Escola Superior de Tecnologia de Tomar

Inteligência Organizacional e Modelos Preditivos Aplicados à Deteção de Perdas em Sistemas de Abastecimento de Água

Dissertação de Mestrado

Ana Célia Freitas Teixeira

Mestrado em Analítica e Inteligência Organizacional

Tomar, julho, 2025

Escola Superior de Tecnologia de Tomar

Inteligência Organizacional e Modelos Preditivos Aplicados à Deteção de Perdas em Sistemas de Abastecimento de Água

Dissertação de Mestrado

Ana Célia Freitas Teixeira

Orientada por:
Professora Doutora Sandra Maria Gonçalves Vilas Boas Jardim - Instituto Politécnico de Tomar

Júri
Professora Coordenadora Sandra Maria Gonçalves de Vilas Boas Jardim
Professor Adjunto Rolando Lúcio Germano Miragaia (Arguente)
Professor Adjunto Vasco Renato Marques Gestosa da Silva

Dissertação apresentada ao Instituto Politécnico de Tomar para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Analítica e Inteligência Organizacional

Sonhar é humano, medir é científico,
transformar dados em conhecimento é
propósito de um mestrado.

AGRADECIMENTOS

Concluir um estudo desta natureza representa não apenas o encerramento de uma etapa académica, mas também a consolidação de um percurso pessoal e profissional marcado por aprendizagens, desafios e superações.

Agradeço, em primeiro lugar, aos docentes e orientadores que me acompanharam ao longo deste trabalho, pelo seu apoio científico, orientação crítica e incentivo constante. As suas contribuições foram fundamentais para a concretização desta investigação.

À Tejo Ambiente, E.I.M., S.A., expresso o meu sincero reconhecimento pela cedência de dados operacionais e pela autorização de utilização da informação, condição essencial para o desenvolvimento prático e aplicabilidade do presente estudo. Agradeço também aos técnicos e colaboradores da entidade, pelo tempo disponibilizado e pelas valiosas contribuições durante as várias fases deste estudo.

Aos meus colegas de trabalho, pela colaboração, partilha de conhecimentos e apoio logístico prestado. A sua disponibilidade permitiu-me conciliar a vida profissional com os requisitos deste desafio académico.

Por fim, deixo uma palavra de gratidão à minha família e amigos, pelo suporte emocional, motivação e compreensão demonstrada ao longo de todo este processo. Sem o seu encorajamento, esta caminhada não teria sido possível.

A todos, o meu muito obrigado.

RESUMO

O presente estudo centrou-se numa análise de padrões de consumo de água, com especial enfoque na identificação de anomalias e ineficiências operacionais, como desperdícios, ruturas e lacunas de faturação. Reconhecendo a água como um recurso vital, limitado e estratégico, tornou-se imperativo aplicar metodologias de análise de dados que promovam uma gestão mais eficiente, sustentável e orientada por evidência.

Neste âmbito, foram utilizadas ferramentas modernas de *Business Intelligence* (BI), com especial destaque para a plataforma *Power* BI da *Microsoft*, no sentido de transformar dados operacionais dispersos em *insights* acionáveis, através da construção de *dashboards* interativos e modelos preditivos.

A investigação permitiu identificar padrões de consumo, comportamentos atípicos, bem como zonas críticas com risco acrescido de perdas, promovendo um novo paradigma de gestão baseada em dados. Foram também aplicadas técnicas estatísticas clássicas e algoritmos de *Machine Learning* (ML) supervisionado, com o objetivo de prever consumos futuros, sinalizar anomalias e apoiar a tomada de decisão estratégica e operacional.

Os resultados obtidos evidenciaram o potencial transformador da analítica de dados para melhorar a fiabilidade da informação, antecipar falhas, reduzir perdas e otimizar os recursos disponíveis. Em complemento, foram apresentadas sugestões para futuras investigações, sublinhando a importância da monitorização contínua e da integração de dados como pilares de uma gestão hídrica inteligente, resiliente e sustentável.

Palavras-chave: *Business Intelligence*; Análise de Dados; Gestão Hídrica; *Machine Learning*

ABSTRACT

This project focuses on an in-depth analysis of water consumption patterns, with particular emphasis on identifying anomalies and operational inefficiencies, such as waste, leaks, and billing gaps. Recognizing water as a vital and limited resource, it becomes essential to implement advanced data analysis methodologies that foster more efficient, evidence-based and sustainable management.

The study made use of modern BI tools, particularly Microsoft's Power BI platform, to transform fragmented operational data into actionable insights through the development of interactive dashboards and predictive models.

The investigation enabled the identification of consumption trends, atypical behaviors, and critical zones with increased risk of water loss. Classical statistical techniques and supervised ML algorithms were applied to forecast future consumption, detect anomalies, and support data-driven decision-making at both strategic and operational levels.

The results highlighted the transformative potential of data analytics to improve information reliability, anticipate failures, optimize resources, and increase the resilience of water supply systems. Recommendations for future research were also presented, reinforcing the relevance of continuous monitoring and data integration as pillars of smart and sustainable water management.

Keywords: Business Intelligence; Data Analytics; Water Management; Power BI; *Machine Learning*

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ÍNDICE DE EQUAÇÕES

ÍNDICE DE ABREVIATURAS

ANF - Água Não Faturada

APA – Agência Portuguesa do Ambiente

API – Application Programming Interface (Interface de Programação de Aplicações)

BI – Business Intelligence

CSV – Comma-Separated Values (Formato de ficheiros de dados separados por vírgulas)

DAX – Data Analysis Expressions (Linguagem de expressões utilizada no Power BI)

EG – Entidade Gestora

EPAL – Empresa Portuguesa das Águas Livres

ERSAR – Entidade Reguladora dos Serviços de Águas e Resíduos

IA – Inteligência Artificial

INE – Instituto Nacional de Estatística

IoT – Internet of Things (Internet das Coisas)

KPI – Key Performance Indicator (Indicador-Chave de Desempenho)

ML – Machine Learning (Aprendizagem Automática)

OT – Ordem de Trabalho

PCA – Principal Component Analysis (Análise de Componentes Principais)

POSEUR – Programa Operacional Sustentabilidade e Eficiência no Uso de Recursos

RASARP – Relatório Anual dos Serviços de Águas e Resíduos em Portugal

RSU – Resíduos Sólidos Urbanos

SAA – Sistema de Abastecimento de Água

SAD – Sistema de Apoio à Decisão

SMOTE – Synthetic Minority Oversampling Technique (Técnica de Sobreamostragem da Classe Minoritária)

SQL – Structured Query Language (Linguagem Estruturada de Consulta)

SST – Segurança e Saúde no Trabalho

SVM – Support Vector Machines (Máquinas de Vetores de Suporte)

TA – Tejo Ambiente – Empresa Intermunicipal de Ambiente do Médio Tejo, S.A.

ZMC – Zona de Monitorização e Controlo

1. INTRODUÇÃO

Num cenário global fortemente marcado pela transformação digital, pelas exigências crescentes de sustentabilidade e pela necessidade de gestão eficiente dos recursos hídricos, torna-se imperativo que as organizações públicas adotem abordagens inovadoras, baseadas em evidência, para garantir a resiliência e eficácia dos seus sistemas operacionais.

A água, enquanto recurso vital, enfrenta atualmente um conjunto de desafios complexos: desde o aumento das perdas nos sistemas de distribuição, passando pela pressão demográfica e urbanística, até à crescente volatilidade climática. Neste contexto, a análise integrada de dados surge como uma ferramenta estratégica para otimizar o desempenho das entidades gestoras, identificar padrões de consumo, antecipar falhas, mitigar ruturas e reduzir desperdícios.

Este estudo insere-se nesta visão, propondo o desenvolvimento de uma plataforma de BI orientada à realidade da Tejo Ambiente – Empresa Intermunicipal de Ambiente do Médio Tejo, S.A. (Tejo Ambiente), uma empresa intermunicipal com responsabilidade territorial, técnica e operacional. Com recurso à ferramenta *Power* BI da *Microsoft*, pretende-se transformar dados operacionais dispersos, provenientes de diversas plataformas tecnológicas como *AquaMatrix*, telegestão, telemetria e outras, em informação visualmente acessível, acionável e preditiva.

Neste capítulo, serão apresentados os conceitos teóricos fundamentais que sustentam esta investigação, bem como os pressupostos e decisões que moldaram a sua estrutura. Iniciaremos com um enquadramento temático sobre a importância da análise de dados na gestão hídrica, seguido da explicitação dos objetivos impulsionadores deste estudo, tanto do ponto de vista técnico como estratégico. Será ainda detalhada a estrutura do relatório, permitindo uma navegação clara e sistematizada ao longo dos diferentes capítulos.

A proposta aqui desenvolvida visa demonstrar não só o potencial das ferramentas analíticas para apoiar a gestão das infraestruturas públicas, mas também a aplicabilidade de modelos de *machine learning* (ML) e de análise estatística avançada na monitorização e tomada de decisão no setor da água. Com base nos dados de consumo reais, obtidos num concelho piloto, foram construídos dashboards e modelos preditivos que permitem à Tejo Ambiente reforçar a sua capacidade de planeamento, de controlo e de resposta em tempo útil.

1.1 ENQUADRAMENTO NACIONAL DA GESTÃO DA ÁGUA E DESEMPENHO DAS ENTIDADES GESTORAS

O setor da água em Portugal enfrenta desafios estruturais significativos, nomeadamente ao nível da eficiência na distribuição e da sustentabilidade económico-financeira das entidades gestoras. Segundo o Relatório Anual dos Serviços de Águas e Resíduos em Portugal (RASARP 2024), publicado pela Entidade Reguladora dos Serviços de Águas e Resíduos (ERSAR), registou-se um volume de perdas de água superior a

184 milhões de metros cúbicos por ano, o que representa aproximadamente 28,3% da água disponibilizada no sistema público de abastecimento (**Error! Reference source not found.**).

O impacto económico associado a estas perdas é estimado em cerca de 152 milhões de euros anuais, traduzindo-se num desperdício financeiro significativo e insustentável. Destaca-se ainda que as entidades de gestão pública direta apresentam níveis de perdas consideravelmente mais elevados do que as concessões privadas, que registam uma média de apenas 15,2% de perdas. Este contraste evidencia não só assimetrias na eficiência operacional, mas também a necessidade urgente de modernização e reforço das competências analíticas por parte das entidades públicas.

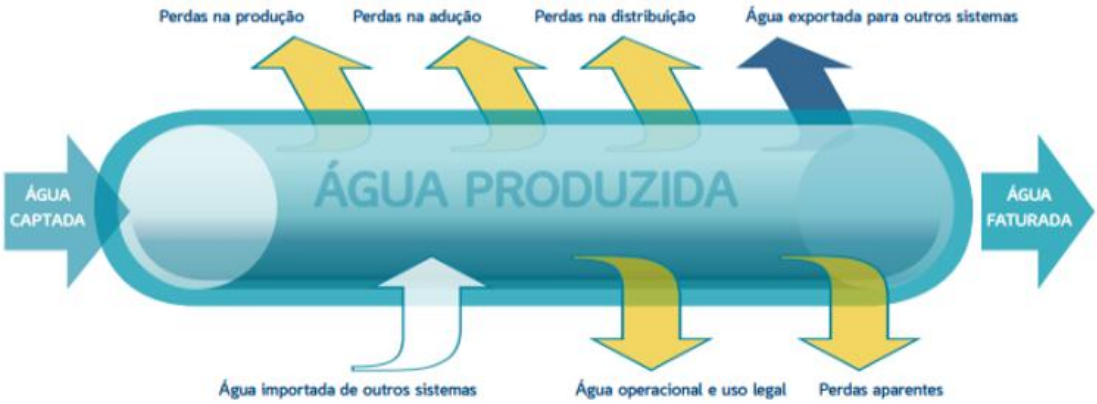


Figura 1: Fluxograma das principais influências das perdas num sistema de abastecimento.

Estes indicadores reforçam a pertinência do presente estudo, sublinhando a importância da análise de dados, da transformação digital e da implementação de ferramentas de BI como meios fundamentais para combater a estagnação do setor. A capacidade de integrar e analisar dados operacionais, de forma a detetar anomalias, antecipar ruturas e otimizar recursos, surge como condição necessária para a melhoria do desempenho das entidades gestoras e para a concretização de uma gestão da água mais eficiente, resiliente e sustentável.

1.2 Objetivos impulsionadores do Processo de Análise de dados

Os objetivos impulsionadores do processo de análise de dados incluem a transformação sistemática e estratégica de dados brutos em informações valiosas, convertendo conjuntos de dados não estruturados, dispersos ou sem tratamento prévio em conhecimento organizacional aplicável. Este processo visa a criação de *insights* que possam ser utilizados de maneira prática e eficaz nos diferentes níveis da gestão.

Além disso, envolve a identificação de padrões, correlações e tendências significativas, utilizando técnicas estatísticas avançadas e modelos de inteligência artificial e aprendizagem automática, com o propósito de descobrir relações ocultas entre variáveis e antecipar comportamentos futuros.

Outro objetivo essencial é o suporte robusto à tomada de decisões informadas, fornecendo informações precisas, contextualizadas e atualizadas para auxiliar a liderança organizacional em decisões estratégicas e

operacionais. Por exemplo, no setor da gestão de água, *dashboards* baseados em *Power* BI podem alertar automaticamente os gestores para aumentos anómalos no consumo em determinada zona, permitindo ações preventivas imediatas.

A melhoria contínua da qualidade dos dados é igualmente fundamental, garantindo que a base analítica esteja limpa, coerente, atualizada e livre de erros. Isto envolve a remoção de ruído, duplicações e inconsistências, bem como a definição de critérios de qualidade que assegurem a fiabilidade da análise. Segundo Batini et al. (2015), uma má qualidade dos dados pode comprometer até 30% da eficácia dos processos de decisão numa organização.

A comunicação clara e eficaz dos resultados analíticos também é um pilar essencial. Os dados devem ser apresentados de forma compreensível e acessível, através de visualizações intuitivas, *dashboards* interativos e relatórios dinâmicos, adaptados aos diferentes públicos-alvo. Ferramentas como o *Power* BI oferecem recursos poderosos neste domínio, permitindo, por exemplo, a criação de relatórios automáticos segmentados por departamento ou área funcional.

Por fim, a análise de dados deve contribuir diretamente para a otimização dos processos e do desempenho organizacional, identificando oportunidades de melhoria, eliminando ineficiências e apoiando a inovação contínua. Na prática, isto pode traduzir-se em ajustes operacionais baseados em indicadores de desempenho, como o tempo médio de resposta a ocorrências ou o custo por litro de água distribuída.

Este estudo integra ainda a aplicação de um modelo de classificação supervisionado, como os algoritmos de regressão logística, *random forest* ou *k-nearest neighbors*, com o objetivo de prever padrões de consumo e antecipar roturas com base no histórico de dados. Modelos de classificação são técnicas de ML que permitem atribuir categorias com base em características previamente observadas.

Estes objetivos norteiam todas as fases do ciclo analítico, desde a recolha e preparação dos dados até à interpretação, comunicação e operacionalização dos resultados, assegurando que a análise de dados seja uma ferramenta efetiva de gestão estratégica e de criação de valor organizacional.

A Tejo Ambiente, enquanto entidade gestora intermunicipal de abastecimento de água, gere milhares de registos de consumo provenientes de diferentes sistemas, como o *AquaMatrix*, telegestão e telemetria.

Através da integração destes dados numa plataforma *Power* BI, é possível construir *dashboards* que monitorizam o consumo por zona, cliente, hora e tipo de utilizador. Com base nesses dados, foi desenvolvido um modelo de previsão de consumos, utilizando algoritmos de regressão e séries temporais, que permite estimar o consumo esperado por zona geográfica em intervalos diários, sinalizar automaticamente desvios superiores a 25% face ao padrão histórico, que podem indicar roturas, fugas ou erros de leitura e gerar relatórios automáticos para os gestores operacionais, destacando as zonas de monitorização e controlo (ZMC) com maior risco de perda.

Este processo melhora significativamente a capacidade de tomada de decisão baseada em dados, reduz o tempo de resposta a ocorrências e apoia a otimização da gestão hídrica, contribuindo para uma maior eficiência operacional e sustentabilidade.

1.3 Estrutura do Relatório

Este relatório organiza-se em sete capítulos principais, estruturados de forma a acompanhar logicamente a evolução do estudo, desde o seu enquadramento teórico até à implementação prática e à análise dos resultados, oferecendo uma visão integrada e aplicada da utilização de BI e técnicas de análise preditiva no contexto da gestão hídrica da Tejo Ambiente. A estrutura foi desenhada para refletir não apenas a dimensão técnica e metodológica da investigação, mas também a sua aplicabilidade real, sustentando-se nos dados e desafios específicos da entidade gestora em estudo.

O capítulo 1 apresenta a introdução geral ao tema, contextualizando a realidade da gestão da água em Portugal e os objetivos impulsionadores do estudo. O capítulo 2 é dedicado aos trabalhos relacionados, oferecendo uma revisão crítica e atualizada da literatura, com especial foco na evolução das ferramentas de visualização de dados, na emergência dos *dashboards* como instrumentos de apoio à decisão, na aplicação de BI ao setor público e na utilização de modelos preditivos no setor hídrico.

No capítulo 3 é feita a apresentação da entidade em estudo, detalhando o planeamento e todas as decisões técnicas subjacentes ao estudo. Esta secção descreve as fontes de dados utilizadas, as dificuldades encontradas na sua integração e pré-processamento, a lógica adotada na seleção de variáveis, a aplicação de técnicas estatísticas e de ML, a validação de anomalias e a implementação de *dashboards* interativos como solução de visualização orientada à decisão. Este capítulo é particularmente importante para compreender a realidade concreta da Tejo Ambiente, incluindo os seus sistemas operacionais, as limitações detetadas e as oportunidades de melhoria.

O capítulo 4 expõe a metodologia de análise utilizada, incluindo a estratégia de controlo de perdas com base nas ZMC. Já o capítulo 5 aborda a fase de implementação, descrevendo o processo de recolha, tratamento e modelação dos dados, bem como a construção dos dashboards em *Power* BI e a preparação dos dados para os modelos preditivos.

O capítulo 6 reúne e analisa os principais resultados do estudo, com subdivisões dedicadas à área comercial (com especial atenção a padrões de faturação e identificação de clientes não faturados), à monitorização de perdas de água (através da agua não faturada (ANF), análises comparativas por zonas e proposta técnica para melhoria do controlo), à análise da antiguidade dos contadores (incluindo o seu impacto nas leituras e recomendações estratégicas para substituição), e finalmente à avaliação do impacto da aplicação de modelos de ML supervisionado na deteção de anomalias. Neste último ponto, destaca-se a aplicação da técnica Synthetic Minority Oversampling Technique (SMOTE) para resolver o desequilíbrio das classes e a obtenção de

melhorias significativas no desempenho dos modelos preditivos, com especial destaque para o *random forest*, validado por métricas como o F1 *Score* e a matriz de confusão.

O capítulo 7 apresenta as conclusões do trabalho, sintetizando os principais contributos e aprendizagens resultantes do estudo. São refletidas as limitações encontradas durante o processo, como a dispersão dos dados, a falta de integração em tempo real e a ausência de *feedback* contínuo dos utilizadores, e são discutidas recomendações para trabalhos futuros, com destaque para a integração da *Internet of Things* (IoT), análise de variáveis complementares como qualidade da água e desenvolvimento de sistemas de alerta inteligentes. Esta conclusão reforça o papel transformador do BI e da ciência de dados na modernização da gestão pública e na promoção de práticas mais sustentáveis, resilientes e baseadas em evidência no setor hídrico.

2. TRABALHOS RELACIONADOS

2.1 EVOLUÇÃO DOS *DASHBOARDS* E VISUALIZAÇÃO DE DADOS

Os *dashboards*, também conhecidos como painéis de controlo, são ferramentas visuais que apresentam dados e informações de forma sintetizada e intuitiva, permitindo uma análise rápida e eficaz por parte dos utilizadores. A história dos *dashboards* está intrinsecamente ligada ao desenvolvimento dos sistemas de informação e à necessidade crescente de visualização analítica, em resposta à massificação dos dados.

A origem conceptual da visualização de dados remonta ao século XVIII, com William Playfair, economista e inventor escocês que criou os primeiros gráficos estatísticos modernos, como o gráfico de linha, o gráfico de barras e o gráfico de setores (Mayer-Schönberger & Cukier, 2013). Embora Playfair não tenha usado o termo "*dashboard*", as suas técnicas de representação visual de dados influenciaram profundamente os fundamentos da visualização moderna (Figura 2).

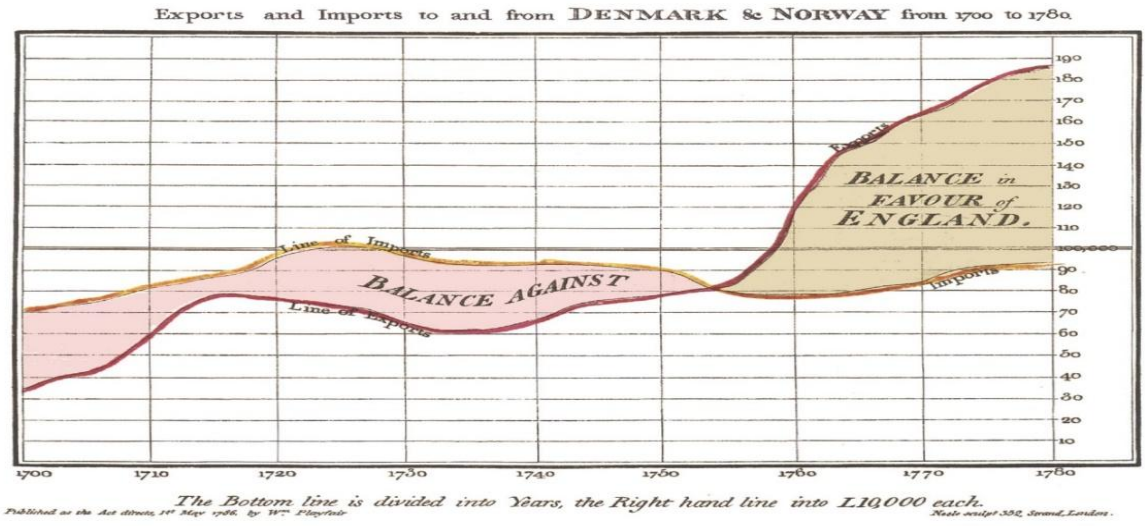


Figura 2: Gráfico de série temporal da balança comercial de *Playfair*, publicado no seu Atlas comercial e político, 1786.

Na década de 1970, surgiram os primeiros sistemas de apoio à decisão (SAD), concebidos para ajudar os gestores a tomar decisões informadas, com base em relatórios e indicadores agregados (Davenport & Harris, 2017). A evolução tecnológica nas décadas seguintes permitiu transitar de ferramentas estáticas para visualizações dinâmicas e interativas.

2.2 *DASHBOARDS* MODERNOS E *BUSINESS INTELLIGENCE*

A década de 1990 trouxe o desenvolvimento de *softwares* mais avançados, como o *Microsoft Excel*, que possibilitaram a manipulação de dados e a construção de relatórios visuais. Na viragem do século XXI, surgiram ferramentas dedicadas ao BI, como *Tableau, QlikView* e, mais recentemente, o *Power BI* da *Microsoft*, lançado em 2015. Esta última veio democratizar o acesso a *dashboards* avançados, com integração de múltiplas fontes de dados, funcionalidades analíticas e elevada personalização (Chen et al., 2019; Gandomi & Haider, 2022).

Os *dashboards* modernos destacam-se pelas seguintes características: Interatividade e visualização em tempo real, integração de dados de fontes heterogéneas, visualização georreferenciada e personalização por utilizador e nível de decisão.

Estudos notáveis incluem o *Real-Time COVID-19 Tracker*, desenvolvido pela *Microsoft* em colaboração com a Universidade de *Washington*, que monitorizou a pandemia em tempo real ou o painel de controlo da *Mecklenburg County* (EUA), que permite alocar recursos públicos com base no uso de serviços municipais (Figura 3).

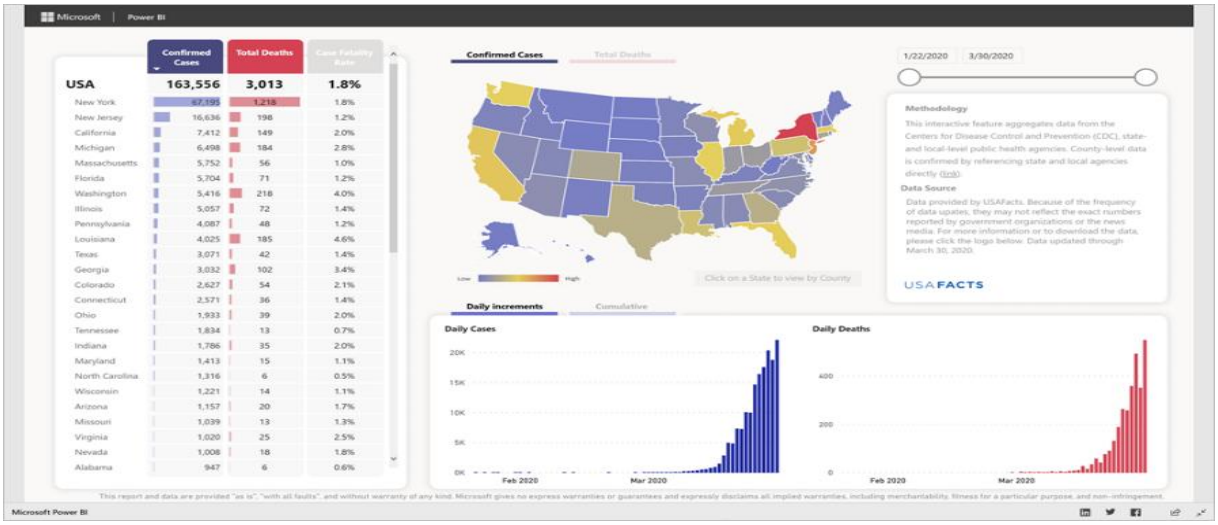


Figura 3: Amostra de Rastreamento de *Covid 19* para Governos Estaduais e Locais Dos EUA.

2.3 Casos de Estudo Internacionais e Nacionais no Setor Hídrico

No setor da água, os *dashboards* passaram de instrumentos meramente reativos (baseados em histórico) para plataformas analíticas com capacidade preditiva, respondendo à necessidade de maior eficiência, sustentabilidade e combate à ANF.

A *Thames Water*, no Reino Unido, recorre a *dashboards* integrados com algoritmos de ML para prever ruturas em tempo real, utilizando sensores distribuídos pela rede que monitorizam pressão e caudal. Sempre que são detetadas anomalias nos padrões registados, o sistema gera alertas automáticos que permitem intervenções preventivas antes da ocorrência de falhas críticas. Na Austrália, a *Sydney Water* implementou painéis analíticos avançados para prever padrões sazonais de consumo e identificar fugas invisíveis com base em desvios estatísticos, o que resultou numa redução superior a 25% das perdas aparentes nos últimos anos. Em Portugal, entidades como a EPAL e as Águas do Porto têm vindo a investir em *dashboards* operacionais ligados a plataformas de ZMC, integrando dados de pressão, caudal e anomalias, o que tem tido impacto direto na deteção de perdas, na priorização de intervenções e na otimização do planeamento da rede de abastecimento.

Estes exemplos reforçam que o setor hídrico está na linha da frente da transformação digital, adotando *dashboards* não apenas como ferramentas de visualização, mas como plataformas inteligentes de suporte à decisão.

2.4 *Dashboards* Preditivos no Contexto da Tejo Ambiente

O presente trabalho insere-se neste panorama internacional, aplicando *dashboards* preditivos à realidade da Tejo Ambiente, entidade intermunicipal responsável pela gestão do ciclo urbano da água em seis municípios do Médio Tejo.

Com base na integração de dados provenientes de plataformas operacionais como o *AquaMatrix,* dados técnicos recolhidos por sistemas de telegestão, sensores e contadores, bem como informação estatística e sociodemográfica dos Censos 2021, foram desenvolvidos dashboards interativos em *Power BI* que permitem identificar zonas críticas com contadores obsoletos ou ausência de leituras, detetar anomalias de consumo com base em padrões históricos, cruzar variáveis como localização, antiguidade do contador, tipo de cliente e número de anomalias, suportar modelos de ML supervisionados, nomeadamente *Random Forest* e Regressão Logística, com capacidade para prever ruturas e perdas aparentes, e ainda priorizar intervenções técnicas com base em critérios objetivos e quantificáveis, reforçando a tomada de decisão baseada em dados.

Estes *dashboards* funcionam como centros de comando digital, permitindo à Tejo Ambiente monitorizar e agir com maior rapidez e eficácia, promovendo uma gestão baseada em evidência, previsibilidade e eficiência.

2.5 Considerações Finais

A investigação realizada evidencia o potencial dos *dashboards* modernos, particularmente os de natureza preditiva, como pilares da transformação digital no setor da água. A sua capacidade de integração de dados, combinada com algoritmos de aprendizagem automática, permite antecipar falhas, otimizar recursos e aumentar a transparência.

O trabalho aqui desenvolvido demonstra que é possível replicar esta abordagem em outras entidades gestoras com diferentes escalas e realidades operacionais. Além do setor hídrico, estas metodologias podem ser aplicadas com sucesso à gestão de energia, resíduos, transportes urbanos, redes de emergência ou logística pública, reforçando o seu caráter transversal.

A conjugação entre BI, ciência de dados e gestão operacional representa um vetor essencial para enfrentar os desafios da sustentabilidade, da escassez de recursos e da crescente exigência dos cidadãos por serviços públicos mais eficientes e inteligentes.

3. ENTIDADE EM ESTUDO

A Tejo Ambiente, fundada em 2019, constitui uma entidade pública de gestão integrada, cuja atuação abrange os municípios de Ferreira do Zêzere, Mação, Sardoal, Tomar, Ourém e Vila Nova da Barquinha. Este modelo de gestão intermunicipal surgiu como resposta à necessidade de coordenar e otimizar os serviços públicos essenciais de abastecimento de água, saneamento de águas residuais urbanas e gestão de resíduos sólidos urbanos, através de uma estrutura organizacional comum e partilhada.

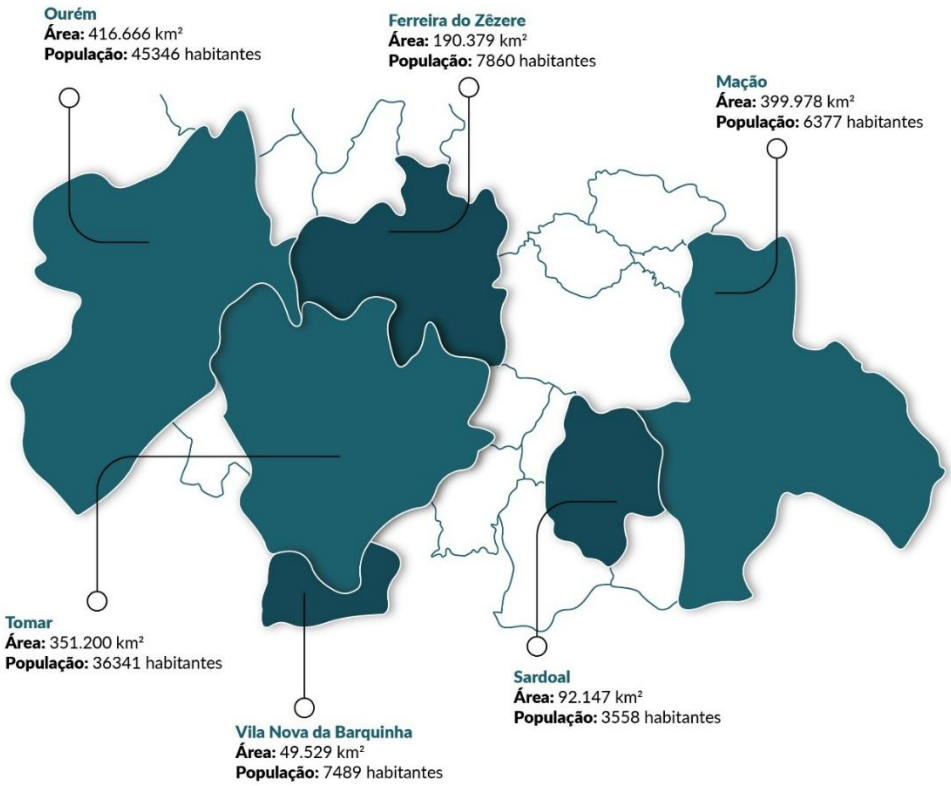


Figura 4: A Tejo Ambiente é uma empresa detida a 100% por 6 Municípios do Médio Tejo: Ferreira do Zêzere, Mação, Ourém, Sardoal, Tomar e Vila Nova da Barquinha.

A criação da Tejo Ambiente resultou da assinatura de um Contrato de Gestão Delegada em outubro de 2019, iniciando a sua atividade operacional em janeiro de 2020 nos municípios de Ourém e Tomar, sendo posteriormente alargada aos restantes quatro concelhos em junho do mesmo ano. A empresa opera numa área geográfica de aproximadamente 1.500 km², prestando serviços públicos essenciais a cerca de 106.000 habitantes. A sua constituição foi impulsionada pelos critérios definidos pelo Governo para acesso a fundos comunitários no âmbito do Programa Operacional Sustentabilidade e Eficiência no Uso de Recursos (POSEUR), sustentada por um estudo que evidenciou a viabilidade económica e financeira da gestão intermunicipal.

A missão da Tejo Ambiente assenta na prestação de serviços ambientais sustentáveis, eficientes e centrados no cidadão, promovendo a proteção dos recursos naturais e a melhoria da qualidade de vida das populações

servidas. No entanto, enquanto organização jovem, a empresa enfrenta diversos desafios estruturais e tecnológicos, derivados da integração de sistemas anteriormente autónomos e da diversidade de procedimentos operacionais herdados de cada município.

A empresa rege-se pela Lei nº 50/2012, que estabelece o regime jurídico da atividade empresarial local, e pelo Código das Sociedades Comerciais. O seu regime financeiro obedece às normas das autarquias locais e entidades intermunicipais, estando sujeito a limites legais de endividamento municipal. A gestão da Tejo Ambiente é orientada pelos princípios de racionalidade, eficiência, eficácia e sustentabilidade, visando a aplicação de um tarifário justo, o aumento dos ganhos de eficiência operacional, a modernização de processos e a excelência no serviço público.

A estrutura organizacional da Tejo Ambiente é funcional e hierarquizada, concebida para assegurar a gestão integrada dos serviços de abastecimento de água, saneamento e resíduos. O capital social da empresa é de 600.000 euros, detido proporcionalmente pelos seis municípios. Os seus órgãos sociais incluem a Assembleia Geral, o Conselho de Administração e o Fiscal Único. A Direção Geral é o órgão executivo máximo, com apoio direto das áreas Jurídica, Sistema de Gestão Integrada, Segurança e Saúde no Trabalho (SST) e Resíduos Sólidos Urbanos (RSU).

A estrutura técnica e administrativa inclui a Direção Comercial, responsável pelo relacionamento com os clientes, comunicação institucional e gestão comercial, subdividida nos polos Este e Oeste; a Direção Administrativa e Financeira, encarregada da contabilidade, dos recursos humanos, do aprovisionamento e da auditoria; e a Direção de Operações, que integra as áreas de engenharia, manutenção, deteção de fugas, gestão de frota, sistemas de informação e qualidade da água.

Um dos principais constrangimentos enfrentados pela organização prende-se com a fragmentação e dispersão dos dados operacionais em múltiplas plataformas, como o *AquaMatrix*, *Filedoc* e Telegestão, o que dificulta a construção de uma visão integrada da operação. Esta realidade limita a eficiência da gestão, dificulta a análise histórica e preditiva e compromete a rapidez na resposta operacional.

Reconhecendo esta fragilidade, a Tejo Ambiente tem vindo a investir progressivamente na digitalização dos processos e na integração de sistemas, destacando-se a instalação de contadores inteligentes, a implementação de sistemas de telemetria e o desenvolvimento de *dashboards* com recurso ao *Power BI*. Estas iniciativas visam mitigar perdas de água, identificar roturas com maior celeridade, priorizar intervenções e apoiar a tomada de decisão baseada em dados.

Neste enquadramento, a escolha da Tejo Ambiente como objeto de estudo justifica-se não só pela sua natureza intermunicipal e dimensão operacional, mas também pela sua aposta na transformação digital e na inteligência organizacional como motores de inovação na gestão pública ambiental. O presente estudo, ao propor a construção de uma plataforma analítica suportada em *Power BI*, pretende ser uma resposta prática e escalável às necessidades reais da organização, contribuindo de forma concreta para a literacia digital, a modernização da administração pública e a sustentabilidade na gestão de recursos hídricos.

3.1 PLANEAMENTO DO ESTUDO

Este capítulo descreve, de forma estruturada, a abordagem metodológica adotada ao longo do desenvolvimento do estudo, tendo como eixo central a aplicação de técnicas de análise de dados à gestão inteligente do consumo de água.

O trabalho foi conduzido no contexto da Tejo Ambiente, entidade intermunicipal responsável pelo ciclo urbano da água, com o objetivo de promover uma gestão mais eficiente, sustentada e orientada por evidência.

O planeamento assentou na construção de um sistema de apoio à decisão baseado em dados, que permitisse analisar padrões de consumo, identificar anomalias e apoiar intervenções operacionais e estratégicas, contribuindo para a redução da ANF e a otimização dos investimentos em infraestrutura.

Entre os principais objetivos delineados para este estudo, destacam-se a identificação de padrões de consumo por freguesia, zona de monitorização e tipo de utilizador (doméstico e não doméstico), a deteção automatizada de anomalias como consumos atípicos, ruturas, ligações indevidas ou falhas de medição, a priorização técnica de intervenções com base em critérios objetivos como a idade dos contadores, frequência de leitura ou histórico de avarias, e o apoio à tomada de decisão estratégica orientada à substituição de equipamentos obsoletos, renovação de condutas e digitalização da rede de leitura.

A metodologia aplicada envolveu a recolha e integração de dados operacionais provenientes de diversas fontes, incluindo o sistema de gestão comercial *AquaMatrix*, registos de telemetria, plataformas de telegestão e dados sociodemográficos dos Censos 2021, o pré-processamento e limpeza dos dados com normalização de formatos, tratamento de valores em falta, identificação de inconsistências e agregação por zonas geográficas.

A construção de *dashboards* interativos em *Power* BI que permitem a visualização dos indicadores de consumo, antiguidade dos contadores, percentagens de leitura e distribuição territorial de anomalias, o desenvolvimento e treino de modelos de ML supervisionado com aplicação de algoritmos como regressão logística, *Support Vector Machines* (Máquinas de Vetores de Suporte – SVM) e *Random Forest* para prever zonas com maior risco de perda aparente ou ineficiência operacional, bem como a avaliação desses modelos com base em métricas estatísticas como *accuracy*, precisão, *recall*, F1 *Score* e análise da matriz de confusão.

Adicionalmente, aplicaram-se técnicas de balanceamento de classes, como o SMOTE, para mitigar o desequilíbrio dos dados e melhorar a sensibilidade dos modelos na deteção de anomalias raras.

Este planeamento metodológico serviu de base para o desenvolvimento de uma abordagem orientada à inteligência organizacional, demonstrando como o tratamento analítico de grandes volumes de dados pode gerar valor real para os serviços públicos ambientais, contribuindo para a modernização da administração local e reforçando a importância da interoperabilidade de sistemas e da cultura de decisão baseada em dados.

3.2 Explicação das decisões

Neste estudo várias decisões cruciais foram tomadas ao longo do processo de análise de dados para garantir a precisão, relevância e utilidade dos resultados.

A seguir, são detalhadas, as principais decisões e os seus fundamentos.

3.2.1 Seleção de Dados e Fontes de Informação

A escolha dos dados foi orientada pela necessidade de compreender plenamente os padrões de consumo de água e identificar anomalias.

Foram selecionados dados de consumo de água, faturação, meios de pagamento, idade dos contadores, localização geográfica e histórico de ruturas. Esta diversidade de fontes permitiu uma análise mais robusta e abrangente, essencial para detenção precisa de padrões de consumo e anomalias.

Na **Error! Reference source not found.** ilustra-se o painel de monitorização geral da área comercial da Tejo Ambiente, entre 01-01-2020 e 29-02-2024, com gráficos de distribuição por classes de consumo, freguesias e ZMC.

Figura 5: Painel de monitorização que permite visualizar dados gerais de clientes, tipos de fatura, modos de pagamento e consumos de água em Tomar**.**

Na **Error! Reference source not found.** apresenta-se um painel de monitorização da evolução das perdas de água de 2022 a 2024, com destaque para diferentes classes de clientes e variações ao longo do tempo, incluindo desvio padrão e média de leitura.

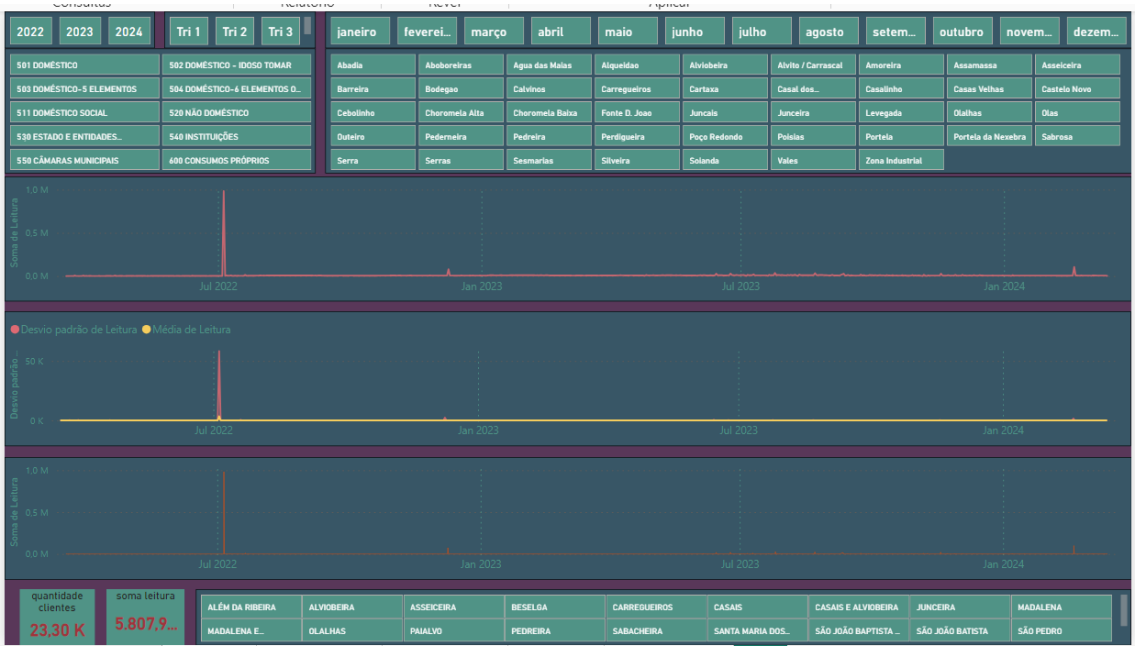


Figura 6: Painel de monitorização Perdas de Água**.**

Na **Error! Reference source not found.** ilustra-se um painel de monitorização comparativo de perdas de água entre quatro ZMC, num período selecionado entre (janeiro de 2024), destacando as diferenças nas perdas ao longo do tempo.

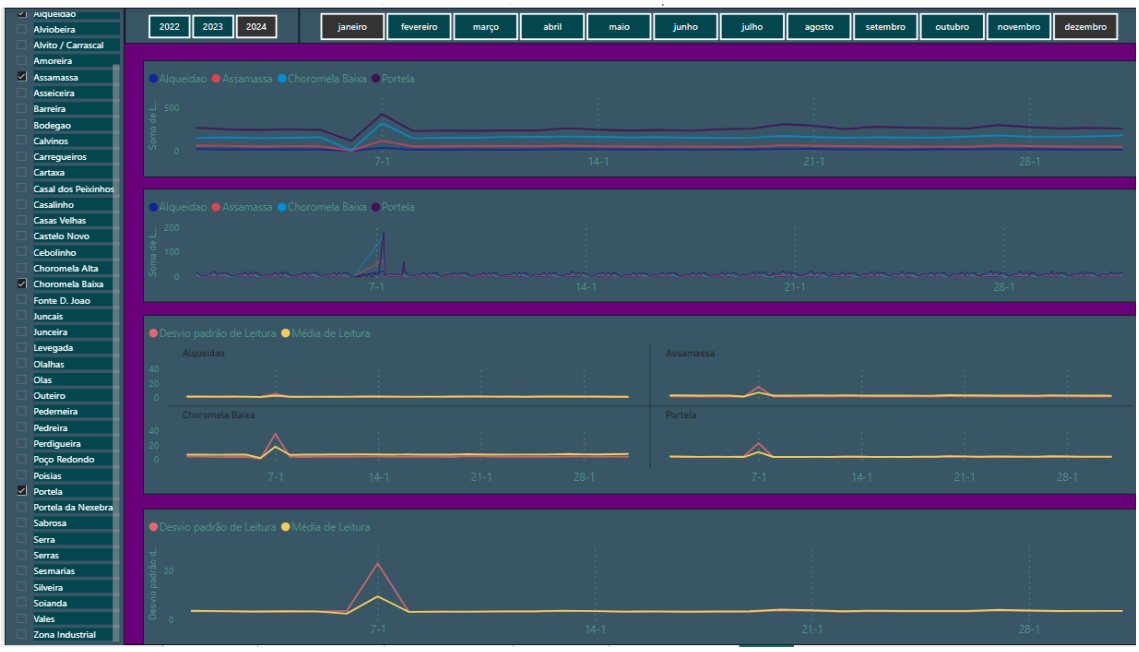


Figura 7: Painel de monitorização comparativo de perdas de água.

3.2.2 Integração e Pré-processamento de Dados

Para este estudo, optou-se por utilizar *Power* BI, uma ferramenta de análise de negócios da *Microsoft* que permite a visualização interativa e a elaboração de relatórios detalhados a partir de diversas fontes de dados.

Esta ferramenta facilita a criação de modelos de dados, gráficos, *dashboards* e relatórios, suportando a tomada de decisões informadas.

O *Power* BI recorre a uma arquitetura baseada em modelos de dados que integra múltiplos elementos analíticos para fornecer visualizações interativas e análises avançadas. O modelo tabular, principal estrutura subjacente, organiza os dados em tabelas relacionais, semelhantes às utilizadas em bases de dados, permitindo a criação de relações entre tabelas, o cálculo de métricas através da linguagem DAX (*Data Analysis Expressions*) e a definição de medidas, colunas calculadas e hierarquias. Complementarmente, a linguagem *Power Query M* assegura a transformação e preparação dos dados. Esta combinação de tabelas, relações, *dashboards*, relatórios e conjuntos de dados torna o *Power* BI uma plataforma robusta e versátil para suportar decisões empresariais orientadas por dados.

As tabelas são estruturas cruciais que apresentam os dados em formato tabular (linhas e colunas), podendo ser importadas de diversas fontes, como bases de dados, ficheiros *Excel* e *Applications Programming Interface* (APIs). As relações entre tabelas facilitam a construção de modelos de dados relacionais, criando ligações entre colunas de diferentes tabelas (**Error! Reference source not found.**).

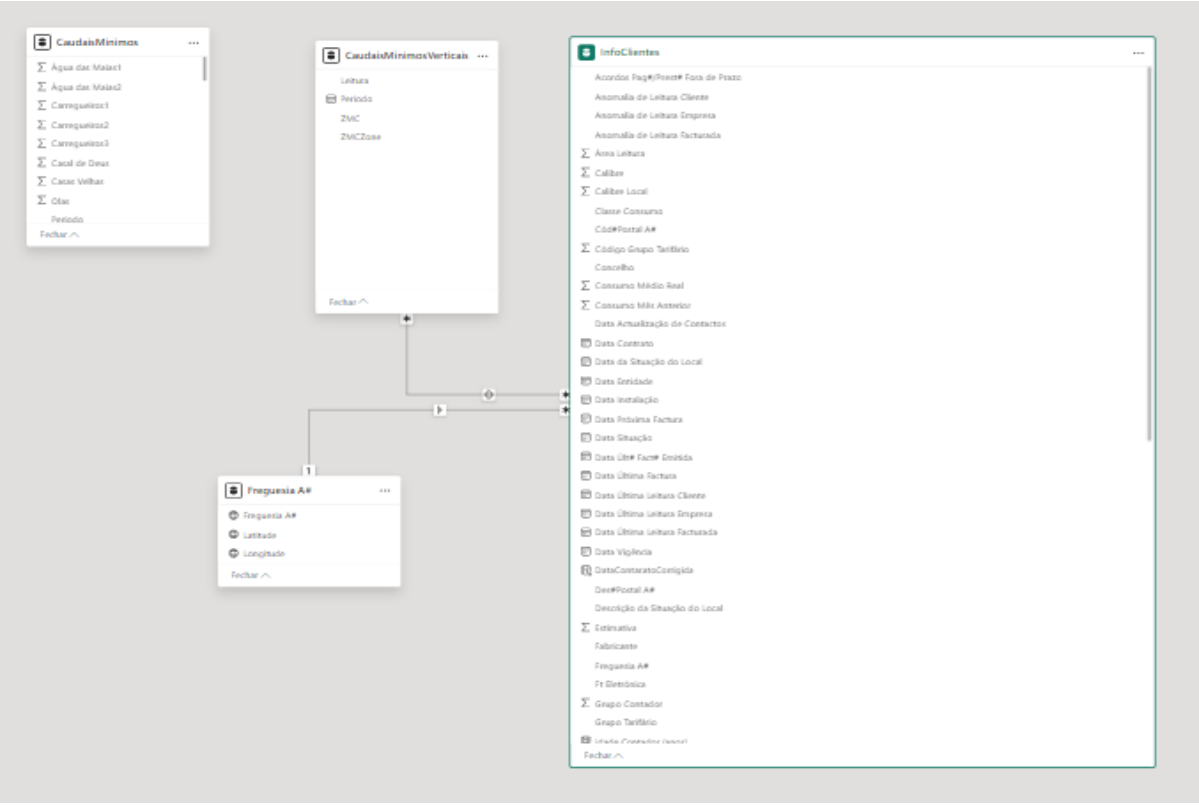


Figura 8: Diagrama do modelo de dados em estudo (*Power* BI).

Para assegurar a integridade e a consistência dos dados utilizados na análise, foi realizada a unificação de todas as fontes disponíveis numa estrutura coerente, eliminando-se redundâncias, discrepâncias e incoerências.

O processo de pré-processamento contemplou a remoção de valores em falta, a correção de erros de digitação e a normalização das unidades de medida.

A integração entre as tabelas “info-clientes” e “caudais mínimos” revelou desafios relevantes, em especial devido à ausência de dados anteriores à adoção do sistema de telegestão em 2022. Embora a empresa tenha sido criada em 2020, muitos contratos existentes precedem a recolha digital sistematizada, o que resultou na existência de cerca de 3000 clientes da região de Tomar com todos os campos necessários preenchidos, exceto a coluna ZMC, que surgia em branco ou designada como “01 indefinido”. Esta limitação comprometia a integração desses clientes na tabela “Caudais Mínimos Verticais” e, consequentemente, a sua inclusão nas análises subsequentes. Para ultrapassar este obstáculo, decidiu-se substituir essas ocorrências por uma designação comum “SEM ZMC”, tanto na coluna ZMC como na coluna ZMC Zone, de forma a uniformizar a informação e garantir a análise integral da base de clientes. Esta adaptação tornou-se indispensável para permitir a participação completa desses registos nas análises comerciais e operacionais. Contudo, ao longo do processo, foram identificadas falhas adicionais que exigiram ajustamentos mais complexos, como documentado na (Figura 9).

Como resposta definitiva, optou-se por introduzir retroativamente dados na tabela “Caudais Mínimos Verticais”, desde o início de atividade da empresa, com data de referência de 1 de janeiro de 2020, ainda que com valores nulos, de modo a garantir a continuidade e completude das séries temporais necessárias à produção de análises robustas.

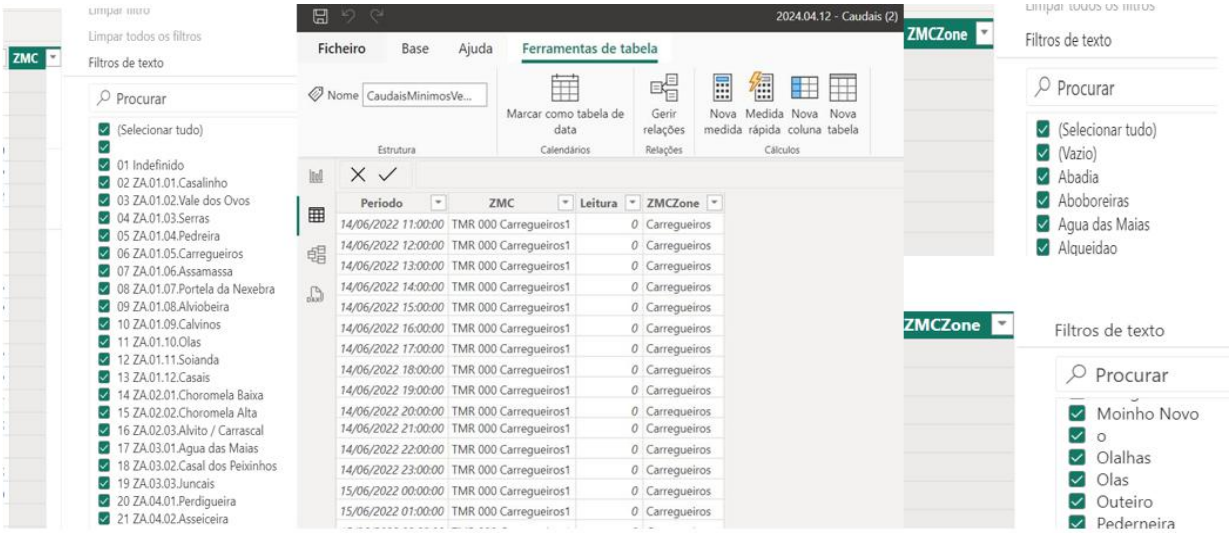


Figura 9: Exemplo de incoerências observadas.

Apesar dessas medidas, persistiam lacunas. Foi necessário estabelecer um método que conferisse credibilidade à pesquisa e incluísse todos os contratos, assumindo que os contratos anteriores à existência da empresa tivessem início em 1 de janeiro de 2020. Assim, foi possível efetuar uma análise inclusiva e fidedigna.

Para facilitar essa conversão, criou-se uma coluna na tabela Info-clientes denominada de DataContaratoCorrigida, utilizando a Equação 1, que permitiu a harmonização dos dados históricos com os atuais, garantindo a integridade e a precisão das análises realizadas.

$$\text{DataContaratoCorrigida} = \text{if}(\text{YEAR}(\text{InfoClientes}[\text{Data Contrato}]) < 2020, \text{date}(2020,1,1), \text{InfoClientes}[\text{Data Contrato}])$$

Equação 1: Fórmula para corrigir a integração de todos os clientes.

3.2.3 Utilização de Técnicas Estatísticas e de ML

Além das análises estatísticas já descritas, foi implementado um modelo de classificação supervisionado, com o objetivo de detetar anomalias operacionais no consumo de água, nomeadamente consumos atípicos, leituras incoerentes ou potenciais ruturas.

Foram utilizados três algoritmos de ML (regressão logística, SVM e *Random forest*).

A variável-alvo (y_anomalia) foi construída com base em critérios operacionais definidos com a Tejo Ambiente, considerando consumos nulos, inconsistências entre leituras da empresa e do cliente, e presença de anomalias conhecidas.

Tabela 1: Resultados antes do balanceamento (sem SMOTE).

Modelo	Accuracy	F1 Score
Regressão Logística	0.92	0.11
SVM	0.93	0.11
Random Forest	0.91	0.16

Apesar da elevada *accuracy*, os valores de *F1 Score* iniciais eram baixos, devido ao desequilíbrio severo nas classes (anomalias eram muito raras). Este cenário refletia um problema comum: os modelos privilegiavam a classe majoritária, resultando em má capacidade de deteção de eventos críticos.

Para resolver o desequilíbrio, foi aplicada a técnica SMOTE. Após o balanceamento dos dados de treino, os modelos foram novamente treinados.

Tabela 2: Resultados do balanceamento após aplicação do SMOTE.

Modelo	F1 Score (após SMOTE)
Regressão Logística	0.94
SVM	0.94
Random Forest	0.99

O modelo *Random Forest* destacou-se como o mais robusto após a aplicação da técnica SMOTE, demonstrando elevado desempenho e posicionando-se como uma ferramenta promissora para auxiliar na deteção de ruturas ou anomalias operacionais.

A aplicação prática de modelos de ML no contexto da Tejo Ambiente permitiu reforçar a capacidade preditiva da organização, com resultados concretos na deteção de ruturas, ligações indevidas e leituras incoerentes, na priorização de zonas críticas para inspeção e na identificação de contadores com eficácia comprometida.

As visualizações criadas incluem gráficos de barras que comparam os *F1 Scores* antes e depois da aplicação do *SMOTE*, matrizes de confusão para cada modelo que ilustram visualmente os erros de classificação, bem como gráficos de importância das variáveis no modelo *Random Forest*, permitindo identificar os fatores com maior impacto na deteção de anomalias.

A abordagem integrada entre *Power BI,* como plataforma de visualização e monitorização operacional, e algoritmos supervisionados de ML, como mecanismo de previsão e apoio à decisão, constitui um exemplo concreto de transformação digital orientada por dados, com impacto direto na redução da ANF e na melhoria da eficiência e sustentabilidade dos serviços públicos prestados pela empresa.

3.2.4 Comparação com Médias Regionais

Adotou-se uma metodologia comparativa robusta para a deteção de anomalias no consumo de água, confrontando o consumo individual dos clientes com a média e o desvio padrão da respetiva zona de abastecimento. Esta abordagem estatística permitiu identificar *outliers* de forma objetiva e padronizada, assegurando que as anomalias detetadas correspondiam a variações estatisticamente significativas face ao padrão regional de consumo.

Com base nos dados operacionais analisados e nas estatísticas do Censos 2021 (Figura 10), foi possível inferir a existência de aproximadamente 3.000 potenciais clientes. No concelho de Tomar, por exemplo, contabilizavam-se, em 2021, cerca de 26.061 habitações suscetíveis de ocupação permanente, excluindo-se ruínas e edifícios em avançado estado de degradação. Comparando com os 23.301 clientes atualmente registados, estima-se uma diferença de aproximadamente 3.000 habitações que poderão estar a consumir água, seja por ausência de ligação formal, consumos não autorizados e utilização de captações próprias para o abastecimento de água, tais como minas, poços e furos.

Este indicador reforça a necessidade de uma estratégia proativa de deteção e integração destas ligações não registadas, contribuindo simultaneamente para a eficiência económica da entidade gestora, a redução de perdas aparentes e a sustentabilidade da rede de abastecimento.

CENSOS DE 2021

CENSOS 2021 POR CONCELHO: EVOLUÇÃO 1960-2021

Veja o que mudou no seu concelho em cada década, entre 1960 a 2021, em diversas áreas como população, famílias, mobilidade, habitação, escolaridade e emprego.

TERRITÓRIO

Tomar

	1960	1981	1991	2001	2011	2021
População residente	44.161	45.672	43.139	43.006	40.677	36.413
Habitações (6) Apartamentos e moradias	-	17.829	21.521	23.871	26.232	26.061
Habitações improvisadas (7) Barracas, casas rudimentares...	-	109	119	150	68	32
Alojamentos coletivos de apoio social, lares, de saúde, prisionais, religiosos e outros	-	21	27	23	45	30
Habitações de uso sazonal (%) (6)	-	15,6	20,8	22,9	24,1	22,9
Habitações vagas (%) (6) para venda, arrendamento e outros casos	7,4	4,5	10,2	11,4	14,4	17,1
Habitações ocupadas pelo proprietário (%) (8)	-	-	73,5	83,3	78,8	76,0
Habitações com duche ou banho (%) (9) (10)	-	55,1	82,1	94,2	97,7	-
Habitações com água canalizada (%) (9) (10)	-	66,6	86,3	97,6	99,1	-
Habitações com instalações sanitárias (%) (9) (10)	-	76,1	87,5	93,1	98,8	-
Habitações com esgoto (%) (9) (10)	-	69,9	85,0	98,3	99,2	-

Figura 10: Distribuição das habitações e identificação de potenciais clientes em Tomar.

3.2.5 Validação de Anomalias

A validação de anomalias é uma etapa crítica no ciclo de monitorização e análise de dados, pois garante que os padrões identificados resultem em ações concretas e justificadas. Neste estudo algumas anomalias foram confirmadas por meio da emissão de ordens de trabalho (OT), enquanto outras foram validadas através do cruzamento de múltiplas fontes de informação, nomeadamente histórico de consumo, dados de faturação, registos de pagamento, antiguidade dos contadores e localização geográfica.

A Figura 11**Error! Reference source not found.** apresenta um painel de monitorização que destaca um aumento significativo no consumo de água registado a 6 de novembro de 2023. Este pico abrupto poderá indicar uma rotura ou evento anómalo que justifica uma intervenção urgente por parte da equipa técnica, com vista à mitigação de potenciais perdas ou riscos operacionais.

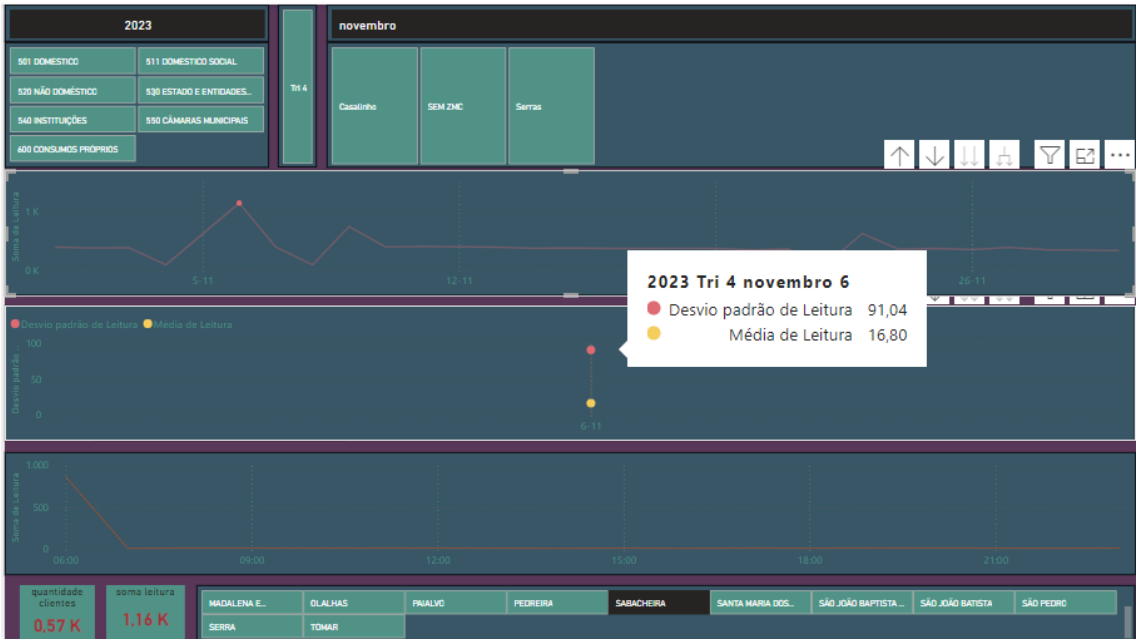


Figura 11: Painel de monitorização: aumento significativo no consumo de água.

A Figura 12 mostra a ordem de trabalho correspondente à situação evidenciada na Figura 11. É importante salientar que os registos administrativos nem sempre refletem com exatidão a data real do evento, existindo, por vezes, alguma discrepância temporal entre a ocorrência e o seu registo nos sistemas operacionais.

AQUA ıbiente

Ordem de trabalho OT_30228

Registo

Código interno	38992
Data do registo	2023-11-15 16:18
Custo total	213.04
Descrição	Ordem de Trabalho

Dados Gerais

Origem da OT	Desktop
Data Criação	2023-11-15 16:18
Referência	OT_30228
Prioridade	Normal
Ref. Origem	
Estado	validada
Data planeada	2023-11-15 16:18
Motivo não conclusão	

Figura 12: Ordem de trabalho originada pelo programa operacional *Aquaworks*.

A Figura 13 ilustra outro painel desenvolvido em Power BI, no qual se analisam os contratos de faturação doméstica da freguesia de Além da Ribeira. Observa-se que existem 391 clientes registados, mas apenas 388 contratos foram efetivamente faturados. Esta discrepância pode sinalizar falhas na integração de novos contratos no sistema de faturação, bem como oportunidades de melhoria na gestão de cobranças e no alinhamento entre dados cadastrais e operacionais.

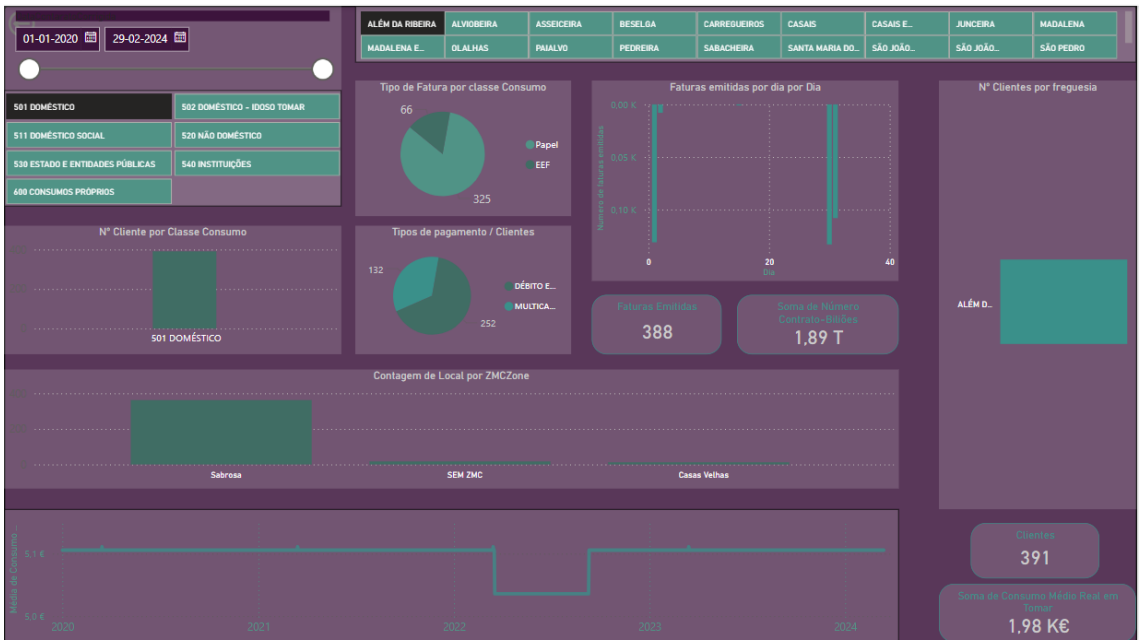


Figura 13: Painel de monitorização: seleção de clientes com tipo de faturação doméstica da freguesia Além da Ribeira.

A Figura 14 detalha a identificação de consumidores não faturados. A análise permitiu localizar um total de 44 clientes sem faturação ativa, dos quais 39 pertencem ao segmento doméstico e 5 ao não doméstico. Esta segmentação evidencia que cerca de 88,6% das anomalias ocorrem no segmento doméstico, o que pode estar associado a diversas causas, como falhas na leitura automática, problemas de comunicação entre sistemas, mapeamento incorreto entre o cliente e a ZMC associada, ou ainda a existência de ligações não formalizadas.

Adicionalmente, os *dashboards* analisados revelaram diversos indicadores-chave: um total de 7.151 clientes, uma soma de consumo médio real em Tomar de 9,23 €, 7.107 faturas emitidas, uma distribuição por freguesia com destaque para Santa Maria dos Olivais, e uma segmentação abrangente por classes de consumo, desde "Doméstico" até "Câmaras Municipais".

Os painéis temporais, localizados na parte inferior dos *dashboards*, permitem acompanhar a evolução das leituras e faturações entre 2020 e 2024, com destaque para momentos de quebra ou subida abrupta, frequentemente associados à adoção de novos ciclos de faturação ou à instalação de contadores inteligentes. A georreferenciação por ZMC evidencia ainda zonas como Choromela Alta e Sem ZMC, onde se registam lacunas ou dispersão nos dados, representando áreas prioritárias para intervenção.

Figura 14: Painel de monitorização: identificação de consumidores não faturados.

Com base nesta análise, destacam-se diversos insights operacionais relevantes, nomeadamente a necessidade de uma auditoria comercial direcionada aos 44 clientes não faturados, cuja situação representa uma perda económica estimada em cerca de 400 euros por mês, o que equivale a aproximadamente 5.000 euros anuais; a concentração de anomalias no segmento doméstico, que evidencia potenciais fragilidades na ligação entre as leituras de consumo e a consequente emissão de faturas; a pertinência da implementação de alertas automáticos em *Power* BI sempre que um cliente com consumo registado não origina a geração de fatura no ciclo seguinte; e, por fim, a importância da adoção de processos de validação cruzada sistemática, nomeadamente auditorias semestrais entre os dados de leitura e de faturação, como medida essencial para reforçar a fiabilidade, a transparência e a eficiência do processo comercial da entidade gestora.

A integração entre BI e os processos operacionais da gestão comercial revela-se, assim, essencial para garantir maior rigor, transparência e eficiência na gestão dos serviços de abastecimento de água, permitindo uma atuação proativa na correção de falhas e na melhoria do desempenho financeiro da entidade gestora.

3.2.6 Implementação de Técnicas de Visualização de Dados

A utilização de técnicas de visualização de dados foi uma etapa estratégica no desenvolvimento do presente estudo, tendo como objetivo facilitar a comunicação dos *insights* extraídos e promover uma tomada de decisão mais ágil, informada e orientada por evidência. A criação de *dashboards* interativos surgiu como a solução mais

eficaz para traduzir grandes volumes de informação em representações visuais intuitivas, acessíveis tanto a utilizadores técnicos como a decisores não especializados.

A visualização através de *dashboards* permitiu condensar dados provenientes de fontes diversas como o sistema *AquaMatrix*, plataformas de telegestão e bases de dados sociodemográficas em painéis dinâmicos que oferecem uma visão abrangente sobre o desempenho da rede, o comportamento de consumo por zona e a identificação de potenciais anomalias.

A Figura 15 apresenta um exemplo de painel de monitorização interativo construído em *Power* BI, cujo disponibiliza uma visão abrangente sobre o consumo de água em diferentes áreas de atuação. A sua estrutura foi desenhada para permitir a identificação imediata de eventos críticos, como ruturas, consumos excessivos, leituras incoerentes ou zonas com baixa cobertura de medição. A interatividade dos *dashboards* permite aplicar filtros por ano, freguesia, classe de consumo, entre outros, tornando o sistema flexível e adaptável às necessidades de análise de cada departamento.

Figura 15: Painel de monitorização de dashboards interativos.

Além da vertente de monitorização operacional, os *dashboards* foram integrados com modelos de ML supervisionado, permitindo visualizar previsões de risco, segmentar clientes com base em perfis de consumo e priorizar intervenções com base em critérios analíticos.

A adoção desta abordagem promoveu maior transparência e eficiência na gestão da rede de abastecimento, ao transformar dados brutos em conhecimento visualmente estruturado e acionável. Esta experiência evidencia o valor estratégico da visualização de dados como componente central de qualquer iniciativa de BI no setor público.

4. METODOLOGIA

Neste capítulo, descreve-se de forma detalhada a abordagem metodológica adotada ao longo do desenvolvimento deste estudo. São apresentados os objetivos operacionais que nortearam a investigação, bem como as fases que compuseram o planeamento, execução e validação do estudo.

A metodologia seguida integra a identificação das fontes de dados, os procedimentos de recolha e pré-processamento, a seleção e aplicação de técnicas analíticas, bem como a implementação de algoritmos de ML supervisionado. Foram também utilizadas ferramentas de BI, nomeadamente a plataforma *Power BI*, para a construção de *dashboards* interativos que suportam a análise e visualização dos dados.

Adicionalmente, são descritas as métricas de avaliação utilizadas para aferir o desempenho dos modelos preditivos, bem como os critérios de validação e os instrumentos estatísticos aplicados para garantir a robustez e fiabilidade dos resultados obtidos. Esta abordagem visa garantir a coerência científica do estudo e a sua aplicabilidade prática no contexto da Tejo Ambiente.

4.1 ESTRATÉGIA DE CONTROLO DE PERDAS E IMPORTÂNCIA DAS ZMC

A água é um recurso essencial, mas as perdas na rede de distribuição continuam a ser um dos principais desafios operacionais das Entidades Gestoras (EG), com impactos diretos na sustentabilidade financeira, ambiental e na qualidade do serviço prestado. O controlo das perdas de água tem vindo, desde há vários anos, a ser reconhecido como uma das atividades mais relevantes para a melhoria da eficiência operacional, sendo, por isso, fundamental definir estratégias de mitigação ajustadas à realidade de cada sistema.

A ANF é um dos principais indicadores da ineficiência nos sistemas de abastecimento E divide-se, tradicionalmente, em três componentes, o consumo autorizado não faturado, que abrange consumos autorizados, mas não cobrados; as perdas reais, resultantes de fugas em condutas, ramais ou reservatórios e as perdas aparentes, associadas a ligações indevidas, erros de medição ou falhas no registo e faturação.

A escassez de água, além de associada à disponibilidade hídrica natural, decorre frequentemente de limitações estruturais dos sistemas de distribuição, como condutas envelhecidas, pressão excessiva ou falhas de manutenção. Quando estes sistemas apresentam níveis elevados de perdas, comprometem a continuidade do serviço e a sustentabilidade da infraestrutura. Neste sentido, a água que se evita perder deve ser considerada como uma fonte alternativa eficaz e económica.

A gestão eficiente dos recursos hídricos exige, assim, uma abordagem baseada em dados e na monitorização contínua da rede, que permita antecipar ruturas e falhas sistémicas, reduzir perdas reais e aparentes, otimizar a substituição de condutas e equipamentos, identificar zonas de risco e priorizar intervenções com base em indicadores objetivos e estimar zonas com população não servida, através do cruzamento de dados operacionais com dados estatísticos, como os Censos do INE.

De acordo com a ERSAR (RASARP 2024), os serviços públicos de abastecimento de água em Portugal perdem anualmente cerca de 184 milhões de metros cúbicos, com 28,3% de perdas em sistemas públicos, contrastando com os 15,2% em sistemas sob gestão privada. Este cenário reforça a necessidade urgente de modernização dos sistemas, nomeadamente através da implementação de tecnologias analíticas, sensorização e contadores inteligentes.

A resposta a este desafio passa pela implementação de ZMC e pela setorização operacional das redes de distribuição. Uma ZMC corresponde a um setor devidamente delimitado, abastecido por uma entrada controlada por medidores de caudal e pressão, permitindo conhecer de forma precisa o volume de água introduzido e consumido, as perdas reais estimadas e os padrões de consumo por zona.

Esta metodologia possibilita a deteção mais célere e precisa de fugas, ruturas ou consumos anómalos, sendo crucial para orientar equipas no terreno de forma direcionada e eficaz.

Para além disso, a integração da análise de dados e inteligência organizacional permite uma abordagem transformadora na gestão das perdas, com *dashboards* interativos em *Power* BI, capazes de apresentar, em tempo real, os volumes, pressões e indicadores de desempenho por zona; modelos preditivos baseados em algoritmos de ML, que estimam o risco de perda em função de variáveis como idade do contador, tipo de zona, histórico de ruturas ou tipo de consumidor; análise integrada de fatores geográficos, demográficos e operacionais, cruzando informação dos sistemas de gestão interna (ex: *AquaMatrix*, telemetria) com dados externos (ex: Censos), permitindo identificar zonas com grande densidade populacional, mas sem registo de clientes ativos, o que poderá indicar ligações ilegais ou situações de não faturação.

De forma complementar, a definição de indicadores operacionais personalizados, para além dos obrigatórios definidos pela ERSAR, pode oferecer uma avaliação mais rigorosa e ajustada à realidade local, permitindo às entidades gestoras agir de forma mais estratégica.

Em suma, a transformação digital na gestão de sistemas de abastecimento não se esgota na recolha de dados. Exige uma estruturação inteligente da informação, ferramentas analíticas avançadas, e um planeamento orientado por evidência empírica, que permitam reduzir perdas, otimizar recursos e aumentar a eficiência global do sistema. O presente estudo explora precisamente como estas estratégias e ferramentas analíticas podem ser integradas na realidade da Tejo Ambiente, contribuindo de forma prática e replicável para uma gestão mais eficiente da água.

4.2. Caso em Estudo

A gestão eficiente dos recursos hídricos é crucial para assegurar a sustentabilidade ambiental e o bem-estar das populações. Este estudo incidiu sobre um concelho pertencente ao universo de atuação da Tejo Ambiente, selecionado com base na disponibilidade, qualidade e consistência dos dados operacionais. A escolha desta área permitiu a aplicação prática de uma metodologia orientada à análise integrada do consumo de água, com recurso à plataforma Power BI e ao cruzamento de múltiplas variáveis.

A análise de dados aplicada à gestão hídrica revelou-se uma ferramenta estratégica indispensável. Permitindo a caracterização de padrões de consumo, a identificação de anomalias e a otimização da tomada de decisão operacional, o estudo propôs uma abordagem mais robusta, que combinou a análise tradicional de dados com a avaliação de fatores complementares, tais como: dados de faturação e meios de pagamento, idade dos contadores e tipo de dispositivo instalado, localização geográfica e densidade habitacional, histórico de ruturas e anomalias reportadas, fatores climáticos e variações sazonais e dados demográficos (Censos do INE).

4.2.1. Importância Estratégica da Análise de Dados na Gestão Hídrica

A análise de dados nos sistemas de abastecimento de água oferece diversas vantagens estratégicas, nomeadamente: a antecipação de ruturas, com base em padrões históricos, a otimização da substituição de redes e equipamentos obsoletos, a redução de perdas físicas e aparentes, minimizando desperdícios e a estimativa da população não servida, através da integração com dados do INE.

De acordo com a ERSAR (2020), contadores com mais de 15 anos apresentam margens de erro significativas, comprometendo a precisão da faturação e a deteção de consumos reais. A substituição proativa desses equipamentos poderá representar ganhos substanciais em eficiência operacional.

Adicionalmente, ao cruzar os dados internos da Tejo Ambiente com os Censos de 2021, foi possível mapear áreas com elevada densidade populacional e ausência de clientes registados, inferindo a existência de aproximadamente 3.000 potenciais ligações. Esta análise revelou oportunidades para expandir a cobertura dos serviços e combater consumos não registados.

A metodologia adotada permitiu, assim, não apenas uma melhor compreensão da realidade operacional, como também o desenvolvimento de modelos preditivos e prescritivos de apoio à decisão, com aplicação direta na definição de prioridades de intervenção, investimento e monitorização da rede.

5. IMPLEMENTAÇÃO

5.1 RECOLHA E PRÉ PROCESSAMENTO DE DADOS

O estudo teve como principal objetivo a recolha e análise de dados relativos aos consumos de água e aos medidores de caudal. Procedeu-se também à obtenção de informações sobre faturação, pagamentos, idade dos contadores, localização geográfica e outros parâmetros relevantes.

O pré-processamento dos dados consistiu na integração de diferentes conjuntos de dados numa estrutura única e coerente, eliminando inconsistências, valores ausentes e erros de digitação. Foram padronizadas unidades de medida e formatos de registo, assegurando a consistência da análise posterior.

Na análise exploratória, recorreu-se a técnicas descritivas para caracterizar os dados: medidas de tendência central, dispersão e análise de frequência. Foram utilizadas visualizações como gráficos e tabelas para detetar padrões, correlações e tendências entre variáveis.

Na análise multivariada, aplicaram-se técnicas como *Principal Component Analysis* (PCA), cluster *analysis* e regressão múltipla, segmentando clientes com base em características semelhantes. Foram integradas técnicas de ML para prever zonas com potencial risco de ruturas e vazamentos.

Na fase de deteção de anomalias, os consumos reais foram comparados com médias e desvios padrão por zona. Os *outliers* detetados foram investigados detalhadamente, cruzando variáveis como histórico de consumo, idade dos contadores, localização, tipo de imóvel e fatores climáticos. As anomalias confirmadas foram validadas por via de OT emitidas no sistema.

O estudo constituiu uma base sólida para o desenvolvimento de uma gestão eficiente dos recursos hídricos, contribuindo para a sustentabilidade ambiental e para o reforço da capacidade de decisão das entidades gestoras. A combinação de estatística, inteligência artificial e visualização de dados permitiu extrair *insights* valiosos, com impacto direto na operacionalização e planeamento estratégico dos sistemas de abastecimento.

6. RESULTADOS

A investigação realizada neste estudo obedece a um conjunto de procedimentos e orientações tendo em vista o rigor e o valor pratico da informação recolhida nas diferentes áreas.

6.1 ÁREA COMERCIAL

A análise da área comercial permitiu uma leitura abrangente do funcionamento da atividade de faturação, perfil dos clientes e práticas de pagamento na Tejo Ambiente. A investigação seguiu uma abordagem sistemática com base em dados recolhidos entre 01/01/2020 e 29/02/2024, o que garante a robustez dos resultados obtidos e a sua relevância para a definição de estratégias operacionais e comerciais futuras.

Foram utilizados *dashboards* interativos desenvolvidos em *Power* BI, que integram múltiplas dimensões de análise, incluindo o volume de faturas emitidas por dia, a tipologia de consumo por cliente (doméstico, não doméstico e entidades públicas), os meios de pagamento mais utilizados, a frequência de faturação e a distribuição dos consumos por freguesias e por ZMC, permitindo assim uma visão integrada da base de clientes e da eficiência do processo de faturação.

6.1.1. ANÁLISE GERAL DA BASE DE CLIENTES

Com base no painel de monitorização global apresentado na Figura 16, apurou-se que a Tejo Ambiente possui atualmente 23.301 clientes ativos e que, no mesmo período, foram emitidas 23.170 faturas. O dia 12 de cada mês destaca-se como o de maior volume de faturação e a maioria dos envios de faturas continua a ser realizada em formato papel. O consumo médio real global situa-se nos 7,82 euros.

Figura 16: Painel de monitorização com *dashboards* de monitorização com informações gerais (faturação, meios de pagamento, consumos).

Este painel destaca uma incongruência entre o número total de clientes e o número de faturas emitidas, sinalizando a existência de casos de não faturação que exigem análise detalhada. Estas situações podem resultar de leituras em falta, ausência de dados válidos para faturação ou ainda de contratos recentemente criados sem histórico de consumo.

6.1.2. Análise Específica por Segmento: Clientes Domésticos

A Figura 17 apresenta o painel de monitorização dedicado exclusivamente ao segmento dos clientes domésticos, permitindo a aplicação de filtros específicos a este grupo. Verifica-se que existem 20.220 clientes domésticos ativos, dos quais 20.110 foram efetivamente faturados, originando um total de 20.110 faturas emitidas. O consumo médio real por cliente situa-se nos 6,71 euros, sendo o débito direto o meio de pagamento predominante. Mais de 80% das faturas continuam a ser emitidas em papel.

Figura 17: Painel de monitorização com foco nos clientes domésticos.

A existência de 110 clientes sem fatura emitida deve ser analisada, uma vez que ao longo do tempo essa discrepância pode representar perdas relevantes para a entidade gestora.

6.1.3. Análise Territorial: Freguesia de Além da Ribeira

A Figura 18 ilustra um painel analítico com foco exclusivo na freguesia de Além da Ribeira. Nesta zona, registam-se 434 clientes ativos e 431 faturas emitidas, o que representa uma taxa de cobertura de faturação de 99,3%. A percentagem de faturas em papel atinge os 82,03% e o débito direto é a forma de pagamento predominante, representando 65,09% das transações. O consumo médio real nesta freguesia é de 5,12 euros. A elevada taxa de faturas em papel revela um potencial de melhoria ao nível da transição digital, que pode traduzir-

se em ganhos substanciais de eficiência administrativa, redução de custos operacionais e aumento da sustentabilidade ambiental da organização.

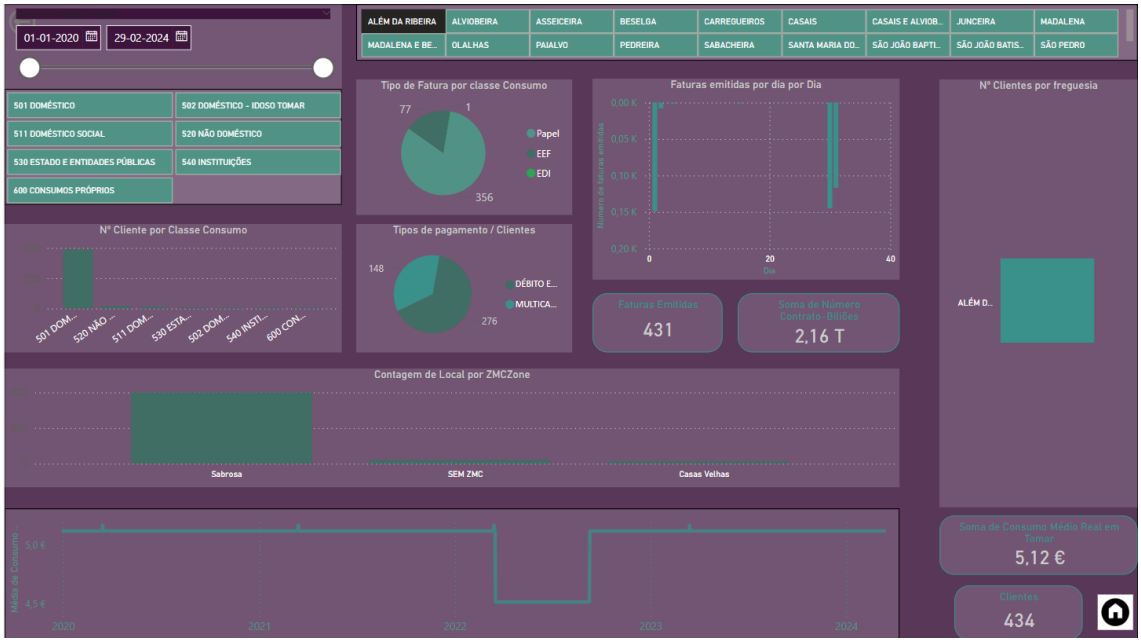


Figura 18: Painel de monitorização com **f**oco nos **c**lientes da freguesia - Além da Ribeira**.**

6.1.4. *INSIGHTS* ESTRATÉGICOS DA ÁREA COMERCIAL

Esta componente do estudo permitiu demonstrar a importância do cruzamento entre dados operacionais e comerciais, revelando oportunidades claras para melhoria do desempenho da faturação, redução de perdas administrativas e promoção da eficiência na gestão da relação com o cliente.

Os painéis desenvolvidos em *Power* BI mostraram-se instrumentos fundamentais para visualizar tendências, identificar exceções e tomar decisões baseadas em dados concretos e atualizados.

Tabela 3: Tabela resumo de *insight*.

Insight	Evidência nos Dashboards	Implicações Estratégicas
Inconsistências na Faturação	Diferença entre clientes ativos e faturas emitidas	Reforçar validação cruzada entre leitura e emissão de faturas
Digitalização Insuficiente	Elevado uso de papel (>80%)	Investir em faturação eletrónica e campanhas de sensibilização
Adoção do Débito Direto	Média de 65% nas freguesias analisadas	Boa prática que pode ser replicada noutras freguesias
Segmentação por Tipo de Cliente	Painéis permitem análises específicas por tipo e zona	Personalização de políticas comerciais e deteção de anomalias
Monitorização em Tempo Real	Painéis com dados atualizados permitem análise dinâmica	Suporte direto à decisão operacional e estratégica

6.2 Monitorização de Perdas de Água

A água constitui um bem essencial, mas simultaneamente limitado, cuja gestão exige cada vez mais eficiência e inovação, sobretudo num contexto marcado pela escassez hídrica, alterações climáticas e envelhecimento das infraestruturas. Uma das principais fragilidades operacionais nas entidades gestoras prende-se com as perdas de água na rede de distribuição, tanto visíveis como invisíveis, físicas ou administrativas.

A gestão eficiente da rede depende do conhecimento técnico detalhado das infraestruturas e da capacidade de resposta célere a anomalias. A criação de ZMC e a setorização da rede são ferramentas críticas que permitem a monitorização contínua dos caudais e pressões, a localização rápida de ruturas, a priorização de intervenções em zonas críticas e o controlo de perdas reais e aparentes de forma sistemática.

Na Figura 19 apresenta-se um exemplo de uma ZMC, sendo apresentada a sua arquitetura e a sua interligação com os sistemas de medição, caudalímetros e sensores de pressão.

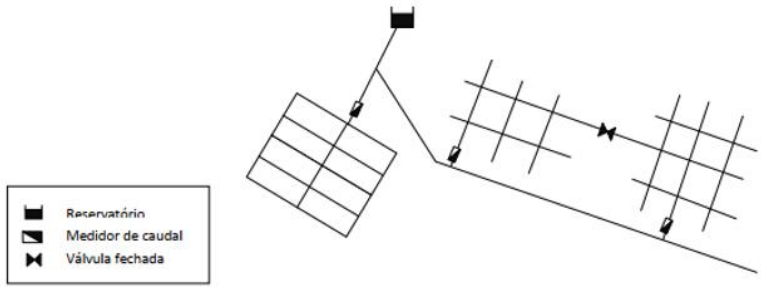


Figura 19: Esquema de zona de monitorização.

6.2.1 Água Não Faturada (ANF)

A (ANF) representa uma parcela significativa do volume total de água produzida que não se traduz em receita para a entidade gestora. Esta pode ser classificada em três grandes categorias: o consumo autorizado não faturado, que inclui situações como isenções institucionais, falhas de leitura ou estimativas administrativas imprecisas; as perdas reais, que derivam de fugas físicas na rede de abastecimento, abrangendo condutas, ramais e reservatórios; e por fim, as perdas aparentes, que resultam de erros de medição, ligações indevidas, consumos ilícitos ou falhas nos processos informáticos e humanos.

A minimização da ANF exige, por isso, uma abordagem multidimensional e integrada, orientada por dados e suportada por tecnologias analíticas como o *Power* BI, associada a sistemas de telemetria e, cada vez mais, à aplicação de técnicas de inteligência artificial, com vista à deteção proativa de ineficiências e à otimização dos processos de gestão hídrica.

6.2.2 Análise Gráfica e Deteção de Ruturas

A Figura 20 permite a visualização integrada das perdas ao longo do tempo e por zona, permitindo detetar variações anómalas de caudal, pressão e consumo. Através da análise temporal, é possível isolar eventos críticos e antecipar falhas.

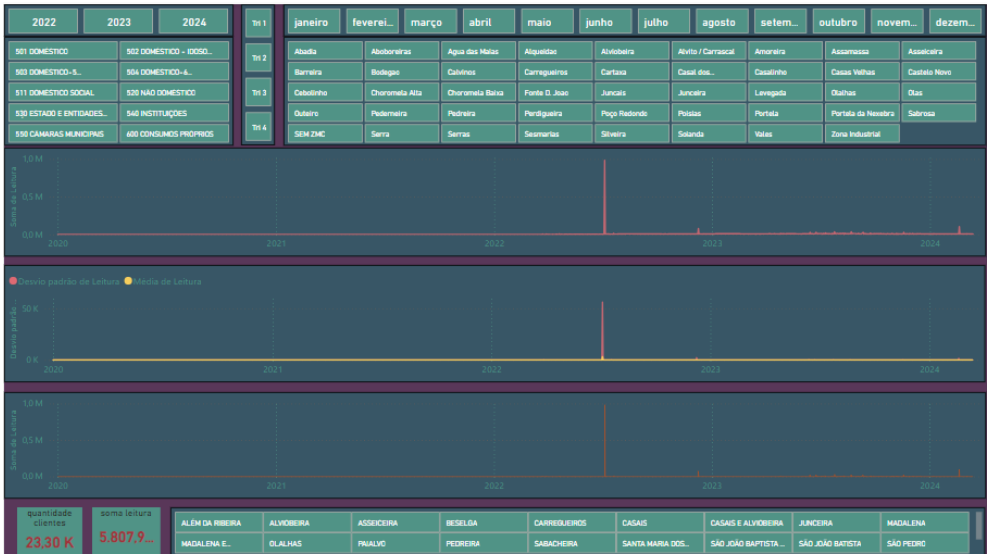


Figura 20: Painel de monitorização detalhada de perdas de água.

A Figura 21 revela dois picos anómalos de consumo, claramente identificados como ruturas, ocorridas em julho de 2023. O uso combinado de consumo médio, desvio padrão e deteção de *outliers* (com algoritmos de regressão e *clustering*), permitiu sinalizar automaticamente as ocorrências, validando a eficácia da análise preditiva.

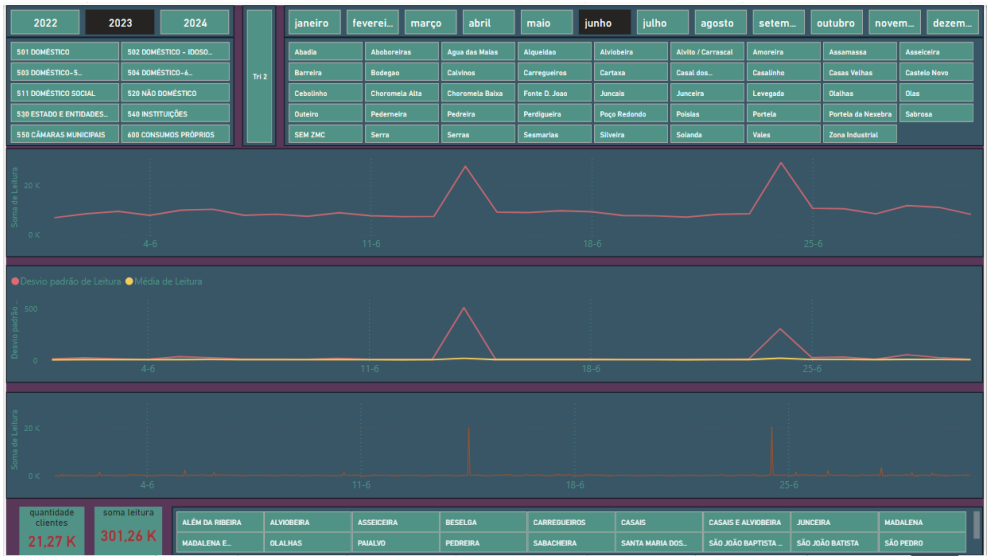


Figura 21: Painel de monitorização de ruturas (Exemplo: julho de 2023).

6.2.3 Comparação de Consumos por Zonas e Proposta Técnica para Melhoria da Monitorização

A análise comparativa dos padrões de consumo em diferentes zonas do concelho ao longo do ano de 2023, conforme ilustra a **Figura 21**Figura 22, revelou diferenças operacionais relevantes entre territórios com perfis aparentemente semelhantes. A freguesia de Carregueiros evidenciou um pico acentuado no segundo semestre, acompanhado por desvios padrão significativamente elevados, o que pode indicar situações anómalas como

ruturas, ligações indevidas ou erros de leitura. Por outro lado, as zonas de Fonte de Dom João, Alviobeira e Água das Maias apresentaram consumos mais estáveis, com variações marginais, traduzindo um elevado grau de fiabilidade nos dados recolhidos.

Estes resultados reforçam a utilidade da análise interzonal como instrumento de apoio à deteção precoce de anomalias, permitindo identificar zonas críticas, priorizar inspeções e estimar perdas potenciais com maior rigor.

Figura 22: Painel de monitorização Comparativo de Consumos: Água das Maias, Alviobeira, Carregueiros e Fonte de Dom João.

A experiência acumulada ao longo deste estudo evidenciou que, apesar da existência de dados operacionais relevantes, subsistem desafios técnicos que dificultam a automação da deteção de perdas.

Propõe-se, como resposta a essa limitação, o desenvolvimento de um programa integrado de monitorização avançada, constituído por cinco módulos interligados.

O primeiro módulo, de deteção de anomalias na rede de distribuição, deverá basear-se nos dados de telegestão (caudal e pressão), com o objetivo de identificar variações abruptas que sinalizem ruturas ou intervenções não planeadas.

O segundo módulo, centrado nas anomalias na rede predial, será alimentado por contadores inteligentes com telemetria (preferencialmente LoRaWAN), permitindo a deteção de consumos anómalos ao nível doméstico ou não doméstico .

Um terceiro módulo garantirá a validação cruzada com leituras manuais extraídas do sistema AquaMatrix, assegurando a confirmação de anomalias previamente sinalizadas. Em paralelo, propõe-se a criação de uma

plataforma mobile de comunicação com o consumidor, responsável por emitir alertas em tempo real sobre ruturas, consumos excessivos ou torneiras abertas.

Por fim, todos os dados provenientes destas fontes serão integrados num modelo analítico unificado, agregando *AquaMatrix*, telegestão e telemetria, de modo a permitir análises em tempo real e maior capacidade preditiva.

Contudo, o processo de desenvolvimento e integração dessas soluções, enfrentará várias dificuldades estruturais que irão afetar a eficácia da monitorização. Entre os principais desafios técnicos e operacionais destaca-se a ausência de correspondência entre as zonas ZMC antigas do *AquaMatrix* e as novas divisões estabelecidas nos sistemas de Telegestão, dificultando a análise longitudinal e territorial. Erros ou ausências de moradas associadas aos contadores, que compromete a localização precisa das anomalias. E a inexistência de integração entre os dados de telemetria (tempo real) e os registos comerciais, que impede a análise cruzada de indicadores críticos como leituras, faturação e histórico de intervenções.

Por fim, as listagens incompletas e desatualizadas no módulo "Infoclientes" do *AquaMatrix* dificultam o mapeamento entre instalações, clientes e consumos reais.

A superação destas limitações exige a implementação de rotinas automáticas de normalização de dados, melhorias na interoperabilidade entre sistemas e, sobretudo, uma colaboração mais estreita entre os departamentos técnicos e comerciais. A construção de uma cultura organizacional orientada por dados torna-se essencial para garantir a eficácia da monitorização em tempo real, o rigor na tomada de decisão e a capacidade de resposta a situações críticas.

Conclui-se que a análise sistemática de perdas, aliada à inteligência artificial e à visualização analítica, constitui uma estratégia eficaz para transformar dados operacionais dispersos em valor organizacional concreto. As propostas aqui delineadas visam reforçar a capacidade da Tejo Ambiente em prevenir ruturas, reduzir perdas e melhorar significativamente a qualidade e a sustentabilidade do serviço prestado à população.

6.3 Análise da Antiguidade dos Contadores de Água

A antiguidade dos contadores de água constitui um dos principais indicadores técnicos a considerar na definição de estratégias de substituição de equipamentos, dada a sua influência direta na fiabilidade da medição e na precisão da faturação.

Com o decorrer do tempo, os dispositivos de medição sofrem degradação mecânica e perdem sensibilidade, conduzindo a erros sistemáticos, tanto por defeito como por excesso, no volume de água registado.

A Figura 23 ilustra um painel de monitorização que permite a análise detalhada da distribuição da idade dos contadores por zona geográfica e classe de consumo, com agrupamento em faixas etárias (<5, 5-10, 11-15, 16-20, >21) e cálculo da média da idade por ZMC. Os dados revelam que a média global de antiguidade dos

contadores situa-se em 10,57 anos, sendo que 34,49% dos dispositivos apresentam mais de 21 anos de serviço, uma faixa considerada crítica.

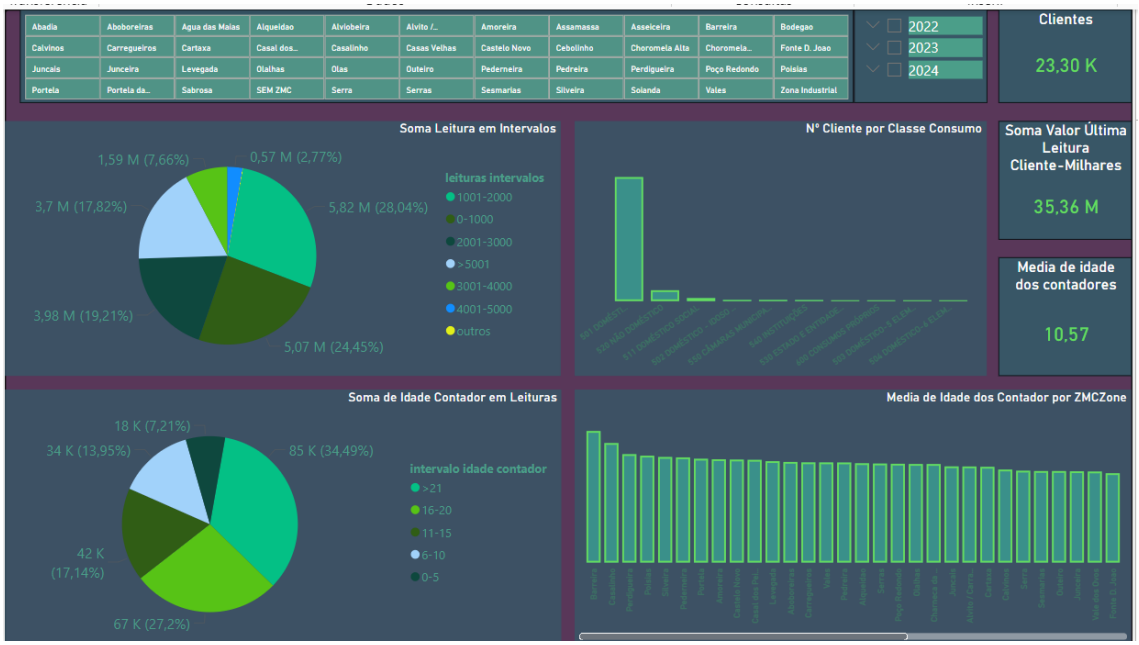


Figura 23: Painel de monitorização de Contadores de Água: Identificação de Áreas Prioritárias de Substituição.

6.3.1 Impacto da Antiguidade na Leitura de Consumos

A Figura 24 evidencia que com recurso à funcionalidade de análise de influenciadores do *Power* BI, a correlação entre a idade dos contadores e os valores médios de leitura, existem zonas com maior antiguidade, como Alvito/Carrascal, registando leituras médias significativamente superiores (índice de influência de 8,2) quando comparadas com zonas com dispositivos mais recentes (índice de 2,36), o que pode indicar desvios na medição real do consumo e provocar sobrevalorização da fatura.

Os desvios identificados poderão decorrer do desgaste progressivo dos contadores mais antigos, cuja deterioração mecânica tende a afetar a precisão das medições ao longo do tempo. Esta degradação compromete não só a justiça na faturação ao cliente, como também a qualidade e fiabilidade dos dados históricos, pondo em causa a robustez de qualquer análise estatística ou modelo preditivo que se baseie nessas leituras.

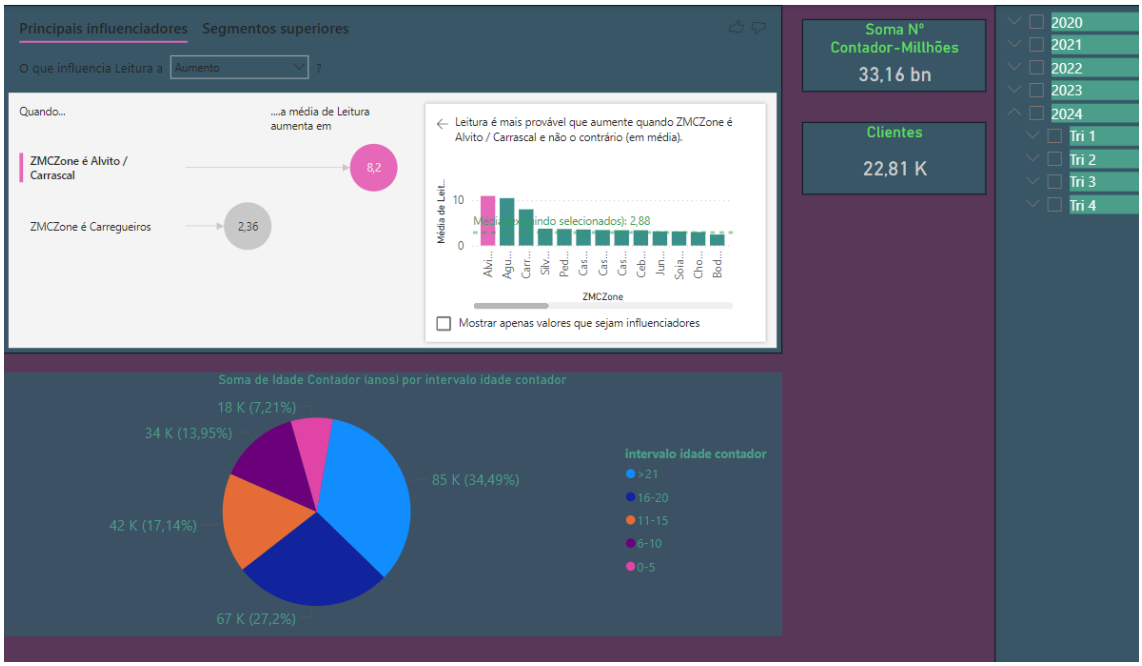


Figura 24: Painel de monitorização: Influência da Antiguidade dos Contadores nas Leituras de Consumo.

6.3.2 Recomendação Técnica e Estratégica

Com base nas evidências recolhidas, torna-se viável propor um plano faseado de substituição de contadores assente em critérios técnicos e operacionais. A priorização deve considerar a idade dos dispositivos (com foco urgente nos contadores com mais de 21 anos), a tipologia de consumo (dando prioridade aos clientes domésticos com elevado volume de consumo), a localização em zonas críticas com incidência de ruturas e anomalias, e o histórico de avarias registadas. A implementação desta estratégia deverá traduzir-se em melhorias mensuráveis na precisão das medições, redução das perdas aparentes, maior equidade na faturação, melhor suporte à decisão estratégica e reforço da eficiência global dos sistemas de abastecimento de água. Trata-se de um investimento orientado por dados e sustentado em análise estatística, com impacto direto na transparência e sustentabilidade do serviço público.

6.3.3 Influência no Ciclo Urbano da Água

A relevância dos dados recolhidos ao nível dos contadores vai muito além da sua função de medição individual. Como ilustrado na Figura 25, esses dados influenciam todas as fases do ciclo urbano da água: desde o planeamento da captação, passando pela distribuição, faturação e até ao reuso e tratamento final. Um erro ou distorção de leitura num único ponto de consumo pode repercutir-se em decisões de planeamento incorretas, afetando a eficiência e a sustentabilidade do sistema como um todo.

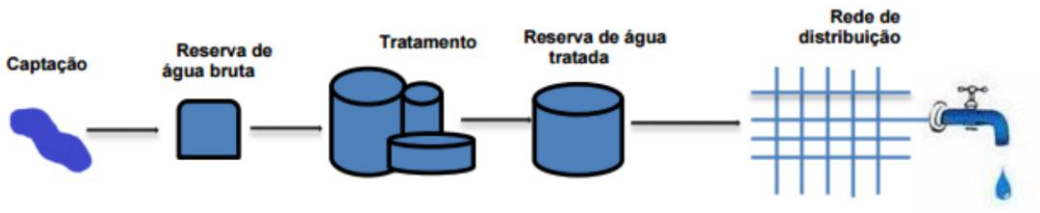


Figura 25: Ciclo urbano da água: da captação ao tratamento final.

A antiguidade dos contadores deve ser encarada não apenas como uma variável técnica, mas como um fator estruturante da qualidade e confiabilidade da gestão hídrica. A substituição orientada por dados, baseada em critérios analíticos robustos e inteligência organizacional, é essencial para garantir um serviço público transparente, resiliente e alinhado com os princípios da sustentabilidade e eficiência operativa.

6.4 Avaliação do Impacto do SMOTE nos Modelos de Classificação

Após a aplicação inicial dos modelos de ML (regressão logística, SVM e *random forest*) aos dados históricos da Tejo Ambiente, observou-se um desempenho significativamente limitado na deteção de anomalias, principalmente devido ao forte desbalanceamento das classes. Para mitigar este problema, foi aplicada a técnica SMOTE, que gera exemplos sintéticos da classe minoritária com base nos seus vizinhos mais próximos. O gráfico apresentado na Figura 26 ilustra comparativamente os resultados obtidos antes e depois da aplicação do SMOTE, com base no F1 Score, métrica que considera simultaneamente a precisão e a sensibilidade dos modelos.

Verifica-se que, após o balanceamento dos dados, todos os modelos apresentaram uma melhoria substancial no F1 *Score*, destacando-se o modelo *random forest*, que passou de um F1 *Score* de 0.16 para 0.61. Este resultado reforça a importância do pré-processamento e do equilíbrio das classes na aprendizagem supervisionada para a deteção de anomalias raras.

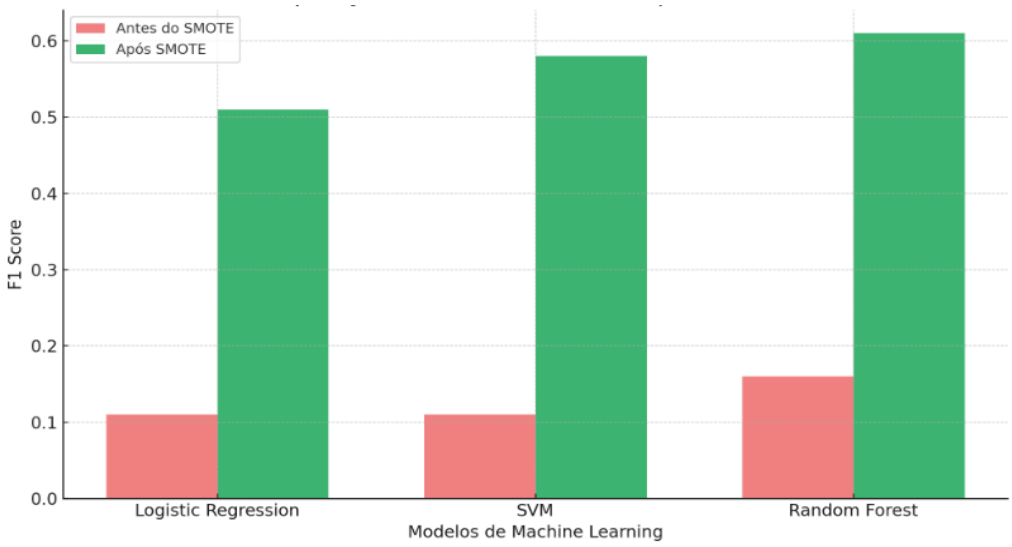


Figura 26: Comparação do F1 *score* antes e depois da aplicação do SMOTE para os modelos l*ogistic regression*, SVM e *random forest*.

Assim, conclui-se que a adoção do SMOTE representou um ganho considerável em termos de capacidade preditiva, oferecendo uma solução eficaz para contextos de elevada assimetria de classes, como é o caso da previsão de ruturas ou consumos anómalos no setor hídrico.

Com base na segmentação por freguesia, foram desenvolvidos *dashboards* analíticos em *Power* BI, focando indicadores operacionais como a idade média dos contadores, frequência de leituras, volume de consumo e percentagem de registos ausentes.

A comparação entre freguesias revelou disparidades significativas (Figura 27). Na Zona Industrial, com apenas 21 clientes, a idade média dos contadores é de 10,38 anos, mas mais de 66% dos registos apresentam ausência de leitura. Em contraste, Carregueiros, com 1.023 clientes, apresenta uma média etária de 16,69 anos nos contadores, sendo que 37% têm mais de 20 anos.

Estes dados permitiram mapear zonas críticas com risco elevado de anomalias, priorizar a substituição de contadores obsoletos com base em critérios objetivos e detetar padrões atípicos que podem estar associados a ligações irregulares ou medições deficientes.

Figura 27: Painéis de monitorização: comparação analítica entre freguesias com base nos contadores instalados e leituras registadas.

Complementarmente, foram aplicadas metodologias de análise estatística e exploratória, como PCA, análise de clusters e regressão múltipla. O PCA permitiu reduzir a dimensionalidade dos dados, identificando variáveis com maior impacto como idade dos contadores, zona geográfica e tipo de cliente. A análise de *clusters* possibilitou o agrupamento de perfis semelhantes de consumo e a segmentação geográfica com base em padrões de leitura. A regressão múltipla confirmou a existência de correlações significativas entre a antiguidade dos contadores e a probabilidade de leitura anómala.

Os *insights* estratégicos derivados desta abordagem incluem a necessidade de gestão diferenciada por freguesia, dada a heterogeneidade territorial observada; a urgência na digitalização da rede de medição, em virtude das elevadas taxas de leitura ausente; a valorização da tomada de decisão baseada em dados, com modelos preditivos que apoiam o planeamento de manutenções e substituições; e a possibilidade de replicar boas práticas de freguesias com melhor desempenho noutras zonas deficitárias.

A conjugação entre *dashboards* analíticos, modelos preditivos e técnicas estatísticas revelou-se uma abordagem integrada e eficaz para a identificação de zonas críticas, deteção de anomalias e otimização da gestão dos recursos.

O modelo *random forest* destacou-se como o mais robusto e fiável, enquanto o SMOTE demonstrou ser uma técnica fundamental para lidar com a assimetria de classes. Esta metodologia representa uma evolução significativa na capacidade analítica da Tejo Ambiente, promovendo ganhos concretos na sustentabilidade operacional e na gestão estratégica da rede de abastecimento de água.

6.5 Considerações Finais

Com base na aplicação de técnicas de BI, estatística avançada e algoritmos de ML ao contexto real da Tejo Ambiente, foi possível validar a eficácia de soluções analíticas na gestão hídrica inteligente.

Através de *dashboards* desenvolvidos em *Power* BI e do cruzamento com dados operacionais e sociodemográficos, foram identificadas zonas críticas com perdas elevadas, contadores obsoletos e possíveis anomalias de leitura. A aplicação dos modelos *Random Forest*, Regressão Logística e SVM confirmou a dificuldade em detetar anomalias devido ao forte desbalanceamento das classes. A técnica SMOTE permitiu melhorar drasticamente o desempenho preditivo, com o F1 Score a subir de 0.16 para 0.61 no melhor modelo.

Estes resultados comprovam o valor da integração de ciência de dados na gestão pública da água: promove decisões baseadas em evidência, aumenta a eficiência operacional e contribui para a sustentabilidade ambiental.

A abordagem é replicável noutros contextos territoriais e reforça a importância da transformação digital como motor de inovação no setor dos serviços ambientais.

7. CONCLUSÕES

Num contexto global fortemente impactado pela intensificação da crise climática, pela escassez crescente de recursos naturais e pela pressão sobre as infraestruturas urbanas, torna-se imperativo adotar modelos de gestão hídrica mais inteligentes, sustentáveis e orientados por dados. Este estudo teve como foco principal a aplicação de soluções de BI ao setor da água, com particular incidência na análise preditiva e na monitorização inteligente dos consumos, utilizando a plataforma *Power* BI.

O objetivo central consistiu no desenvolvimento de um sistema analítico capaz de transformar grandes volumes de dados operacionais, provenientes de plataformas como *AquaMatrix*, Primavera, *Filedoc*, telegestão e telemetria, em informação estruturada, visual e acionável para apoiar a tomada de decisão técnica e estratégica. A metodologia baseou-se na recolha rigorosa, pré-processamento e integração de dados históricos e atuais, complementada pela aplicação de algoritmos de ML supervisionado, análise estatística e visualizações interativas.

Entre os principais contributos e resultados obtidos destacam-se a segmentação inteligente de clientes, considerando variáveis como tipo de consumo, localização, idade dos contadores e histórico de anomalias; a identificação de contadores obsoletos com mais de 15 anos e com desvios de leitura superiores a 20%, comprometendo a fiabilidade da medição e a justiça tarifária; o mapeamento de potenciais ligações ausentes, estimadas em cerca de 3.000, a partir do cruzamento entre dados internos e os Censos 2021, o desenvolvimento de modelos preditivos para deteção de anomalias e previsão de consumos, com destaque para o modelo *random forest*, que demonstrou elevada eficácia após aplicação da técnica de balanceamento SMOTE; e a criação de *dashboards* interativos e funcionais em *Power* BI, capazes de integrar múltiplos indicadores operacionais e apoiar a definição de prioridades de investimento e substituição de equipamentos.

A análise de dados revelou-se, assim, um instrumento estratégico crucial para otimizar a gestão da rede, reduzir perdas, tanto aparentes como reais, melhorar a eficiência dos serviços e fomentar uma gestão pública mais transparente, resiliente e orientada para resultados. Comprovou-se que a inteligência analítica é mais do que uma ferramenta técnica: é um verdadeiro catalisador de inovação, sustentabilidade e transformação digital na administração pública.

Apesar dos resultados promissores, o estudo enfrentou algumas limitações importantes. A incompletude e dispersão dos dados operacionais dificultaram a construção de uma base unificada e coesa, a falta de normalização e padronização de dados históricos, especialmente nos registos manuais e nos dados relativos à antiguidade dos contadores, gerou desafios adicionais ao nível da qualidade da informação, a integração limitada com sistemas em tempo real, como a telegestão e a telemetria, impediu a criação de *dashboards* com atualização contínua e a ausência de feedback sistemático por parte dos utilizadores finais limitou a capacidade de adaptação iterativa da solução desenvolvida.

Ainda assim, o estudo representa um contributo prático e científico relevante para a literacia analítica na gestão pública local. Demonstra, de forma concreta, como ferramentas de BI e técnicas de ML podem ser integradas num modelo operacional aplicável ao setor da água. A solução proposta é escalável, replicável e pode ser adaptada a outras entidades gestoras, mesmo em contextos territoriais e tecnológicos distintos.

No futuro, e em alinhamento com os princípios da gestão inteligente da água, destaca-se como recomendação estratégica a integração com sistemas IoT, sensores de caudal e pressão para deteção automática de fugas e ruturas, o alargamento da análise a variáveis complementares como qualidade da água, pressão e perdas físicas por zona, e o desenvolvimento de alertas inteligentes para o consumidor final, promovendo uma gestão mais participativa e consciente do recurso hídrico.

Em suma, esta investigação comprova que o BI, aliado à ciência de dados, constitui um vetor de transformação essencial para o setor hídrico. A combinação entre dados, tecnologia e decisão representa um caminho robusto e promissor para enfrentar os desafios da escassez, garantir equidade no acesso à água e assegurar a sustentabilidade ambiental das gerações futuras.

REFERÊNCIA BIBLIOGRÁFICAS

REFERÊNCIAS ACADÉMICAS E CIENTÍFICAS

Ahrens, J., Geveci, B., & Law, C. (2014). ParaView: An End-User Tool for Large-Data Visualization. In Visualization Handbook (pp. 717–732). Elsevier.

Cao, L. (2017). Data Science: A Comprehensive Overview. ACM Computing Surveys, 50(3), 43.

Chen, H., Chiang, R. H., & Storey, V. C. (2019). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165–1188.

Davenport, T. H., & Harris, J. G. (2017). Competing on Analytics: The New Science of Winning. Harvard Business Review Press.

Dasgupta, S., Datta, R., & Ghosh, S. (2022). Ethical Considerations in Big Data Analytics. Journal of Business Ethics, 179(3), 555–572.

Gandomi, A., & Haider, M. (2022). Beyond the Hype: Big Data Concepts, Methods, and Analytics. International Journal of Information Management, 63, 102141.

Géry, M., Roustant, P., & Clémençon, M. (2013). Data Mining and Statistical Inference. John Wiley & Sons.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hofmann, E., & Rutschmann, E. (2023). Sustainable Data Analytics: Towards Responsible and Environmentally Friendly AI. Business Process Management Journal, 29(3), 924–949.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and Its Technical Challenges. Communications of the ACM, 57(7), 86–94.

Knaflic, C. N. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley.

Marini, F., Facchinetti, S., & Ebecken, N. (2023). O impacto da inteligência artificial e da aprendizagem automática em dashboards: Uma revisão da literatura. Journal of Business Intelligence, 56(2), 243–260.

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.

Pan, Z., & Ghani, J. (2023). Real-time Data Dashboards for Business Intelligence: A Comprehensive Review and Taxonomy. ACM Computing Surveys, 56(2), 1–38.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.

Ren, J., Hu, D., & Schwabe, G. (2023). Leveraging Data Analytics for Sustainable Business Transformation. Business Horizons, 66(2), 241–251.

Stocker, M., & Korak, D. (2024). Data-Driven Decision Making with Modern Dashboards: A Practical Guide for Managers. Routledge.

Wang, L., Li, X., & Zhuang, Y. (2024). Explainable Artificial Intelligence for Trustworthy Data Analytics. IEEE Transactions on Knowledge and Data Engineering, 36(2), 443–457.

Yau, N. (2021). Data Visualization: A Practical Introduction. Harvard University Press.

Zheng, Y., Ardolino, M., & Bacchetti, A. (2024). Prescriptive Analytics in Manufacturing: A Systematic Literature Review. Computers & Industrial Engineering, 149, 107810.

FONTES ONLINE E RELATÓRIOS TÉCNICOS

Tableau (2024). Dashboard Design Best Practices. Disponível em: https://help.tableau.com/current/pro/desktop/en-us/dashboards_best_practices.htm

Microsoft (2023). Power BI: Creating Effective Dashboards. Disponível em: https://learn.microsoft.com/en-us/power-bi/create-reports/service-dashboard-create

DataCamp (2022). Introduction to Data Visualization with Python. Disponível em: https://www.datacamp.com/tracks/data-visualization-with-python

BIGroup (2024). Tendências em Dashboards para 2024. Disponível em: https://pixelplex.io/blog/business-intelligence-trends-2024/

FONTES INSTITUCIONAIS E TÉCNICAS CONSULTADAS

Tejo Ambiente – Empresa Intermunicipal de Ambiente do Médio Tejo, S.A.
https://www.tejoambiente.pt

ERSAR – Entidade Reguladora dos Serviços de Águas e Resíduos
https://www.ersar.pt
RASARP – Relatório Anual do Setor de Águas e Resíduos em Portugal

INE – Instituto Nacional de Estatística
https://www.ine.pt
Censos 2021

APA – Agência Portuguesa do Ambiente
https://apambiente.pt
Indicadores de eficiência no uso da água

EPAL – Empresa Portuguesa das Águas Livres
https://www.epal.pt
Soluções AQUAmatrix®, Billmeter®, Waterbeep

JANZ – SmartIO
https://www.janz.pt
Contadores inteligentes e soluções LoRaWAN

Siconia (Diehl Metering)
https://www.diehl.com/metering
Tecnologia de contadores ultrassónicos

Filedoc – Gestão documental
https://www.filedoc.pt

Primavera BSS
https://www.primaverabss.com

Microsoft Power BI
https://powerbi.microsoft.com
Ferramenta utilizada para construção dos dashboards analíticos

Python Software Foundation
https://www.python.org
Linguagem de programação utilizada nos modelos

Scikit-learn Documentation
https://scikit-learn.org
Implementação dos algoritmos de Machine Learning (Random Forest, SMOTE, Logistic Regression, SVM)

Google Colab
https://colab.research.google.com
Ambiente de desenvolvimento na cloud utilizado para treino, validação e visualização dos modelos preditivos

ChatGPT (OpenAI)
https://chat.openai.com
Ferramenta de apoio técnico e metodológico, usada para esclarecimento de dúvidas, organização de ideias, validação de código Python e estruturação da redação científica da tese

BIGroup – Tendências em Dashboards para 2024
https://pixelplex.io/blog/business-intelligence-trends-2024/

Microsoft Learn – Power BI Documentation
https://learn.microsoft.com/en-us/power-bi/

Tableau – Dashboard Best Practices
https://help.tableau.com/current/pro/desktop/en-us/dashboards_best_practices.htm

DataCamp – Visualização de Dados com Python
https://www.datacamp.com/tracks/data-visualization-with-python

ANEXOS

I. COLAB

Criação de Modelos Preditivos

Parte I - Préprocessamento e Exploração dos Dados

```
# Instalações iniciais
!pip install imbalanced-learn
!pip install pandas numpy scikit-learn
```

```
Requirement already satisfied: imbalanced-learn in /usr/local/lib/python3.11/dist-packages (0.13.0)
Requirement already satisfied: numpy<3,>=1.24.3 in /usr/local/lib/python3.11/dist-packages (from imbalanced-learn)
(2.0.2) Requirement already satisfied: scipy<2,>=1.10.1 in /usr/local/lib/python3.11/dist-packages (from imbalanced-
learn) (1.15.3) Requirement already satisfied: scikit-learn<2,>=1.3.2 in /usr/local/lib/python3.11/dist-packages (from
imbalanced-learn) (1.6.1) Requirement already satisfied: sklearn-compat<1,>=0.1 in /usr/local/lib/python3.11/dist-
packages (from imbalanced-learn) (0.1.3) Requirement already satisfied: joblib<2,>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from imbalanced-learn) (1.5.1) Requirement already satisfied:
threadpoolctl<4,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from imbalanced-learn) (3.6.0) Requirement already
satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages
(2.0.2) Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.11/dist-packages (1.6.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas)
(2.9.0.post0) Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas)
(2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas)
(2025.2) Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-packages (from scikit-
learn) (1.15.3) Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from
scikit-learn) (1.5.1) Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn) (3.6.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas)
(1.17.0)
```

```
# Importações
necessárias
import pandas
as pd import
numpy as np
from sklearn.model_selection import
train_test_split from sklearn.linear_model
import LogisticRegression from
sklearn.ensemble import
RandomForestClassifier from sklearn.svm
import SVC
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix,
classification_report from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import LabelEncoder, StandardScaler # Para pré-
processamento import warnings


# Ignorar warnings para melhor legibilidade
na saída warnings.filterwarnings('ignore')
```

Carregamento e visualização dos dados

```
# Carregar os dados
df = pd.read_csv('/content/sample data/AnaTeixeira BD.csv')

# Análise Inicial do DataFrame (importante para entender
os dados) print("Informações iniciais do DataFrame:")
df.info()
print("\nPrimeiras 5 linhas do DataFrame:")
print(df.head())
```

```
Informações iniciais do DataFrame:
<class
'pandas.core.frame.DataFr
```

```
ame'>  RangeIndex:  88351
entries, 0 to 88350 Data
columns     (total     62
columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   SubGrupo                              88351 non-null  int64
 1   Data Entidade                         88351 non-null  object
 2   Situação da Entidade                  88351 non-null  object
 3   Número Contrato                       88351 non-null  int64
 4   Data Contrato                         88351 non-null  object
 5   Tipo de Fatura                        88351 non-null  object
 6   Data Vigência                         88115 non-null  object
 7   Último Dia Estimado                   28256 non-null  object
 8   Último dia Real                       51605 non-null  object
 9   Último Dia Facturado                  51789 non-null  object
 10  Data Última Factura                   51781 non-null  object
 11  Data Próxima Factura                  51775 non-null  object
 12  Data Situação                         88351 non-null  object
 13  Estimativa                            77301 non-null  float64
 14  Consumo Médio Real                    51476 non-null  float64
 15  Grupo Contador                        70417 non-null  float64
 16  Nº Contador                           70417 non-null  float64
 17  Selos do Contador                     7665 non-null   float64
 18  Data Instalação                       70417 non-null  object
 19  Número Fabricante                     70417 non-null  object
 20  Fabricante                            70417 non-null  object
 21  Modelo do Contador                    70417 non-null  object
 22  Calibre                               70417 non-null  float64
 23  Freguesia A#                          88266 non-null  object
 24  Concelho                              88266 non-null  object
 25  Cód#Postal A#                         88115 non-null  object
 26  Des#Postal A#                         88115 non-null  object
 27  Nº Prédio                             88351 non-null  int64
 28  Nº da Instalação                      86445 non-null  float64
 29  Ramal                                 88351 non-null  int64
 30  ZMC                                   88351 non-null  object
 31  Classe Consumo                        88351 non-null  object
 32  Tipo Consumo                          88351 non-null  object
 33  Situação                              88351 non-null  object
 34  Zona Leiutura                         88351 non-null  int64
 35  Área Leitura                          88351 non-null  int64
 36  Local                                 88351 non-null  int64
 37  Nº Seq# Leitura                       88351 non-null  int64
 38  Calibre Local                         88351 non-null  int64
 39  Modalidade Pagamento                  88351 non-null  object
 40  Data Últ# Fact# Emitida               79228 non-null  object
 41  Situação Fecho                        464 non-null    object
 42  Tipo Utilização                       88351 non-null  object
 43  Data Última Leitura Facturada         68983 non-null  object
 44  Valor Última Leitura Facturada        68983 non-null  float64
 45  Anomalia de Leitura Facturada         88351 non-null  object
 46  Data Última Leitura Empresa           72371 non-null  object
 47  Valor Última Leitura Empresa          72371 non-null  float64
 48  Anomalia de Leitura Empresa           88351 non-null  object
 49  Data Última Leitura Cliente           67968 non-null  object
 50  Valor Última Leitura Cliente          67968 non-null  float64
```

```
# Pré-processamento de Dados

# Tratamento de Datas: Converter colunas de data para
datetime date_columns = [
    'Data Entidade', 'Data Contrato', 'Data Vigência', 'Último Dia
    Estimado', 'Último dia Real', 'Último Dia Facturado', 'Data Última
    Factura',
    'Data Próxima Factura', 'Data Situação', 'Data
    Instalação', 'Data Últ# Fact# Emitida', 'Data Última
    Leitura Facturada', 'Data Última Leitura Empresa',
    'Data Última Leitura Cliente', 'Data da Situação do
    Local', 'Data Actualização de Contactos'
]


for col in
    date_colum
    ns: if col
```

```
    in
    df.columns
    :
        # Usar errors='coerce' para converter valores inválidos para NaT (Not
        a Time) df[col] = pd.to_datetime(df[col], errors='coerce')



# Engenharia de Features de Tempo
# Calcular a diferença de dias desde uma data de referência (pode ser a data mais recente
nos dados) # Ou, calcular tempo entre leituras
df['Tempo_Ultima_Leitura_Facturada'] = (df['Data Última Factura'] - df['Data Última Leitura Facturada']).dt.days
df['Tempo_Ultima_Leitura_Empresa'] = (df['Data Última Factura'] - df['Data Última Leitura Empresa']).dt.days
df['Tempo_Ultima_Leitura_Cliente'] = (df['Data Última Factura'] - df['Data Última Leitura Cliente']).dt.days



# CRIAÇÃO DA VARIÁVEL ALVO (Y) - ESTE É O PASSO MAIS IMPORTANTE!

# Premissas para a criação de 'y_anomalia':
# 1. Uma rotura geralmente resulta num AUMENTO SÚBITO e INESPERADO do consumo.
# 2. Podemos usar o 'Consumo Mês Anterior' como base e comparar com o 'Consumo Médio Real' ou 'Valor Última Leitura
Facturada'. # 3. As colunas 'Anomalia de Leitura Facturada', 'Anomalia de Leitura Empresa', 'Anomalia de Leitura
Cliente'
#    são CADASTRADAS, mas precisam de ser inspecionadas para ver quais valores indicam
uma rotura. #    Vamos inspecionar os valores únicos dessas colunas para identificar
termos relevantes.

print("\nValores únicos nas colunas de Anomalia de Leitura:")
print("Anomalia de Leitura Facturada:", df['Anomalia de Leitura
Facturada'].unique()) print("Anomalia de Leitura Empresa:", df['Anomalia de
Leitura Empresa'].unique()) print("Anomalia de Leitura Cliente:", df['Anomalia de
Leitura Cliente'].unique())

# --- ESTRATÉGIA PARA CRIAR 'y' (Anomalia/Rotura) ---
# foi preciso ajustar os critérios abaixo com base no que  encontramos nos `unique()` e
`value_counts()` # das colunas de anomalia e no conhecimento do domínio.

# 6.1. Tentativa 1: Baseado em um LIMIAR de consumo anormalmente alto (ex: 3x o consumo médio ou
mês anterior) # Substituir NaNs em colunas de consumo por 0 ou a média, dependendo da estratégia
df['Consumo  Mês  Anterior'].fillna(0,
inplace=True) df['Valor Última Leitura
Facturada'].fillna(0, inplace=True) df['Consumo
Médio Real'].fillna(0, inplace=True)

# Criar uma feature de diferença de consumo para
identificar picos
df['Diferenca_Consumo'] = df['Valor Última Leitura Facturada'] - df['Consumo Mês Anterior']

# Definir um critério para anomalia:
# Exemplo 1: Consumo atual é X vezes maior que o consumo do mês anterior (se o anterior
não for zero) # Exemplo 2: Ocorre uma diferença de consumo muito alta E o consumo do mês
anterior não era nulo.
# Exemplo 3: Se houver termos específicos nas 'Anomalia de Leitura' que indiquem rotura (e.g.,
'Vazamento') #   Vamos assumir 'VAZAMENTO' como um termo, mas você precisa VERIFICAR os seus
dados reais!

# Para esta correção, criou-se 'y_anomalia' combinando um limiar
de consumo # e a presença de termos de anomalia de leitura se
existirem.
# ASSUMIMOS que um 'consumo muito alto' E/OU uma 'anomalia de leitura' indicam uma rotura.

# Primeiro, vamos identificar anomalias textuais que possam ser
roturas. # Ajuste estes termos conforme o que encontrar nas saídas de
.unique()
termos_rotura = ['CONSUMO ANORMAL', 'VAZAMENTO', 'LEITURA ALTA', 'CONSUMO EXCESSIVO'] # Adicione ou remova termos
df['Anomalia_Leitura_Combinada'] = df['Anomalia de Leitura Facturada'].astype(str) + ' ' + \
                                   df['Anomalia de Leitura Empresa'].astype(str) + ' ' +
                                   \ df['Anomalia de Leitura Cliente'].astype(str)

df['y_anomalia'] = 0 # Inicializa todas as linhas como 'não anomalia' (0)

# Condição 1: Aumento percentual significativo no consumo (evitar
divisão por zero) # Um aumento de 200% (3x o consumo anterior) pode ser
```

```
um sinal
limiar_aumento_percentual = 2.0 # Significa 200% de aumento (novo consumo é 3x o anterior)
df['Percentual_Aumento_Consumo'] = df.apply(lambda row: (row['Valor Última Leitura Facturada'] - row['Consumo Mês Anterior'])
/ row['Co
                                            if row['Consumo Mês Anterior'] > 0 else np.nan, axis=1)

df.loc[df['Percentual_Aumento_Consumo'] > limiar_aumento_percentual, 'y_anomalia'] = 1

# Condição 2: Consumo absoluto muito alto (mesmo que o mês anterior fosse
zero ou baixo) # Definimos um limiar absoluto para consumo (ex: 500 unidades
cúbicas)
limiar_consumo_absoluto_alto = 500 # Ajuste este valor conforme o histórico de roturas da sua
empresa df.loc[df['Valor Última Leitura Facturada'] > limiar_consumo_absoluto_alto,
'y_anomalia'] = 1

# Condição 3: Anomalias de leitura que indicam rotura (baseado nos
termos_rotura) # Combine com as anomalias de leitura se a descrição
delas realmente indicar rotura for termo in termos_rotura:
    df.loc[df['Anomalia_Leitura_Combinada'].str.contains(termo, case=False, na=False), 'y_anomalia'] = 1


# --- VERIFICAÇÃO CRÍTICA DA NOVA VARIÁVEL ALVO ---
y = df['y_anomalia']
print(f"\nValores únicos e contagens da nova variável alvo 'y_anomalia':")
print(np.unique(y, return_counts=True))

if np.unique(y).size < 2:
    print("\nAVISO: A variável alvo 'y_anomalia' ainda tem menos de 2 classes. Ajuste os critérios de
    criação!") print("Por favor, reveja as condições para `y_anomalia` na Célula 6.")
    # Se isso acontecer, o restante do código para classificação ainda falhará ou
    será inútil. # Você precisará ajustar os limiares e termos de anomalia na
    Célula 6.
else:
    print("\nSucesso! 'y_anomalia' tem múltiplas classes. Podemos prosseguir com a classificação.")
```

```
 Valores únicos nas colunas de Anomalia de Leitura:
 Anomalia de Leitura Facturada: ['0 - SEM ERRO' ' - ' '2 - CONSUMO = 0 E
  MÉDIA <> 0' '10 - DESVIO DE MÉDIA' '11 - DESVIO DE MÉDIA' '12 - DESVIO
  DE MÉDIA'
  '31 - LEITURA INFERIOR À ULTIMA' '3 -
  DESVIO DE MÉDIA' '1 - CONSUMO <> 0 E MÉDIA
  = 0' '5 - DESVIO DE MÉDIA' '4 - DESVIO DE
  MÉDIA' '6 - DESVIO DE MÉDIA'
  '33 - INEXISTÊNCIA DE PRIMEIRA
  LEITURA/LEITURA BASE' '7 - DESVIO DE
  MÉDIA' '39 - CONSUMOS NEGATIVOS'
  '32 - LOCAL C/ ÁGUA FECHADA E C/ CONSUMO'
  '40 - IMPOSSÍVEL CALCULAR DIAS DE ÁGUA
  FECHADA/ABERTA' '41 - CLIENTE SEM MÉDIA DE
  CONSUMO'
  '45 - CLIENTE SEM DATA DE INÍCIO DE VIGÊNCIA']
 Anomalia de Leitura Empresa: ['0 - SEM ERRO' ' - ' '2 - CONSUMO = 0 E
  MÉDIA <> 0' '31 - LEITURA INFERIOR À ULTIMA' '10 - DESVIO DE MÉDIA'
  '11 - DESVIO DE MÉDIA' '43 - LEITURA ANTERIOR À BASE'
  '32 - LOCAL C/ ÁGUA FECHADA E C/ CONSUMO' '12 - DESVIO
  DE MÉDIA' '1 - CONSUMO <> 0 E MÉDIA = 0' '5 - DESVIO
  DE MÉDIA'
  '4 - DESVIO DE MÉDIA' '6 - DESVIO DE MÉDIA' '3 -
  DESVIO DE MÉDIA' '39 - CONSUMOS NEGATIVOS'
  '33 - INEXISTÊNCIA DE PRIMEIRA LEITURA/LEITURA BASE'
  '41 - CLIENTE SEM MÉDIA DE CONSUMO' '7 - DESVIO
  DE MÉDIA' '45 - CLIENTE SEM DATA DE INÍCIO DE
  VIGÊNCIA'
  '40 - IMPOSSÍVEL CALCULAR DIAS DE ÁGUA FECHADA/ABERTA']
 Anomalia de Leitura Cliente: ['0 - SEM ERRO' ' - ' '10 - DESVIO
  DE MÉDIA' '2 - CONSUMO = 0 E MÉDIA <> 0' '4 - DESVIO DE
  MÉDIA'
  '31 - LEITURA INFERIOR À ULTIMA' '3 -
  DESVIO DE MÉDIA' '1 - CONSUMO <> 0 E MÉDIA
  = 0' '12 - DESVIO DE MÉDIA'
  '6 - DESVIO DE MÉDIA' '5 -
  DESVIO DE MÉDIA' '43 - LEITURA
  ANTERIOR À BASE'
  '33 - INEXISTÊNCIA DE PRIMEIRA LEITURA/LEITURA BASE'
  '11 - DESVIO DE MÉDIA' '32 - LOCAL C/ ÁGUA FECHADA E
```

```
      C/ CONSUMO' '40 - IMPOSSÍVEL CALCULAR DIAS DE ÁGUA
      FECHADA/ABERTA'
      '39 - CONSUMOS NEGATIVOS' '7 - DESVIO
      DE MÉDIA' '45 - CLIENTE SEM DATA DE
      INÍCIO DE VIGÊNCIA']

     Valores únicos e contagens da nova variável alvo 'y_anomalia':
     (array([0, 1]), array([47364, 40987]))

     Sucesso! 'y_anomalia' tem múltiplas classes. Podemos seguir com a classificação.


# Seleção de Features (X) e Tratamento de Nulos nas Features

# Colunas que serão usadas como features (X) - selecione as mais relevantes
# Remoção IDs, colunas de data originais (usaremos as features de
tempo derivadas) # e a coluna alvo que acabamos de criar.
features_to_exclude = [
    'SubGrupo', 'Data Entidade', 'Data Contrato', 'Data
    Vigência', 'Último Dia Estimado', 'Último dia Real',
    'Último Dia Facturado', 'Data Última Factura', 'Data
    Próxima Factura', 'Data Situação', 'Número Contrato',
    'Nº Contador', 'Número Fabricante', 'Nº Prédio', 'Nº da
    Instalação', 'Ramal', 'Nº Seq# Leitura', 'Local',
    'Data Instalação', 'Data Últ# Fact# Emitida',
    'Data Última Leitura Facturada', 'Data Última
    Leitura Empresa', 'Data Última Leitura Cliente',
    'Data da Situação do Local',
    'Data Actualização de Contactos', 'y_anomalia', # Excluir a variável alvo
    'Anomalia_Leitura_Combinada', # Esta foi usada para criar y, mas pode ser excluída de X se codificar os termos originais
    'Percentual_Aumento_Consumo', # Já é uma feature criada, pode ser incluída ou não. Se tiver muitos NaNs, tratar.
    # Colunas com muitos NaNs que podem ser difíceis de imputar ou não muito
    relevantes 'Selos do Contador', 'Situação Fecho', 'Ft Eletrónica'
]


# Selecionar as colunas que não estão na lista de exclusão
X = df.drop(columns=[col for col in features_to_exclude if col in df.columns], errors='ignore')

# Tratamento de NaNs em X (features)
# Para colunas numéricas, imputar com a média ou mediana
# Para colunas categóricas, imputar com a moda ou uma categoria
'Desconhecido' for col in X.columns:
    if X[col].dtype == 'float64' or X[col].dtype == 'int64':
        X[col].fillna(X[col].median(), inplace=True) # Mediana é
    robusta a outliers elif X[col].dtype == 'object':
        X[col].fillna('UNKNOWN', inplace=True) # Para categóricas, preencher com 'UNKNOWN'


# Codificação de Variáveis Categóricas em X
# Identificar colunas categóricas (object ou category
dtype) categorical_cols =
X.select_dtypes(include=['object']).columns

# Usar One-Hot Encoding para evitar problemas de ordenação
X = pd.get_dummies(X, columns=categorical_cols, drop_first=True) # drop_first evita multicolinearidade



# Normalização/Padronização de Features Numéricas (Importante para SVM e Regressão
Logística) scaler = StandardScaler()
# Aplicar o scaler apenas às colunas numéricas que não foram resultado de one-
hot encoding # (isto é, as colunas numéricas originais e as features de tempo
criadas)
numeric_cols = X.select_dtypes(include=['int64', 'float64']).columns
X[numeric_cols] = scaler.fit_transform(X[numeric_cols])



# Divisão dos dados em treino e teste (COM STRATIFY)
# Certifique-se de que y foi definido corretamente na Célula 6
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

print(f"\nTamanho de X_train: {X_train.shape}, y_train:
{y_train.shape}") print(f"Tamanho de X_test: {X_test.shape},
```

```
y_test: {y_test.shape}") print(f"Valores únicos e contagens em
y_train (após split):") print(np.unique(y_train,
return_counts=True))
```

```
Tamanho de X_train: (61845, 2938), y_train: (61845,)
Tamanho de X_test: (26506, 2938),
y_test: (26506,) Valores únicos e
contagens em y_train (após split):
(array([0, 1]), array([33154, 28691]))
```

```
# Balanceamento dos dados com SMOTE (se a classe minoritária for muito pequena)
# Só aplicamos SMOTE se houver mais de uma classe e se houver um desequilíbrio significativo.


# Se a contagem da classe 1 em y_train for muito baixa, SMOTE é essencial.
if np.unique(y_train).size > 1 and np.unique(y_train, return_counts=True)[1].min() > 1: # SMOTE precisa de pelo menos 2
    amostras na min print("\nAplicando SMOTE para balancear os dados de treino...")
    smote = SMOTE(random_state=42)
    X_resampled, y_resampled = smote.fit_resample(X_train, y_train)
    print(f"Tamanho de X_resampled: {X_resampled.shape}, y_resampled:
    {y_resampled.shape}") print(f"Valores únicos e contagens em y_resampled (após
    SMOTE):") print(np.unique(y_resampled, return_counts=True))
else:
    print("\nSMOTE não aplicado: Não há desequilíbrio ou não há classes suficientes para o SMOTE.")
    X_resampled, y_resampled = X_train, y_train # Usar os dados originais de treino
```

```
Aplicando SMOTE para balancear os dados de treino...
Tamanho de X_resampled: (66308, 2938), y_resampled:
(66308,) Valores únicos e contagens em y_resampled
(após SMOTE): (array([0, 1]), array([33154,
33154]))
```

```
# Treinar e avaliar
os modelos models =
{
    "Logistic   Regression":   LogisticRegression(max_iter=2000,   solver='liblinear'),   #
    Aumentado   max_iter   "SVM":   SVC(probability=True),   #   probability=True   permite
    predict_proba      (útil      para      certas      análises)      "Random      Forest":
    RandomForestClassifier(n_estimators=100, random_state=42)
}


results = []

for name, model in
    models.items():
    print(f"\n--- Treinando
    {name} ---")
    model.fit(X_resampled, y_resampled) # Usar os dados balanceados para treino

    y_pred = model.predict(X_test) # Prever no conjunto de TESTE (não
    balanceado) acc = accuracy_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred) # F1-Score é crucial para classes desequilibradas

    results.append((name, acc, f1))

    print(f"{name} - Acurácia: {acc:.4f}")
    print(f"{name} - F1-Score: {f1:.4f}")
    print(f"Matriz de Confusão para {name}:\n{confusion_matrix(y_test, y_pred)}")
    print(f"Relatório de Classificação para {name}:\n{classification_report(y_test, y_pred, zero_division=0)}") #
    zero_division=0 para

results_df = pd.DataFrame(results, columns=["Modelo", "Acurácia",
"F1-Score"]) print("\n--- Resultados Finais dos Modelos ---")
print(results_df)
```

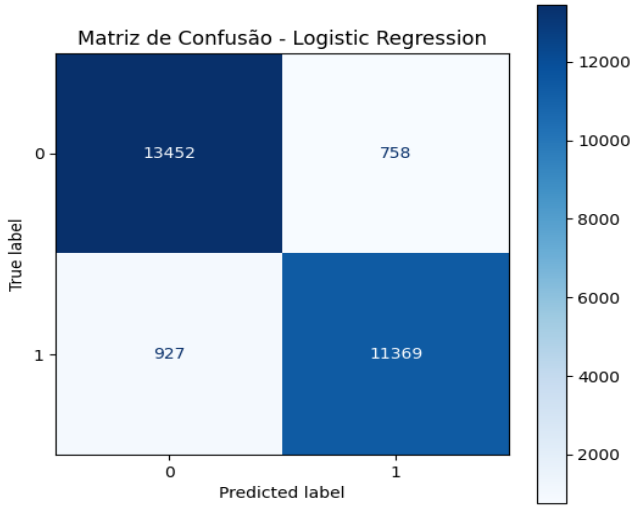
```
--- Treinando Logistic
Regression --- Logistic
Regression - Acurácia: 0.9364
Logistic Regression - F1-Score:
0.9310 Matriz de Confusão para
Logistic Regression: [[13452
          758]
 [  927 11369]]
```

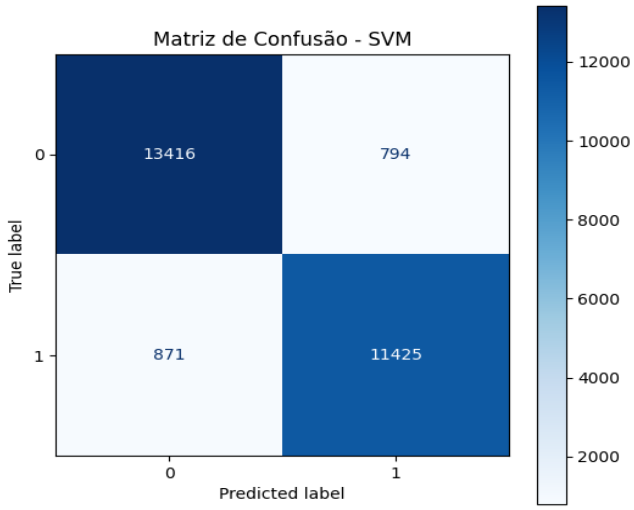
```
Relatório de Classificação para Logistic Regression:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94     14210
           1       0.94      0.92      0.93     12296

    accuracy                           0.94     26506
   macro avg       0.94      0.94      0.94     26506
weighted avg       0.94      0.94      0.94     26506


---
Treinando
SVM ---
SVM -
Acurácia:
0.9372 SVM
- F1-
Score:
0.9321
Matriz de
Confusão para
SVM: [[13416794]
 [  871 11425]]
Relatório de Classificação para SVM:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94     14210
           1       0.94      0.93      0.93     12296

    accuracy                           0.94     26506
   macro avg       0.94      0.94      0.94     26506
weighted avg       0.94      0.94      0.94     26506


--- Treinando Random
Forest --- Random
Forest - Acurácia:
0.9893 Random Forest
- F1-Score: 0.9885
Matriz de Confusão para Random Forest:
[[14020   190]
 [   93 12203]]
Relatório de Classificação para Random Forest:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     14210
           1       0.98      0.99      0.99     12296

    accuracy                           0.99     26506
   macro avg       0.99      0.99      0.99     26506
weighted avg       0.99      0.99      0.99     26506


--- Resultados Finais dos Modelos ---
                Modelo  Acurácia  F1-Score
0  Logistic Regression  0.936429  0.931008
1                  SVM  0.937184  0.932082
2        Random Forest  0.989323  0.988537
```

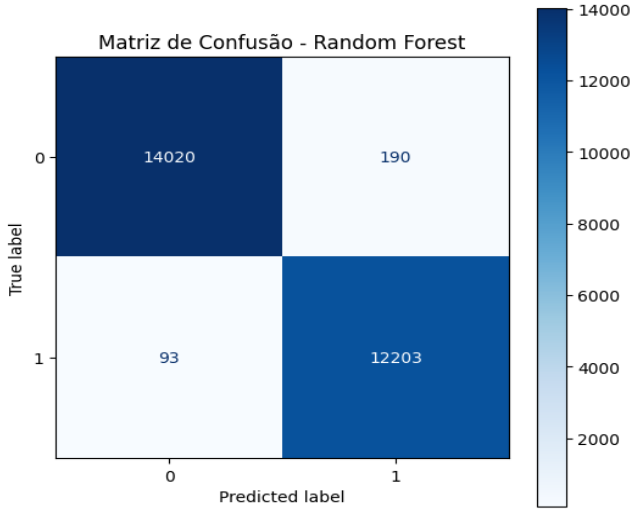
```
import
matplotlib.pyplot as
plt import seaborn as
sns
from sklearn.metrics import

ConfusionMatrixDisplay print("Visualização

das Matrizes de Confusão:")

for name, model in
    models.items(): #
    Prever no conjunto
    de TESTE y_pred =
    model.predict(X_tes
    t)
```

```
fig, ax = plt.subplots(figsize=(6, 6))
ConfusionMatrixDisplay.from_estimator(model, X_test, y_test, cmap=plt.cm.Blues,
ax=ax) ax.set_title(f'Matriz de Confusão - {name}')
plt.show()
```

```
Visualização das Matrizes de Confusão:
```

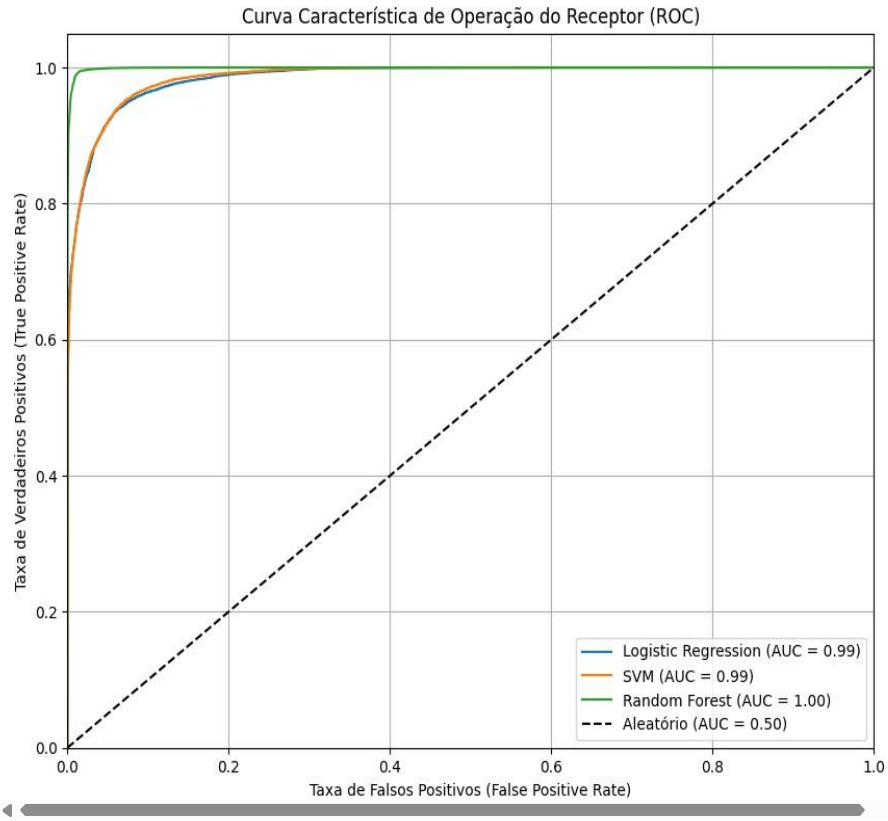






```
from sklearn.metrics import roc_curve, auc

print("\nVisualização das Curvas ROC:")


plt.figure(figsize=(10, 8))

for name, model in models.items():
    if hasattr(model, "predict_proba"): # Verificamos se o modelo tem predict_proba (nem todos têm)
        y_prob = model.predict_proba(X_test)[:, 1] # Probabilidade da classe positiva (1)
        fpr, tpr, thresholds = roc_curve(y_test, y_prob)
        roc_auc = auc(fpr, tpr)
        plt.plot(fpr, tpr, label=f'{name} (AUC = {roc_auc:.2f})')
    else:
        print(f"Modelo {name} não suporta predict_proba para Curva ROC.")

plt.plot([0, 1], [0, 1], 'k--', label='Aleatório (AUC = 0.50)')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Taxa de Falsos Positivos (False Positive Rate)')
plt.ylabel('Taxa de Verdadeiros Positivos (True Positive Rate)')
plt.title('Curva Característica de Operação do Receptor (ROC)')
plt.legend(loc="lower right")
plt.grid(True)
plt.show()
```

```
Visualização das Curvas ROC:
```




```
from sklearn.model_selection import GridSearchCV

print("\nOtimização de Hiperparâmetros com GridSearchCV:")

# Exemplo para Random Forest
# Definir o range de hiperparâmetros a testar
param_grid_rf = {
    'n_estimators': [50, 100, 200], # Número de árvores
    'max_depth': [None, 10, 20],    # Profundidade máxima da árvore
    'min_samples_split': [2, 5],    # Mínimo de amostras para dividir um nó
    'class_weight': [None, 'balanced'] # Ajusta o peso das classes
}
```

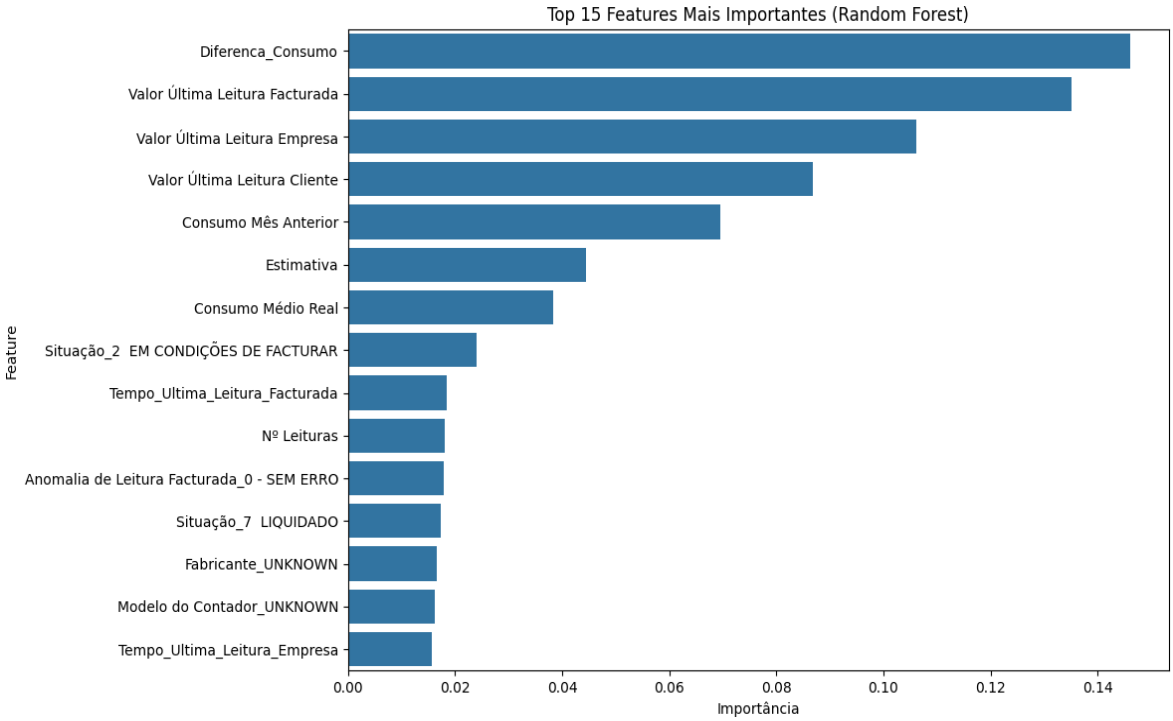
```
# Criar o modelo base
rf_model = RandomForestClassifier(random_state=42)

# Criar o GridSearchCV
grid_search_rf = GridSearchCV(estimator=rf_model, param_grid=param_grid_rf,
                              scoring='f1', # Otimizar para F1-Score devido ao desequilíbrio
                              cv=5, verbose=2, n_jobs=-1) # cv=5 para validação cruzada, n_jobs=-1 para usar todos os núcleos

# Treinar o GridSearchCV (usando os dados balanceados para treinamento) print("\nIniciando GridSearchCV para Random Forest...")
grid_search_rf.fit(X_resampled, y_resampled)

print(f"\nMelhores parâmetros para Random Forest: {grid_search_rf.best_params_}")
print(f"Melhor F1-Score na validação cruzada para Random Forest: {grid_search_rf.best_score_:.4f}")

# Avaliar o melhor modelo encontrado no conjunto de teste original
best_rf_model = grid_search_rf.best_estimator_
y_pred_best_rf = best_rf_model.predict(X_test)

print("\nResultados do Melhor Random Forest no conjunto de Teste (otimizado):")
print(f"Acurácia: {accuracy_score(y_test, y_pred_best_rf):.4f}")
print(f"F1-Score:  {f1_score(y_test,  y_pred_best_rf):.4f}")
print(f"Matriz de Confusão:\n{confusion_matrix(y_test, y_pred_best_rf)}")
print(f"Relatório de Classificação:\n{classification_report(y_test, y_pred_best_rf, zero_division=0)}")

# Podemos repetir para Logistic Regression e SVM, ajustando o param_grid. # Exemplo para
Logistic Regression:
# param_grid_lr = {'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}
# grid_search_lr = GridSearchCV(LogisticRegression(solver='liblinear', random_state=42), param_grid_lr, scoring='f1', cv=5, verbose=2,
# grid_search_lr.fit(X_resampled, y_resampled)
```

```
Otimização de Hiperparâmetros com GridSearchCV:

Iniciando GridSearchCV para Random Forest...
Fitting 5 folds for each of 36 candidates, totalling 180 fits

Melhores parâmetros para Random Forest: {'class_weight': None, 'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
Melhor F1-Score na validação cruzada para Random Forest: 0.9894

Resultados do Melhor Random Forest no conjunto de Teste (otimizado):
Acurácia: 0.9895
F1-Score: 0.9887
Matriz de Confusão:
[[14022   188]
 [   91 12205]]
Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     14210
           1       0.98      0.99      0.99     12296

    accuracy                           0.99     26506
   macro avg       0.99      0.99      0.99     26506
weighted avg       0.99      0.99      0.99     26506
```

```
# Apenas para modelos que fornecem feature_importances_ (como RandomForestClassifier)
if 'Random Forest' in models and hasattr(models['Random Forest'], 'feature_importances_'):
    print("\nImportância das Features (Random Forest):")
    # Se usou GridSearchCV, use o best_rf_model
    if 'best_rf_model' in locals(): # Verifica se o GridSearchCV foi executado
        feature_importances = best_rf_model.feature_importances_
    else:
        feature_importances = models['Random Forest'].feature_importances_ # Usa o modelo sem otimização

    features_df = pd.DataFrame({'Feature': X.columns, 'Importance': feature_importances})
    features_df = features_df.sort_values(by='Importance', ascending=False)

    print(features_df.head(10)) # Mostra as 10 features mais importantes

    # Visualização
    plt.figure(figsize=(12, 7))
    sns.barplot(x='Importance', y='Feature', data=features_df.head(15))
    plt.title('Top 15 Features Mais Importantes (Random Forest)')
    plt.xlabel('Importância')
    plt.ylabel('Feature')
    plt.tight_layout()
    plt.show()
```

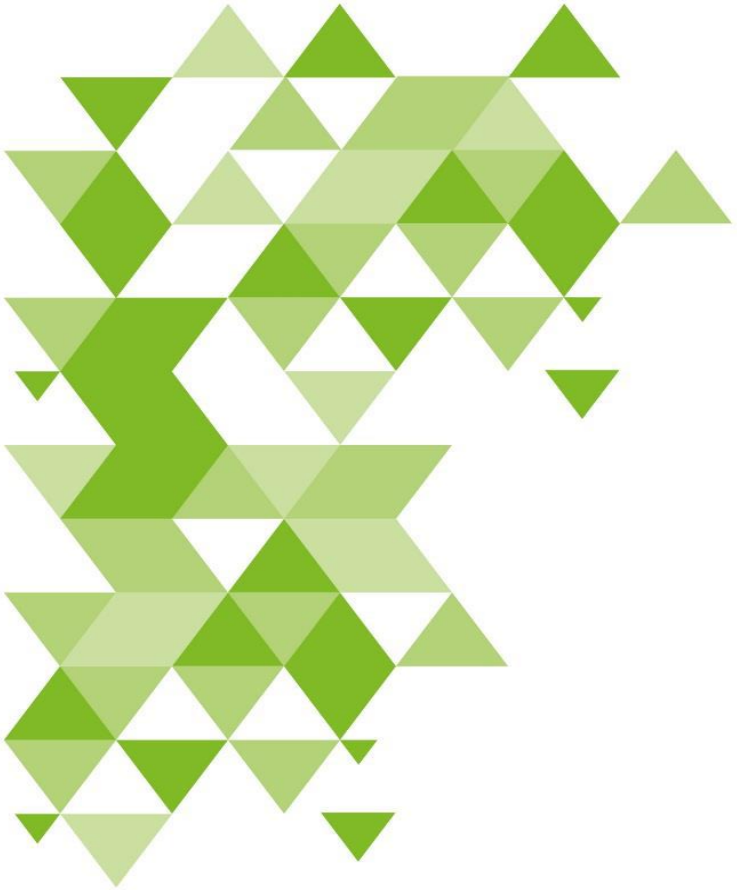
```
# Para Regressão Logística, pode-se olhar os coeficientes
if 'Logistic Regression' in models and hasattr(models['Logistic Regression'], 'coef_'):
    print("\nCoeficientes das Features (Regressão Logística - indicam importância relativa):")
    # Se usou GridSearchCV, use o best_lr_model se tiver sido otimizado
    lr_model_coef = models['Logistic Regression'].coef_[0]

    coef_df = pd.DataFrame({'Feature': X.columns, 'Coefficient': lr_model_coef})
    coef_df['Abs_Coefficient'] = np.abs(coef_df['Coefficient'])
    coef_df = coef_df.sort_values(by='Abs_Coefficient', ascending=False)

    print(coef_df.head(10)) # Mostra os 10 coeficientes com maior magnitude
```

```
Importância das Features (Random Forest):
                                    Feature  Importance
17                         Diferenca_Consumo    0.146043
7            Valor Última Leitura Facturada    0.135155
8              Valor Última Leitura Empresa    0.106069
9              Valor Última Leitura Cliente    0.086856
10                    Consumo Mês Anterior    0.069545
0                              Estimativa    0.044420
1                         Consumo Médio Real    0.038273
2851   Situação_2  EM CONDIÇÕES DE FACTURAR    0.024013
14           Tempo_Ultima_Leitura_Facturada    0.018382
11                               Nº Leituras    0.018153
```




```
Coeficientes das Features (Regressão Logística - indicam importância relativa):
                                           Feature  Coefficient  \
17                                 Diferenca_Consumo    22.240146
7                    Valor Última Leitura Facturada    22.234406
8                      Valor Última Leitura Empresa     3.866606
2851          Situação_2  EM CONDIÇÕES DE FACTURAR     2.944218
2862   Anomalia de Leitura Facturada_0 - SEM ERRO     2.894769
2864  Anomalia de Leitura Facturada_10 - DESVIO DE M...     2.705145
1334                          Cód#Postal A#_2240-122     1.956906
2801                        ZMC_32 ZA.05.09.Bodegao    -1.845273
1300                          Cód#Postal A#_2240-031    -1.761049
548             Situação da Entidade_V 2022-04-06     1.733667

      Abs_Coefficient
17          22.240146
7           22.234406
8            3.866606
2851         2.944218
2862         2.894769
```

II. BASE DE DADOS DO MICROSOFT SQL SERVER



ipt.pt

Gestão Hídrica Inteligente com Power BI - Instituto Politécnico de Tomar