INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Big Data Analytics na Gestão de uma Rede de Distribuição de Água

João Miguel Grácio Alves

Mestrado em Business Analytics

Orientadora:
Professora Doutora Patrícia Filipe, Professora Auxiliar, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Novembro 2022

BUSINESS
SCHOOL

Big Data Analytics na Gestão de uma Rede de Distribuição de Água

João Miguel Grácio Alves

Mestrado em Business Analytics

Orientadora:
Professora Doutora Patrícia Filipe, Professora Auxiliar, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Novembro 2022

Agradecimento

Quero agradecer à Infraquinta, pela confiança e disponibilização dos dados, sem eles não seria possível realizar esta Dissertação.

À minha orientadora e Professora Patrícia Filipe por todo o apoio prestado, conhecimento e motivação, para a finalização deste estudo, como para a apresentação do meu trabalho num encontro científico.

Quero agradecer aos meus pais e irmã pelo apoio e incentivo para que conseguisse determinar este projeto.

Por fim quero agradecer aos amigos que me deram algumas ideias no seguimento desta investigação, e também pela partilha de conhecimento nesta área.

Resumo

A água é um bem essencial e escasso nos dias de hoje, pelo que é necessário conservá-lo e distribuí-lo de maneira eficiente e sustentável. Com a ajuda dos novos medidores digitais é possível medir os consumos com uma cadência horária, e através de técnicas de *Machine Learning*, produzir informações sobre padrões de consumo, comportamento dos consumidores, deteção de eventos nas redes de distribuição de água e feedback aos clientes.

A deteção de anomalias, como fugas de água domésticas e na rede de saneamento, avaria dos medidores e fraude, são algumas das causas que provocam desperdícios de água e que muitas autarquias não têm como detetar de forma eficaz. A presente dissertação tem como objetivo desenvolver uma metodologia, baseada em séries temporais, para a previsão de consumos de água e a deteção de anomalias, recorrendo a dados dos consumidores domésticos, de uma zona situada na Quinta do Lago, Algarve. Para isso, são utilizados métodos de *Clustering* para agrupar os consumidores com padrões de consumo idênticos, e modelos de *Machine Learning,* como o ARIMA, *Neural Networks e Random Forest*, utilizados para prever os consumos de água. São apresentados os resultados de vários modelos para previsão de consumos e, por fim, são desenvolvidos *dashboards* que auxiliam na deteção de possíveis anomalias.

Palavras-chave: Consumo de Água*; Data Analytics*; *Machine Learning;* Perdas de Água
Classificação JEL: M10, M15

Abstract

Water is an essential and limited resource these days, so it is necessary to conserve and distribute it efficiently and sustainably. With the help of the new digital meters, it is possible to measure consumption on an hourly basis, and through Machine Learning techniques, produce information on consumption patterns, consumer behavior, detection of events in water distribution networks and feedback to customers.

The detection of anomalies, such as domestic water leaks and in the sanitation network, meter failure and fraud, are some of the causes that cause water waste and that many municipalities are unable to detect effectively. This dissertation aims to develop a methodology, based on time series, for forecasting water consumption and detecting anomalies, using data from domestic consumers in an area located in Quinta do Lago, Algarve. For this, Clustering methods are used to group consumers with identical consumption patterns, and Machine Learning models such as ARIMA, Neural Networks and Random Forest, used to predict water consumption. The results of several models for forecasting consumption are presented, and finally, dashboards are developed to help detect possible anomalies.

Keywords: Data Analytics; Machine Learning; Water consumption; Water Losses.
JEL Classification: M10, M15

Índice

Índice de Figuras

Índice de Tabelas

Glossário de Siglas e Acrónimos

AR - *Auto Regressive Model*
ARIMA - *AutoRegressive Integrated Moving Average*
CNN - *Convolutional Neural Networks*
ES - *Entropy Scorer*
GBT - *Gradient Boosted Tree*
KNN - *K - Nearest Neighbors*
LSTM - *Long Short Term Memory*
MAE – *Mean Absolute Error*
MAPE - *Mean Absolute Percentage Error*
ML - *Machine Learning*
MLP - *Multilayer Perceptron*
NN - *Neural Networks*
RF - *Random Forest*
RMSE - *Root Mean Square Error*
RNN - *Recurrent Neural Networks*
SC - *Silhouette Coefficient*
SVM - *Support Vector Machine*
SVR - *Support Vector Regression*
XGBoost – *Extreme Gradient Boosting*

Capítulo 1. Introdução

1.1 Tema e a sua Importância

A água doce é um bem extremamente limitado, representando apenas 0.014% de toda a água do planeta (Rahim et al., 2020). É necessário conservá-la e utilizá-la de maneira eficiente. Neste aspeto a instalação de medidores digitais de água em conjunto com modelos de *Machine Learning,* podem desempenhar papéis importantes e fundamentais para gerar novas descobertas em tempo real, para dar *feedback* aos consumidores e encorajá-los a adotar novos hábitos de consumo.

Atualmente já existem países, como os Estados Unidos da América, que possuem portais online *end-to-end* que dão feedback aos clientes em tempo real, como monitorizar os consumos e alertas de potenciais fugas de água, o que se traduz numa minimização de perdas de água (H. Li et al., 2011; Schultz et al., 2018).

Segundo um estudo realizado no norte da Califórnia, cerca de quarenta porcento dos consumidores de uma zona residencial tiveram perdas de água, em média de meio litro a um litro por minuto, o que perfaz cerca de 340 litros por residência por dia. Estas perdas foram detetadas pelos contadores digitais de água (Chastain-Howley & Wallenstein, 2007).

Em Portugal, devido à falta de investimento na área de análise de dados aplicado a esta indústria, foi criado o projeto WISDOM (2022), financiado pela Fundação para a Ciência e Tecnologia (FCT), tendo por base uma equipa de várias instituições académicas e parceiras, como o Instituto Politécnico de Setúbal, o Instituto Superior Técnico , o Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa, a Infraquinta, a Câmara Municipal do Barreiro e a Empresa Municipal de Água e Saneamento de Beja. O objetivo deste projeto é criar ferramentas no suporte às operações e gestão dos serviços de fornecimento de água, como análise e exploração de dados, previsão de consumos, ferramentas de suporte à decisão, identificação de eventos anómalos e identificação de roturas.

A Infraquinta (um dos parceiros do projeto WISDOM) é um exemplo de uma empresa nacional que realiza a gestão de infraestruturas e da rede de distribuição de água, na Quinta do Lago. A Infraquinta já possui uma interface para monitorização e análise dos consumos na rega dos espaços verdes, e um sistema de alerta de perdas de água, na qual os clientes são notificados. Estes sistemas são essenciais nos dias de hoje, tanto para controlar custos, para haver um maior controlo na gestão das infraestruturas, como para reduzir desperdícios de água e tornar o consumo o mais eficiente possível. Os dados fornecidos por esta empresas estão na base desta investigação.

Os estudos realizados neste contexto podem ainda caracterizar-se em cinco áreas distintas: *feedback* de uso da água, categorização de eventos (incluindo deteção de anomalias), previsão de consumo de água, análise socioeconómica e análise de comportamento do consumidor (Rahim et al., 2020). Esta pesquisa foca-se na previsão de consumos, análise de comportamento e deteção de eventos.

Sendo a aplicação de *Data Analytics* e modelos de *Machine Learning (ML)* recentes neste setor, existe um *gap* no conhecimento e em estudos empíricos nesta área, pelo que esta investigação vem preencher algumas lacunas existentes no conhecimento e, em simultâneo, responder à pergunta de investigação. Resumindo, o objetivo deste estudo é criar um modelo de *Big Data Analytics* através da aplicação de técnicas de *Machine Learning* para previsão de consumos de água e deteção de eventos (perdas de água, anomalia de sensores de água ou fraudes) nas redes de distribuição de água, contribuindo com conhecimento nesta área que possa ser útil em estudos futuros.

1.2 Problema e Questão de Investigação

Devido às alterações climáticas e aos períodos de seca crescente no nosso planeta, a preservação da água doce é cada vez mais importante nos dias de hoje. É necessário criar novos mecanismos e explorar novas formas de tornar o consumo de água mais sustentável e reduzir os desperdícios.

Os dados nos quais se baseiam esta investigação foram fornecidos pela Infraquinta, uma empresa que faz a gestão das infraestruturas e de toda a rede de abastecimento de água na Quinta do Lago, no Algarve. Esta empresa, apesar de ser *Data-driven,* já possui uma solução para deteção de fugas em tempo real denominada de *IQAlert* e encontra-se à procura de novas soluções para a deteção de problemas que ocorram na rede de águas. A empresa possui dados de consumo de água de todos os seus consumidores, de onde é possível criar padrões de consumo e descobrir consumos anómalos. O problema que a Infraquinta enfrenta é conseguir detetar possíveis anomalias nos medidores hídricos ou rede doméstica, através dos padrões de consumo gerados pelos consumos horários de cada consumidor. Para esse efeito, a empresa lançou o desafio para o desenvolvimento desta investigação - criar um modelo capaz de detetar anomalias/eventos na sua rede de distribuição de água, utilizando dados reais do consumo dos seus consumidores. Estes dados têm presentes os consumos de todos os consumidores (domésticos e não domésticos), e outras informações, com uma cadência horária.

Os padrões de consumo residenciais e comerciais são bastante diferentes (H. Li et al., 2011), pelo que se optou pelo foco no consumo residencial.

Neste sentido, a principal questão de investigação que este estudo se propõe a resolver é: Como modelar consumos de água, de modo a prevê-los futuramente e detetar possíveis falhas que possam ocorrer nos medidores hídricos dos consumidores domésticos?

A resposta a esta questão dependerá do cumprimento dos objetivos definidos no subcapítulo seguinte e dos resultados dos modelos de ML aplicados para este efeito.

1.3 Objetivos e Contributos

Tendo em conta a questão de investigação descrita no subcapítulo anterior, esta dissertação tem como objetivo principal criar um modelo analítico para previsão de consumos de água e deteção de eventos (anomalias) nos medidores de consumo de água doméstico.

Para a concretização do objetivo principal, foram definidos os seguintes objetivos específicos:

- Identificar e caracterizar os padrões de consumo dos consumidores da rede;
- Prever o consumo de água dos consumidores ao longo do tempo;
- Comparar o consumo previsto e o consumo real dos consumidores;
- Criar *dashboard* para apoio à deteção de anomalias.

Esta investigação vem contribuir para o *gap* de conhecimento nesta área, utilizando modelos de ML, tal como dar a conhecer os tipos de algoritmos mais eficazes nesta temática aplicada às séries temporais. Este estudo apresenta uma metodologia que pretende contribuir para a deteção de eventos incomuns na rede em tempo real, e assim gerir o recurso de uma forma mais eficiente.

A utilização de métodos de *Clustering* para agrupar indivíduos com padrões de consumo semelhantes também pode ser um contributo importante para as empresas que fazem a gestão das redes de água, uma vez que permite aplicar diferentes taxas consoante os padrões de consumo dos clientes, e assim incentivar a poupança de água.

1.4 Organização da Dissertação

A metodologia desta investigação é baseada no CRISP-DM (Cross-Industry Standard Process for *Data Mining*), bastante utilizada em várias indústrias, principalmente em projetos relacionados com *Data Mining*. A dissertação está organizada tendo por base as etapas descritas por esta metodologia, que se apresentam da seguinte forma:

O capítulo dois apresenta a revisão literária sobre a temática de deteção de eventos e previsão de consumos nas redes de distribuição de água. O capítulo três é descrita a metodologia seguida e abordada a direção tomada pela dissertação. No capítulo quatro é

apresentado resumidamente o negócio da empresa que forneceu os dados para este estudo e feita uma caracterização dos dados disponibilizados, envolvendo a sua preparação e recorrendo a análise exploratória dos dados. No capítulo cinco são aplicados os algoritmos de ML adequados com vista a atingir os objetivos propostos, desde técnicas de *Clustering* a modelos preditivos aplicados aos consumos de água. Apresenta também uma proposta de dois *dashboards*, um direcionado para o para o consumidor e outro para a empresa, para apoio à deteção de anomalias. Por fim no capítulo seis são apresentadas as conclusões do estudo, contributos, limitações e sugestões para continuação dos trabalhos de investigação nesta área.

Capítulo 2. Revisão da Literatura

2.1 Contextualização

Primeiramente neste estudo é importante clarificar alguns conceitos e dar um enquadramento geral, de forma que os resultados sejam percetíveis.

Os conjuntos de dados que são analisados nos estudos utilizados, são gerados a partir de leituras efetuadas por medidores digitais com uma cadência horária, sendo possível mudar a sua cadência consoante a necessidade. De seguida, estas leituras são enviadas automaticamente para uma base de dados de *IoT (Internet of Things)*. Alguns sensores deste tipo têm a capacidade de alertar o seu estado, como saudável ou anómalo (Lee et al., 2021).

A deteção de anomalias está relacionada com a procura de observações que não são comuns na rede ou nos padrões de consumo, através de valores muito altos ou muito baixos nas leituras de consumo, resultantes, por exemplo, de perdas de água e de deficiência nos medidores (Merta & Fikejz, 2019).

Para além de servirem para a deteção de anomalias, os medidores digitais poderão ser uma ferramenta bastante importante para os consumidores serem taxados de forma justa, bem como para ajudar a combater o consumo excessivo de água (Ray & Goswami, 2020). Neste aspeto estão a ser estudadas novas ideias de “multi integração” de todos os serviços de consumo de energia numa só plataforma, onde é possível o consumidor ver o seu histórico e consumos em tempo real, tal como alertas para falhas na rede, através dos medidores inteligentes (Wu et al., 2020).

Todos os medidores de água possuem uma percentagem de erro das leituras dependente do fabricante, mas em média, possuem uma taxa de erro na leitura da água consumida de dois porcentos, com temperaturas de água inferiores a trinta graus celsius (Clinciu & Clinciu, 2017). No estudo de Kainz et al. (2021), os autores referem que podem existir fugas de água com uma velocidade de escoamento tão pequena, que não permite ao medidor aferir esse consumo extra, mas que ao longo de alguns dias poderá ser possível identificar essa fuga através de uma análise aos consumos registados.

Já existem alguns estudos pertinentes cujo objetivo é transformar medidores analógicos (convencionais) em digitais, recorrendo a uma câmara, da utilização de *IoT* e de pequenos processadores. Deste modo, conseguem, com a ajuda de modelos de *Machine Learning,* como *Random Forest* e *Convolutional Neural Networks* (CNN), identificar os consumos dos medidores analógicos e enviar os consumos com uma cadência horária para uma base de dados (Lall et al., 2021).

Para a deteção de eventos nas redes de distribuição de água veremos que são utilizados, maioritariamente, modelos supervisionados, como Redes Neuronais, *Support Vector Regression*, *Linear Regression*, *Random Forest*, entre outros (Rahim et al., 2020).

Os setores dos recursos hídricos têm bastante potencial para serem *Data-driven* e fazer parte da nova revolução industrial (Indústria 4.0). A aplicação da IoT (E. Y. Li et al., 2017) para conectar vários medidores à rede, juntamente com técnicas de análise preditiva (*Advanced Analytics*) traz redução de custos e melhora a gestão das redes de água. Para além disso, estas tecnologias permitem medir a pressão de água que é enviada para a rede mais eficientemente, identificar fugas de água, realizar uma manutenção prescritiva da rede e maior transparência nos consumos (Alabi et al., 2019).

Esta revisão sistemática da literatura será dividida em enquadramento (será efetuada uma pequena abordagem ao tema e conceitos associados), metodologia de pesquisa (processo de pesquisa e filtragem dos estudos que compõem esta investigação), resultados (apresentação dos resultados obtidos da pesquisa), discussão (resposta às perguntas da investigação e avaliação de qualidade) e conclusões finais.

2.1.1 Métricas de desempenho abordadas

Em grande parte dos estudos científicos selecionados, foram utilizadas métricas de avaliação para analisar o desempenho dos modelos na previsão dos casos (Lee et al., 2021; Zese et al., 2021).

No caso de aplicação de modelos de classificação, mais frequentemente aplicados na deteção de anomalias, foram identificadas como métricas usuais, a *Precision* (precisão), *Recall* (revocação ou sensibilidade), *Accuracy* (PECC), *F-measure* ou *F-score*, *MCC (Matthews Correlation Coefficient)*, *ROC (Receiver Operating Characteristic)*, *AUC (Area Under Curve)*. De entre estas, a curva ROC é das métricas mais comuns para problemas de classificação binária. Consiste num gráfico de duas dimensões em que o eixo do x representa o rácio de falsos positivos e o eixo do y o rácio de verdadeiros positivos. Já a AUC é a área abaixo da curva ROC, na qual quanto mais perto de 1 for o seu valor, melhor o modelo se ajusta à amostra, onde um bom classificador deve ter um valor superior a 0.5 (Fawcett, 2006). Para além da curva ROC é de referir que Accuracy, a Precision e o Recall também são bastante utilizadas.

Para este estudo em questão, pelo facto do consumo de água se tratar de uma variável *target* numérica, são utilizadas métricas como o *R Square*, *Mean Absolute Percentage Error* (MAPE), *Root Mean Square Error* (RMSE) e o Mean Absolute Error (MAE).

Representando os valores observados por Y_i, os valores previstos pelos modelos por $\hat{Y}_i$, para i=1, 2, …, n; por $\bar{Y}$ a média das observações, descrevem-se as métricas referidas.

O *R Square* indica a proporção da variação do *target* explicado pelo modelo (Xu et al., 2021), e representa-se da seguinte forma:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

O MAPE consiste na média do valor absoluto da diferença (em percentagem) entre os valores previstos e os valores observados em cada período. Quanto menor é o seu valor, melhor o desempenho do modelo previsto (Candelieri, 2017), representando-se por:

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{\left|Y_i - \hat{Y}_i\right|}{|Y_i|}$$

O RMSE calcula a raiz quadrada da média dos erros entre os valores observados e os valores previstos, com a particularidade de que são atribuídos diferentes pesos aos erros (Chen & Boccelli, 2018). O RMSE irá aumentar consideravelmente se existir um erro superior aos restantes:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}$$

Por fim, o MAE calcula o erro absoluto médio entre os valores observados e os valores previstos:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right|$$

Para problemas de regressão como o que será apresentado mais à frente, dos estudos extraídos, estas foram as métricas de desempenho mais utilizadas.

2.2 Metodologia de pesquisa

Uma revisão sistemática da literatura tem como objetivo fazer uma análise da literatura de trabalhos já existente sobre um determinado tema, ao qual colocamos perguntas que terão de ser respondidas ao longo da investigação. A revisão sistemática da literatura, que será apresentada, tem por base os conselhos escritos pelo autor Kitchenham (2004).

Para chegar aos resultados desejados e cumprir os objetivos, foi criada uma estratégia baseada na metodologia de Kitchenham, representada na figura 1.

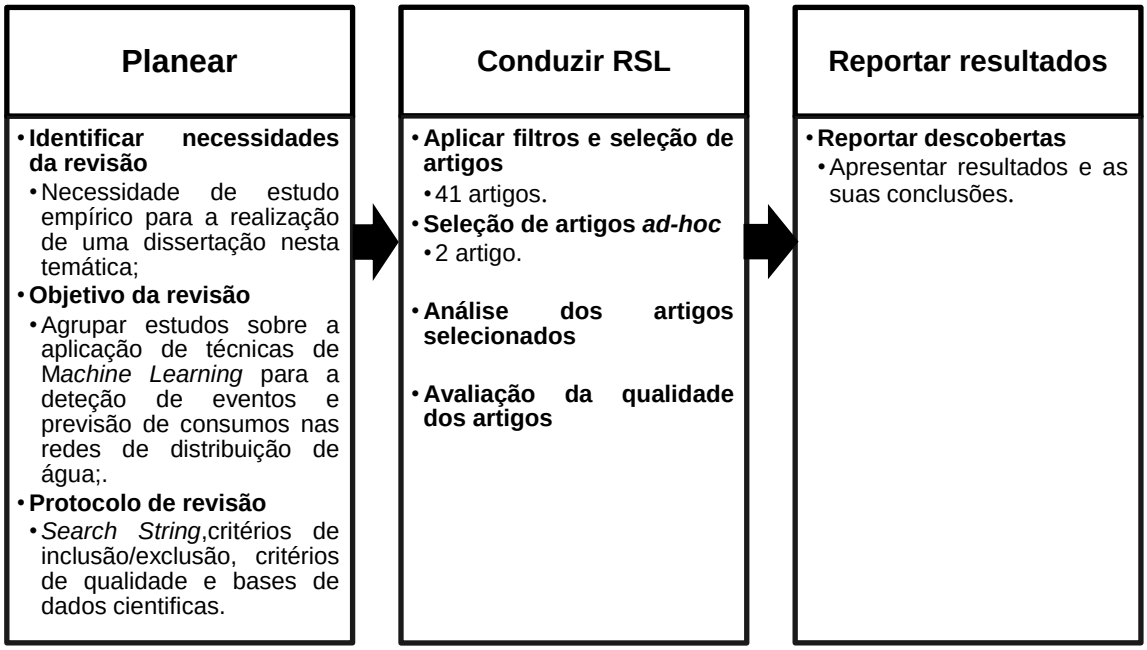

Figura 1: Estágios da revisão sistemática da literatura

Esta revisão sistemática da literatura centra-se na resposta à seguinte questão principal: Como aplicar modelos de *Machine Learning* para a deteção de eventos e previsão de consumos/perdas nas redes de distribuição de água?

Depois de definido o objetivo e a pergunta principal para a revisão sistemática de literatura, propõem-se as seguintes perguntas específicas:

1. Qual a variável *target* mais referida nos estudos?;
2. Que variáveis são utilizadas como *inputs* nos modelos?;
3. Quais os modelos de *Machine Learning* mais utilizados para a deteção de eventos ou previsão de consumos?;
4. Que métricas são utilizadas para avaliar os resultados obtidos?

As bases de dados utilizadas para a realização desta revisão literária foram as seguintes:

- *Web of Science*;
- *Scopus*;

A escolha destas bases de dados científicas prende-se com o facto de serem as mais populares na área de pesquisa científica, e devido ao facto de o motor de pesquisa procurar estudos noutras bases de dados como a ACM, IEEE Xplore, MDPI, entre outras.

O protocolo de revisão inicia-se com a criação de uma *Search String*, utilizada nas bases de dados escolhidas para fazer uma primeira seleção dos estudos com potencial interesse para esta revisão. Dado que o método de pesquisa destas bibliotecas é idêntico, a *Search String* mantém-se praticamente inalterada. Esta *query* foi criada através de pesquisas de

estudo nas bases de dados acima mencionadas, de onde foram extraídas as expressões presentes nas palavras-chave, nos tópicos e títulos, realizando validações incrementais dos resultados originados. A S*earch String* utilizada na base de dados *Web of Science* e *Scopus* foi a seguinte: ("Digital Water Meter*" OR "Water meter*" OR "smart water meter*" OR “Medidor* Digital de Água” OR “Medidor* de Água” OR “Medidor* Inteligente de Água”) AND ("Data Analytics" OR "data analysis" OR "Big Data Analytics” OR “Big Data Analysis” OR “DA” OR “BDA” OR “Análise de Dados” OR "Machine Learning" OR "Artificial Intelligence" OR “ML” OR “AI” OR “Inteligência Artificial” OR “IA”) AND (("detection*") OR ("event*") OR ("Anomaly Detection") OR ("Nonstandard Situation Detection") OR ("Event* Detection") OR (“Fault$ Detection”) OR (“Water Consumption”) OR (“water loss*”) OR ("water Efficiency") OR ("leak detection") OR ("Leak*") OR ("failure") OR ("failure detection") OR ("Deteç*") OR ("Evento*") OR ("Deteção de Anomalia*") OR (“Deteção de Falha*”) OR (“Consumo de Água”) OR (“Perda* de Água”)).

Inicialmente foi realizada uma pesquisa com a *Search String* acima mencionada com a pesquisa em todos os campos, na qual resultaram 41 artigos. De seguida, foi aplicada uma pesquisa por tópico (título, resumo e palavras-chave), pelo facto de se verificar um número considerável de estudos que não se enquadravam na pesquisa, de onde resultaram 28 artigos. O passo seguinte foi aplicar os critérios de inclusão e exclusão, presentes na tabela 1. Começou-se por filtrar pelos anos de 2016 a 2021, que resultou em 25 artigos. Este filtro de inclusão deve-se ao facto de os modelos de *Machine Learning* terem uma aplicação recente nesta área, e por ser uma tecnologia em constante evolução. De seguida filtrou-se pelo idioma em português e inglês, que deu origem a 24 artigos, e por fim, os duplicados, terminando em 23 artigos. Para terminar, foi efetuada uma análise ao resumo de cada artigo para garantir que os estudos estavam relacionados com o tema, de onde foram escolhidos 20 artigos, todos eles disponíveis na base de dados. É possível verificar o processo de filtragem na figura 2.

Tabela 1: Critérios de inclusão e exclusão

Critérios de Inclusão	Artigos em português e inglês;
	Artigos científicos publicados em conferências ou revistas científicas;
	Artigos que abordam a deteção de eventos na rede;
	Artigos desde o ano de 2016 até 2021;
Critérios de exclusão	Artigos que não estão disponíveis através das bases de dados;
	Artigos de acesso antecipado, jornais e livros;
	Artigos em outros idiomas que não inglês ou português;
	Artigos duplicados;

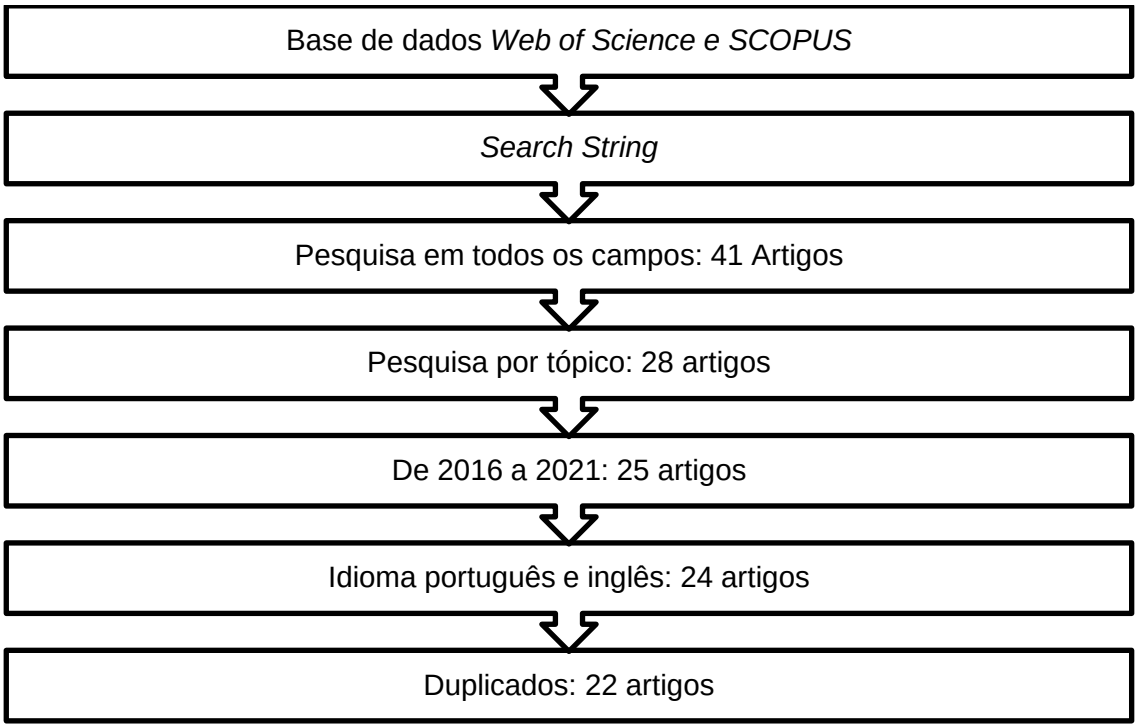

Figura 2: Processo de filtragem dos estudos

Depois da filtragem e seleção, foram extraídos 20 artigos, que se encontram listados na tabela 2. Adicionalmente, os estudos dos autores Candelieri (2017) e Moeeni H et al. (2017) foram retirados das referências de outros artigos e incorporados nesta RSL em *Ad-hoc* devido à sua relevância nesta investigação, perfazendo assim um total de 22 artigos.

Após a análise dos resultados, foi realizada uma avaliação de qualidade a cada artigo na secção das conclusões desta RSL, de forma a verificar a sua qualidade, segundo o objetivo abordado nesta RSL. Os critérios de qualidade são apresentados na tabela 3.

Artigo	Ano	Título	autores	Jornal/Conferência	Quartil
1	2021	A cost-effective CNN-LSTM-based solution for predicting faulty remote water meter reading devices in AMI systems	Lee, J.; Choi, W.; Kim, j.	Sensors, (2021), 21(18).	Q2
2	2021	Neural Network techniques for detecting intra-domestic water leaks of different magnitude	Zese, R.; Bellodi, E.; Luciani, C.; Alvisi, S.	IEEE Acess, (2021).	Q2
3	2021	An alternative approach to dimension reduction for pareto distributed data: a case study	Roccetti, M.; Delnevo, G.; Casini, L.; Mirri. S.	Journal of Big Data, (2021).	Q1
4	2020	Machine learning and data analytic techniques in digitalwater metering: A review	Rahim, M.; Nguyen, k.; Stewart, R.; Giurco, D.; Blumestein, M.	Water(Switzerland), (2020).	Q2
5	2019	Detection of non-standard situation in smart water metering	Kainz, O.; Michalko, M.; Karpiel, E.; Petija, R.; Jakab,F.	2019 IEEE 15TH Internacional Scientific Conference on Informatics, (2019).	NT
6	2019	Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures	Roccetti, M.; Delnevo, G.; Casini, L.; Cappiello, G.	Journal of Big Data, (2021).	Q1
7	2019	Utilization of machine Learning to detect sudden water leakage for smart water meter	Merta, J.; Fikejz, J.	2019 29TH Internacional Conference Radioelektronika	NT
8	2017	Clustering and support vector regression for water demand forecasting and anomaly detection	Candelieri, A.	Water(Switzerland), (2017).	Q2
9	2017	Statistical analysis of the measurement erros in an installation of water meters. Study on the volume of the wayer loss in the installation	Clinciu, M.; Clinciu R.	4th International Conference on Computing and Solutions in Manufacturing Engineering (CoSME) (2017)	NT
10	2019	A Paradox in ML Design: Less data for a smarterwater metering cognification experience	Roccetti, M.; Delnevo, G.; Casini, L.	ACM International Conference Proceeding Series (2019)	NT
11	2020	A Cautionary Tale for Machine Learning Design: why we Still NeedHuman-Assisted Big Data Analysis	Roccetti, M.; Delnevo, G.; Casini, L.; Salomoni, P.	Mobile Networks and Applications (2020)	Q2
12	2017	Anomaly Detection in Smart Water Metering Networks	Kanyama, M.; Nyirenda, C.; Temaneh-Nyah, C.	The 5th International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2017)	NT
13	2017	Machine Learning based Models for Fault Detectionin Automatic Meter Reading Systems	Kou, Y.; Cui, G.; Chen, J.; Li, w.	2017 International Conference on Security, Pattern Analysis,and Cybernetics (SPAC)	NT
14	2020	IoT and Cloud Computing based Smart Water Metering System	Ray, A.; Goswami, S.	International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control, PARC (2020)	NT
15	2016	Fraud detection in energy consumption: A supervised approach	Coma-Puig. B.; Carmona, J.; Gavalda, R.; Alcoverro, S.; Martin, V.;	3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016	NT
16	2017	Adopting IoT Technology to Optimize Intelligent Water Management	Li, E.; Wang, W.; Hsu, Y.;	ICEB 2017 Proceedings (2017)	NT
17	2011	Usage Analysis for Smart Meter Management	Li, H.; Fang, D.; Mahatma, S.; Hampapur, A.	2011 8th International Conference and Expo on Emerging Technologies for a Smarter World (2011)	NT
18	2021	Measurement of Water Consumption based on Image Processing	Kainz, O.; Michalko, M.; Dujava, M.	IEEE 19th World Symposium on Applied Machine Intelligence and Informatics(2021)	NT
19	2018	Forecasting Hourly Water Demands with Seasonal Autoregressive Models for Real Time Application	Chen, J.; Bocceli, D.	Water Resources Research (2018)	Q1
20	2021	Two-stage Framework for Seasonal Time Series forecasting	Xu, Q.; Wen, Q.; Sun, L.	ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing	NT
21	2017	Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction	Moeeni, H.; Bonakdari, H.; Fatemi, S.	Journal of Hydrology (2017)	Q1
22	2020	An empirical investigation of water consumption forecasting methods	Karamaziotis, P.; Raptis, A.; Nikolopoulos, k.; Litsiou, K.; Assimakopoulos, V.;	International Journal of Forecasting (2020)	Q1

Tabela 2: Artigos incluídos na revisão sistemática da literatura

Nota: NT – não tem quartil

Tabela 3: Critérios de qualidade para avaliação de artigos

	Critérios para a avaliação dos artigos
P1	Q1: Aborda qual o *target* do estudo?
	Q2: Descreve o *target*?
P2	Q3: Identifica os modelos de ML utilizados na deteção de eventos?
	Q4: Justifica a escolha dos modelos?
	Q5: Compara os diferentes modelos aplicados?
P3	Q6: Menciona as variáveis utilizadas no estudo?
	Q7: Justifica a escolha das variáveis?
P4	Q8: Identifica os critérios de avaliação de qualidade/performance dos modelos?
	Q9: Descreve os critérios de avaliação?
P5	Q10: A metodologia está bem desenvolvida?
	Q11: São apresentadas as limitações do estudo?

2.3 Resultados e Discussão

2.3.1 Características da base de dados e tipos de eventos observados

Na tabela 4 são apresentados os resultados extraídos dos artigos selecionados nesta RSL relativamente à variável *target*.

Os tipos de eventos abordados nos estudos selecionados são as perdas de água e anomalias nos medidores de água, exceto o artigo número 4, que se baseia numa RSL sobre as possíveis aplicações de *Data Analytics* e *Machine Learning* na análise dos dados provenientes de medidores digitais de água (Rahim et al., 2020). O estudo número 8 foi o único que abordou genericamente a deteção de fraude como um possível evento, e o estudo 14, o único que aborda a temática do consumo excessivo de água, o que pode significar consequentemente uma anomalia. Neste estudo não é aplicado nenhum modelo de ML. O estudo 8 e os estudos do 18 ao 22 abordam a previsão de consumos de água através de algoritmos de ML, utilizando séries temporais e algoritmos para regressão, como por exemplo o ARIMA (Moeeni et al., 2017).

Na coluna das variáveis utilizadas na aplicação dos modelos, é comum em todos os estudos a utilização dos consumos de água, com cadência horária em horas ou intervalos de minutos, a data das medições, e hora a que ocorrem. Existem ainda outras variáveis

pertinentes como os dados meteorológicos no estudo 14, e informações dos clientes como a idade, número de agregados familiares e características do medidor, presentes no estudo 6 e 17. Por fim, no estudo 18 são usadas imagens capturadas ao consumo de um medidor analógico. A utilização destas variáveis dependerá principalmente dos dados presentes na metodologia desenvolvida.

Tabela 4: Variáveis e *target* dos estudos

Artigo	Variáveis	Variável Target
1	Hora da medida, diâmetro do tubo, consumo de água, volts da bateria do medidor, consumo acumulativo, consumo corrente	anomalia nos medidores de água
2	Consumo de água, consumo de água acumulativo	Perdas de água
3	ID do medidor, leituras dos consumos, data dos consumos, ID do material, ID da marca, tipo de uso, rotulo (defeituoso, não defeituoso)	Anomalia nos medidores de água
4	NE	Perdas de água
5	Consumo de água, tipo de água, hora, data	Anomalia nos medidores de água
6	Consumos de água, leitura anterior, data da leitura anterior, número de série, ID do material, ID tipo de medidor, ano de construção, categoria de uso	Anomalia nos medidores de água
7	Consumo de água, hora, data	Perdas de Água
8	Consumo de água, data, hora	Deteção de perdas de água, deteção de anomalia, fraude, previsão de consumos
9	Volume de água registado, média do vol de água, hora, vol de água perdida, vol de água perdida por hora	Perdas de Água
10	Leituras dos consumos por hora (2014 a 2018)	Anomalia nos medidores
11	Leituras dos consumos por hora (2014 a 2018)	Anomalia nos medidores
12	Leituras dos consumos de água minuto a minuto	Deteção de anomalias
13	Leitura dos consumos por hora	Anomalia nos medidores
14	Leitura dos consumos	Consumo excessivo de água
15	Leituras dos consumos de gás e eletricidade mensal, tarifas, histórico de fraudes, tipo de leitor, meteorologia, localização, informações de faturação	Deteção de fraude
16	Consumos de diários, semanais e mensais e de hora a hora	Perdas de água
17	Consumos horários, dados meteorológicos diários, tipo de cliente, idade do leitor, material do leitor, número de agregados familiares	Anomalia nos medidores
18	Imagens dos consumos dos medidores (hora a hora)	Anomalia nos medidores
19	Consumos de água de hora a hora	Previsão de consumos
20	Consumos de água de 12 em 12 horas	Previsão de consumos
21	Fluxo de água mensal	Previsão de consumos
22	Consumo de água mensal (2007 a 2013)	Previsão de consumos

NE: não especifica

Começando pela primeira questão – Qual a variável *target* mais referida nos estudos? - Os eventos identificados na amostra de estudos extraídos, são as perdas de água provocadas por roturas na rede de distribuição de água, e por anomalias que possam ocorrer nos

medidores de água, que provocam leituras erradas do consumo de água. No conjunto de estudos escolhidos, apenas o estudo 15 trata o tema da deteção de fraudes, não sendo este diretamente relacionado com os recursos hídricos (Coma-Puig et al., 2016). O estudo de Candelieri (2019), também fez uma pequena abordagem à deteção de fraudes.

A resposta à segunda pergunta - Que variáveis são utilizadas como *inputs* nos modelos? - São essencialmente leituras dos consumos de água, hora das leituras, data, e eventualmente características dos medidores, como por exemplo, marca, ano de construção, modelo, e variáveis compostas como o consumo acumulativo.

2.3.2 Aplicação de modelos de *Machine Learning*

Como é possível verificar na tabela 5, ao longo dos diversos artigos existe uma grande diversidade de algoritmos de ML, no entanto as técnicas escolhidas têm por base a mesma família em grande parte deles. A técnica mais utilizada pertence à família dos modelos supervisionados, sendo ela as *Neural Networks*, em específico as *Recurrent Neural Network* (Roccetti et al., 2021), as *Convolutional Neural Networks* (Lee et al., 2021; Zese et al., 2021), juntamente com séries temporais, e as *Feed Forward Neural Networks* (Roccetti et al., 2020).

Por outro lado, são também utilizadas técnicas de regressão, como é o caso da *Support Vector Regression* (Candelieri, 2017), nas séries temporais. Apesar de esta família de técnicas ser bastante considerada nos estudos, foi pouco implementada, principalmente porque uma anomalia pode ser mais facilmente detetada utilizando um modelo de classificação (desde que existam exemplos para treinar os modelos). O conceito de séries temporais é inerente a este tema, pelo que todos os estudos apresentados neste RSL foram realizados tendo em conta este contexto.

No estudo 18, o *Teesaract*, *OCropus e GOCR* são ferramentas de *Open Source* (no caso do *Tessaract* tem uma licença pertencente ao *Apache*), que têm por base algoritmos como o MLP (*OCropus* e *GOCR)*, para fazer a leitura dos caracteres das imagens e transformá-los em dados. O estudo não descreve aprofundadamente os algoritmos considerados em cada ferramenta, pelo que na tabela estão presentes os nomes das ferramentas consideradas (Kainz et al., 2021). O estudo número 13 utiliza um algoritmo *Adaboost,* que resulta da fusão entre RNN e *Decision Trees*. Aplicando os três algoritmos, o *Adaboost* apresenta maior *Accuracy, sendo* superior aos restantes, com 0.990 nos resultados (Kou Yinggang et al., 2017). O estudo 17 utiliza a técnica ARIMA (*Auto-Regressive Integrated Moving Average*), que se baseia numa regressão linear e médias móveis, com o objetivo de prever resultados futuros influenciados por resultados do passado. É uma técnica que pode oferecer bons resultados se a série temporal for estacionária e/ou se apresentar sazonalidade na amostra, ou seja, uma repetição de padrões ao longo de um intervalo de tempo.

Também foram utilizadas técnicas da família não-supervisionada como *a K-Nearest Neighbors*, devido ao facto de os dados disponíveis não estarem rotulados (Nyah et al., 2017). Adicionalmente, foram extraídos estudos tendo em vista o *target* de previsão de consumos de água. Pode verificar-se na tabela 5 os estudos 8, 17, 19, 20, 21 e 22, que abordam este *target*. Todos os modelos dos estudos acima mencionados têm por base o *Autoregressive Model* na previsão de consumos de água, como por exemplo o ARIMA ((Autoregressive Integrated Moving Average). Existe ainda um algoritmo híbrido (utilização de dois algoritmos) que agrega um algoritmo NN(MLP) e um *Autoregressive Model* (*Multi horizon Autoregressive Model*)(Xu et al., 2021).

Na terceira pergunta - Quais os modelos de *Machine Learning* mais utilizados para a deteção de eventos ou previsão de consumos? - Podemos concluir que existem várias técnicas que podem ser aplicadas, e que o sucesso da aplicação dos modelos vai depender do tipo de variáveis de que dispomos e da sua qualidade (preparação dos dados) e quantidade (Roccetti, Delnevo, Casini, Zagni, et al., 2019). Segundo os artigos analisados, as técnicas mais comuns são a *Recurrent Neural Network (RNN) e a Convolutional Neural Network (CNN),* pertencentes à família das NN. Os algoritmos para previsão de consumos de água mais utilizados foram os correspondentes aos modelos de séries temporais, como o ARIMA e SARIMA.

Artigo	Técnica usada	Técnicas consideradas	Objetivo do Estudo
1	Neural Networks (CNN-LSTM)	*Random Forest, GMM*	Prever a falha de um medidor de água através da aplicação de uma rede neuronal convolucional e séries temporais.
2	Neural Networks (CNN), séries temporais	*Random Forest, Recurrent Neural Network, hybrid convolutional-recurrent Neural Networks*	Detetar perdas de água de diferentes magnitudes com a ajuda de técnicas supervisionadas.
3	Recurrent Neural Network (RNN)	NE	Deteção de falhas em medidores mecânicos através de técnicas de *deep learning,* utilizando diferentes abordagem nas variáveis categóricas.
4	NE	NN *SVR, RFt, multivarite Adaptative Regression Spliners, Projection Pursuit Regression, k-Means clustering*	Mostrar o que foi desenvolvido na aplicação de *machine learning e data analyrics* na leitura digital de dados de água.
5	*XGBOOST: XGBRegressor*	*K-Means clustering*	Aplicação de um algoritmo para a deteção de situações não habituais no consumo de água.
6	*Recurrent Neural Network (RNN)*	*Linear regression, Lasso, Classification and Regression Tree(CART),* SVR, KNN, *Adaptaive boosting(AB), Gradient Boosting(GB),* RF, MLP	Desenhar um algoritmo de *machine learning* que consiga detetar falhas nos medidores de água.
7	*Symbolic Regression, Genetic Programing*	*NN*	Deteção de perdas de água em vários intervalos de tempo possíveis.
8	*Support Vector Regression(SVR),* séries temporais	*NN*	Previsão de consumo de água numa rede e deteção de anomalias com a aplicação da técnica de *support vector regression.*
9	*Least Square*, *maximum likelihood*	NE	Contagem do número de litros de água perdidos utilizando parâmetros estatísticos e vários medidores de água.
10	*Feed forward NN e Recurrent NN*	NE	O objetivo deste tudo é prever quando um medidor falha através de modelos de classificação, utilizando menos dados.
11	*Feed forward NN e Recurrent NN*	*Naïve data semantics*	Prever medidores anómalos utilizando um grande volume de dados (15 milhões linhas) e modelos de classificação
12	*K-NN, Cluster-based local outlier factor, histograma-based outlier score*	NE	Detetar registos anómalos de consumos de água através de técnicas não supervisionadas.
13	*Adaboost*	*Reccurrent Neural network com LSTM, Decision Tree,*	Apresentar alguns modelos supervisionados para a deteção de falhas nos medidores e seleção do melhor.
14	SVM e regressão Logística	Regressão Linear	Deteção de consumo excessivo de água em residências e estabelecimentos comerciais através de ferramentas de *Machine learning.*
15	*Gradient Boosting e Decision Tree*	Naive Bayes, KNN, C4.5, CART, NN, RF, Adaboost, SVM	Deteção de fraude na rede de gás e elétrica de uma cidade em Espanha, com recurso aos consumos mensais, dados faturação e histórico de fraude, utilizando modelos de ML.
16	Redes neuronais, SCADA	**NE**	Através da tecnologia SCADA conectar vários medidores a uma infraestrutura para gerir os recursos hídricos, incluindo detetar perdas de água
17	ARIMA	NE	Este estudo propõe uma abordagem analítica para a deteção de anomalias nos medidores e previsão de consumo de água
18	*Tessaract* e *OCropus: Multi-layer perceptron NN*	*Tessaract, OCropus, GOCR*	Colecionar automaticamente as leituras de medidor analógico e processar os dados para descobrir situações fora do comum
19	*ARI (Autoregressive integrated)*	*NN*	Prever a procura de água através do algoritmo ARI e tendo em conta a sazonalidade da série temporal
20	*MLP + MAR(multi-horizon Autoregressive Model)*	*SARIMA*	Prever o consumo de água nas 12 horas seguintes utilizando dois estágios de sazonalidade
21	*ARMA*	ARIMA	Prever o fluxo de água mensal de alguns rios dos Estados Unidos, tendo em conta vários tipos de sazonalidade
22	*ARIMA*	*ETS, Theta Method, Optimized Theta, MLP, Naive Method, MAPA*	Prever os consumos mensais de água de agrupamentos de consumidores de diferentes zonas

Tabela 5: Técnicas utilizadas e consideradas, e respetivo objetivo de pesquisa por artigo

Nota: NE - Não especifica

2.3.3 Métricas de Avaliação

Na tabela 6 estão presentes os resultados das métricas de desempenho aplicados ao modelo de cada estudo.

Nos estudos que apresentaram métricas de desempenho para a avaliação dos modelos, destacam-se a ROC e a AUC em 6 dos 22 estudos. Foram também utilizadas as métricas MAPE, RMSE, *R Square* e MAE.

Tabela 6: Métricas de desempenho com o respetivo algoritmo aplicado

Artigo	Técnica usada	Métricas de desempenho
1	Redes Neuronais (CNN-LSTM)	*Precision, recall, f-measure, MCC, ROC*
2	Redes neuronais (CNN), séries temporais	Accuracy, Precision, recall, AUC, ROC
3	Recurrent Neural Network (RNN)	AUC, ROC
4	NE	NE
5	*XGBOOST: XGBRegressor*	NE
6	*Recurrent Neural Network (RNN)*	AUC, ROC
7	*Symbolic Regression, Genetic Programing*	NE
8	*Support Vector Regression(SVR),* séries temporais	MAPE (Mean Absolute Percentage Error)
9	*Least Square*, *maximum likelihood*	*Kolmogrov-Smirnov test, normality goodness of fit test*
10	*Feed forward NN e Recurrent NN*	AUC, ROC
11	*Feed forward NN e Recurrent NN*	AUC, ROC
12	*K-NN, Cluster-based local outlier factor, histograma-based outlier score*	*False positives rate (FPR), detection rate (DR)*
13	*Adaboost*	*Accuracy*
14	SVM e regressão Logística	NE
15	Gradient Boosting e Decision Tree	Precision, Recall, AUC
16	Redes neuronais, SCADA	NE
17	ARIMA	*Accuracy*
18	*Tessaract* e *OCropus: Multi-layer perceptron NN*	*Accuracy*
19	*ARI*	*MAPE, R Square, RMSE, PICP, Correlation Coefficient*
20	*MLP + MAR*	*MAPE, RMSE, RMSPE, MAE*
21	*ARMA*	*R Square, MAPE, RMSE, RMSRE, MAE, VAF, SI, MRE, BIAS*
22	*ARIMA*	*MAE, MASE, SMAPE, RMSE*

NE: não especifica

Por fim, é possível responder à quarta questão de investigação - Que métricas são utilizadas para avaliar os resultados obtidos? – O ROC, AUC e Accuracy são métricas de desempenho, utilizadas nos estudos onde são desenvolvidos modelos de classificação binária. Nos estudos 8, 17, 19, 20, 21 e 22, nos quais o objetivo principal é a previsão de consumos de água (modelos de regressão), foram utilizadas outras métricas de desempenho. Neste tipo de abordagem as métricas mais observadas são o R Square, o MAPE, o MAE e o RMSE. Estas são as métricas aplicadas para a avaliar o desempenho dos modelos no capítulo dos resultados e discussão.

2.4 Avaliação de Qualidade dos artigos científicos

Uma revisão sistemática da literatura tem por base o estudo e a avaliação de artigos relacionados com um tema em específico, neste caso as técnicas de *Machine Learning*, utilizadas para a previsão de consumos e deteção de eventos nas redes de distribuição de água. Foram desenvolvidos critérios para avaliar a qualidade dos estudos selecionados para esta RSL. Para cada critério foi atribuída uma classificação de 0 (zero), como não cumpre os critérios; 1, como cumpre o critério e 0.5, cumpre parcialmente o critério, sendo a pontuação máxima 11. Esta Avaliação está presente na tabela 7.

Dos artigos extraídos para esta RSL, conclui-se que o artigo com maior pontuação é o número 6 (Roccetti, Delnevo, Casini, & Cappiello, 2019), em segundo o número 1 (Lee et al., 2021), com uma pontuação de 10.5, e o critério de qualidade mais importante (melhor qualidade) foi o Q1. Em terceiro na classificação está o estudo 22 que aborda a previsão de consumos utilizando algoritmos de ML e *Clustering* para agrupar os consumidores por localização (Karamaziotis et al., 2020).

O estudo que apresenta menor classificação é o número 17 (H. Li et al., 2021) que, apesar de apresentar uma avaliação baixa na sua qualidade, continua a ser uma contribuição pertinente para o estudo deste tema. O critério de qualidade com menor pontuação é o Q11, com 6.5 pontos de 0 a 11.

Existem outros estudos com boas pontuações, como é o caso do número 12 (Nyah et al., 2017) com 9.5, seguido do número 2 (Zese et al., 2021), e do número 3 (Roccetti et al., 2021) apresentando uma pontuação de 9 em 11. Para a investigação em questão, os estudos 1, 2, 3, 12 e 22 são ótimos contributos, pois abordam o *target* da deteção de eventos e previsão de consumos, e a aplicação de *Clustering* para agrupar consumidores. Para além destes, o estudo 8 (Candelieri, 2017), apesar de não ter uma pontuação tão boa, é o que se aproxima mais do objetivo desta dissertação e, por isso, serve de base para a construção da metodologia desenvolvida.

Tabela 7: Avaliação da qualidade dos artigos analisados

	P1		P2			P3		P4		Aval		
Número	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Total**
1	1	1	1	1	1	1	1	1	1	1	0.5	**10.5**
2	1	1	1	1	1	0.5	0.5	1	1	1	0	**9**
3	1	0.5	1	1	1	1	1	1	0.5	1	0	**9**
4	1	1	1	0	0.5	0	0	0	0	1	1	**5.5**
5	1	1	1	0.5	0	0.5	0	0	0	0.5	0.5	**5**
6	1	1	1	1	1	1	1	1	1	1	1	**11**
7	1	0.5	1	0.5	0	1	0	1	0.5	1	0	**6.5**
8	1	1	1	0.5	0.5	0.5	0	1	1	1	1	**8.5**
9	1	0.5	0	0	0	1	0	1	0.5	1	0.5	**5.5**
10	1	1	1	0	0	1	1	1	0.5	1	0	**7.5**
11	1	1	1	0.5	0	0.5	0	1	0.5	1	0	**6.5**
12	1	1	1	1	1	1	0	1	1	0.5	1	**9.5**
13	1	0.5	1	1	1	1	0	1	0	1	0	**7.5**
14	1	1	1	0.5	0	0.5	0	0	0	1	0	**5**
15	1	1	1	1	1	1	0.5	1	0	1	0	**8.5**
16	1	1	1	0	0	1	0	0	0	1	0	**5**
17	1	0.5	0.5	0	0	1	0.5	0	0	1	0	**4.5**
18	1	1	1	0.5	1	1	1	1	0	1	0	**8.5**
19	1	1	1	1	0	1	1	1	0.5	1	0	**8.5**
20	1	1	1	1	0	1	0	1	0	1	0.5	**7.5**
21	0.5	1	1	0	1	1	0	1	0	1	0	**7.5**
22	1	1	1	1	1	1	1	1	0	1	1	**10**
Total	**21.5**	**19.5**	**20.5**	**13**	**11**	**18.5**	**8.5**	**17**	**8**	**21**	**6.5**	

Capítulo 3. Metodologia e Exploração dos Dados

3.1 Metodologia

Neste subcapítulo é explicada a metodologia adotada para a realização deste estudo. A metodologia é baseada na metodologia apresentada no estudo do autor Candelieri. A. (2017), onde o autor utiliza métodos de *Clustering* para agrupar consumidores consoante os seus consumos de água e faz uma previsão do consumo de água desse grupo utilizando um modelo de regressão.

O primeiro passo foi realizar uma análise descritiva e exploratória dos dados para entender que atributos existem à disposição e quais devem ser descartados na análise, pelo facto de não se enquadrarem nos objetivos deste estudo. Depois de alinhar os dados temporalmente e filtrar a informação, procedeu-se à análise de *missing values*. São eliminados os consumidores que ultrapassem o limite estabelecido de *missings,* e os valores em falta são preenchidos através da técnica de interpolação linear.

Tendo como amostra vários clientes, o passo seguinte foi a aplicação de uma técnica de *Clustering* para agrupar os consumidores através dos seus comportamentos, dando origem ao consumo médio de cada grupo. Essas técnicas são o *K-Means* e o *DBSCAN*. A escolha da técnica a utilizar foi influenciada pelo resultado da métrica de desempenho escolhida, que foi o *Silhouette Coefficient.*

De seguida foram aplicados modelos de regressão ao grupo de consumidores escolhido para dar seguimento ao estudo, tendo em conta a estacionaridade e sazonalidade da série, e as características e limitações de cada modelo e, por fim, escolhido o modelo com o melhor desempenho nas métricas selecionadas. Graças ao RSL apresentado anteriormente, foi possível conhecer alguns dos algoritmos mais utilizados na temática das séries temporais, sendo utilizados neste estudo, como o ARIMA e o MLP. Para analisar os resultados dos algoritmos foram usadas as métricas de desempenho, utilizadas também nos estudos apresentados na revisão da literatura, na qual se destacam o MAPE, o MAE e o RMSE.

Na etapa final, a informação foi analisada e foram criados dois *dashboards.* O primeiro *dashboard* direcionado para o consumidor final, enquanto o segundo, direcionado para a empresa, de modo a representar a análise de possíveis anomalias. Através dos valores reais de consumo de cada cliente, cruzados com os possíveis desvios do padrão de consumo de água que o grupo de consumidores escolhido apresenta. No decurso da realização da investigação foi utilizado a plataforma *KNIME 4.3.4* e o *Power Bi Desktop.*

3.2 Análise e Exploração dos Dados

Este subcapítulo inicia-se com uma pequena abordagem e descrição da empresa que disponibilizou os seus dados para a realização desta dissertação. Os subcapítulos seguintes apresentam as principais características da amostra e uma análise exploratória inicial dos dados.

3.2.1 Estudo de caso: A Infraquinta

A Infraquinta é uma empresa localizada na Quinta do Lago, em Loulé, e a sua principal função é fazer a gestão das infraestruturas deste espaço (como por exemplo, jardins) e fornecimento e gestão da rede de distribuição de água. O seu principal objetivo é realizar quaisquer trabalhos e adaptação das infraestruturas, na Quinta do Lago, de forma a permitir a gestão sustentável da rede de infraestruturas, manter a qualidade urbanística do espaço e um bom relacionamento com proprietários, residentes, utentes e organismos privados e públicos. A sua missão consiste em promover a qualidade e a boa gestão do espaço urbano e infraestruturas públicas.

A Infraquinta pode ser considerada uma empresa *Data-driven* e mais avançada tecnologicamente comparativamente com outras empresas do setor em Portugal, pelo facto de ter infraestruturas que permitem a receção de milhares de dados diariamente que necessitam ser analisados. Estes dados são provenientes de medidores digitais de água presentes nas propriedades dos clientes, como de caudalímetros espalhados pela rede e medidores à entrada de cada ZMC (Zona de Medição e Controlo).

Esta empresa já possui sistemas como o *IQ Alert* com o objetivo de avisar o cliente, no caso de existir uma possível fuga na sua residência. O sistema analisa os consumos dos clientes e, caso os seus consumos sejam superiores a um determinado valor estabelecido por hora (a cada 24 horas), o sistema envia um e-mail automático ao cliente com o respetivo aviso. Este sistema demora cerca de dois dias a fazer a deteção, não havendo ainda uma solução para uma deteção em tempo real.

A utilização de um software SCADA (Supervisory Control and Data Acquisition) é também essencial para fazer a gestão e monitorização de toda a rede de abastecimento, através das informações recebidas pelos caudalímetros espalhados pela rede. É utilizado ainda um sistema de utilização interna, para fazer a gestão do parque de contadores digitais (IQSig) e dos medidores à entrada de cada ZMC.

Em relação aos medidores digitais nas propriedades, estes comunicam por ondas de rádio estabelecendo uma rede *mesh*, que envia diariamente as leituras horárias num determinado momento do dia para uma base de dados que concentra todas as leituras do respetivo dia.

Depois é executada uma *query* em *SQL (Structured Querie Language)* para inserir as leituras no *Data Warehouse* para análise. Todas as comunicações entre medidores e sistemas de gestão é realizada através de telemetria. Existem algumas limitações quanto à origem dos consumos de água, sendo esta uma zona onde existe bastante turismo e propriedades de luxo, onde o consumo pode ser bastante superior à média de uma propriedade unifamiliar normal, uma vez que muitas das propriedades têm uma grande área. com jardins e piscinas, e a maioria não possui um contador específico para a zona exterior da propriedade, não sendo, por isso, possível classificar um aumento de consumo. Devido a estes fatores, torna-se difícil fazer uma deteção eficaz de anomalias nas propriedades domésticas.

3.2.2 Caracterização da base de dados

Os dados fornecidos pela Infraquinta para a realização desta dissertação focam-se na tabela das leituras, que correspondem aos consumos de água, à data e hora de cada leitura e ao ID do consumidor (snno).

A tabela dos contratos, que apresenta informações sobre os contratos de cada instalação, com a chave primária, a tarifa (comércio e indústria, hotéis e aparthotéis, rega não doméstica, obras, doméstica, rega doméstica e interno), o estado do contrato (inativo, cortado e ligado) e o tipo de cliente.

A tabela instalações indica as informações das instalações, através da chave primária do consumidor, o ID da ZMC, o ID das coordenadas, o número de série do medidor e o tipo. Desta tabela apenas foi utilizado o ID da ZMC para agrupar os consumidores nas respetivas ZMC.

As três tabelas mencionadas acima são as utilizadas para o desenvolvimento deste estudo e relacionam-se através do ID dos consumidores.

A tabela das leituras tem no total 131.991.919 linhas de leituras retiradas dos medidores de cada propriedade. As leituras variam entre um mínimo (-45,5 m^3/h) e 2.097 m^3/h, uma média de leituras de 0.265 m^3/h e um desvio padrão 3.97m^3/h. Numa primeira análise não há qualquer valor em falta (*missing values).* No atributo da data de leitura, o primeiro registo de leitura foi a 2 de novembro de 2013, às zero horas, e a data da última leitura foi a seis de setembro de 2021, às 23 horas. Foram detetados cerca de 1.792.642 *outliers* na amostra.

Na dimensão dos contratos, cada linha corresponde ao número de instalações presentes na rede (que se tem acesso às informações de contrato), o que significa que existe um total de 2384 contratos. Destes, foram considerados os clientes com contratos ativos (2316).

A empresa distingue os clientes através da variável do tipo de cliente, que se divide em clientes não domésticos e domésticos, no qual o mais representativo são os clientes domésticos, com 2048 consumidores, como se pode verificar na figura 3.

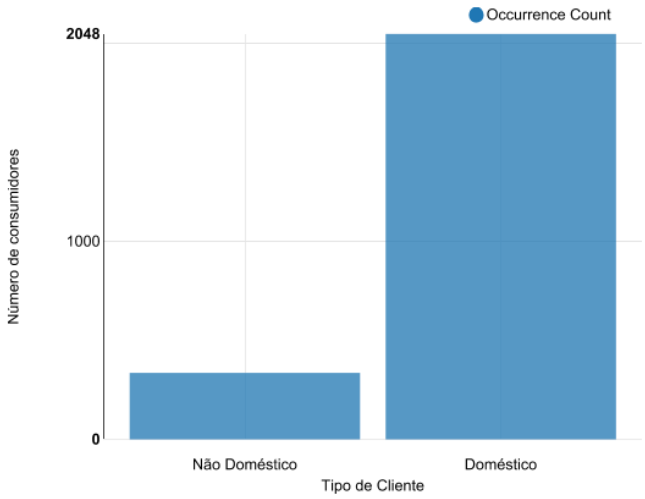

Figura 3: Número de consumidores por tipo de cliente

No atributo Tarifa, descrito na figura 4, estão presentes seis categorias: doméstica, comércio e indústria, rega doméstica (rega D), interno, obras, hotéis e aparthotéis, e rega não doméstica (rega ND). As tarifas mais utilizadas são a doméstica, com 1933 contratos, seguida do comércio e indústria, com 189 contratos, a rega doméstica, com 115 contratos e, por fim, o uso interno, com 105. Em menor número existem a tarifa de obras, com 21, hotéis, com 14 e, finalmente, a Rega ND (rega não doméstica), com 16 consumidores. No atributo tipo de cliente são considerados clientes domésticos aqueles que têm uma tarifa doméstica e rega doméstica. Não existem valores omissos aparentes nesta tabela e os dados foram filtrados apenas para a tarifa doméstica.

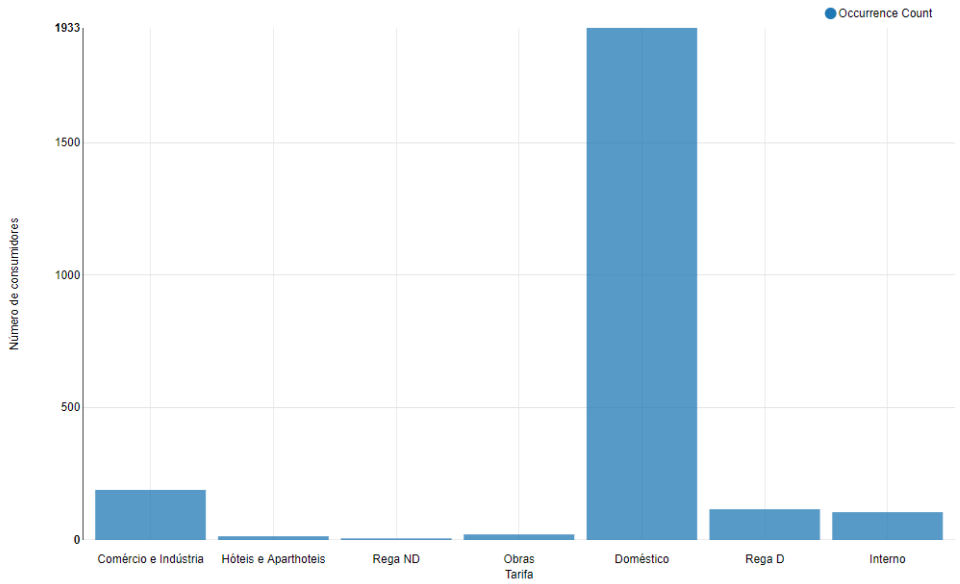

Figura 4: Número de consumidores por tipo de tarifa

Na tabela das instalações, cada linha corresponde a uma instalação registada na rede (2474), existindo uma linha por ID de consumidor, e o ID de cada ZMC correspondente. A variável número de série é única para cada equipamento, não havendo valores repetidos. Das 2474 instalações presentes, existem 21 tipos de medidores na amostra, sendo o mais popular o AQUADIS20, com 1910 exemplares. Existem no total 156 valores omissos, tanto na variável série, como no tipo. A série e o tipo de medidor não são tidos em conta neste estudo.

3.2.3 Seleção da Amostra

Tendo em conta o objetivo principal desta dissertação, descrito anteriormente na introdução, foi necessário filtrar algumas variáveis e tabelas que não pertenciam ao contexto. A primeira etapa na seleção dos dados foi filtrar a amostra com a variável o estado do contrato, mantendo apenas os consumidores cujo estado é *ligado*.

Sendo um dos objetivos, a deteção de possíveis anomalias na rede doméstica, a segunda etapa realizada foi filtrar pelo tipo de cliente doméstico e, de seguida, pela tarifa doméstica, de modo a garantir que os dados a ser tratados são exclusivamente de consumidores domésticos. Dado que as leituras estão presentes na tabela dos consumos, e as variáveis filtradas na dimensão dos contratos, foi necessário fazer uma agregação entre tabelas (através do nó *Join* no *KNIME*), utilizando o ID dos consumidores. O mesmo foi efetuado para os consumidores que pertencem a cada ZMC. Um excerto do *workflow* relativo à seleção dos dados pode ser visualizado na figura 5.

Na última etapa, foi selecionada a ZMC a utilizar para este estudo. Foi escolhida a ZMC 2, pelo facto de possuir um maior número de consumidores (203 consumidores) e, por consequência, maior variedade de perfis de consumidores, como está presente na figura 6.

Em relação ao intervalo de tempo considerado para análise, os dados foram limitados de 1 de janeiro de 2015 a 31 de dezembro de 2019. Apesar de existirem mais dados para além do fim de 2019, estes não foram considerados por representarem um período atípico de consumo provocado pelo início da situação pandémica.

É de realçar que o processamento dos dados do estudo foram efetuados numa máquina local, e que para agilizar o tempo de desenvolvimento e processamento dos dados, não foi possível realizar a investigação com todos os dados disponibilizados, o que explica o foco apenas numa ZMC da base de dados.

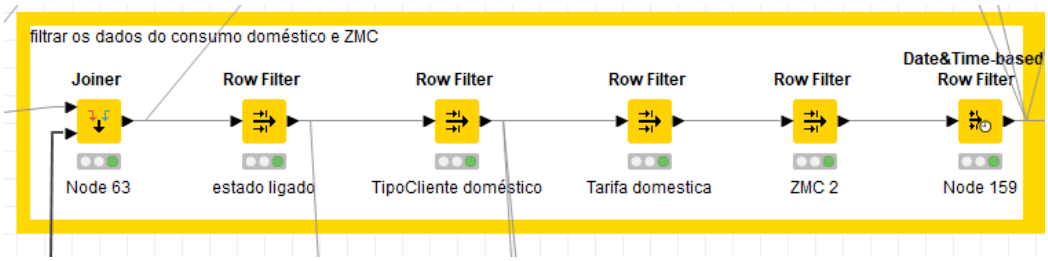

Figura 5: Excerto de *workflow* na filtragem dos dados

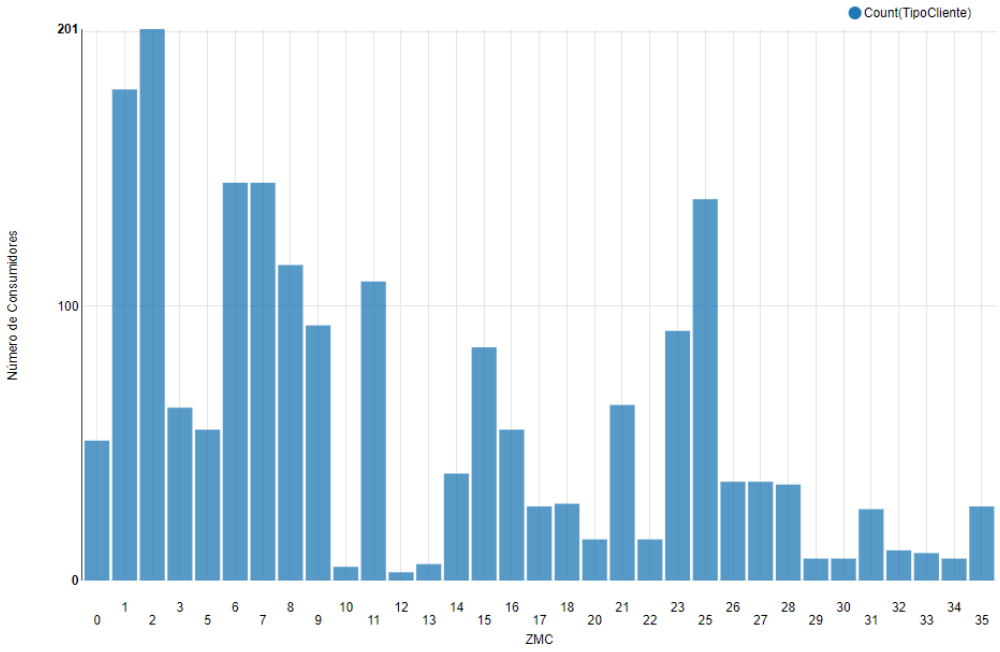

Figura 6: Número de Consumidores por ZMC

Capítulo 4. Resultados e Discussão

Este capítulo pretende dar continuidade à metodologia e à exploração dos dados, expondo os resultados e interpretações destes, tal como discuti-los conforme os objetivos traçados.

4.1 Aplicação de Métodos de Agrupamento

Antes da aplicação das técnicas de agrupamento, foi realizada uma verificação para apurar se existiam horas em falta na série temporal, na qual se confirmou a inexistência de algumas horas, acabando por verificar que existem *missing values* na amostra. Inicialmente foi realizado um alinhamento na série, inserindo os períodos em falta para verificar o número de *missing values* por ID de consumidor na amostra. Foi estabelecido um *threshold* de cinco por cento, onde o grupo de clientes desta ZMC passou a ser 180 consumidores (anteriormente 201 consumidores), pelo facto de ultrapassar o limite de *missing values* que foi estabelecido.

De seguida, foi aplicada a técnica de interpolação linear para preencher os restantes valores de leituras omissas dos consumidores (Ottosen & Kumar, 2019). Nesta investigação foram utilizados métodos de *Clustering* para agrupar os clientes segundo os seus comportamentos de consumo. Neste sentido, com recurso às leituras dos medidores, foram construídas as seguintes variáveis: consumo médio anual, consumo médio mensal, consumo médio semanal, consumo medio diário, consumo médio por hora, percentagem do consumo médio para cada dia da semana e percentagem do consumo médio dividido em 4 segmentos do dia.

Utilizando a plataforma *KNIME,* foram aplicadas duas técnicas de *Clustering*, sendo elas *K-Means* (no Ambiente Spark) e *DBSCAN.* O *K-Means foi aplicado* no estudo do autor Candelieri (2017). De modo a dar seguimento a esta investigação, para selecionar a técnica a utilizar, foi definida uma métrica de desempenho, sendo ela o *Silhouette Coeffcient (SC).* Quanto maior for o valor de SC, melhor será o desempenho da técnica (Shi & Zeng, 2014). O *threshold* definido inicialmente nas técnicas desta investigação, tem de ser superior ou igual a 70%. Os resultados destas métricas estão representados na tabela 8.

Tabela 8: Comparação dos resultados das técnicas de *clustering*

	Silhouette coefficient
K-Means	0.807
DBSCAN	0.653

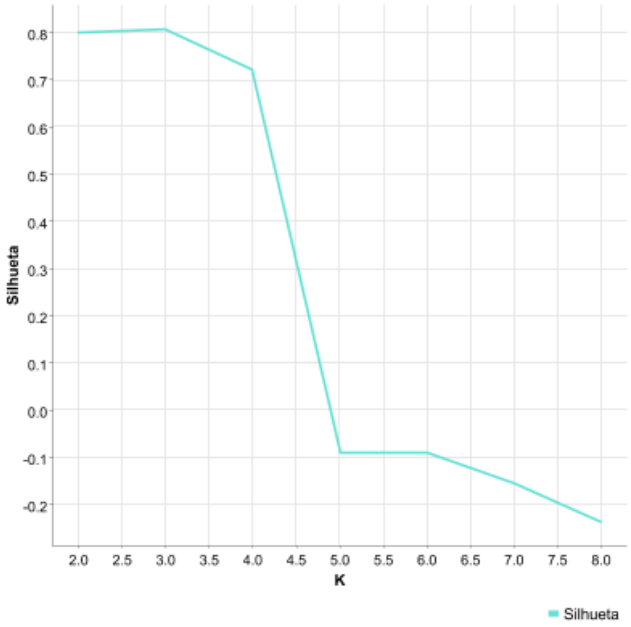

Figura 7: Valor do SC por *K*

Na tabela 9 podemos verificar que a técnica que obteve melhor desempenho nas métricas escolhidas foi o *K-Means*. Na figura 8 temos presente um excerto do *workflow* desenvolvido para o *K-means*. É de referir que o *K-Means* realizou-se num ambiente *Big Data* utilizando o *Spark*. Para a escolha do número de grupos *K* no *K-Means*, as técnicas foram testadas para vários valores de 2 a 8. Dos resultados presentes na figura 7, o *K* igual a 3 obteve o melhor desempenho de todos os *K.*

Já no DBSCAN, os parâmetros que apresentaram melhor desempenho foram o *Epsilon* igual a 8.5 e o número mínimo de pontos igual a 4. A matriz de distância utilizada antes da aplicação das técnicas foi a *Z-Score*.

Depois da análise das métricas de desempenho de cada técnica, foi escolhido o *K-Means* e analisado o perfil de cada *cluster*, que se encontram descritos nas tabelas 9 e 10.

Tabela 9: Perfil de consumo de cada *cluster*

Cluster	Tamanho	Média por hora	Média diário	Média semanal	Média mensal	noite	manhã cedo	manhã	tarde	anoitecer
0	138	0.016	0.376	2.619	11.437	12.558%	23.882%	26.08%	23.283%	14.199%
1	8	0.419	10.054	70.032	305.807	37.72%	25.993%	7.526%	6.476%	22.345%
2	34	0.161	3.872	26.97	117.77	28.913%	24.877%	9.554%	8.86%	27.796%

Tabela 10: Características de consumo de cada cluster por dia da semana

Cluster	Tamanho	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
0	138	15.932%	14.968%	15.12%	14.701%	15.305%	12.519%	11.456%
1	8	14.661%	13.338%	14.166%	14.96%	14.802%	14.167%	13.907%
2	34	14.589%	14.081%	14.355%	14.5%	14.914%	14.018%	13.542%

Na tabela 9 está referido o número de consumidores por *cluster*, as médias de consumo por cada período (em metros cúbicos), e as percentagens médias de consumo por cada período do dia. É possível averiguar que o *cluster* 0 é o que apresenta a maior parte da amostra, com 138 consumidores. Este *cluster* também é o que representa uma distribuição de consumo mais uniforme ao longo do dia. Verifica-se que os consumidores do *cluster* 1, consomem maior quantidade de água ao longo do intervalo de tempo definido (média por hora, diária, semanal e mensal), seguido do cluster 2 e 1. Durante a noite (das 0h às 5h) e manhã cedo (das 5h às 9h), o cluster 1 é o que gasta mais água, com cerca de 64% do consumo do dia. Os consumidores do *cluster* 0 consomem mais durante a manhã (das 9h às 12h) e à tarde (das 13h às 14h), contrariamente aos restantes grupos. Por fim, o *cluster* 2 destaca-se por consumir mais ao anoitecer (das 18h às 23h).

Na tabela 10 estão presentes os consumos médios de cada cluster ao longo da semana, em percentagem. Não é possível verificar um padrão acentuado durante a semana para nenhum cluster em específico. A distribuição de consumo ao longo da semana é algo uniforme, à exceção do domingo, no qual o consumo é um pouco menor do que nos restantes dias da semana.

Em resumo, o *cluster* 0 pode ser caracterizado pelos consumidores comuns, que não possuem grandes propriedades com piscina e jardins, e que levam a consumos mais altos, como o *cluster* 1 e 2. Para além disso, o consumo é mais significativo a partir do período da manhã, possivelmente quando os consumidores saem de casa para trabalhar. O *cluster* 1 pode apelidar-se de *madrugadores*, pelo facto de os seus consumos serem mais altos durante a noite e madrugada. E por fim, o *cluster* 2, que tem uma distribuição de consumo idêntica ao *cluster* 1, no entanto distribui-se mais ao anoitecer e durante a noite. É percetível que os *clusters* 1 e 2 pertencem aos consumidores que têm propriedades com uma grande área com jardins e piscina, e que acabam por ter consumos muito superiores à média. O aumento dos consumos nos períodos da noite também pode dever-se ao facto de os jardins serem regados durante o anoitecer/noite para poupança de água.

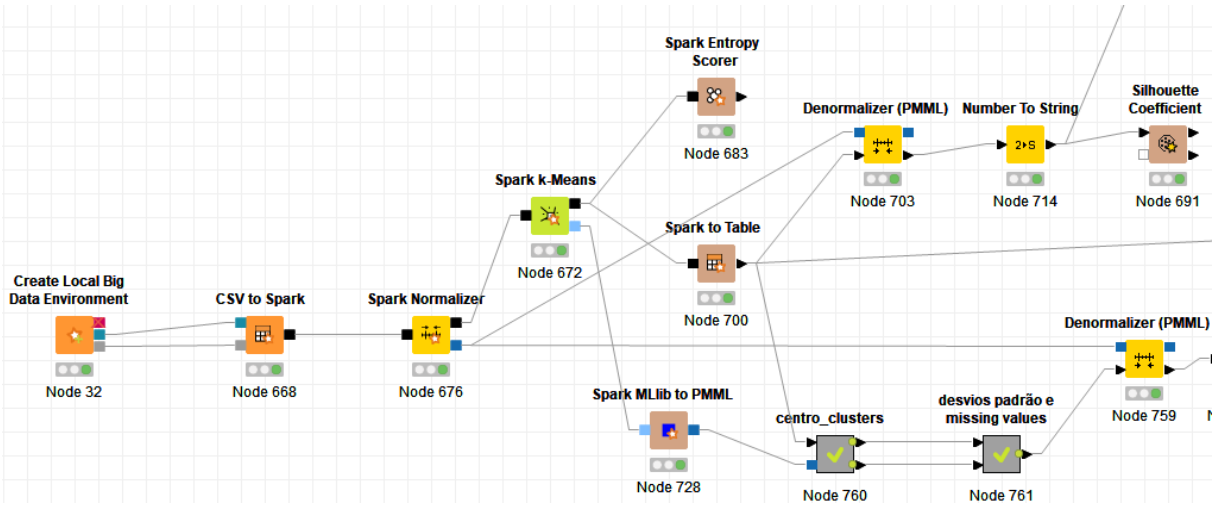

Figura 8: Excerto do *workflow* referente à aplicação dos *k-Means*.

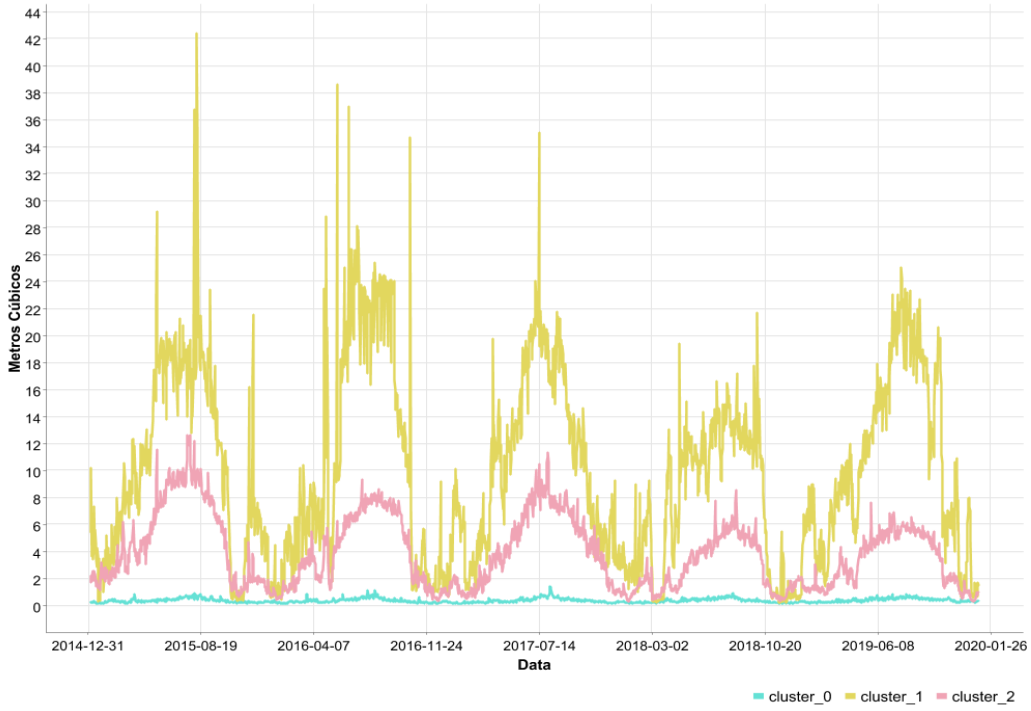

Figura 9: Comparação entre as séries temporais dos consumos médios diários dos três clusters

Na série temporal de cada agrupamento presente na figura 9, pode verificar-se os consumos médios diários em metros cúbicos. Existe notavelmente uma sazonalidade anual, com os consumos a aumentarem até aos meses de julho e agosto, voltando a descer acentuadamente nos meses seguintes. Sendo a Quinta do Lago uma zona turística, é natural que os consumos aumentem bastante durante o verão. Existem ainda, em alguns casos,

pequenos picos em dezembro, que se pode dever à época festiva do Natal e Ano Novo. A sazonalidade é abordada no subcapítulo seguinte.

4.2 Previsão da Série Temporal de Consumos de Água

Neste subcapítulo são aplicadas técnicas de *Machine Learning* para prever o consumo diário durante o ano de 2019.

Foi realizada uma partição dos dados em treino e teste (80% para treino, e o restante para teste), sendo utilizado para treino os consumos de 2015 a 2018 (valores retirados do topo até ao final de 2018), e para teste o ano de 2019. O *cluster* escolhido para a aplicação dos modelos foi o *cluster* 2, devido a ter um consumo intermédio entre os 3 *clusters* e também pelo facto de se caracterizar por consumidores que registam um consumo mais regular de água.

Efetuado o alinhamento da série para garantir que não existiam dias em falta e com os *missing values* tratados anteriormente, foi necessário verificar a estacionaridade e sazonalidade da série. Uma série é estacionária se as suas propriedades não dependerem dos tempos em que a série é observada, ou seja, se uma série não apresentar padrões como tendências e/ou sazonalidade ao longo do tempo. Dado que existe uma sazonalidade anual, verificada na figura 9, é necessário torná-la estacionária para a aplicação dos *Autoregressive Models* (ARIMA). É utilizada uma técnica de diferenciação para tornar a série estacionária*,* que consiste na diferenciação entre observações consecutivas (Hyndman & Athanasopoulos, 2021)*.*

De seguida foi utilizado um componente criado pelo *KNIME*, denominado *inspect seasonality,* onde é dado como *inputs* o número de desfasamentos temporais que se quer analisar (*lag*), o valor de corte da auto correlação (*seasonality cut off*), o número de movimentos da janela do ACF (*auto correlation function*) e, por fim, a coluna a inspecionar (*value column*), com o objetivo de verificar se a série continua a apresentar sazonalidade. A função de auto correlação verifica a correlação entre os valores *lag* da série temporal (Hyndman & Athanasopoulos, 2021), neste caso, em dias. O gráfico da ACF não apresentou valores de correlação superiores a 0.45, oscilando no intervalo entre -0.2 e 0.2. Isto significa que não existe uma sazonalidade após a série ser transformada em estacionária.

A análise visual dos gráficos das funções de auto correlação parcial (PACF) é a mais utilizada para descobrir qual poderá ser o melhor modelo ARIMA. No entanto o *KNIME* possui um componente chamado *Auto-ARIMA Learner* que calcula o melhor conjunto de *p* e *q*. O componente faz a seleção do melhor modelo com base no AIC (Critério de Informação de *Akaike*) ou no BIC (critério de informação *Bayesiano*).

Tendo como referência a revisão literária realizada no capítulo 2 deste estudo e os modelos disponibilizados no *KNIME*, foram considerados cinco algoritmos de ML: o MLP, o RF (*Random Forest*), o GBT (*Gradient Boosted Tree*), o XGBoost (*Extreme Gradient Boosting*), e o ARIMA.

Não foi necessário transformar a série em estacionária antes de aplicar certos modelos, como é o caso do MLP. No caso dos modelos autorregressivos, como o ARIMA, foi necessário a série ser estacionária e/ou retirar a sazonalidade, caso exista (Hamzaçebi, 2008).

Foi efetuada uma aplicação de todos os modelos, com a série estacionária, e sem qualquer transformação da série (série simples). Chegou-se à conclusão de que, comparativamente à série não estacionária, os modelos como MLP, RF, GBT e XGBoost não obtiveram um aumento significativo de performance ao transformar a série em estacionária. Como indicado no estudo do autor Hamzaçebi, C. (2008), nestes modelos não são realizadas transformações na série temporal.

O input dado foi exatamente o mesmo para todos os algoritmos, à exceção do ARIMA, que a série teve de ser transformada para se tornar estacionária.

Segundo a revisão literária realizada, como métricas de desempenho foram escolhidas o MAPE, o MAE, e o RMSE como pode-se averiguar na tabela 11. Apesar de a revisão de literatura referir a utilização do *R Square* como métrica de desempenho, esta não é incluída na avaliação de desempenho dos modelos, devido ao facto de não se adequar aos casos em que existe dependência entre as observações da amostra (dependência temporal), podendo não apresentar resultados fidedignos.

Tabela 11: Resultados de desempenhos dos modelos

Modelos	MAPE	MAE	RMSE
MLP	0.146	0.335	0.456
RF	0.148	0.335	0.461
GBT	0.17	0.406	0.565
XGBoost	0.146	0.346	0.499
ARIMA	0.143	0.364	0.507

Podemos verificar na tabela 11 que os modelos escolhidos obtiveram bons resultados, ressalvando como melhor resultado o *R Prop MLP*. Este foi o modelo que se adaptou melhor à serie de teste, o MAE e o RMSE mais baixo (0.335 e 0.456, respetivamente), no entanto todos os algoritmos tiveram desempenhos idênticos, com o pior resultado do CBT. O modelo GBT foi o que obteve pior desempenho. O ARIMA destacou-se por apresentar o MAPE mais

baixo dos resultados (0.143) e o RF teve um desempenho muito idêntico ao do MLP. O ambiente *Big Data* em *Spark* foi apenas utilizado no processamento do GBT e AR, pelo facto de estes modelos estarem disponíveis no ambiente *Spark*. O modelo elegido como o que teve melhores resultados é o MLP. É possível visualizar na figura 10 a comparação da série real (a amarelo), e da série prevista (a azul), verificando que os resultados foram satisfatórios para este *cluster* (Alves & Filipe, 2022).

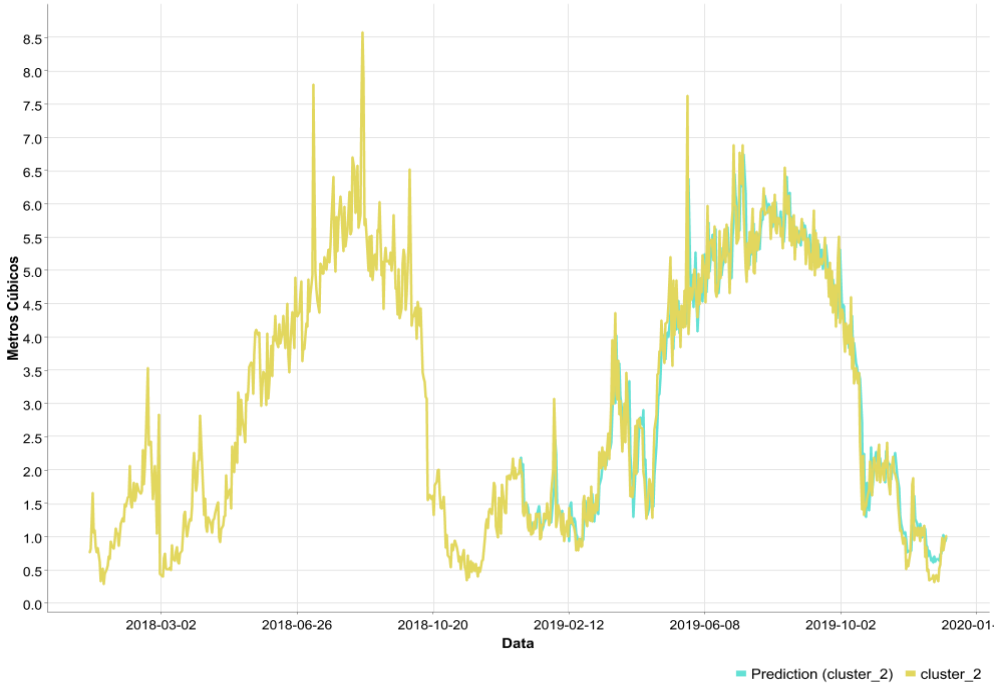

Figura 10: Comparação entre a série de consumos diários real e a prevista pelo MLP

4.3 *Dashboards* para a Deteção de Anomalias

Neste último subcapítulo dos resultados, são apresentadas duas opções de *dashboards* com o objetivo de contribuir para a deteção de anomalias: a primeira opção na ótica do cliente e a segunda na perspetiva da empresa, neste caso a Infraquinta. No segundo *dashboard*, o objetivo foi cruzar as possíveis anomalias dos consumos dos consumidores com possíveis desvios de padrão detetados pela diferença entre a série real e a série prevista.

Dado que não existem registos de anomalias na base de dados, não foi possível existir uma forma de reconhecer se houve realmente uma anomalia para confrontar os resultados, existindo apenas um exemplo de funcionamento do *IQ Alert,* explicado no subcapítulo 4.1 *Estudo de caso: A Infraquinta.* Esta abordagem teria de ser testada com a empresa no futuro.

O *dashboard* na figura 11 foi desenvolvido na perspetiva do cliente, através de um componente no *KNIME*. O componente teve como *input,* o ID do cliente e a data de início a partir do qual o cliente pretendia visualizar as possíveis anomalias. Aplicados os filtros, o

componente apresentou um *line plot* com os consumos diários para o período definido no início. Abaixo do gráfico da figura 11 é apresentada uma tabela que especifica a data da ocorrência da anomalia, indicando o estado como anomalia ou anomalia - leitura negativa. Para ser considerada anomalia tem de haver um consumo superior a 100 litros em todas as horas (durante 24 horas). No caso de anomalia com leitura negativa, basta existir um consumo negativo numa das horas do dia para despoletar um estado de anomalia. Em propriedades domésticas, uma das principais razões para leituras negativas é uma avaria no medidor de água.

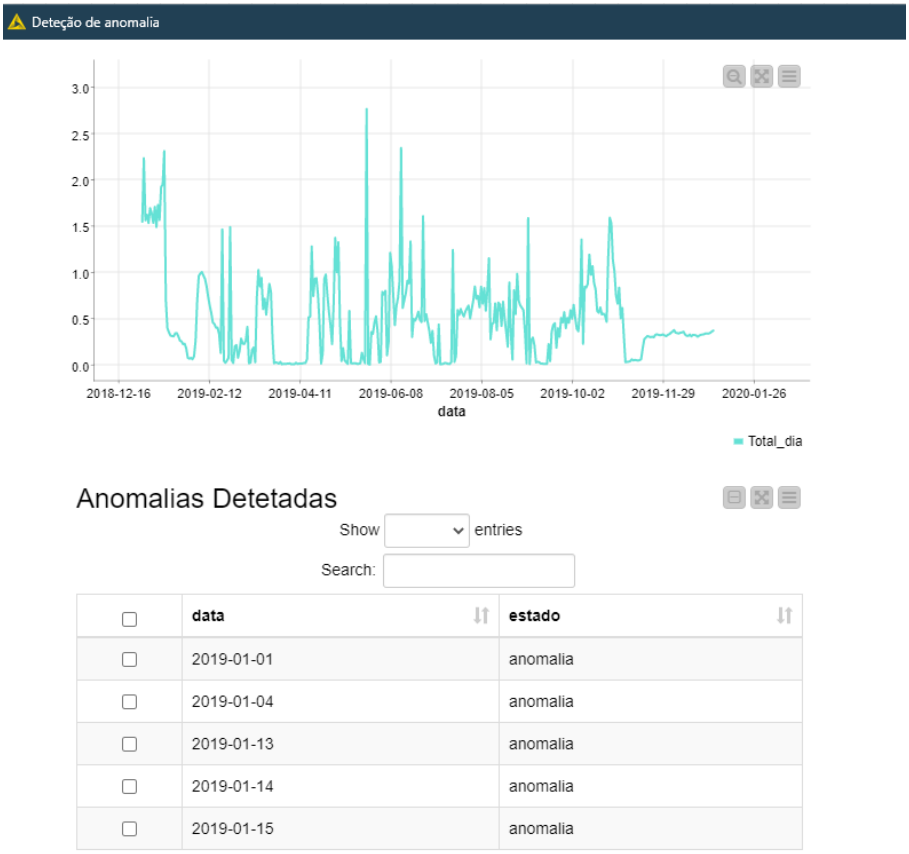

Figura 11: *Dashboard* para a deteção de anomalias na perspetiva do consumidor

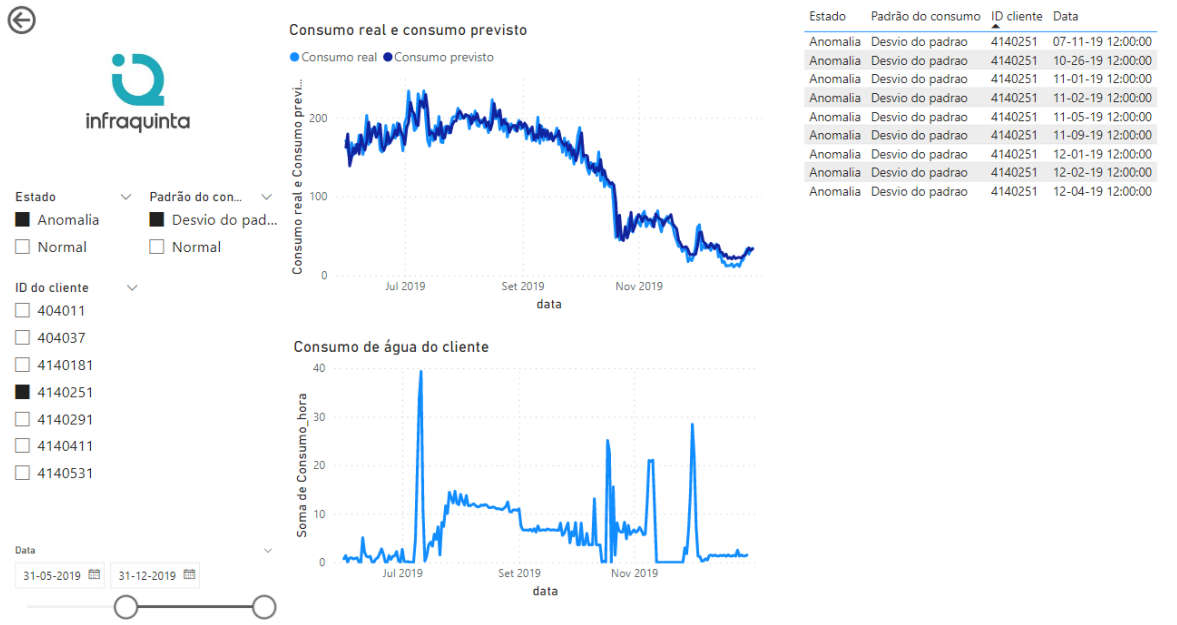

Figura 12: *Dashboard* para a deteção de anomalias na perspetiva da empresa

O *dashboard* na perspetiva da empresa foi realizado em *Power BI,* pelo facto do KNIME interagir com esta ferramenta e poder mostrar como construir o *dashboard* aproveitando a sua maior facilidade para tornar o *dashboard* mais interativo e dar a possibilidade de filtrar a informação por cliente, estados e desvios do padrão de consumo. Este *dashboard* pode ver visualizado na figura 12. Os dados foram exportados diretamente do *KNIME* para um ambiente de trabalho no *Power BI* através do nó *send to Power BI.* Depois dos dados importados no *Power BI,* não é possível transformá-los, o que limita um pouco a representação da informação no *KNIME*. Apesar disso, ambas as ferramentas se complementaram e o *KNIME* tem a potencialidade de poder integrar facilmente outras ferramentas de *analytics,* como é o caso do *Power BI.*

No lado esquerdo do *dashboard* é possível filtrar a informação do estado, em relação às leituras dos clientes, o padrão do consumo, referente à previsão da série, e um *slider* para escolher o intervalo de tempo desejado. O raciocínio para a deteção de anomalias foi o mesmo que foi utilizado no *dashboard* para o cliente. Já o desvio do padrão de consumo foi calculado para cada dia em que existe diferença entre a leitura média nesse dia e a leitura prevista no modelo. Se o resultado da diferença for superior a 300 litros, é considerado um desvio do padrão de consumo. Por sua vez, ao centro da *dashboard* estão representados um gráfico com previsão dos consumos do *cluster 2 por dia*, e um gráfico que demonstra o consumo do cliente escolhido. Do lado direito está representada uma tabela com a data da ocorrência, o estado, o padrão de consumo e o ID do cliente. A tabela apresentada permite identificar a possível origem do desvio do padrão através do ID do Consumidor.

Capítulo 5. Conclusões

As alterações climáticas são um tema que preocupa a comunidade mundial, principalmente em Portugal, devido ao aumento da temperatura, que origina secas um pouco por todo o país.

Sendo a água potável um bem essencial para a sobrevivência, é necessário as autarquias implementarem novas medidas e projetos para a redução dos desperdícios de água.

A aplicação de *Data Analytics* e técnicas de *Machine Learning* revelam-se uma mais-valia para aumentar o controlo das redes de distribuição, em caso de ocorrência de algum fenómeno de anomalia, bem como para tornar a medição dos consumos mais eficiente.

O estudo desenvolvido pretende apresentar uma forma possível para a gestão mais eficiente do parque de medidores de uma zona, com o objetivo de deteção de anomalias de grande capacidade na rede, nomeadamente fugas de água.

A RSL desenvolvida foi fundamental para adquirir conhecimentos essenciais para a realização da dissertação, como por exemplo, técnicas utilizadas, variáveis de entrada, métricas de avaliação e metodologias criadas para alcançar os resultados finais.

Os dados foram, inicialmente, filtrados pelo tipo de cliente e pela tarifa dos consumidores, de modo a assegurar que dizem respeito apenas a consumidores domésticos. Inicialmente não foram detetados *missing values*, mas com a realização de alinhamento temporal foi verificado que existiam datas e os respetivos consumos em falta na base de dados. Procedeu-se à eliminação de alguns consumidores que tinham muitos valores em falta e ao preenchimento dos restantes consumos através da técnica de interpolação linear.

A aplicação de técnicas de *Clustering* e de técnicas de ML à amostra sortiram bons resultados para o agrupamento escolhido a analisar. O *K-Means* gerou bons resultados, o que permitiu distinguir os consumidores em três grupos, caracterizados pela quantidade de água consumida.

Os resultados gerados pelos algoritmos de ML escolhidos na previsão do consumo de água no *cluster* 2, tiveram bom desempenho na sua maioria, com MLP e RF a terem resultados muito idênticos nas métricas de desempenho, sendo o GBT o modelo com pior resultado comparativamente aos restantes modelos.

Os objetivos para esta investigação foram concluídos, à exceção do quarto objetivo, que foi parcialmente cumprido: criação de *dashboard* para a deteção de anomalia. Apesar da criação de dois *dashboards*, não foi possível averiguar a veracidade dos resultados, pelo facto de não existirem registos de exemplos de ocorrências de anomalias no período analisado de 2015 a 2019. Dado este inconveniente, a metodologia é válida e pode ser testada em trabalhos futuros.

Em conclusão, este estudo permite responder à questão principal da investigação, como uma nova forma de deteção de anomalias nas redes domésticas mais eficiente e rápida e contribuir com novas ideias relativamente à temática abordada.

5.1 Limitações

No que diz respeito às limitações, a falta de exemplos de ocorrência de anomalias reais limitou um pouco a metodologia e o desenvolvimento do estudo, pois não permitiu realizar um modelo de classificação binária para deteção de anomalias, como explorado na revisão literária, e sendo essa a ideia inicial desta investigação. A ausência de exemplos reais de anomalias, afetou também a veracidade dos resultados finais, principalmente nas demonstrações dos *dashboards,* uma vez que não existe forma de se confirmar a fiabilidade dos resultados. Apenas uma implementação futura e conjunta com a empresa determinará se esta abordagem poderá ser uma mais-valia para a deteção mais eficaz da localização da anomalia.

Outro fator limitador no estudo foi na criação dos *dashboards*. O *KNIME* ainda não possui muitas das funcionalidades disponibilizadas em outros *softwares* de *Analytics,* como o caso do *Power BI,* o que também limitou um pouco a criação do *dashboard* na perspetiva do cliente e passou a ser realizado no *Power BI*, devido à necessidade da interatividade e da aplicação de filtros aos dados.

5.2 Contributos

Dado os objetivos estabelecidos no início desta dissertação, este estudo vem contribuir para colmatar alguns *Gap* no meio académico, existentes no tema de utilização de técnicas de ML na gestão de redes de distribuição de água. Vem também ajudar a que, futuramente, as empresas que gerem os recursos hídricos implementem novas soluções que permitam economizar mais água.

O RSL desenvolvido na área de deteção de eventos na rede pode vir a auxiliar no desenvolvimento de novas investigações nesta área, bem como perceber o que já foi desenvolvido neste âmbito, para além de dar as principais referências para o estudo do tema abordado.

A parte empírica desta investigação também revela alguns contributos importantes, nomeadamente o agrupamento dos consumidores por comportamento, o que permitiu identificar os padrões de consumo mais comuns, através de algumas técnicas de *Clustering*, como o *K-Means* e o *DBSCAN.* A utilização de algoritmos de ML revelaram-se interessantes na abordagem da previsão de consumos, com bons resultados nas métricas de desempenho,

levando em consideração que cada modelo possa ter resultados diferentes consoante a situação, ou seja, estes modelos tiveram um bom resultado, aplicados a este agrupamento, mas o mesmo pode não se verificar em outros agrupamentos de clientes, com diferentes padrões de consumo. Como comprovado pela literatura, as NN obtiveram os melhores resultados, mas tanto o XGBoost, como o RF obtiveram resultados razoáveis.
A criação dos *dashboards* é um bom contributo para as empresas deste setor poderem ter uma ideia de como criar *dashboards* nas perspetivas abordadas e para tornar a análise dos dados mais intuitiva e fácil.

5.3 Trabalhos Futuros

Para trabalhos futuros é recomendado o teste da metodologia criada, comparando os resultados gerados, com exemplos reais de anomalias que tenham surgido em datas coincidentes, e que tenham sido detetadas pela empresa. Recomenda-se também, na previsão de consumos futuros, a experimentação com outros modelos de ML, principalmente da família das *Neural Networks*, como por exemplo as LSTM. Por fim, sugere-se ainda a testagem em propriedades cujas tarifas correspondam a comércio e indústria, hotéis e aparthotéis.

Referências Bibliográficas

Alabi, M., Telukdarie, A., & Rensburg, N. (2019). INDUSTRY 4.0: INNOVATIVE SOLUTIONS FOR THE WATER INDUSTRY. *American Society for Engineering Management 2019 International Annual Conference*.

Alves, J., & Filipe, P. (2022). Big Data Analytics in Water Consumption. Em *VIII Workshop on Computacional Data Analysis and Numerical Methods* (pp. 67–68). Instituto Politécnico de Tomar. http://www.wcdanm.ipt.pt/2022

Candelieri, A. (2017). Clustering and support vector regression for water demand forecasting and anomaly detection. *Water (Switzerland)*, *9*(3). https://doi.org/10.3390/w9030224

Chastain-Howley, A., & Wallenstein, D. (2007). *Using an AMR System to Aid in the Evaluation of Water Losses: A Small DMA Case Study at East Bay Municipal Utility District, USA Water Prospecting and Resource Consulting*. 394.

Chen, J., & Boccelli, D. L. (2018). Forecasting Hourly Water Demands With Seasonal Autoregressive Models for Real-Time Application. *Water Resources Research*, *54*(2), 879–894. https://doi.org/10.1002/2017WR022007

Clinciu, M., & Clinciu, R. (2017, Novembro 4). Statistical analysis of the measurement errors in an installation of water meters. Study on the volume of the water loss in the installation. *4th International Conference on Computing and Solutions in Manufacturing Engineering (CoSME)*. https://doi.org/10.1051/

Coma-Puig, B., Carmona, J., Gavalda, R., Alcoverro, S., & Martin, V. (2016). Fraud detection in energy consumption: A supervised approach. *3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016*, 120–129. https://doi.org/10.1109/DSAA.2016.19

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Hamzaçebi, C. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences*, *178*(23), 4550–4559. https://doi.org/10.1016/j.ins.2008.07.024

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting Principles and Practice* (Otexts, Ed.; 3rd ed.). https://otexts.org/fpp2/.

Kainz, O., Dujava, M., Petija, R., Michalko, M., & Jakab, F. (2021). Measurement of Water Consumption based on Image Processing. *SAMI 2021 - IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 33–37. https://doi.org/10.1109/SAMI50585.2021.9378611

Karamaziotis, P. I., Raptis, A., Nikolopoulos, K., Litsiou, K., & Assimakopoulos, V. (2020). An empirical investigation of water consumption forecasting methods. *International Journal of Forecasting*, *36*(2), 588–606. https://doi.org/10.1016/j.ijforecast.2019.07.009

Kou Yinggang, Cui Gaoying, Fan Jie, Chen Xiao, & Li Wei. (2017). Machine Learning based Models for Fault Detectionin Automatic Meter Reading Systems. *2017 International Conference on Security, Pattern Analysis,and Cybernetics (SPAC)*, 684–689.

Lall, A. K., Khandelwal, A., Bose, R., Bawankar, N., Nilesh, N., Dwivedi, A., & Chaudhari, S. (2021). Making Analog Water Meter Smart using ML and IoT-based Low-Cost Retrofitting. *Proceedings - 2021 International Conference on Future Internet of Things and Cloud, FiCloud 2021*, 157–162. https://doi.org/10.1109/FiCloud49777.2021.00030

Lee, J., Choi, W., & Kim, J. (2021). A cost-effective CNN-LSTM-based solution for predicting faulty remote water meter reading devices in AMI systems. *Sensors*, *21*(18). https://doi.org/10.3390/s21186229

Li, E. Y., Wang, W.-H., Hsu, Y.-S., Li, E. Y. ;, Wang, W.-H., Hsu, Y., Li, E. Y., Wang, W. H. &, & Hsu, Y. S. (2017). *Adopting IoT Technology to Optimize Intelligent Water Management*. *12*, 38–46. http://aisel.aisnet.org/iceb2017http://aisel.aisnet.org/iceb2017/6

Li, H., Fang, D., Mahatma, S., & Hampapur, A. (2011). Usage analysis for smart meter management. *2011 8th International Conference and Expo on Emerging Technologies for a Smarter World, CEWIT 2011*. https://doi.org/10.1109/CEWIT.2011.6135871

McHugh, M. L. (2012). The Chi-square test of independence. *Biochemia Medica*, *23*(2), 143–149. https://doi.org/10.11613/BM.2013.018

Merta, J., & Fikejz, J. (2019). Utilization of Machine Learning to Detect Sudden Water Leakage for Smart Water Meter. *2019 29TH INTERNATIONAL CONFERENCE RADIOELEKTRONIKA (RADIOELEKTRONIKA)*, 340–344. https://www.webofscience.com/wos/woscc/full-record/WOS:000492026100062

Moeeni, H., Bonakdari, H., & Fatemi, S. E. (2017). Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction. *Journal of Hydrology*, *547*, 348–364. https://doi.org/10.1016/j.jhydrol.2017.02.012

Nyah, C. T., Kanyama, M. N., Nyirenda, C. N., & Temaneh-Nyah, C. (2017, novembro). *Anomaly Detection in Smart Water Metering Networks*. https://www.researchgate.net/publication/320961110

Ottosen, T. B., & Kumar, P. (2019). Outlier detection and gap filling methodologies for low-cost air quality measurements. *Environmental Science: Processes and Impacts*, *21*(4), 701–713. https://doi.org/10.1039/c8em00593a

Rahim, M. S., Nguyen, K. A., Stewart, R. A., Giurco, D., & Blumenstein, M. (2020). Machine learning and data analytic techniques in digitalwater metering: A review. *Water (Switzerland)*, *12*(1). https://doi.org/10.3390/w12010294

Ray, A., & Goswami, S. (2020). IoT and Cloud Computing based Smart Water Metering System. *International Conference on Power Electronics and IoT Applications in Renewable Energy and Its Control, PARC 2020*, 308–313. https://doi.org/10.1109/PARC49193.2020.236616

Roccetti, M., Delnevo, G., Casini, L., & Cappiello, G. (2019). Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0235-y

Roccetti, M., Delnevo, G., Casini, L., & Mirri, S. (2021). An alternative approach to dimension reduction for pareto distributed data: a case study. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00428-8

Roccetti, M., Delnevo, G., Casini, L., & Salomoni, P. (2020). A Cautionary Tale for Machine Learning Design: why we Still Need Human-Assisted Big Data Analysis. *Mobile Networks and Applications*, *25*(3), 1075–1083. https://doi.org/10.1007/s11036-020-01530-6

Roccetti, M., Delnevo, G., Casini, L., Zagni, N., & Cappiello, G. (2019). A paradox in ML design: Less data for a smarter water metering cognification experience. *ACM International Conference Proceeding Series*, 201–206. https://doi.org/10.1145/3342428.3342685

Schultz, W., Javey, S., & Sorokina, A. (2018). *Peer Reviewed Smart Water Meters and Data Analytics Decrease Wasted Water Due to Leaks*.

Shi, W., & Zeng, W. (2014). Application of k-means clustering to environmental risk zoning of the chemical industrial area. *Frontiers of Environmental Science and Engineering*, *8*(1), 117–127. https://doi.org/10.1007/s11783-013-0581-5

Wu, Q., Kang, L., Cao, B., Shang, Y., Zhang, M., Liu, X., Zhou, H., & Liu, W. (2020). Intelligent Diagnosis of Acquisition Equipment Failure Promote «multi-integration» Application. *Journal of Physics: Conference Series*, *1486*(7). https://doi.org/10.1088/1742-6596/1486/7/072044

Xu, Q., Wen, Q., & Sun, L. (2021). Two-stage framework for seasonal time series forecasting. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2021-June*, 3530–3534. https://doi.org/10.1109/ICASSP39728.2021.9414118

Zese, R., Bellodi, E., Luciani, C., & Alvisi, S. (2021). Neural Network Techniques for Detecting Intra-Domestic Water Leaks of Different Magnitude. *IEEE Access*, *9*, 126135–126147. https://doi.org/10.1109/ACCESS.2021.3111113