

RELATÓRIO TRABALHO 3 - MACHINE LEARNING

Sumário

Análise Exploratória dos Dados	1
Seleção de Atributos	5
Geração de Modelo	6
Conclusão	9

Análise Exploratória dos Dados

Ao observar a tabela de *Feature Statistics* na Figura 1 abaixo, é possível notar alguns padrões sobre os dados sendo analisados:

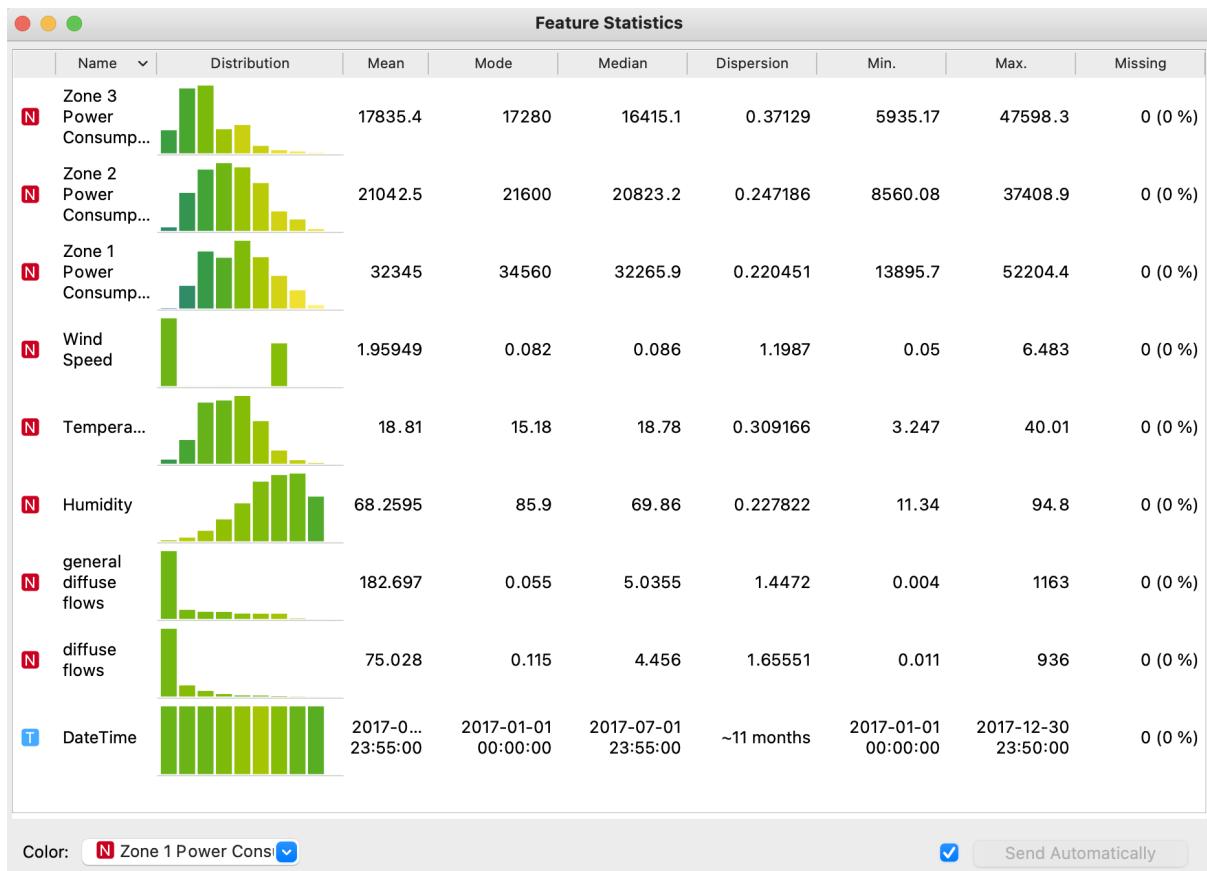


Figura 1: Tabela de *Feature Statistics*.

Nota-se que o consumo nas zonas 1 e 2 tem uma distribuição mais próxima da normal, com concentração em valores centrais, enquanto a zona 3 tem níveis de

consumo frequentes mais próximos do mínimo. Isso pode indicar que a zona 3 tem mais outliers potenciais em valores altos.

Para entender o padrão de consumo ao longo do ano, podemos considerar scatter plots dos níveis de consumo ao longo do ano. Esses gráficos (abaixo), mostram que as zonas 1 e 2 têm consumos mais constantes, enquanto a zona 3 apresenta um pico de consumo no meio do ano.

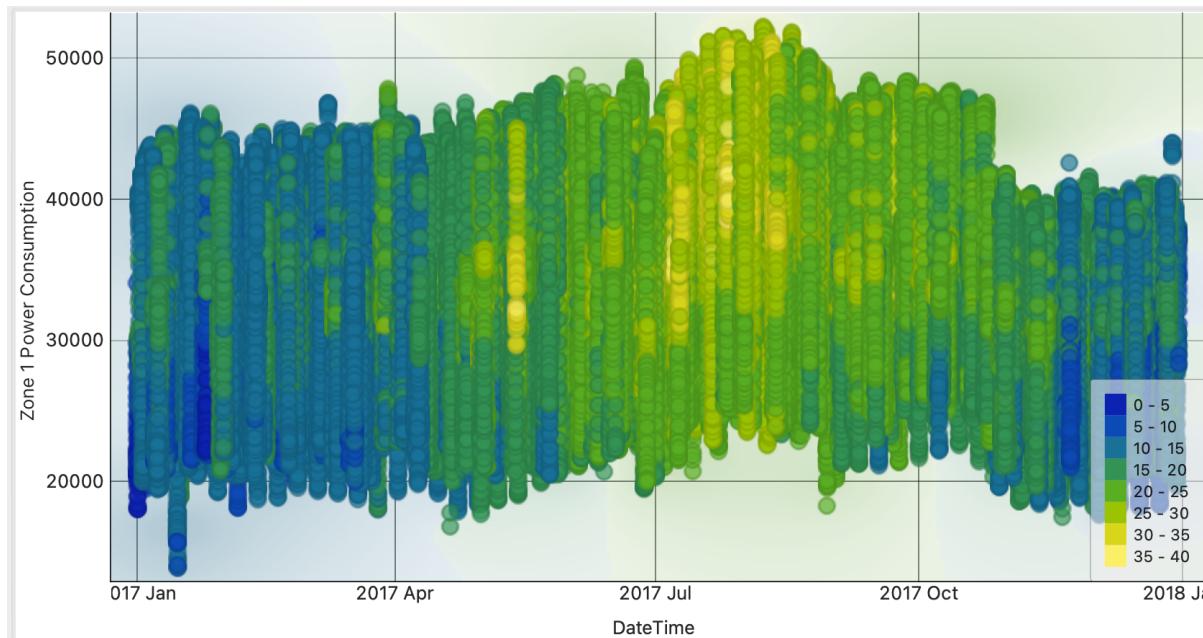


Figura 1: Scatter Plot entre o consumo de energia da Zona 1 e o tempo, colorido por temperatura.

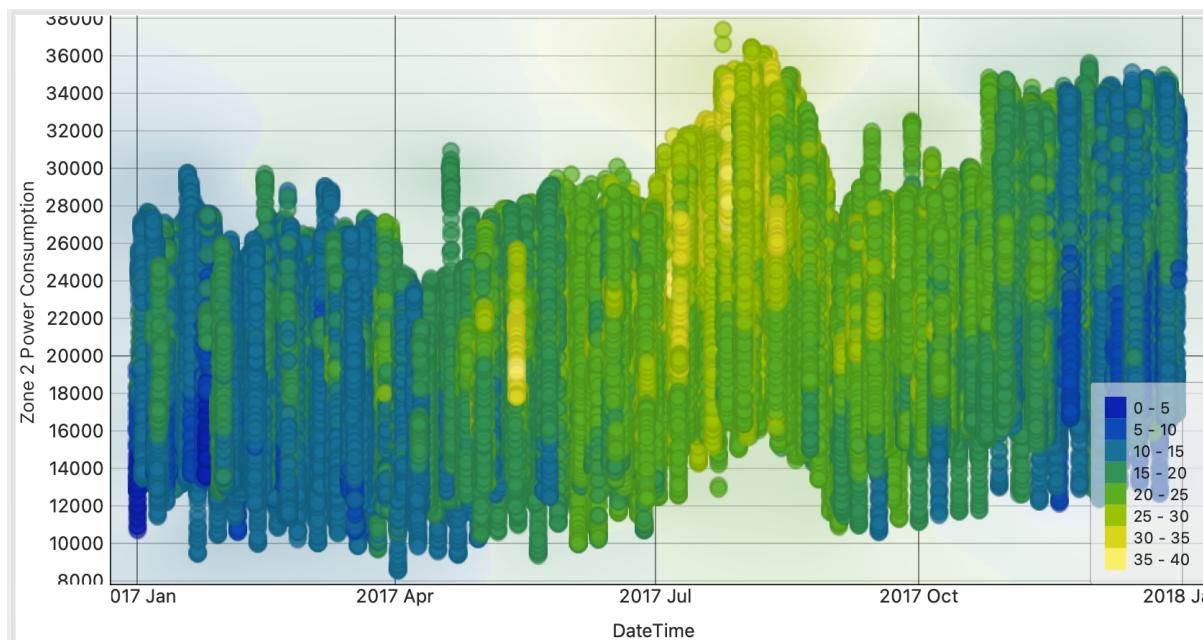


Figura 2: Scatter Plot entre o consumo de energia da Zona 2 e o tempo, colorido por temperatura.

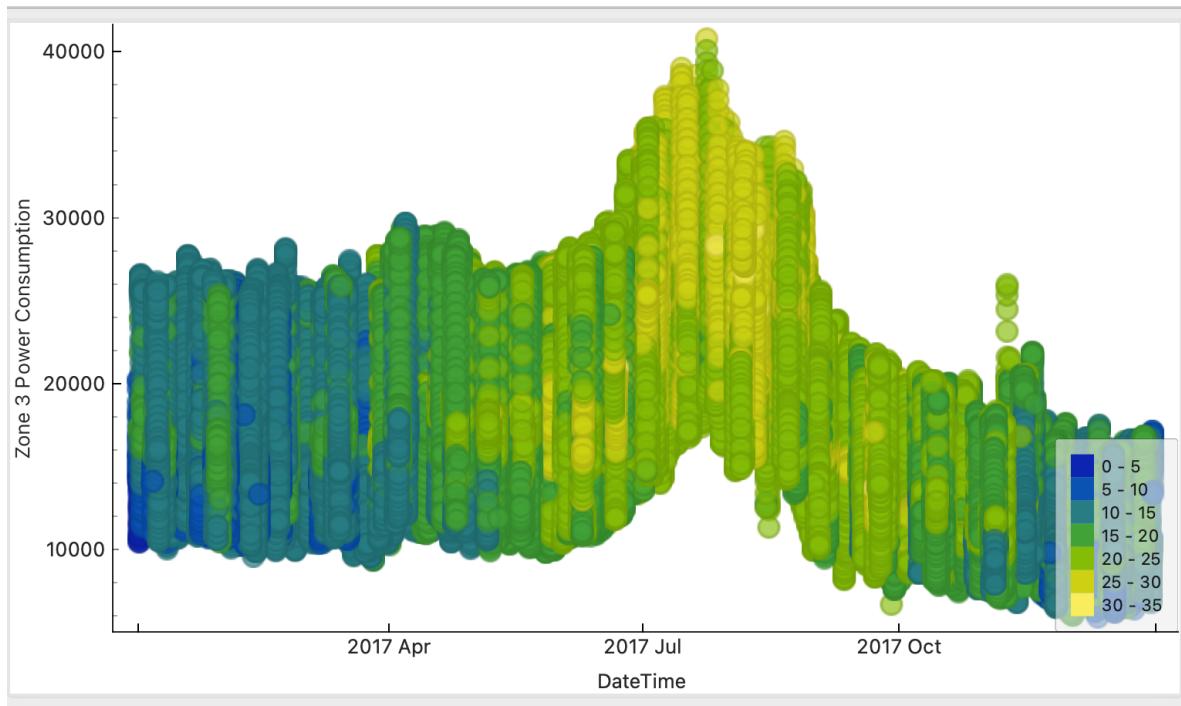


Figura 3: *Scatter Plot* entre o consumo de energia da Zona 3 e o tempo, colorido por temperatura.

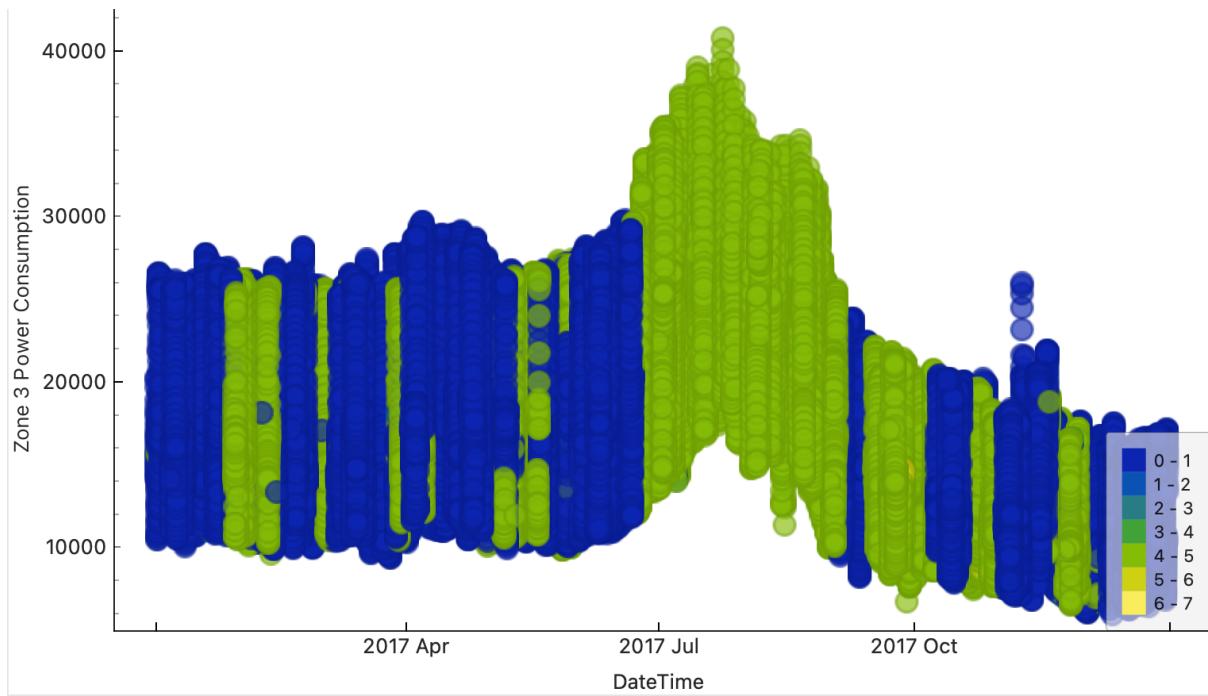


Figura 4: *Scatter Plot* entre o consumo de energia da Zona 3 e o tempo, colorido por velocidade do vento.

Esse padrão de consumo constante das zonas 1 e 2 e com pico na zona 3 pode indicar que as primeiras duas zonas operam próximas de suas capacidades máximas e que a zona 3 é ativada apenas em momentos de maior consumo. Pode ser também que esse pico em demanda seja o resultado de um pico na oferta de

energia. Como ele ocorre em épocas de maior insolação e mais vento, pode ser que a Zona 3 seja uma usina de energia eólica ou solar.

Além disso, a relação exibida na Figura 2 também demonstra um padrão onde o consumo de energia cai durante a primavera e o outono, quando comparados ao verão, mas o menor período de consumo de energia para a Zona 3 aparenta ser durante o inverno.

Considerando os dados da tabela de estatísticas (Figura 1), que mostram que a amplitude térmica registrada na região foi entre 3 e 40 graus, é curioso que o consumo de energia caia junto com as temperaturas durante o período de inverno, podendo indicar então que talvez tenha-se uma preferência por aquecedores à gás ao invés de elétricos, e que o consumo se mantenha alto no verão pelo uso de aparelhos de ar condicionado.

Por último, tomando a tabela de correlações entre as colunas da base de dados (na Figura 4 a seguir), temos que os parâmetros que mais se destacam com relação ao consumo de energia são a temperatura e a umidade, que possuem os maiores coeficientes de correlação (tanto entre si, quanto com o consumo de energia nas três zonas).

5	+0.490	Temperature	Zone 3 Power Consumption
6	+0.477	Temperature	Wind Speed
7	-0.468	Humidity	general diffuse flows
8	+0.460	Temperature	general diffuse flows
9	-0.460	Humidity	Temperature
10	+0.440	Temperature	Zone 1 Power Consumption
11	+0.382	Temperature	Zone 2 Power Consumption
12	+0.325	DateTime	Zone 2 Power Consumption
13	-0.295	Humidity	Zone 2 Power Consumption
14	-0.287	Humidity	Zone 1 Power Consumption
15	+0.283	DateTime	Temperature
16	+0.279	Wind Speed	Zone 3 Power Consumption
17	-0.257	Humidity	diffuse flows
18	-0.234	DateTime	Zone 3 Power Consumption
19	-0.233	Humidity	Zone 3 Power Consumption

Figura 5: *Correlation* para todos os atributos do banco de dados.

Essa relação é melhor vista para o gráfico do consumo de energia da Zona 1, na Figura 5 abaixo:

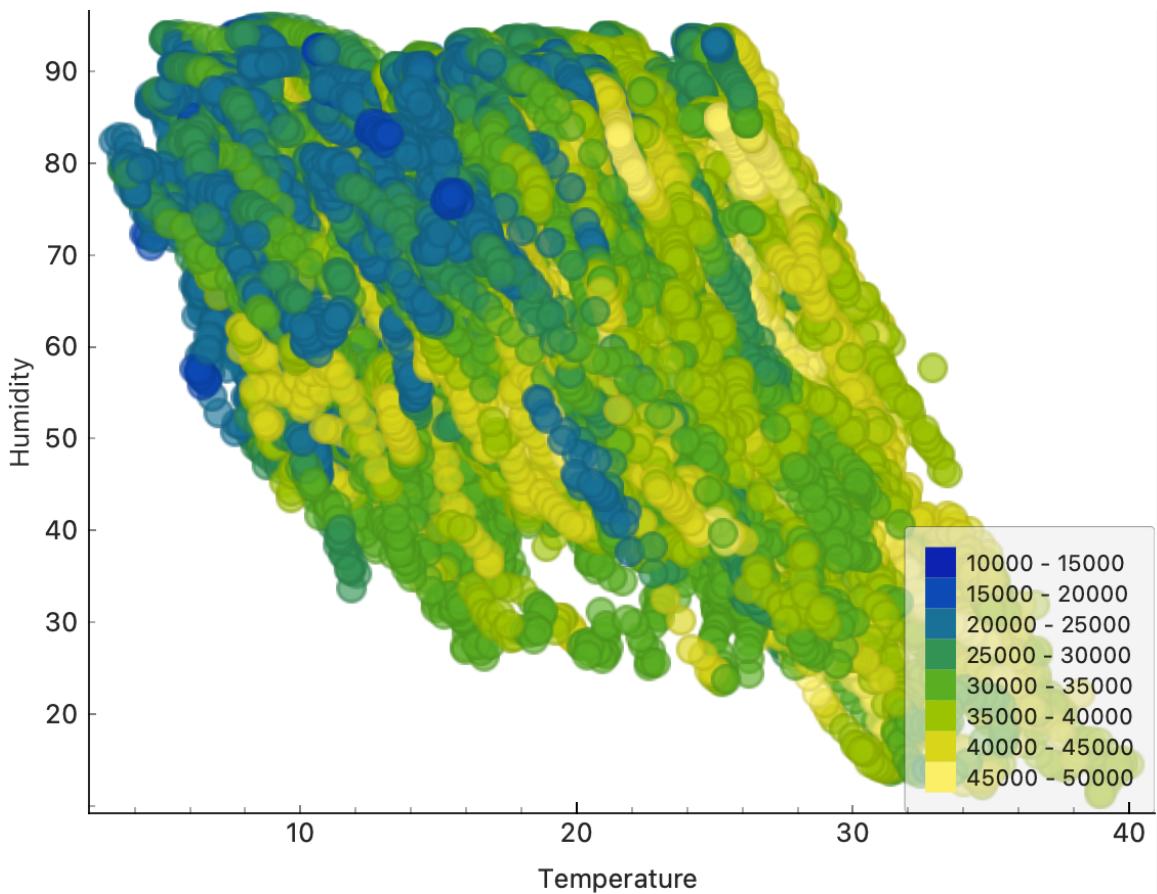


Figura 6: *Plot Outliner* entre temperatura e umidade colorido com o consumo da Zona 1.

Nota-se que a temperatura tem um relacionamento inversamente proporcional à umidade (como visto pelo coeficiente de correlação na Figura 4), e além disso, conforme a temperatura aumenta e a umidade cai, o consumo de energia tende a aumentar. Esse gráfico mostra que a cidade tem invernos úmidos e frios e tem verões quentes e secos.

Seleção de Atributos

Para fazer a predição de consumo, duas estratégias foram usadas: (1) usar as features do modelo sem tratamento prévio e (2) usar como feature o resultado de um PCA. Executando o PCA com apenas dois componentes finais, os componentes propostos já explicam 99% da variância dos dados e vemos que o componente 1 é praticamente apenas a data.

	components	variance	Humidity	Wind Speed	diffuse flows	general diffuse flow	Temperature	DateTime
1	PC1	1	2.40325e-09	-1.98649e-08	1.45193e-08	2.04069e-09	-3.11738e-08	-1
2	PC2	2.80432e-14	-0.472309	0.2636	0.413877	0.546574	0.487239	-1.4436e-08

Figura 7: Resultado do PCA com dois componentes.

Como os dados são intervalados de 10 minutos, a data e hora acabam influenciando muito o resultado. Para evitar que o modelo dependa apenas disso, data e hora foram removidas do dataset antes de se treinar os modelos e o PCA foi repetido buscando uma explicação de 95% da variância. Para isso, o PCA precisou usar todas as 5 variáveis, obtendo uma explicação de 99%. Reduzindo para 4 componentes, a variância explicada é de 93%, então tomou-se esse resultado para efeito de comparação.

	components	variance	Humidity	Wind Speed	diffuse flows	general diffuse flow	Temperature
1	PC1	0.466041	-0.469146	0.283585	0.392975	0.536093	0.507599
2	PC2	0.238623	-0.0417809	-0.695948	0.536014	0.293513	-0.374767
3	PC3	0.143922	0.728689	0.436217	0.511568	0.105125	-0.0772926
4	PC4	0.0833011	-0.497062	0.43506	0.325241	-0.362602	-0.571306

Figura 8: Resultado do PCA com quatro componentes.

Geração de Modelo

Os modelos testados foram Árvore de Decisão, Random Forest, Rede Neural e Regressão Linear. O último foi usado apenas como referência para propósito de comparação, visto que os dados não são lineares e espera-se que todos os outros modelos tenham uma performance melhor que a este. Esses modelos foram executados com e sem passar pelo PCA como pré-processamento, e também foi testado um SVM sem pré-processamento.

Test and Score						
Model	MSE	RMSE	MAE	R2	v	CVRMSE
Random Forest	23234306.244	4820.198	3340.574	0.538		14.949
Neural Network	36055374.162	6004.613	4743.368	0.283		18.622
Linear Regression	40119078.748	6333.962	5232.176	0.202		19.644
Tree	44822249.108	6694.942	4131.736	0.109		20.763
SVM	59632823.228	7722.229	6305.887	-0.186		23.949

Figura 9: Comparação dos algoritmos sem pré-processamento ordenador por R2 decrescente.

Test and Score (1)						
Model	MSE	RMSE	MAE	R2	CVRMSE	
Random Forest (1)	23275779.025	4824.498	3347.302	0.537	14.962	
Neural Network	36055374.162	6004.613	4743.368	0.283	18.622	
Tree (1)	44822249.108	6694.942	4131.736	0.109	20.763	

Figura 10: Comparação dos algoritmos com pré-processamento ordenador por R2 decrescente.

O SVM teve uma performance muito ruim. A configuração que consta na tabela, com R2 -0.182, usa um kernel RBF. Também foram testados um kernel linear (R2 -0.604), um polinomial (R2 -0.281) e um kernel sigmoid (R2 -0.528). O fato de todos os valores de R2 serem negativos significa que os SVMs foram tão ruins que uma reta horizontal igual à média de consumo explica os dados melhor que esses modelos.

As Redes Neurais foram configuradas com solver Adam, função de ativação ReLu e duas camadas ocultas de 50 neurônios cada. Pode-se observar que o impacto de usar ou não PCA no pré-processamento é nulo. Outras configurações de camadas foram testadas, como uma camada com 100 neurônios, três camadas com 20 neurônios cada, camadas 20-100-100-20, e duas camadas 50-50, mas nenhuma obteve resultado melhor que a apresentada na tabela. Experiências com outras funções de ativação seriam interessantes, mas o programa usado (Orange Datamining) falhou todas as vezes que tais mudanças foram feitas, impossibilitando a obtenção desses resultados.

Em ambos os casos, com e sem PCA, o Random Forest obteve o melhor resultado, com um R2 de 53,7%. Esse resultado foi obtido com 20 árvores, com folhas contendo até 20 datapoints e sem limite de profundidade. Essa falta de limite, no entanto, pode levar a overfitting. Para combater essa possibilidade, foram testadas variações de outras duas configurações: variação no número máximo de datapoints por folha e a profundidade máxima da árvore.

Número máximo de datapoints por folha	PCA	Profundidade máxima da árvore	R2
20	Sim	2	0.194
20	Sim	4	0.266
20	Sim	15	0.508
20	Sim	-	0.519
20	Não	2	0.195
20	Não	4	0.265
20	Não	15	0.507
20	Não	-	0.519
144	Sim	2	0.194
144	Sim	4	0.257
144	Sim	4	0.431
144	Sim	-	0.433
144	Não	2	0.197
144	Não	4	0.257
144	Não	4	0.428
144	Não	-	0.432
999	Sim	2	0.192
999	Sim	4	0.262
999	Sim	15	0.329
999	Sim	-	0.326
999	Não	2	0.197
999	Não	4	0.264
999	Não	15	0.329
999	Não	-	0.329

Tabela 1: R2 do *Random Forest* com diferentes parâmetros

Pode-se observar com os resultados acima que à medida que a profundidade da árvore aumenta, o R2 também aumenta, já que a árvore passa a poder subdividir cada vez mais os resultados. Já com o número de datapoints por folha, a relação é inversa, mas pelo mesmo motivo: menos datapoints em uma folha implica em uma árvore mais especializada. Pode-se observar, finalmente, que o impacto do PCA não é significativo para este dataset.

Outro ponto interessante é que a pior configuração do Random Forest, com 999 datapoints por folha, com PCA e com limite de profundidade 2, gera um R2 de 0.192, que já é melhor que a árvore de decisão sozinha. Além disso, quando configurado sem limite de profundidade, sem PCA e com limite de 2 datapoints por folha, ou seja, uma configuração que tende a overfit, o R2 do Random Forest foi de 0.583. Isso mostra que sem a informação da data e hora há um alto nível de imprevisibilidade no dataset. A variação de consumo ao longo dos meses no ano se deve por mudanças de tempo (temperatura, umidade, ...), que são informações ainda disponíveis no dataset. As variações observadas podem, no entanto, depender de outros fatores, como horário no dia, com consumo de madrugada provavelmente menor que ao longo do dia, o dia da semana, que influencia principalmente o consumo de prédios comerciais e indústrias, e época de alta temporada para turistas, que pode impactar significativamente o número de pessoas na cidade.

Conclusão

A análise exploratória dos dados se mostrou muito importante para trazer mais contexto para o problema, entender o dataset e suas features, e fazer suposições para a organização dos dados. Um ponto muito relevante que foi observado é a importância da data e da hora do dia no nível de consumo de energia. Esse fator surgiu novamente quando decidindo como aplicar o PCA.

Como os dados obtidos são numerosos, mas todos dentro de um único ano, os padrões observados não podem ser generalizados, e ficam muito dependentes ao ano observado. Talvez se mais anos fossem observados seria possível prever, por exemplo, o máximo global de consumo e a data em que ele começaria a reduzir, o que pode ser uma informação valiosa ao se considerar políticas de preços de energia.

Outro fator que poderia permitir a utilização da data no treino seria decompô-la por mês (ou trimestre), por dia da semana, ou por hora do dia. A separação dessas informações em features diferentes poderia permitir o aprendizado de tendências em escalas maiores (como a alta de consumo no verão) ao comparar dias da semana e horários iguais, ou de identificar tendências em escala intermediária (ao longo da semana), ou ainda em escala diária (vendo a variação de manhã, tarde e noite, por exemplo). Essa análise não foi feita no entanto porque não foi possível encontrar uma forma prática de pré-processar as variáveis de tipo *datetime* como proposto dentro de um tempo razoável.

Esses pontos de exploração poderiam ser tratados em trabalhos futuros. Outras pistas de evolução possível seriam tentar melhorar a performance das redes neurais através de mais pré-processamento ou do uso de arquiteturas mais modernas e mais adaptadas a séries temporais; tentar prever outros fatores, como a data de início do aumento do consumo na zona 3, ou, com mais informações, os impactos de possíveis mudanças em políticas de preço ou do investimento em isolamento térmico de casas e apartamentos.