2.595822	5.603652

Figura 18 – Previsão semanal de Potência Ativa — LSTM vs Baseline (Cluster GMM 0)

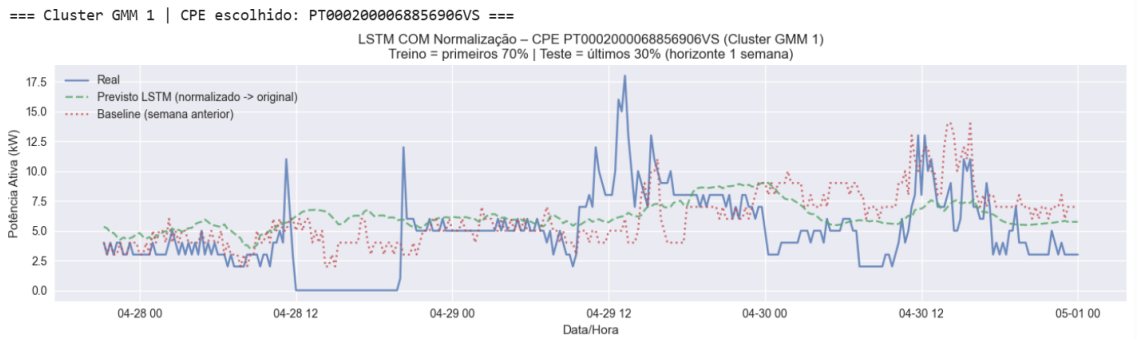
```
Clusters GMM existentes: [np.int64(0), np.int64(1), np.int64(2), np.int64(3), np.int64(4)]

=== Cluster GMM 0 | CPE escolhido: PT0002000032942455NH ===
```



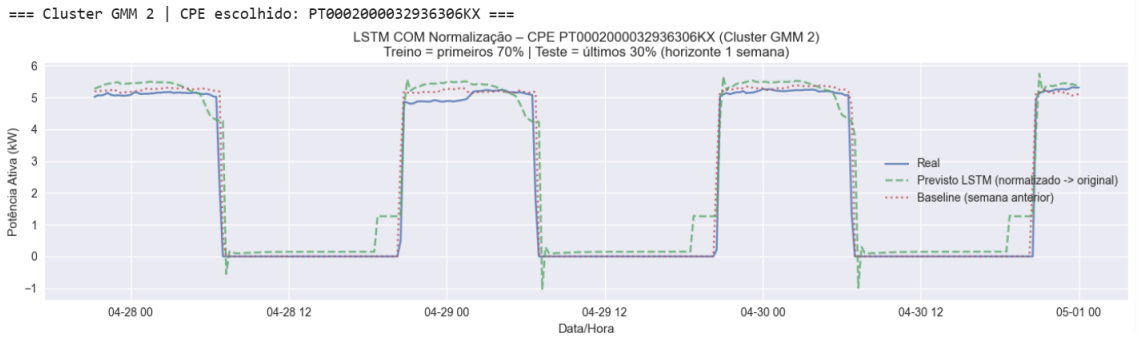
O Cluster GMM 0, caracterizado por consumos baixos e comportamento altamente regular, apresentou um dos melhores desempenhos do modelo LSTM. Tanto com como sem normalização, o LSTM obteve valores de erro muito reduzidos, próximos do baseline. Em termos de MAE, o baseline manteve uma ligeira vantagem, enquanto em RMSE o LSTM apresentou melhor desempenho, indicando maior capacidade para reduzir erros de maior magnitude e acompanhar variações intra-semana.

Figura 19 – Previsão semanal de Potência Ativa — LSTM vs Baseline (Cluster GMM 1)



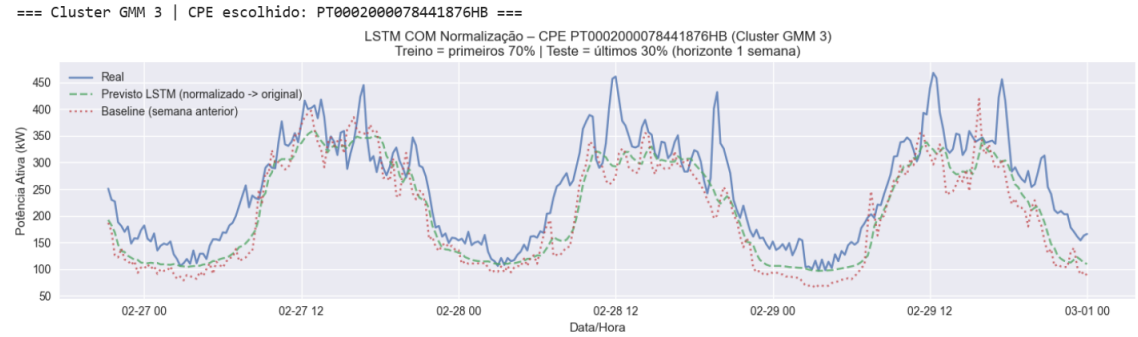
No Cluster 1, caracterizado por consumo médio e maior variabilidade, o LSTM supera claramente o baseline, apresentando reduções consistentes de MAE e RMSE, especialmente quando a série é normalizada. Este resultado indica que o modelo consegue aprender padrões temporais que não são captados pela simples repetição semanal.

Figura 20 – Previsão semanal de Potência Ativa — LSTM vs Baseline (Cluster GMM 2)



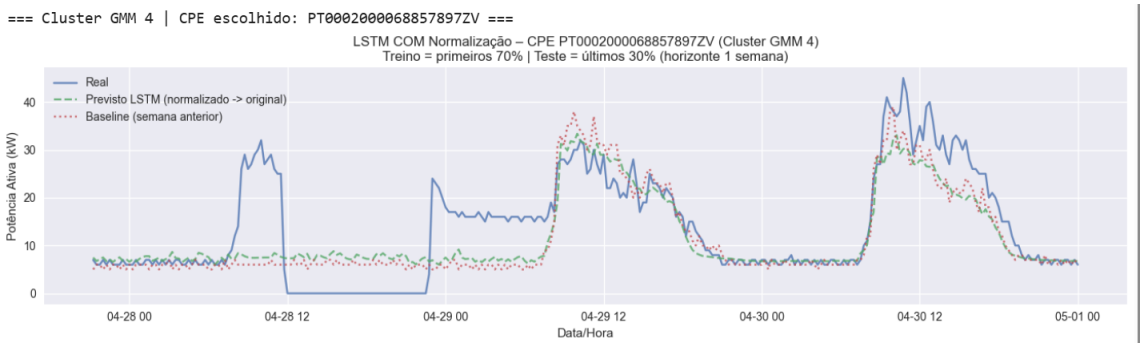
No Cluster 2, que apresenta um perfil extremamente regular, o baseline tem desempenho superior ao LSTM. Neste caso, o comportamento quase determinístico da série favorece a abordagem baseada na semana anterior, tornando o uso de um modelo neural menos vantajoso.

Figura 21 – Previsão semanal de Potência Ativa — LSTM vs Baseline (Cluster GMM 3)



O Cluster 3, com consumo muito elevado e grande variabilidade, evidencia de forma clara a importância da normalização. Sem normalização, o LSTM apresenta erros muito elevados; contudo, após normalização, o modelo melhora substancialmente e passa a superar o baseline, reduzindo significativamente tanto o MAE como o RMSE. Isto confirma que, para séries de grande escala, a normalização é essencial para permitir uma aprendizagem eficaz.

Figura 22 – Previsão semanal de Potência Ativa — LSTM vs Baseline (Cluster GMM 4)



No Cluster 4, com consumo elevado e padrão sazonal bem definido, o desempenho do LSTM é próximo do baseline, com diferenças pouco expressivas. A normalização melhora a estabilidade do modelo, mas o baseline continua a ser altamente competitivo devido à forte repetição semanal do consumo.

Efeito da normalização

A normalização da série temporal teve impacto relevante em clusters com consumo elevado e alta variabilidade (clusters 1, 3 e 4), permitindo ao LSTM aprender padrões relativos em vez de amplitudes absolutas. Nos clusters de baixo consumo e maior estabilidade (clusters 0 e 2), a diferença entre usar ou não normalização foi mínima

5.4. Conclusão

A comparação entre os modelos ARIMA e LSTM, avaliados contra o baseline da "semana anterior" e aplicados a séries temporais representativas de cada cluster GMM, demonstra que o desempenho dos modelos depende fortemente do tipo de perfil de consumo.

O ARIMA mostrou-se eficaz apenas em perfis simples e de baixa variação, como no Cluster GMM 0, mas revelou grandes limitações em clusters com elevada variação, picos acentuados ou mudanças estruturais (Clusters 3 e 4). Nestes casos, o modelo falhou em acompanhar a dinâmica real da série, apresentando erros elevados mesmo quando foi aplicada normalização.

O LSTM, sobretudo quando treinado sobre séries normalizadas, apresentou um desempenho globalmente superior. O modelo conseguiu aprender padrões temporais complexos, captando ciclos diários, rampas de subida e descida e variações sazonais. Isto é particularmente evidente nos Clusters GMM 1, 3 e 4, onde o LSTM reduziu de forma significativa os erros relativamente ao baseline e ao ARIMA. Apenas no Cluster GMM 2, caracterizado por um comportamento quase determinístico, o baseline simples da semana anterior permaneceu mais eficaz.

Em síntese, os resultados confirmam que:

- ARIMA é adequado apenas para perfis simples e estáveis;
- LSTM é mais robusto e escalável para padrões reais de consumo energético, sobretudo quando combinado com normalização;
- A segmentação prévia por clusters GMM é fundamental, pois permite aplicar modelos de previsão mais ajustados a cada tipo de perfil.

Estes resultados validam a abordagem híbrida de Clustering + Deep Learning como uma estratégia eficaz para previsão de consumo energético heterogéneo.

6. Questão 3 – Feature-based Prediction (RF / MLP)

6.1. Data Preparation

Nesta questão, o objetivo é prever o consumo médio de energia de uma semana futura de cada edifício (CPE) com base em características agregadas do seu histórico, em vez de utilizar diretamente a série temporal completa. Para isso, foi construído um dataset supervisionado ao nível do CPE, garantindo ausência de data leakage e consistência entre os diferentes modelos (Random Forest, MLP).

6.1.1. Construção das séries temporais por CPE

Para cada edifício (CPE), foi construída uma série temporal regularizada a 15 minutos, incluindo a potência ativa e, quando disponível, as potências reativas:

- Registos duplicados no mesmo instante foram agregados por média;
- A série foi ordenada temporalmente;
- A frequência foi regularizada para 15 minutos usando forward-fill.

Isto assegura séries contínuas e comparáveis entre edifícios.

6.1.2. Separação temporal treino / teste

Para cada CPE, a série foi dividida cronologicamente em:

- 70% treino
- 30% teste

Esta divisão preserva a causalidade temporal e evita qualquer uso de informação futura no cálculo das variáveis explicativas.
Foram apenas considerados edifícios com pelo menos 3 semanas de dados, garantindo histórico suficiente para cálculo das features, definição do target e construção do baseline

6.1.3. Definição do target e do baseline

A previsão é formulada como um problema supervisionado:

- Target (y_true_future_week)
  Média da potência ativa na primeira semana do período de teste.
- Baseline (y_base_future_week)
  Média da potência ativa na última semana do período de treino
  (equivalente a assumir que a próxima semana será igual à semana anterior).

Este baseline permite avaliar se os modelos realmente aprendem padrões úteis.

6.1.4. Extração das features (usando apenas o treino)

Todas as features foram calculadas exclusivamente a partir dos dados de treino, prevenindo *data leakage*. Para cada CPE foram extraídas as mesmas selected_features usadas no clustering, assegurando coerência entre as abordagens.

6.1.5. Construção do dataset supervisionado

Cada linha do dataset final representa um edifício (CPE) e contém:

- As features calculadas a partir do treino;
- O target: consumo médio da semana seguinte;
- O baseline: consumo médio da semana anterior.

O resultado é uma matriz do tipo:

$$\text{Features do CPE} \rightarrow \text{Consumo médio da próxima semana}$$

6.1.6. Preparação específica para cada modelo

Random Forest

- Utilizam diretamente o dataset X_rf e y_rf;
- Não requerem normalização obrigatória das features.

MLP

- As features foram normalizadas (StandardScaler);
- O scaler foi ajustado apenas no conjunto de treino e aplicado ao teste, evitando *leakage*;
- O *split* treino/teste foi feito ao nível de CPE, garantindo generalização entre edifícios.

Este processo transforma um problema de séries temporais num problema de regressão supervisionada baseada em perfis de consumo, permitindo que RF, MLP aprendam relações entre características históricas dos edifícios e o seu consumo futuro. A utilização de targets semanais e de um baseline explícito fornece uma base sólida para avaliação comparativa dos modelos.

6.2. Modelling

Para a previsão do consumo médio da semana seguinte, foram utilizados dois modelos supervisionados baseados em *features* agregadas por CPE: Random Forest (RF) e Multi-Layer Perceptron (MLP).

O Random Forest foi aplicado diretamente às *selected_features*, explorando relações não lineares entre padrões de consumo, horários, variabilidade e variáveis reativas. A sua robustez a escalas diferentes e a ruído torna-o adequado para este tipo de dados tabulares.

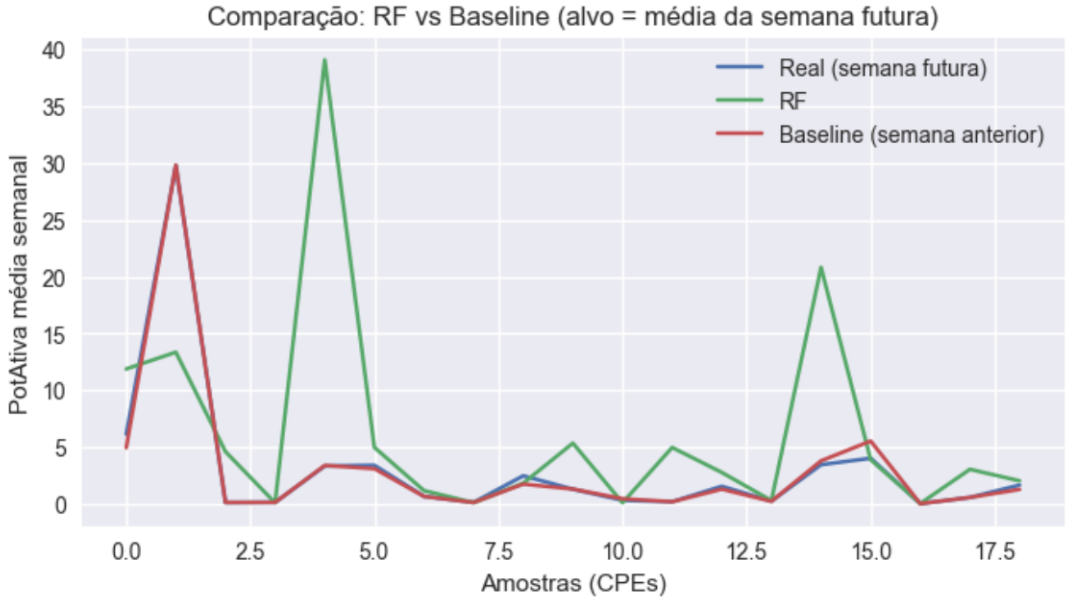
O MLP foi treinado sobre as mesmas *features*, após normalização com StandardScaler, ajustado apenas no conjunto de treino para evitar *data leakage*. A divisão treino/teste foi feita ao nível dos CPEs, assegurando generalização entre edifícios.

Ambos os modelos foram treinados para prever a média de consumo da semana futura, permitindo comparar uma abordagem baseada em árvores (RF) com uma abordagem de rede neuronal (MLP) no contexto de previsão baseada em características.

6.3. Evaluation

6.3.1. Random Forest

Figura 23 – RF vs Baseline (alvo = média da semana futura)



A Figura "RF vs Baseline (alvo = média da semana futura)" compara, para cada CPE, o valor real da média da semana seguinte, a previsão do Random Forest e o baseline, definido como a média da semana anterior.

Os resultados globais confirmam que o baseline é claramente superior ao Random Forest neste problema:

- Random Forest
  MAE = 5.05 kW
  RMSE = 10.15 kW
- Baseline (semana anterior)
  MAE = 0.26 kW
  RMSE = 0.50 kW

Estes valores mostram que o Random Forest apresenta erros quase 20 vezes superiores ao baseline, indicando fraca capacidade de generalização para a previsão da média semanal futura.

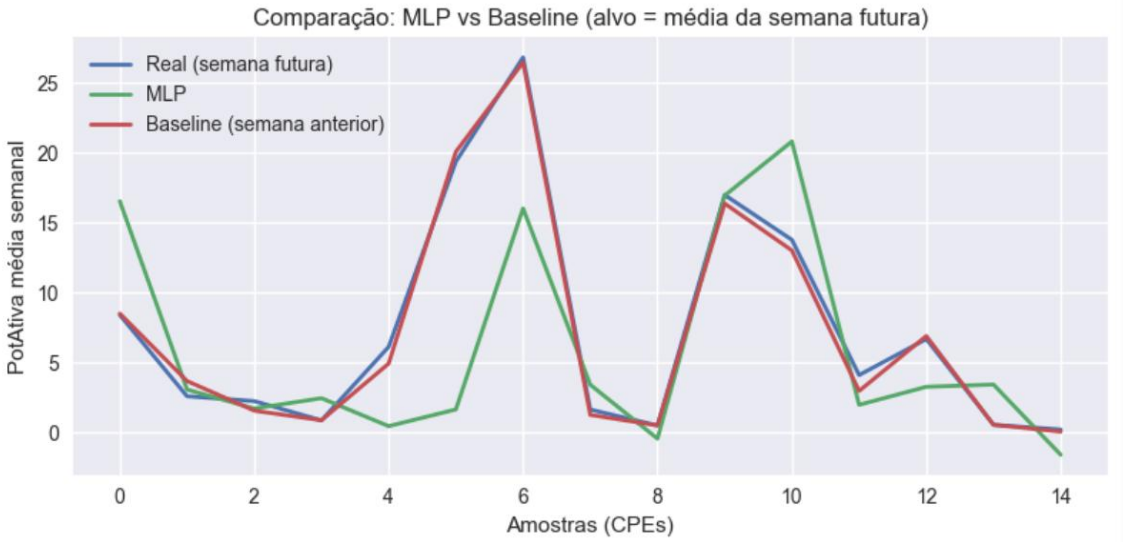
A análise visual do gráfico reforça esta conclusão. Enquanto o baseline acompanha de forma muito próxima o valor real em quase todos os CPEs, o Random Forest apresenta grandes desvios pontuais, com sobrestimações extremas em alguns edifícios (por exemplo, picos acima de 30–40 kW), que explicam o valor elevado do RMSE.

Este comportamento indica que o modelo está a ser influenciado por combinações de features que não são suficientemente estáveis para este horizonte de previsão (média semanal), levando a previsões instáveis e pouco robustas.

Em síntese, para este problema e este conjunto de features, o Random Forest não consegue superar uma abordagem simples baseada na persistência semanal, confirmando que a dinâmica do consumo é fortemente dependente da repetição do padrão da semana anterior.

6.3.2. MLP

Figura 24 – MLP vs Baseline (alvo = média da semana futura)



A Figura 23 apresenta a comparação entre o modelo MLP e o baseline para a previsão da média do consumo da semana futura, considerando diferentes CPEs do conjunto de teste. O baseline corresponde à média do consumo da semana imediatamente anterior, enquanto o MLP utiliza apenas features agregadas calculadas a partir do histórico de cada consumidor.

Os resultados mostram uma diferença muito clara entre os dois métodos. O MLP apresenta erros médios relativamente elevados (MAE de 4.33 kW e RMSE de 6.40 kW), enquanto o baseline, baseado simplesmente no valor da semana anterior, consegue resultados muito mais precisos (MAE de 0.49 kW e RMSE de 0.64 kW). Na prática, isto significa que o MLP erra várias vezes mais do que o baseline, apesar de ser um modelo mais complexo.

A análise visual reforça esta conclusão. Observa-se que a curva do baseline acompanha de forma muito próxima os valores reais da semana futura em praticamente todos os CPEs, enquanto o MLP apresenta desvios relevantes, sobretudo nos pontos de maior consumo, onde tende a subestimar ou sobrestimar os picos.

Este comportamento indica que, para este problema, a dinâmica temporal do consumo — capturada implicitamente pelo baseline através do valor da semana anterior — é muito mais informativa do que as features agregadas utilizadas pelo MLP. Como o MLP não incorpora diretamente a continuidade temporal da série, perde a principal fonte de informação preditiva, resultando num desempenho inferior ao de um método extremamente simples.

6.4. Conclusão

Nesta terceira abordagem, o objetivo foi prever o consumo médio da semana futura de cada CPE a partir de features agregadas extraídas do histórico (médias, picos, variabilidade e distribuição temporal do consumo), recorrendo a Random Forest (RF) e Multi-Layer Perceptron (MLP). O desempenho foi sempre comparado com um baseline simples, que usa a média da semana anterior como previsão.

Figura 25 – Distribuição do erro absoluto – RF vs MLP vs Baseline

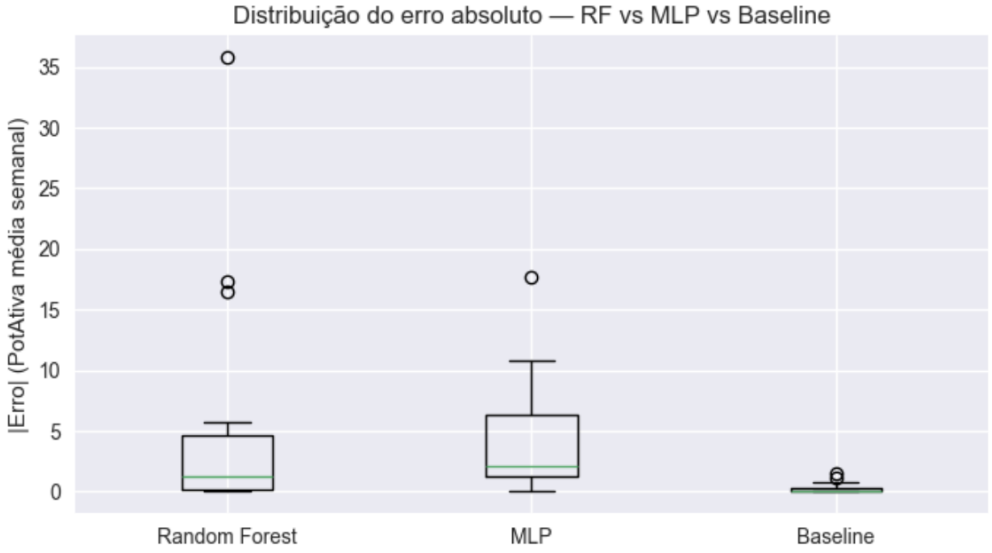


Tabela 5 – Comparação de desempenho – Feature-based Prediction

Modelo	MAE (kW)	RMSE (kW)	Interpretação
Baseline (semana anterior)	0.26	0.50	Referência: extremamente preciso devido à forte inércia temporal do consumo
MLP	4.33	6.40	Melhor que RF, mas ainda muito pior que o baseline
Random Forest	5.05	10.15	Pior desempenho; forte sensibilidade a outliers

Os resultados mostram que nenhum dos dois modelos supervisionados conseguiu superar o baseline.
No caso do Random Forest, os erros foram substancialmente superiores aos do baseline (MAE ≈ 5.05 kW e RMSE ≈ 10.15 kW, contra MAE ≈ 0.26 kW e RMSE ≈ 0.50 kW do baseline). A análise gráfica e a distribuição do erro absoluto evidenciam a presença de outliers severos, indicando que o RF é sensível a casos atípicos e não generaliza bem quando o consumo varia abruptamente entre semanas.

O MLP apresentou um comportamento mais estável do que o RF, com erros mais baixos (MAE ≈ 4.33 kW, RMSE ≈ 6.40 kW), mas ainda claramente inferiores ao baseline, que manteve MAE ≈ 0.49 kW e RMSE ≈ 0.64 kW. As boxplots mostram que, apesar de o MLP reduzir alguns erros extremos, a sua variabilidade continua muito superior à do baseline.

Em termos práticos, estes resultados indicam que, para a tarefa de prever a média semanal futura, o consumo apresenta uma forte inércia temporal, tornando a média da semana anterior um preditor extremamente eficaz. As features agregadas não capturam informação adicional suficiente para justificar modelos mais complexos.

Assim, conclui-se que, neste contexto e com este conjunto de dados, a abordagem feature-based (RF e MLP) não é adequada para previsão semanal, sendo claramente dominada pelo baseline simples. Isto contrasta com os modelos de séries temporais (ARIMA e LSTM), que exploram diretamente a dinâmica temporal e mostraram maior capacidade para modelar padrões de consumo.

7. Conclusão Final

Ao longo deste trabalho foi aplicada a metodologia CRISP-DM ao conjunto de dados de consumo energético do Município da Maia, combinando técnicas de clustering, previsão por séries temporais e modelos supervisionados baseados em features para explorar o potencial analítico do dataset.

Na Questão 1 (Clustering), verificou-se que métodos mais simples como o K-Means e o DBSCAN tiveram dificuldade em lidar com a elevada heterogeneidade dos edifícios e com a presença de outliers. O Gaussian Mixture Model (GMM), especialmente após normalização, destacou-se por conseguir formar clusters mais coerentes e interpretáveis, separando de forma clara diferentes perfis de consumo, como edifícios residenciais, comerciais e grandes consumidores. Esta segmentação revelou-se fundamental para estruturar a análise preditiva que se seguiu.

Na Questão 2 (Previsão por Séries Temporais), foram comparados os modelos ARIMA e LSTM em cada cluster, usando como referência um baseline simples baseado no valor da mesma hora da semana anterior. Os resultados mostraram que o ARIMA só foi realmente competitivo nos clusters mais estáveis, em particular no Cluster 0. Já o LSTM, sobretudo quando treinado com séries normalizadas, conseguiu capturar melhor os padrões não lineares e os ciclos diários de consumo, apresentando melhor desempenho nos clusters de consumo mais regular e intermédio. Nos clusters mais voláteis, como o Cluster 3, ambos os modelos tiveram dificuldades, mas o LSTM ainda assim conseguiu reduzir significativamente o erro face ao ARIMA.

Na Questão 3 (Previsão baseada em features), os modelos Random Forest e MLP foram usados para prever a média da semana seguinte a partir de estatísticas do histórico. No entanto, em ambos os casos, o baseline (média da semana anterior) teve um desempenho claramente superior. Isto mostra que, para este horizonte temporal, o consumo apresenta uma forte continuidade, e que as features agregadas utilizadas não foram suficientes para acrescentar poder preditivo relevante sem informação externa adicional, como dados meteorológicos, ocupação dos edifícios ou tipo de utilização.

Em conjunto, os resultados mostram que este dataset é muito adequado para segmentação e caracterização de perfis de consumo e útil para previsões de curto prazo quando se usam modelos de séries temporais, especialmente combinados com clustering. Por outro lado, para previsões agregadas baseadas apenas em features históricas, o ganho face a um baseline simples é limitado. Isto sugere que o maior valor destes dados está na análise detalhada do comportamento energético e na monitorização operacional, e que previsões mais robustas exigirão a integração de variáveis externas e modelos mais ricos no futuro.

Apesar dos resultados obtidos, existem algumas limitações que importa referir. Em primeiro lugar, os dados disponíveis não incluem informação externa, como condições meteorológicas, tipo de edifício ou níveis de ocupação, que são fatores importantes para explicar o consumo de energia. Além disso, os CPEs apresentam comportamentos muito diferentes entre si e, em alguns casos, séries bastante irregulares, o que dificulta o desempenho dos modelos, sobretudo nos clusters mais instáveis. Por fim, a avaliação foi feita apenas com um CPE por cluster, o que não permite garantir que os resultados sejam totalmente representativos de todos os edifícios.

No futuro, este trabalho pode ser melhorado avaliando vários CPEs por cluster, de forma a obter resultados mais robustos. A integração de dados externos, como meteorologia, níveis de ocupação e características dos edifícios, poderá também aumentar significativamente a qualidade das previsões. Para além disso, a utilização de modelos mais avançados ou combinações de vários modelos poderá ajudar a captar melhor os diferentes padrões de consumo e reduzir os erros de previsão.