
# DETEÇÃO AUTOMÁTICA DE SITUAÇÕES DE FALHA EM CONDUTAS DE ABASTECIMENTO DE ÁGUA

Mestrado em Engenharia Eletrotécnica

Henrique João Castanho Esteves Patrício

Leiria, março de 2024

escola superior de tecnologia e gestão
instituto politécnico de leiria

# DETEÇÃO AUTOMÁTICA DE SITUAÇÕES DE FALHA EM CONDUTAS DE ABASTECIMENTO DE ÁGUA

Mestrado em Engenharia Eletrotécnica

Henrique João Castanho Esteves Patrício

Dissertação realizada sob a orientação do Professor Doutor Luís Miguel Ramos Perdigoto e do Professor Doutor João Miguel Charrua de Sousa

Leiria, março de 2024

# Originalidade e Direitos de Autor

A presente dissertação é original, elaborada unicamente para este fim, tendo sido devidamente citados todos os autores cujos estudos e publicações contribuíram para a elaborar.

# Dedicatória

Aos meus Pais, à minha Esposa e aos meus Filhos.

# Agradecimentos

À minha Mãe (in memorian), ao meu Pai e ao meu Irmão pelo apoio que me deram ao longo de toda a minha vida.

À Filipa, pelo apoio incondicional, incentivo e paciência ao longo desta caminhada.

Ao Martim e à Maria, que tantas vezes ficaram para segundo plano para que esta caminhada fosse possível.

À EPAL pela disponibilização da informação necessária ao desenvolvimento do estudo aqui apresentado.

Aos Colegas do Departamento de Manutenção pelo apoio ao longo dos últimos anos, nomeadamente Eng. Miguel Fernandes, Eng. António Félix, Bruno Bernardino e Eng. Vítor Patrício.

Aos orientadores deste trabalho, Professores Doutores João Sousa e Luís Perdigoto, pelo apoio, disponibilidade, aconselhamento e estímulo.

# Resumo

Uma percentagem significativa das falhas em sistemas de abastecimento de água são consequência de roturas nas condutas de adução e distribuição. Quando essas falhas ocorrem em sistemas de abastecimento municipais, são afetados apenas os munícipes que são abastecidos pelo ramal em causa. Já numa situação em que tal acontece num sistema que abastece diversos reservatórios municipais, existe o risco de o número de clientes lesados subir exponencialmente. Quando ocorre uma rotura, há que detetá-la o mais rapidamente possível para poder tomar medidas e assim minimizar as perdas de água, que numa conduta de grandes dimensões podem atingir custos consideráveis.

O presente trabalho tem como objetivo a exploração de métodos que permitam detetar situações de rotura e alterações de regime em condutas de abastecimento de água nos processos de transporte até aos reservatórios municipais, podendo no futuro ser integrada por exemplo num sistema SCADA (Sistema de Supervisão e Aquisição de Dados) para informar sobre situações anómalas detetadas. Em particular, são analisadas técnicas de aprendizagem automática supervisionada.

Durante o trabalho foram adotados modelos de deteção baseados nos algoritmos *Random Forest, XGBoost, Support Vector Machines* e *Artificial Neural Networks*. Estes algoritmos foram aplicados para deteção de situações de rotura e de alterações ao regime normal de funcionamento. Foram utilizados dados de duas condutas reais, sendo uma de abastecimento gravítico e outra de abastecimento elevatório. Verificou-se que os modelos que obtiveram melhores resultados em ambas as situações foram os modelos baseados em *Random Forest*.

Foi também analisada a diminuição de desempenho dos modelos na situação de substituição de componentes hidráulicos e alteração de regime típico de funcionamento dos sistemas de distribuição.

Nos casos estudados obtiveram-se melhores resultados para o sistema de abastecimento elevatório, facto justificável pela presença de informações relativas ao local de destino, nomeadamente caudal de entrada e nível do reservatório.

**Palavras-chave:** Deteção de roturas em condutas de abastecimento, Aprendizagem Automática, Otimização

# Abstract

A significant percentage of supply failures are the result of ruptures in adduction and distribution pipes. When these failures occur in municipal supply systems, only those residents who are supplied by the branch in question are affected. In a situation where this happens in a system that supplies several municipal reservoirs, there is a risk that the number of injured customers will rise exponentially. When a rupture occurs, it must be detected as quickly as possible in order to take measures and thus minimize water losses, which in a large pipeline can reach a considerable size.

The present work aims to create an application that detects rupture situations and regime changes in water supply pipes in transport processes to municipal reservoirs, which could in the future be integrated into, for example, the SCADA system to report on detected anomalous situations. In particular, supervised learning techniques are considered.

During the work, detection models based on *Random Forest, XGBoost, Support Vector Machines* and *Artificial Neural Networks* algorithms were adopted. These algorithms were applied to detect ruptures and changes to the normal operating regime. Data from two real pipelines were used, one of them being a gravity supply system, and the other is a lifting supply system. The models that obtained the best results in both situations were based on the *Random Forest* algorithm.

It was also analysed the performance decrease in situations of hydraulic components replacement and changes to the normal operating regime.

In the studied cases, a better performance was obtained for the lifting supply system. This can be justified due to the use of additional features available in the destination reservoir, specifically inlet flow and reservoir level.

**Keywords:** Rupture detection in water supply pipes, Machine Learning, Optimization

# Índice

# Lista de Figuras

# Lista de Tabelas

# Lista de siglas e acrónimos

| | |
|---|---|
| AA08a | Água não faturada em “alta” |
| AA08b | Água não faturada em “baixa” |
| Adução | Transporte de água entre locais, como por exemplo entre uma captação de água e uma Estação de Tratamento de Água, uma Estação de Tratamento de água e a rede de distribuição ou mesmo entre reservatórios |
| AdVT | Águas do Vale do Tejo |
| ANN | *Artificial Neural Networks* |
| BiLSTM | *Bidirectional LSTM* |
| CNN | *Convolutional Neural Network* |
| DN | Diâmetro Nominal |
| DNN | *Deep Neural Network* |
| DT | *Decision Trees* |
| EE | Estação Elevatória |
| EG | Entidade(s) Gestora(s) |
| EPAL | Empresa Portuguesa das Águas Livres |
| EPANET | *Software* utilizado para modelação de sistemas de abastecimento de água potável |
| ERSAR | Entidade Reguladora dos Serviços de Águas e Resíduos |
| ETA | Estação de Tratamento de Água |
| IA | Inteligência Artificial |
| IBM | International Business Machines |
| LDA | *Linear Discriminant Analysis* |
| LOF | *Local Outlier Factor* |
| LR | *Logistic Regression* |
| LSTM | *Long Short-Term Memory* |

| | |
|---|---|
| MAG | Modelo Aditivo Generalizado |
| MAXIMO | Software de gestão da IBM |
| MEMS | Sistemas microeletromecânicos sem fios |
| PE | Ponto de Entrega |
| PVC | Policloreto de vinila |
| RF | *Random Forest* |
| RISE | *Random Interval Spectral Ensemble* |
| RNN | *Recurrent Neural Networks* |
| SCADA | Sistema de Supervisão e aquisição de dados |
| SCB | *Spatial Coordinates Based* |
| SOM | *Self-Organizing Maps* |
| SVM | *Support Vector Machines* |
| *XGBoost* | *eXtreme Gradient Boosting* |
| ZMC | Zonas de Monitorização e Controlo |

# 1. Introdução

Numa análise aos sistemas de produção e distribuição de água constata-se que a parte mais significativa dos investimentos está alocada às redes de transporte. Sabe-se ainda que uma percentagem significativa das falhas de abastecimento são consequência de roturas na rede, ou seja, nas condutas de adução e distribuição.

A avaliação das causas das roturas permite identificar alguns fatores primordiais como sejam erros de projeto, execuções e implementações incorretas, inadequada operação, utilização para além do tempo de vida útil, roturas provocadas, etc.

Na maioria dos casos, assume vital importância a Entidade Gestora (EG) ter *know-how* adequado por forma a avaliar o funcionamento dos sistemas, na expectativa de poder minimizar os erros que possam ter existido até à entrada em serviço de condutas e equipamentos.

A caracterização dos sistemas passa naturalmente por uma avaliação hidráulica, reconhecendo-se desde logo duas grandezas fundamentais: o caudal e a pressão. Nesse sentido é importante conhecer ou calcular as evoluções temporais destas duas grandezas em diferentes pontos constituintes dos sistemas. É sabido que, em regime permanente, o caudal é a consequência natural da solicitação que é feita por parte do consumo, sendo a pressão a consequência do caudal e da linha de energia do sistema, onde naturalmente os perfis altimétricos das condutas têm particular relevância.

Não sendo naturalmente possível alterarem-se as características dos regimes permanentes, é relativamente simples de se pensar que a margem de otimização está exatamente nos regimes transitórios. São estes regimes que podem assumir significativa relevância no que diz respeito à diminuição da fadiga desnecessária, que pode ser imposta a materiais e a órgãos hidráulicos. Neste capítulo assumem particular importância as válvulas seccionadoras em regimes gravíticos, (sejam estas válvulas redutoras, limitadoras, etc) e todo o seu comportamento na mudança de estado aberto/fechado/aberto. Em sistemas elevatórios acrescem ainda os transitórios hidráulicos provocados pelo arranque de grupos eletrobomba e, mais gravosamente, as suas paragens intempestivas por corte da alimentação de energia elétrica, podendo promover o famigerado golpe de Aríete.

## 1.1. Motivação

Segundo o Relatório de Sustentabilidade da EPAL (Empresa Portuguesa das Águas Livres)/AdVT (Águas do Vale do Tejo ), ao longo do ano de 2022, 10,9% da água captada e tratada pela EPAL não foi faturada (Departamento de Sustentabilidade Empresarial da EPAL, 2023-b). Toda a água que não é faturada tem um custo associado uma vez que é captada, tratada, transportada, armazenada e, em determinadas situações, distribuída.

Segundo a *International Water Association*, citada por (Sardinha et al., 2017), a água não faturada num sistema é a diferença entre o volume total de água que entra no mesmo e o volume total de água consumida faturada.

A água não faturada pode ser resultado de consumos autorizados não faturados, perdas aparentes ou perdas reais de água em condutas ou em ramais, conforme se pode verificar na Figura 1.

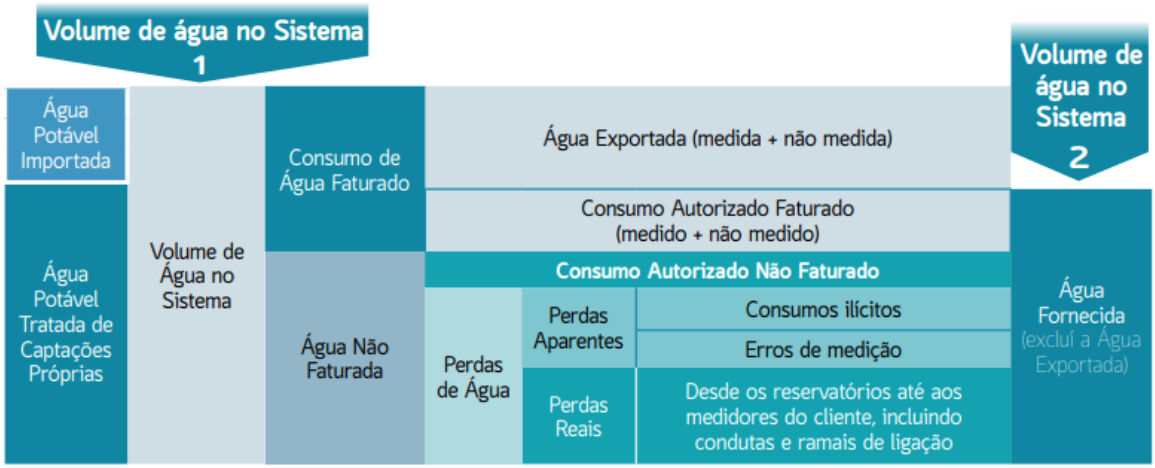


**Figura 1 - Balanço Hídrico, de acordo com a definição da *International Water Association*. Retirado de (Sardinha et al., 2017)**

Para que haja uma melhor compreensão dos dados abaixo representados é necessário referir que o abastecimento pode ser feito em "alta" ou em "baixa". O abastecimento em "alta" compreende captação, tratamento, armazenamento e transporte de água, geralmente até um reservatório municipal. Nesse reservatório municipal inicia-se o abastecimento em "baixa", que compreende a distribuição aos consumidores finais.

Entre 2017 e 2021 a percentagem de água não faturada pelos municípios (abastecimento em "baixa") esteve na ordem dos 28 a 30%, conforme dados da Entidade Reguladora dos

Serviços de Águas e Resíduos (ERSAR) (ERSAR, 2023) que se encontram na forma de gráfico na Figura 2, adaptada desse mesmo documento.

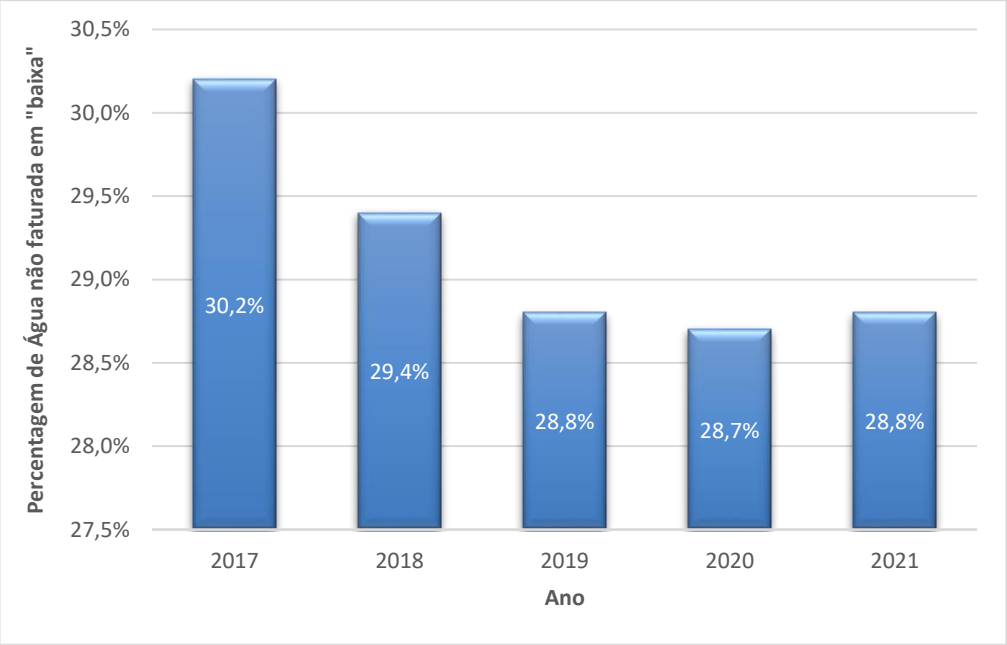


**Figura 2 - Evolução média do indicador de água não faturada em "baixa" 2017-2021. Adaptado de (ERSAR, 2023)**

Segundo dados recentes apurados na imprensa (Diário de Notícias, 2023), em 2021 não foram faturados 237 milhões de metros cúbicos de água em "baixa", o que corresponde a 347 milhões de euros não faturados.

Relativamente às perdas de água em "alta", os valores de água não faturada são substancialmente mais baixos, tendo entre 2017 e 2021 rondado uma gama entre os 4,8 e os 5,7%, de acordo com o gráfico da Figura 3, adaptado do documento da ERSAR já referido (ERSAR, 2023).

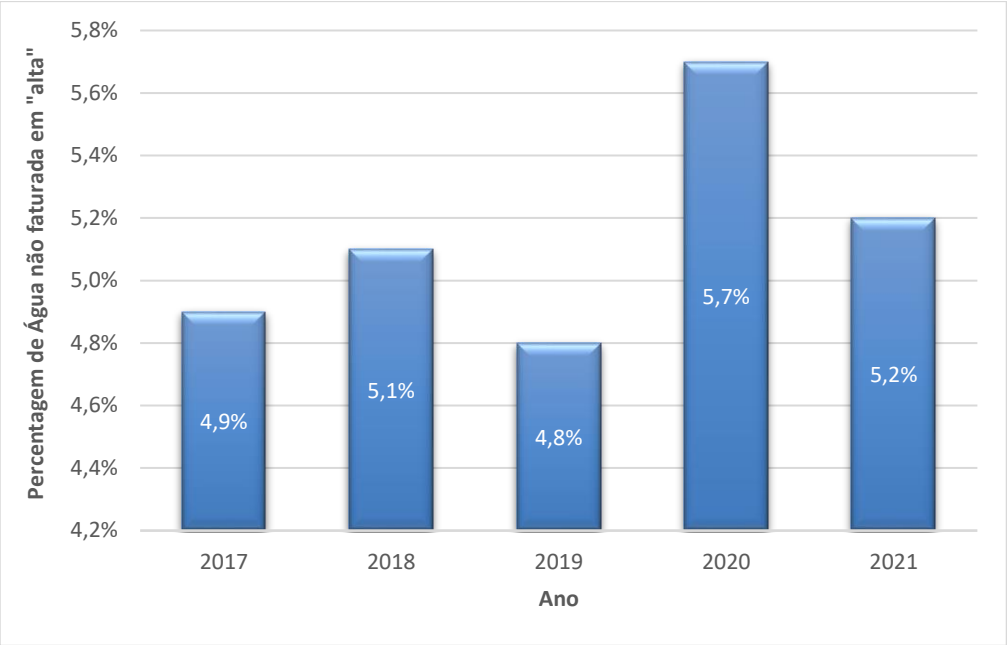


**Figura 3 - Evolução média do indicador de água não faturada em "alta" 2017-2021. Adaptado de (ERSAR, 2023)**

Tendo em conta a falta de água que se tem verificado ao longo dos últimos anos e que tem causado a redução dos caudais dos rios, dos níveis das barragens e dos lençóis freáticos, torna-se cada vez mais urgente baixar os volumes de água não faturada que correspondem aos consumos ilícitos, aos erros de medição e às perdas reais.

As roturas nas condutas nos sistemas de abastecimento são uma das causas das perdas reais, podendo ser causadas por efeitos transitórios originados por arranques e paragens de grupos elevatórios e/ou por válvulas de seccionamento com tempos de abertura/fecho demasiado curtos, válvulas redutoras de pressão ou reguladoras de caudal mal afinadas e manobras realizadas de forma incorreta. A análise de séries temporais de dados relativos a caudais e pressões permite perceber o comportamento dos equipamentos existentes ao longo de uma conduta, sendo possível perceber se algum equipamento se encontra com funcionamento deficiente e, a partir daí, desenvolver ações corretivas atempadamente por forma a evitar um aumento da fadiga da conduta e consequentes problemas no futuro.

## 1.2. Objetivos

Face ao exposto na secção anterior, definiu-se para este trabalho o objetivo de explorar métodos que possibilitassem a verificação a título preventivo de efeitos transitórios associados ao funcionamento dos sistemas de abastecimento e a deteção prematura de roturas em condutas, explorando técnicas de aprendizagem automática aplicadas a um conjunto de dados históricos reais, nomeadamente níveis, pressões e caudais, que se encontram armazenados num servidor da EPAL. O foco do trabalho recaiu, em particular, sobre a utilização de técnicas de aprendizagem supervisionada.

Dado que existem duas condutas no Sistema de Abastecimento em "alta" da Zona Oeste em que é usual a ocorrência de roturas, optou-se por realizar o estudo necessário ao desenvolvimento da aplicação nessas mesmas condutas.

## 1.3. Estrutura da Dissertação

No Capítulo 2 apresenta-se uma revisão bibliográfica realizada aquando do início do desenvolvimento do trabalho, por forma a percecionar quais os caminhos já percorridos nesta área e assim perceber quais os passos a dar para se obterem melhores resultados e quais os desafios que se poderiam esperar. Abordam-se também as temáticas das perdas de água nas redes e estratégias clássicas de deteção.

No Capítulo 3 apresenta-se o Sistema de Abastecimento da EPAL na generalidade e os sistemas elevatórios que foram utilizados na realização do presente trabalho.

No Capítulo 4 é apresentado um conjunto de algoritmos de aprendizagem supervisionada, que foram utilizados neste estudo. Abordam-se também alguns passos dados na aquisição e preparação de dados para que fosse possível aplicar os algoritmos de deteção.

No Capítulo 5 apresenta-se o primeiro caso de estudo, "Deteção de perdas de água numa conduta de abastecimento gravítico".

No Capítulo 6 apresenta-se o segundo caso de estudo, "Deteção de perdas de água numa conduta de abastecimento elevatório".

No Capítulo 7 encontra-se a Discussão de resultados e por fim no Capítulo 8 as Conclusões e perspetivas de trabalho futuro.

# 2. Revisão bibliográfica

A aprendizagem automática é um dos ramos ou subconjuntos da Inteligência Artificial (IA) (Helm et al., 2020) e consiste em algoritmos que dotam os computadores de capacidade para aprender sem uma programação específica. Ao analisar um conjunto de dados, o algoritmo deteta padrões que lhe permitem não só avaliar produtos ou eventos (algoritmos de classificação) mas também prever situações futuras (algoritmos de previsão).

Os três principais tipos de aprendizagem automática são a aprendizagem supervisionada, não supervisionada e por reforço (Camila Waltrick, 2020).

No caso da aprendizagem supervisionada, utilizada para classificação e regressão, é fornecido ao algoritmo um conjunto de dados previamente classificados ou para os quais se conhece o(s) valor(es) da(s) variável(is) dependente(s). Esse conjunto é composto por variáveis de entrada que podem ser por exemplo valores lidos de sensores e pela(s) variável(is) de saída que é/são o resultado da classificação/regressão.

A regressão tem como objetivo a previsão de variável(is) contínua(s) a partir de um modelo resultante da análise de dados previamente existentes. O algoritmo analisa alterações nas variáveis de entrada com o objetivo de detetar padrões/tendências que sejam responsáveis pela alteração da(s) variável(is) de saída. Após ser devidamente treinado e parametrizado, o modelo deve ser capaz de prever novos registos apenas com base nas variáveis de entrada. Utiliza-se para previsão e modelagem em diversas áreas, como por exemplo vendas, marketing, análise de clientes, análise financeira, saúde, controlo de qualidade, ciências sociais, otimização de processo, preços de ações, participação de mercado, avaliação de risco, análise do mercado financeiro, deteção de fraudes, deteção de anomalias, energia, desporto e ambiente(Sarvandani, 2023). Pode também prever a evolução de doenças contagiosas, como aconteceu aquando da pandemia Covid-19 (Shakeel, 2021).

No caso da classificação, o algoritmo analisa as entradas que geraram uma dada saída (evento) com o objetivo de encontrar padrões nos dados de entrada fornecidos que ajudem a classificar corretamente novos dados. Ao longo do processo de treino do algoritmo, este afina os parâmetros (que são atribuídos habitualmente de forma aleatória no início) para aumentar a capacidade de classificar corretamente os eventos de acordo com os padrões detetados. Após o processo de treino, o algoritmo deve ser capaz de classificar corretamente novos dados.

No caso da aprendizagem não supervisionada, esta utiliza apenas os dados a analisar, sem necessidade de classificação prévia. Ao analisar os dados, o algoritmo encontra padrões que permitem agrupá-los de acordo com características comuns (baseado em critérios de proximidade), formando conjuntos (*clusters*) distintos. A aprendizagem por reforço consiste no treino do algoritmo por tentativa e erro, resultando numa estratégia de incentivos em caso de sucesso ou de penalizações em caso de fraco desempenho.

No presente trabalho são abordados métodos de aprendizagem supervisionada.

Qualquer que seja o tipo de aprendizagem utilizado há que ter em conta a necessidade de reunir dados para que o algoritmo possa ser treinado de forma eficiente, levando a que as classificações e/ou previsões a realizar tenham credibilidade. Esses dados são utilizados como variáveis de entrada no processo de treino do algoritmo. Caso se trate de séries temporais, os sinais devem ser registados com uma periodicidade que permita ao algoritmo percecionar alteração de padrões. Por exemplo se se pretender analisar alterações de pressão causadas por golpes de Ariete em redes de transporte e/ou distribuição de água, a periodicidade entre leituras deverá ser baixa para registar as variações decorrentes dos efeitos transitórios. Verificou-se noutros trabalhos já realizados neste âmbito que os dados utilizados como variáveis de entrada foram obtidos de três formas diferentes:

- Sensores colocados em redes de abastecimento (dados reais);
- Simulações em computador;
- Protótipos criados em laboratório.

Nesse conjunto de trabalhos, que tinham objetivos similares ao deste estudo, verificou-se que (Vivas et al., 2021), (Fares et al., 2023), (André de Oliveira, 2020), (Nascimento, 2021), (Fang et al., 2019) e (Blázquez-García et al., 2021) utilizaram no seu estudo apenas dados reais, enquanto (Mashhadi et al., 2021) e (Costa, 2020) utilizaram não só dados reais, mas também dados obtidos através de simulações em EPANET (*Software* utilizado para modelação de sistemas de abastecimento de água potável). Já (Caputo et al., 2003), (Liu et al., 2019), (Fan et al., 2021) e (Kammoun et al., 2023) utilizaram apenas dados provenientes de simulações em computador, embora em (Fan et al., 2021) e em (Kammoun et al., 2023) sejam criadas simulações baseadas em redes de distribuição reais e dados reais. Os autores de (Alves Coelho et al., 2020) e de (Şahin et al., 2023) realizaram simulações com protótipos criados em laboratório para obterem os dados para análise. Num outro trabalho que tinha

como objetivo a avaliação de qualidade da água (Muharemi et al., 2019) foram utilizados apenas dados reais.

Já relativamente aos tipos de dados utilizados para análise nos trabalhos acima mencionados, verifica-se que na maior parte dos casos foram utilizadas medições de caudal, seja totalizado num dado período de tempo, instantâneo ou ainda na forma de aceleração do líquido em trânsito (Alves Coelho et al., 2020; André de Oliveira, 2020; Blázquez-García et al., 2021; Caputo et al., 2003; Costa, 2020; Fang et al., 2019; Kammoun et al., 2023; Liu et al., 2019; Mashhadi et al., 2021; Muharemi et al., 2019; Nascimento, 2021; Şahin et al., 2023; Vivas et al., 2021). Também a pressão foi utilizada numa boa parte dos trabalhos consultados (Caputo et al., 2003; Costa, 2020; Fan et al., 2021; Kammoun et al., 2023; Mashhadi et al., 2021; Şahin et al., 2023). Existem ainda trabalhos que tiveram em consideração outras variáveis, como dados de manutenção, extensão, material e diâmetro das condutas, distância a vias, ferrovias e rede hidrográfica, elevação, declive, pH, densidade populacional, consumo de água, ocupação do solo e geologia do terreno, temperatura da água, quantidade de dióxido de cloro, potencial redox, condutividade e turvação (André de Oliveira, 2020; Blázquez-García et al., 2021; Muharemi et al., 2019; Nascimento, 2021). Os autores (Fares et al., 2023; Liu et al., 2019) analisaram dados acústicos, que no primeiro caso foram recolhidos na rede de abastecimento de *Hong Kong* e no segundo caso resultaram de uma simulação de rede de abastecimento. Na Figura 4, construída com os dados acima mencionados, é possível verificar a predominância de utilização dos dados de caudal e pressão nos trabalhos consultados (de um total de 15 artigos analisados).

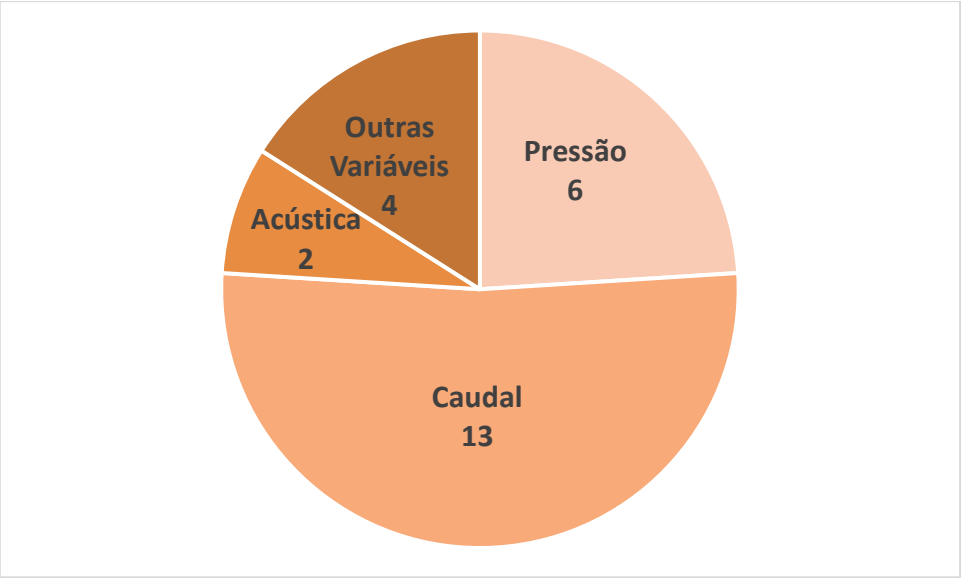


**Figura 4 – Número de trabalhos em que são utilizadas as variáveis "Pressão", "Caudal", "Acústica" e "Outras Variáveis" de um total de 15 trabalhos analisados**

Nem sempre foi utilizada apenas uma variável de entrada. No caso de (Muharemi et al., 2019), foram selecionadas as variáveis mais influentes (aquelas com maior relação entre si e/ou com a saída) a partir da totalidade de dados disponíveis. Essa seleção foi automatizada através de um algoritmo de seleção. Na Figura 5 é possível ver que na maior parte dos trabalhos foram utilizadas duas ou mais variáveis de entrada.

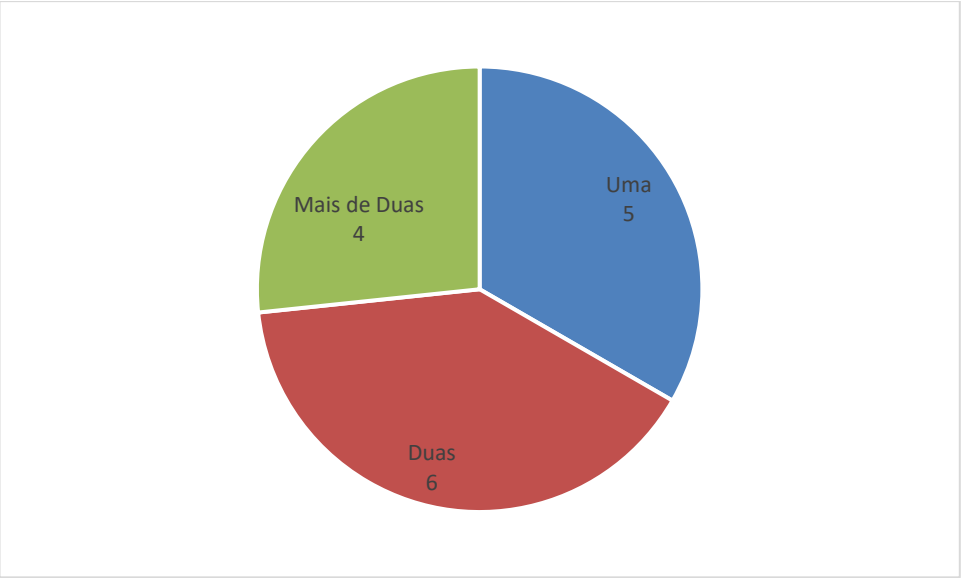


**Figura 5 – Número de trabalhos em que são utilizadas "uma", "duas" e "mais de duas" variáveis de entrada de um total de 15 trabalhos analisados**

Segundo (Muharemi et al., 2019), um grande número de variáveis de entrada não é sinónimo de maior precisão, pelo que devem ser selecionadas apenas as variáveis essenciais para evitar informação desnecessária e/ou redundante.

Para além da seleção das variáveis de entrada também é de grande importância o pré-processamento dos dados a utilizar, que na maioria das vezes chegam com erros, corrompidos ou incompletos. Este procedimento é considerado fundamental por (Zhang et al., 2003), que considera também que a sua falta pode prejudicar a deteção de padrões importantes e causar baixo desempenho do algoritmo. Para os autores esta situação pode dever-se a dados incompletos, ruidosos ou com valores inconsistentes. Já (Huang et al., 2015) refere que, embora muitos investigadores considerem o pré-processamento de dados como uma etapa fundamental, existem poucos estudos relativamente aos efeitos das técnicas utilizadas. Em (Saias et al., 2018) são referidos dois processos a executar no pré-processamento de dados, sendo estes a limpeza e a transformação dos dados. O primeiro processo inclui a eliminação de erros, a mitigação de valores em falta e a verificação de consistência dos dados. O segundo prende-se com a forma como os dados são "apresentados" ao algoritmo, sugerindo-se por exemplo a normalização de dados. O pré-tratamento dos dados a analisar é de preponderante importância para obter os melhores resultados possíveis, seja qual for o algoritmo utilizado.

No subcapítulo 2.1 abordam-se os algoritmos de aprendizagem supervisionada mais utilizados nos trabalhos analisados, no subcapítulo 2.2 aborda-se a temática das perdas de água nas redes e no subcapítulo 2.3 apresenta-se um pouco da história da evolução das estratégias de deteção de fugas de água em redes de transporte e distribuição desde os primórdios até aos dias de hoje.

## 2.1. Algoritmos de aprendizagem mais utilizados

Conforme já referido, os algoritmos mais utilizados para problemas de classificação são os de aprendizagem supervisionada e não supervisionada. Não se encontrando no âmbito do presente trabalho a utilização de algoritmos de aprendizagem não supervisionada, mas sendo estes utilizados em termos comparativos em alguns dos trabalhos consultados, serão ainda assim aqui referidas algumas conclusões relativas ao seu desempenho e utilização.

Relativamente às vantagens e desvantagens da utilização dos algoritmos de aprendizagem supervisionada e não supervisionada, os autores de (Mashhadi et al., 2021) compararam os resultados de deteção de fugas de água utilizando os dois tipos de algoritmo, tendo concluído que os métodos de aprendizagem não supervisionada têm dificuldade na localização das perdas de água devido à sobreposição de clusters.

Na Tabela 1 podem consultar-se os algoritmos de aprendizagem automática utilizados num conjunto de trabalhos realizados por outros autores no âmbito da deteção de fugas de água e também de monitorização de qualidade da água.

**Tabela 1 - Algoritmos de aprendizagem automática utilizados noutros Trabalhos**

| **Trabalho** | **Aplicação** | **Âmbito** | **Métodos de aprendizagem automática utilizados** |
|---|---|---|---|
| (Vivas et al., 2021) | Rede de Abastecimento de Água | Deteção de fugas em sistema de abastecimento | *Artificial Neural Networks (*ANN*),* Modelo Aditivo Generalizado (MAG) |
| (Fares et al., 2023) | Rede de Abastecimento de Água de Hong Kong | Deteção de fugas em sistema de abastecimento | *Decision tree (DT), Giant boosted trees, Random Forest (RF),* ANN*, Deep learning, Naïve Bayes, Naïve Bayes kernel, Support Vector Machines (*SVM*), Rule induction, Linear regression* |
| (Caputo et al., 2003) | Simulação de Rede de Abastecimento de líquidos perigosos | Deteção de fugas em sistema de abastecimento | *Artificial Neural Networks* |
| (Liu et al., 2019) | Simulação de Rede de Abastecimento | Deteção de fugas em sistema de abastecimento | *Support Vector Machines* |
| (Alves Coelho et al., 2020) | Sistema protótipo criado em laboratório | Deteção de fugas em sistema de abastecimento | *Support Vector Machines, Decision Trees, Random Forest, Neural Networks, XGBoost* |

| **Trabalho** | **Aplicação** | **Âmbito** | **Métodos de aprendizagem automática utilizados** |
|---|---|---|---|
| (Mashhadi et al., 2021) | Dados simulados com EPANET e dados reais Sistema de abastecimento do Campus da Universidade de Lille | Deteção de fugas em sistema de abastecimento | *Logistic Regression, Decision Trees, Random Forest, Hierarchical Classification, K-Means, Artificial neural network* |
| (Costa, 2020) | Dados simulados com EPANET e dados reais Sistema de distribuição de água da Quinta do Lago | Deteção de fugas em sistema de abastecimento | CNN-LSTM, CNN-BiLSTM, SCB-LSTM, *Stacked* BiLSTM *e Stacked* LSTM |
| (André de Oliveira, 2020) | Rede de distribuição de água de Vila Nova de Gaia | Deteção de fugas em sistema de abastecimento | *Decision Trees, Random Forest* |
| (Nascimento, 2021) | Sistema de distribuição de água de Curitiba, estado do Paraná, Brasil | Deteção de fugas em sistema de abastecimento | *Local Outlier Factor* (baseada em densidade), *Self-Organizing Maps* (baseada em ANN), *Standard Score* (Z-Score, abordagem estatística), *AutoML* |
| (Fan et al., 2021) | Dados simulados com EPANET/simulação de rede de abastecimento | Deteção de fugas em sistema de abastecimento | *Artificial Neural Networks, Autoencoder Neural Networks* |

| **Trabalho** | **Aplicação** | **Âmbito** | **Métodos de aprendizagem automática utilizados** |
|---|---|---|---|
| (Muharemi et al., 2019) | Dados reais da rede pública de abastecimento de Thüringer Fernwasserversorgung, Alemanha | Deteção de falhas na qualidade da água | *Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Artificial Neural Network, Deep Neural Network, Recurrent Neural Network, Long Short-Term Memory* |
| (Fang et al., 2019) | Rede de Abastecimento de Água | Deteção de fugas/leituras erradas em sistema de abastecimento | *One-Class Support Vector Machine* |
| (Blázquez-García et al., 2021) | Cenário A - Rede privada de abastecimento em Azkoitia, Espanha; Cenário B - Dados de rede de abastecimento da Yorkshire Water (Inglaterra) | Deteção de fugas em sistemas de abastecimento | *Random Interval Spectral Ensemble* |
| (Kammoun et al., 2023) | Dados simulados com base em dados da rede de distrubuição de Hanoi | Deteção de fugas em sistemas de abastecimento | LSTM *autoencoder, Recurrent Neural Network* |
| (Şahin et al., 2023) | Sistema protótipo criado em laboratório | Deteção de fugas em sistemas de abastecimento | SVM, *Graph Convolutional Neural Network* |

Constatou-se que a maioria dos autores utilizou algoritmos baseados em redes neuronais, conforme se pode verificar na Figura 6.

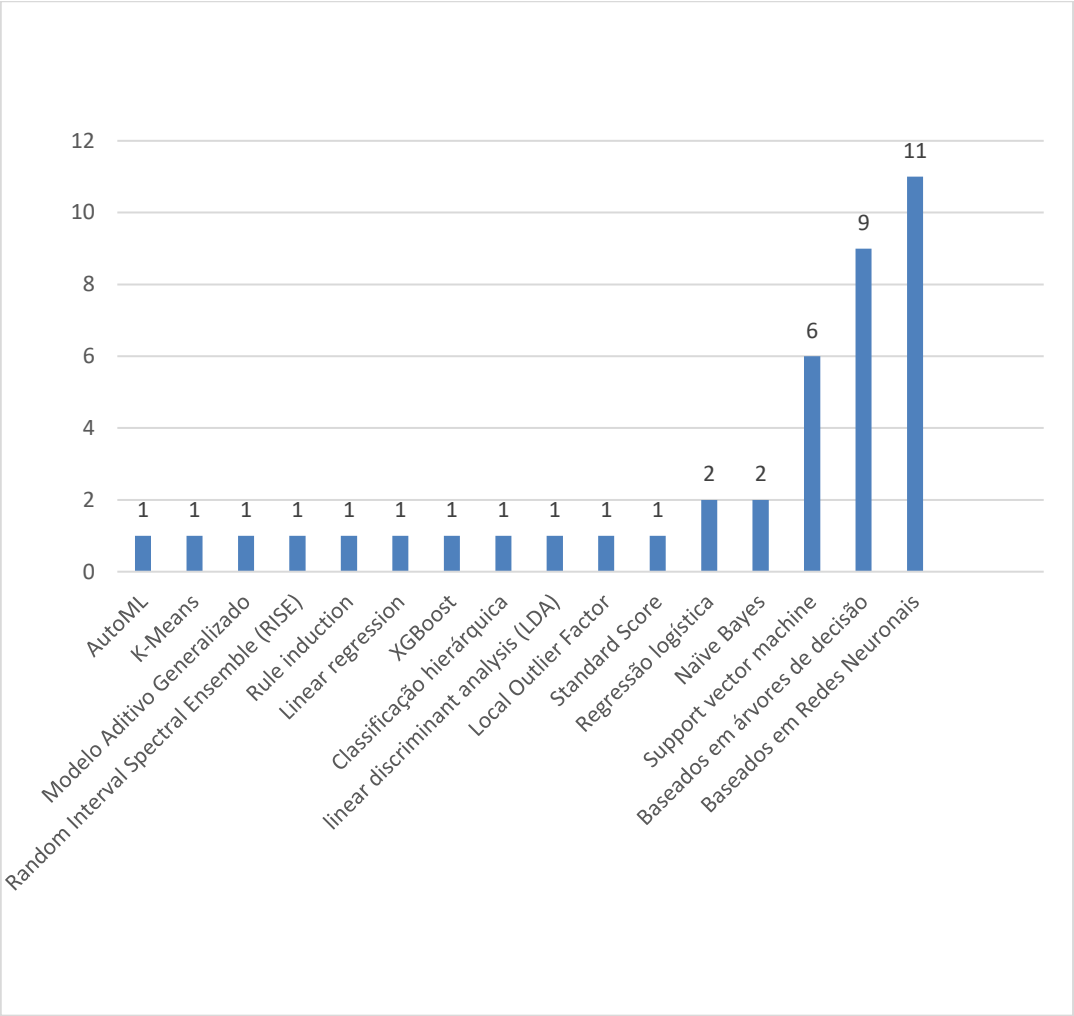


**Figura 6 – Utilização de algoritmos de aprendizagem automática nos trabalhos consultados**

Os algoritmos baseados em Redes Neuronais são os que mais aparecem nos trabalhos consultados. Também os algoritmos baseados em árvores de decisão (*decision trees*, *random forest* e *giant boosted trees*) e em *Support Vector Machines* (SVM e *One Class* SVM*)* são utilizados em grande parte das referências consultadas. Os algoritmos de aprendizagem supervisionada baseados em *Support Vector Machines* são considerados eficientes para classificação supervisionada por (Şahin et al., 2023) e (Valero-Carreras et al., 2023). De entre os algoritmos baseados em *Support Vector Machines*, é importante referir que o *One Class* SVM é um algoritmo de aprendizagem não supervisionada.

Alguns dos autores realizaram um estudo comparativo de métodos de aprendizagem. Um desses casos foi (Mashhadi et al., 2021), que utilizou um conjunto de variantes de aprendizagem supervisionada e não supervisionada, não baseadas em redes neuronais artificiais, e comparou os resultados obtidos por estes com os resultados obtidos por um outro algoritmo baseado em redes neuronais artificiais (aprendizagem supervisionada), tendo concluído que todos os algoritmos de aprendizagem supervisionada (incluindo as redes neuronais artificiais) apresentaram excelentes resultados. Já os algoritmos de aprendizagem não supervisionada apresentaram piores resultados na classificação, situação que os autores atribuem à sobreposição de clusters. Já (Alves Coelho et al., 2020) concluiu que, dos algoritmos que comparou (*Support Vector Machine, Decision Trees, Random Forest, Neural Networks, XGBoost*), o melhor foi baseado em *Decision Trees*. Constatou ainda que com um número maior de variáveis conseguia melhorar a exatidão de deteção. Também (Fares et al., 2023) comparou algoritmos de aprendizagem supervisionada, tendo concluído que os algoritmos baseados em *Support Vector Machines*, *Artificial Neural Networks* e *Deep Learning* têm uma capacidade bastante promissora de deteção em novos dados.

(Fan et al., 2021) comparou os resultados obtidos com *Artificial Neural Networks* e com *Autoencoder neural network* para verificar se este último, por não necessitar de um conjunto de dados equilibrados, forneceria melhores resultados, tendo concluído que o modelo *Autoencoder neural network* funcionava bem quando as fugas ocorriam perto dos sensores, caso contrário a precisão de deteção descia. Utilizando múltiplas tentativas independentes de deteção, a precisão subiu. Também (Muharemi et al., 2019) estudou a influência do desequilíbrio de dados, neste caso com o objetivo de encontrar o melhor algoritmo para monitorização da qualidade da água, tendo concluído que todos os algoritmos estudados são vulneráveis a esta situação, sendo *Support Vector Machines*, *Artificial Neural Networks* e *Logistic Regression* os menos afetados.

Por vezes são utilizados dois ou mais algoritmos de aprendizagem automática para melhorar a qualidade de deteção, tal como acontece em (Vivas et al., 2021), onde foram utilizadas *Artificial Neural Networks* coordenadas com Modelo Aditivo Generalizado (MAG). As ANN foram utilizadas para detetar fugas de grandes quantidades de água, que são mais facilmente detetáveis, tendo obtido resultados satisfatórios. Já o MAG foi utilizado para detetar pequenas perdas que apenas são identificáveis a partir da análise do aumento das perdas de base, que se reflete numa alteração da tendência de perdas. Os autores consideram o algoritmo baseado em MAG uma mais-valia. No caso de (Costa, 2020) foram comparados

métodos baseados em *Recurrent Neural Networks*, tendo ainda sido utilizadas combinações de algoritmos. A autora constatou dificuldade na deteção de situações anómalas utilizando algoritmos de aprendizagem não supervisionada, que foi também referida por (Mashhadi et al., 2021).

## 2.2. O problema das perdas de água nas redes de transporte e distribuição

Desde os primórdios da construção de sistemas de abastecimento que as perdas de água devido a roturas se tornaram um problema, pois a sua ocorrência gerava falhas no abastecimento (Sardinha et al., 2017).

As perdas de água classificam-se, segundo (Sardinha et al., 2017), em perdas reais e perdas aparentes. As perdas aparentes podem advir de consumos indevidos e de erros de medição. Os consumos indevidos são muitas vezes realizados através de *bypass* ao contador, derivação de ramal e ligação direta de um ramal clandestino à rede de distribuição (Rodrigues, 2021), situações ilustradas na Figura 7.

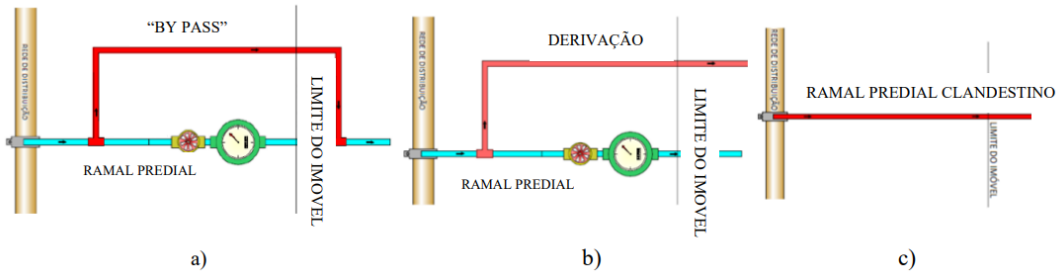


**Figura 7 - a) Ligação por "bypass" ao ramal; b) Ligação por derivação de ramal; c) Ligação direta. Retirado de (Rodrigues, 2021)**

Segundo (Rodrigues, 2021), os erros de medição podem ser causados por:

- Adulteração dos contadores ou por desgaste dos mesmos, seja devido à idade ou às características da água;
- Presença de sólidos suspensos na água;
- Dimensionamento errado e/ou posição de montagem do contador;
- Erros humanos na leitura dos contadores;

- Erros informáticos.

Já as perdas reais ou físicas nas redes de transporte e distribuição podem dever-se a roturas em condutas ou fugas nas ligações entre tubos (Nascimento, 2021). Essas perdas podem ser facilmente detetáveis através de alteração notória dos valores de caudal e pressão na conduta quando as fugas são de grandes dimensões e podem mesmo ser localizadas a olho nu, através do aparecimento de água em estradas, caminhos de terra batida ou terrenos em dias sem humidade. No entanto quando as fugas são de pequenas dimensões ou a água se infiltra em lençóis de água subterrâneos a deteção pode demorar bastante tempo.

Estando apresentadas as causas das perdas físicas de água, apresenta-se no subcapítulo 2.3 a evolução das estratégias de deteção dessas mesmas perdas.

## 2.3. Estratégias clássicas de deteção de fugas de água

Dados os constrangimentos que podem advir das fugas de água no transporte e distribuição, há bastante tempo que se tenta colmatar estas situações, tendo sido desenvolvidas diversas tecnologias ao longo dos anos no âmbito da localização de roturas (El-Zahab & Zayed, 2019; Sardinha et al., 2017). Segundo (Sardinha et al., 2017) em 1850 a deteção era feita utilizando uma vara de madeira que permitia escutar as fugas existentes em condutas enterradas e/ou ramais (varinha de escuta). Este foi o início da deteção acústica, que evoluiu e se mantém atualmente. Já nessa altura este método tinha resultados aceitáveis, mas a sua aplicação era difícil quando havia necessidade de escutar condutas e ramais de grandes dimensões.

Na Figura 8 encontra-se a evolução das técnicas de deteção de fugas ao longo dos anos, entre 1850 e 2005 de acordo com (Sardinha et al., 2017).

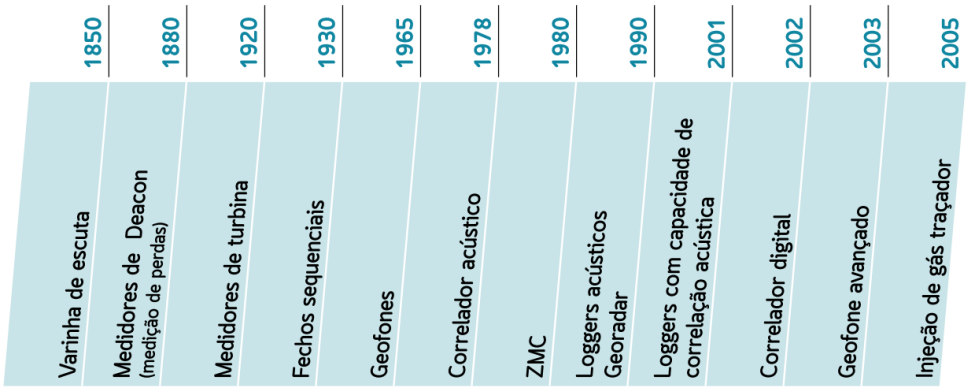


**Figura 8 – Evolução das técnicas de deteção de fugas desde 1850 até 2005. Retirado de (Sardinha et al., 2017)**

Na opinião dos autores houve um desenvolvimento notório quando se adotou a divisão do processo de localização de roturas em macro e micro-localização, que levam à identificação do local da fuga. A macro-localização permite identificar áreas com possíveis fugas de água dentro de Zonas de Monitorização e Controlo, onde são monitorizados os caudais de entrada e saída, com possibilidade de fecho progressivo de válvulas para análise do consumo de uma área específica, normalmente durante a noite quando o caudal é mais baixo. A micro-localização analisa essas áreas para identificar o local da rotura. Um possível método de micro-localização é a deteção acústica realizada através de correladores acústicos que permitem localizar as fugas de água com bastante fiabilidade, que evoluiu com a introdução do tratamento digital do ruído em 2002. Estes equipamentos utilizam microfones e geofones que amplificam o som e permitem uma localização mais eficaz.

Para os casos de grande dificuldade na deteção/localização de roturas pelos métodos acústicos utiliza-se, segundo (Sardinha et al., 2017), a injeção de gás inerte, que embora exija mais recursos (tempo e dinheiro) apresenta uma boa fiabilidade.

Para além da utilização de equipamentos acústicos, loggers, injeção de gás e georadar ou radar de penetração no solo, (El-Zahab & Zayed, 2019) refere ainda:

- Termografia com infravermelhos;
- Robôs de deteção;
- Sistemas microeletromecânicos sem fios como acelerómetros, sensores acústicos e sensores térmicos, que permitem uma monitorização contínua.
- Métodos baseados em análise de dados.

(Ferreira, 2017) classificou um conjunto de métodos de localização de acordo com o tipo de localização que permitem (aproximada ou exata). Essa classificação encontra-se na Tabela 2.

**Tabela 2 - Classificação de métodos de localização aproximada e exata. Retirado de (Ferreira, 2017)**

| Aproximada | Exata |
|---|---|
| *Step-test* (fecho progressivo de válvulas) | Sensores acústicos |
| *Noise logging* (Loggers acústicos) | Radar de penetração no solo |
| Modelação hidráulica | Injeção de traçadores |
| | Termografia |
| | Inspeção visual (solo e infraestruturas) |

O autor introduz mais dois métodos para além dos já abordados que são a modelação hidráulica, que permite simular a rede para identificar situações de rotura de forma aproximada e a inspeção visual do solo e das infraestruturas.

Parte destes métodos ou dos seus princípios foram utilizados pelos autores dos trabalhos consultados para aquisição dos dados que foram depois analisados com recurso às técnicas de Machine Learning, como por exemplo a modelação hidráulica, em (Caputo et al., 2003; Costa, 2020; Fan et al., 2021; Kammoun et al., 2023; Liu et al., 2019; Mashhadi et al., 2021), a divisão em zonas, (Mashhadi et al., 2021; Nascimento, 2021; Vivas et al., 2021), loggers acústicos sem fios em (Fares et al., 2023) e pequenos sensores de caudal combinados com tecnologias IoT, que permitem o envio de dados para um servidor (Alves Coelho et al., 2020).

# 3. O Sistema em Estudo

A EPAL, S.A. é uma empresa do grupo Águas de Portugal que tem como função o abastecimento público de água para consumo humano. O processo de abastecimento engloba captação, tratamento, transporte, armazenamento e distribuição de água, que no caso da EPAL são feitos em "alta" e em "baixa". O abastecimento em "alta" engloba captação, tratamento, armazenamento e transporte de água até um reservatório municipal. O abastecimento em "baixa" compreende a distribuição aos consumidores finais a partir de um reservatório municipal. Para além dos seus ativos, a EPAL é responsável (por concessão) pela gestão e exploração das instalações pertencentes à AdVT, que inclui abastecimento e saneamento com exceção da Zona Oeste, onde apenas é concessionária do sistema de abastecimento. Não se abordará no presente trabalho a parte do saneamento, uma vez que o estudo se resume ao sistema de abastecimento. (Departamento de Sustentabilidade Empresarial da EPAL, 2023-a; ERSAR, 2022)

A EPAL abastece em "baixa" o município de Lisboa, onde tem aproximadamente 350 mil clientes. É também responsável pelo abastecimento em "alta" de dois sistemas multimunicipais pertencentes à AdVT e à Águas do Ribatejo e ainda de mais 17 municípios, totalizando assim 34 municípios em "alta" (Pereira et al., 2022). Já a AdVT capta e trata água para abastecimento em "alta" a 70 municípios (Pereira et al., 2022).

Ao longo do presente capítulo será utilizado diversas vezes o termo "adução", que segundo (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2011), que cita o Glossário da Águas do Algarve, significa o transporte de água entre locais, como por exemplo entre uma captação de água e uma Estação de Tratamento de Água, uma Estação de Tratamento de água e a rede de distribuição ou mesmo entre reservatórios. No Sistema de Abastecimento da EPAL existem alguns adutores muito importantes, que permitem o transporte diário de grandes quantidades de água entre as Estações de Tratamento de Água e as redes de distribuição, sendo estes os Adutores de Castelo do Bode, do Tejo, do Alviela, Vila Franca de Xira-Telheiras e Circunvalação. Os adutores permitem em determinados pontos a passagem de água entre si para reforço de caudal em caso de necessidade e no caso dos adutores do Alviela e do Tejo podem ainda receber o apoio de captações existentes na Ota, em Alenquer e nas Lezírias de Vila Franca.

Nos próximos subcapítulos será feita uma breve apresentação do sistema de abastecimento da EPAL e Oeste, bem como dos sistemas de abastecimento que foram utilizados no presente trabalho.

## 3.1. O Sistema de Captação, Transporte e Distribuição de água da EPAL

O sistema de captação, transporte e distribuição da EPAL abastece, tal como já foi descrito, um vasto conjunto de municípios, conforme se pode verificar na Figura 9.

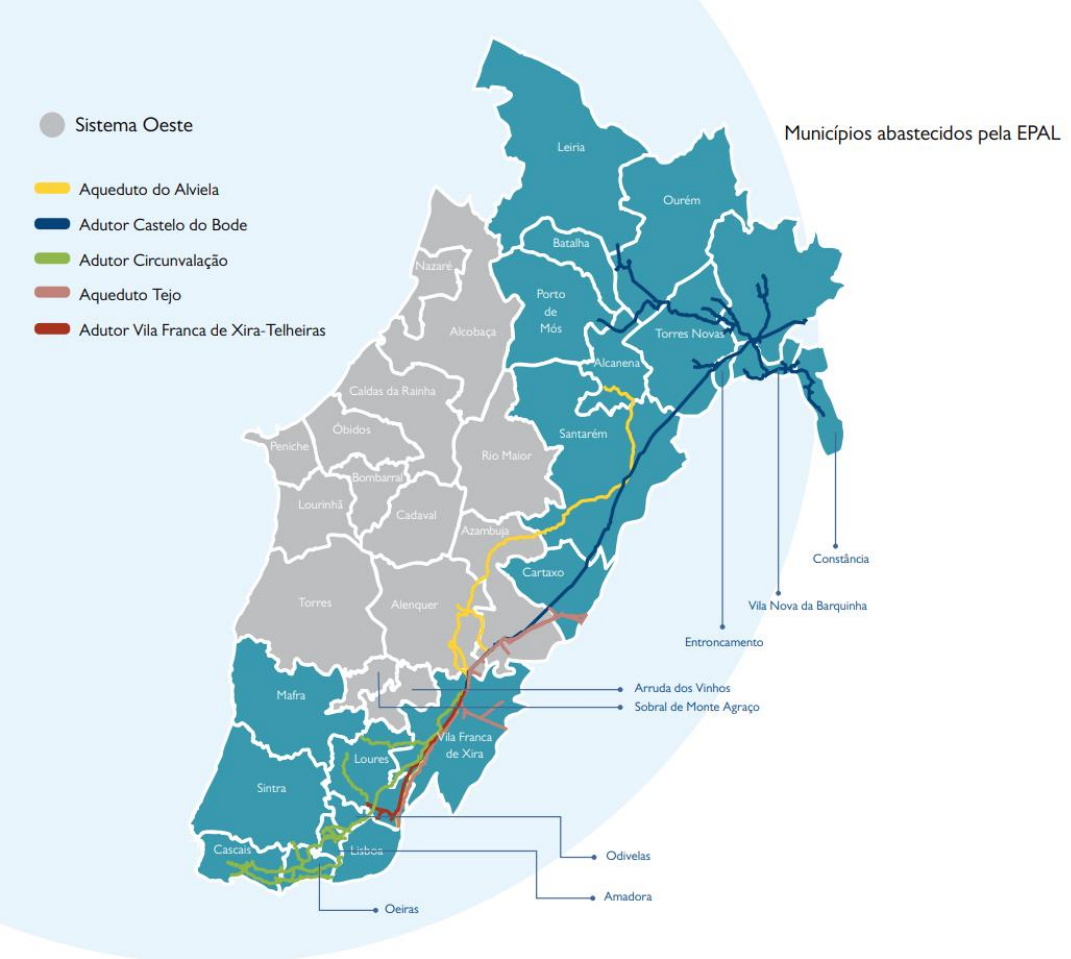


**Figura 9 - Municípios abastecidos pela EPAL. Retirado de (Departamento de Sustentabilidade Empresarial da EPAL, 2023-b)**

A EPAL teve em 2022 um volume total de água produzida (tratada) de 224,6 milhões de metros cúbicos ($Mm^3$), segundo (Departamento de Sustentabilidade Empresarial da EPAL, 2023-b), tendo essa água sido captada de duas formas, superficial e subterrânea.

As duas captações superficiais encontram-se na Barragem de Castelo do Bode, localizada na bacia hidrográfica do rio Zêzere, e no rio Tejo na localidade de Valada. Já as principais captações subterrâneas situam-se na Lezíria de Vila Franca de Xira. Existem ainda captações locais que permitem o abastecimento autónomo à zona onde se encontram, como por exemplo as captações existentes na Quinta da Vassala e em Alcoentre (Sistema Oeste). As captações locais são também utilizadas para reforçar os sistemas de abastecimento principais, como por exemplo as captações existentes em Reguengo Grande, que permitem um apoio precioso ao sistema de abastecimento da Zona Oeste durante o verão. No total, segundo (Departamento de Sustentabilidade Empresarial da EPAL, 2023-b), o Sistema EPAL e Oeste contou em 2022 com 113 captações de água, 1 069 quilómetros de condutas em "alta" e 1 449 em "baixa", permitindo que a água produzida chegasse a 63 estações elevatórias e a 65 reservatórios em "alta" e ainda a 11 estações elevatórias e 13 reservatórios em "baixa".

No Anexo III encontra-se uma apresentação do sistema de abastecimento da EPAL e Oeste, com referência às principais Captações e Estações de Tratamento de Água da EPAL e aos principais adutores, que permitem que a água chegue aos vários sistemas municipais (rede de abastecimento em "baixa") e à cidade de Lisboa.

## 3.2. Objeto de estudo do presente trabalho

Conforme já referido, o presente trabalho tem como objetivo a exploração de métodos que permitam detetar automaticamente perdas de água e funcionamento anómalo de equipamentos que possa colocar em causa a integridade das condutas de abastecimento, utilizando algoritmos baseados em aprendizagem supervisionada. Nos trabalhos consultados foram analisadas redes de abastecimento ao cliente final (abastecimento em "baixa"), não se tendo encontrado referências a trabalhos realizados em redes de abastecimento em "alta".

No presente trabalho a análise realizou-se em duas condutas de abastecimento em "alta", ou seja, abastecimento a reservatórios municipais que depois distribuem a água aos clientes

finais através de redes de abastecimento em "baixa". Quando ocorre uma fuga numa rede de abastecimento em "baixa", são afetados apenas os clientes afetos ao ramal de abastecimento onde a mesma acontece. Já numa fuga em rede de abastecimento em "alta" são afetados os reservatórios localizados a jusante da fuga e os localizados a montante até à válvula de seccionamento mais próxima, podendo uma situação destas afetar vários municípios. Sempre que é detetada uma rotura numa conduta é necessário seccioná-la, ou seja, fechar a válvula de seccionamento mais próxima a montante da fuga para que não se perca mais água e se possa reparar a anomalia. Desta forma corta-se não só a alimentação à fuga, mas também o abastecimento a todos os reservatórios a jusante da válvula. Pelas razões enumeradas é possível perceber que uma rotura num sistema de abastecimento em "alta" pode lesar um maior número de pessoas relativamente a uma rotura num sistema de abastecimento em "baixa". O tempo necessário para reparar a conduta pode ser também superior, uma vez que é necessário retirar a água do troço afetado para que a reparação seja possível, sendo necessário carregar novamente a conduta no final dos trabalhos.

Para realizar o estudo aqui apresentado foram selecionadas duas condutas de abastecimento, nas quais têm ocorrido algumas situações de rotura ao longo dos últimos anos. Uma delas é de abastecimento gravítico e a outra de abastecimento elevatório. A conduta de abastecimento gravítico realiza o transporte de água vinda do Sistema Norte-Centro (que se apresenta no Anexo III) até um conjunto de pontos de entrega/reservatórios pertencentes ao Município do Bombarral. Tem início no Reservatório do Alto da Serra e é constituída por tubo em PVC que nesse ponto tem diâmetro nominal (DN) 300mm. Esta conduta leva água a quatro pontos de entrega designados Pó R1, Pó Baixa, Baraçais e Caniceira. A cerca de 304 metros do Reservatório do Alto da Serra localiza-se uma derivação, onde o diâmetro da conduta passa para 250mm, mantendo-se o material de construção (PVC). Este troço com DN 250 tem uma extensão total de 8,21 quilómetros e é responsável pelo transporte de água até ao ponto de entrega (PE) de Baraçais, localizado a cerca de 5,38 quilómetros, e ao ponto de entrega da Caniceira, localizado a cerca de 8,21 quilómetros. Imediatamente a jusante da derivação encontra-se uma válvula redutora de pressão, que garante uma descida da pressão no troço de DN250 de cerca de 5 bar para valores que rondam os 3,8 a 4 bar. As leituras de pressões e caudal utilizadas para o estudo da conduta de abastecimento gravítico foram retiradas junto a essa válvula, através de um transdutor de pressão instalado a montante da mesma e outro a jusante. O caudalímetro encontra-se a jusante da válvula. O ponto de entrega da Caniceira, que abastece o reservatório de Vale do Leito, encontra-se muitas vezes fora de

serviço uma vez que pode ser abastecido através de outro ponto de entrega. Na Figura 10 encontra-se a representação esquemática da conduta.

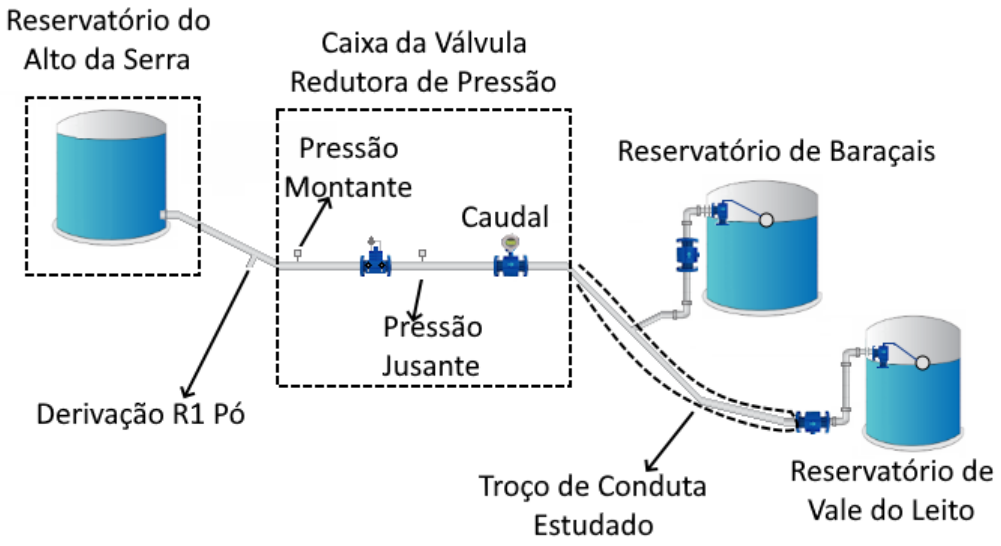


**Figura 10 – Representação esquemática da conduta de abastecimento gravítico**

As variáveis disponíveis para este caso de estudo foram apenas as referentes às leituras realizadas na caixa da válvula redutora de pressão. Presentemente encontra-se em processo de instalação um caudalímetro de entrada no reservatório municipal de Baraçais e já existe um medidor de nível no reservatório. Está também projetada a instalação de um caudalímetro no PE Caniceira. Futuramente, estas implementações poderão ajudar a melhorar as deteções de situações anómalas no troço de abastecimento.

A conduta de abastecimento elevatório analisada pertence aos sistemas ditos autónomos (apresentados no Anexo III). Inicia-se na Estação Elevatória de Virtudes e transporta água até ao Ponto de Entrega Pista de Corrida e ao Ponto de Entrega de Casais da Lagoa, numa extensão total de aproximadamente 3 quilómetros. Tem DN250 e a sua estrutura é em fibrocimento. O Ponto de Entrega Pista de Corrida tem consumo residual e funciona muito raramente. Esse consumo é apenas registado num totalizador mecânico, não havendo registos na base de dados do SCADA. A Estação Elevatória possui duas bombas que funcionam em regime de redundância. Na Figura 11 encontra-se a representação esquemática da conduta.

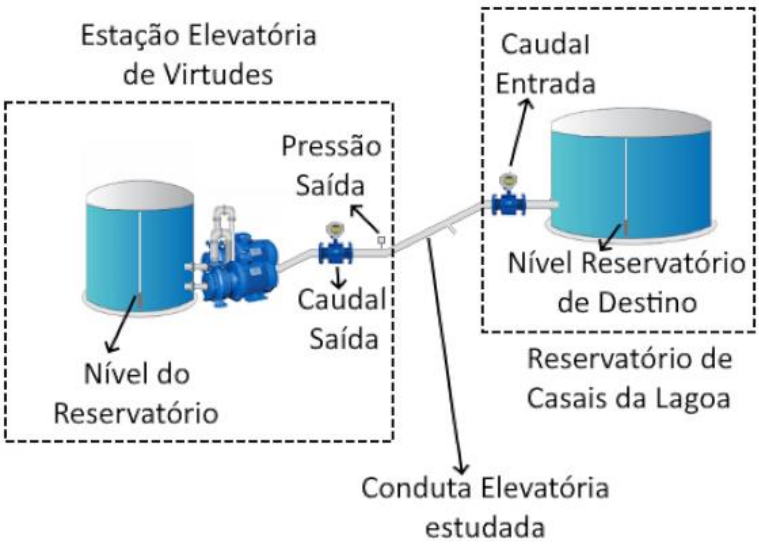


**Figura 11 – Pormenor da conduta de abastecimento elevatório**

Para que melhor se possa compreender a informação que se segue é necessário referir que o troço de aspiração de uma bomba corresponde ao troço de admissão de água da bomba, que neste caso é a tubagem existente entre o reservatório da Estação Elevatória e a bomba. Já o troço de compressão corresponde à tubagem de saída da bomba.

Para este caso existiu um conjunto mais alargado de variáveis disponíveis para análise. Relativamente à Estação Elevatória foram disponibilizados o nível do reservatório, o caudal de saída e a pressão na compressão. O nível do reservatório não foi utilizado, uma vez que esta instalação abastece de forma gravítica o Ponto de Entrega de Virtudes e, portanto, esta variável pode alterar-se por razões externas ao sistema em estudo. Já relativamente ao Ponto de Entrega de Casais da Lagoa foram disponibilizados o caudal de entrada e o nível do reservatório. Todos os valores disponibilizados são instantâneos.

Estando apresentados os casos de estudo, serão introduzidos no Capítulo 4 os algoritmos de aprendizagem supervisionada utilizados.

# 4. Algoritmos de Aprendizagem Automática Supervisionada

Os algoritmos de aprendizagem automática são presentemente bastante utilizados para diversos fins, como por exemplo diagnósticos médicos, cibersegurança, reconhecimento de imagens, reconhecimento de voz, deteção de fraudes e *chatbots* para apoio a clientes, por exemplo nos meios bancários (Varone et al., 2020; Wittel, 2022). Um exemplo de chatbot é o *Chat-GPT (Generative Pre-trained Transformer)* (Sharma et al., 2023), que possui capacidade de interpretação para criar respostas devidamente contextualizadas com as questões que lhe são colocadas (Deng & Lin, 2022). Esta ferramenta pode também criar resumos de obras literárias ou redigir uma carta um texto sobre um determinado tema (Sharma et al., 2023).

Para que um algoritmo de aprendizagem supervisionada obtenha os melhores resultados é imperativo que se realize uma pré-classificação dos dados a analisar. Num contexto de deteção de anomalias esse passo é preponderante para que o algoritmo possa aprender os padrões apresentados em situações normais e em situações anómalas. Desta forma será possível detetar anomalias quando lhe são apresentados novos dados. Há que ressalvar que uma anomalia cujos padrões sejam significativamente diferentes dos existentes nas anomalias apresentadas ao algoritmo no processo de treino provavelmente não será detetada, uma vez que o algoritmo apenas aprende a identificar os padrões que lhe são apresentados ao longo desse processo.

Tipicamente uma conduta de abastecimento de água funciona sem qualquer anomalia na maior parte do tempo, registando pontualmente situações de rotura, sendo expectável que os conjuntos de dados que caracterizam o seu funcionamento sejam fortemente desequilibrados, com grande parte das amostras a representar situações de funcionamento normal e uma pequena parcela de amostras a representar anomalias (Fan et al., 2021). Esta situação pode ter consequências mais ou menos graves, consoante a vulnerabilidade dos algoritmos que sejam utilizados. (Muharemi et al., 2019) conclui pelos testes que realizou que os algoritmos *Logistic Regression* (LR), *Linear Discriminant Analysis*, *Support Vector Machines* (SVM), *Artificial Neural Network* (ANN), *Deep Neural Network* (DNN), *Recurrent Neural Network* (RNN) e *Long Short-Term Memory* (LSTM) são influenciados pelo desequilíbrio, sendo SVM, ANN e LR os menos afetados. Também (Fan et al., 2021) refere que o algoritmo ANN

é afetado por esta situação. No presente trabalho foram realizados alguns testes utilizando reamostragem aleatória dos dados de treino. Utilizou-se a sobreamostragem, que segundo (Brownlee, 2021) consiste em duplicar aleatoriamente dados da(s) classe(s) minoritária(s), que são posteriormente adicionados aos dados de treino para equilibrar o número de amostras das diferentes classes existentes. Foi ainda utilizada a subamostragem, que segundo (Brownlee, 2021) consiste na exclusão de amostra(s) da(s) classe(s) maioritária(s). Não se obteve uma melhoria de performance com estas abordagens, pelo que os resultados desses testes não são aqui apresentados.

## 4.1. Aquisição, análise, seleção e visualização de Dados

Para que se possa treinar um algoritmo é necessário ter um conjunto de dados para analisar. No caso do presente trabalho esse conjunto foi importado de um servidor que guarda o histórico de informação das instalações da zona Oeste e permite a realização de gráficos históricos no sistema SCADA existente. A periodicidade de gravação desses dados é de 1 minuto e o número de variáveis lidas para cada instalação depende da sua tipologia e dos equipamentos instalados. Para o caso dos pontos de entrega existem normalmente duas variáveis que poderão ser importantes para análise, a pressão a montante da válvula redutora de pressão (pressão na conduta) e o caudal de entrada, que poderão servir para encontrar problemas de regulação das válvulas. Podem ainda existir outros valores de pressão intermédia, percentagem de abertura da válvula reguladora de caudal e medição de nível do(s) reservatório(s) municipal(is), caso existam equipamentos para medição desses valores no terreno e o servidor registe esses dados. Já no caso das estações elevatórias pode haver registos de pressão e caudal de entrada, nível do reservatório, pressões e caudais de saída e ainda percentagem de abertura das válvulas. Dependendo do número de grupos elevatórios e do número de destinos independentes para onde a instalação eleve, podem existir várias medições de caudais e pressões de saída.

Existem ainda algumas instalações que servem para seccionamento, derivação ou *bypass*, onde normalmente se medem pressões e caudais.

Neste trabalho foram analisados dois casos específicos, onde se verificaram várias roturas ao longo dos anos.

No primeiro caso trata-se de uma conduta de abastecimento gravítico situada no Sistema Centro, que liga o reservatório do Alto da Serra aos Reservatórios de Baraçais e Caniceira (como descrito anteriormente na secção 3.2). Os dados a analisar foram recolhidos numa caixa localizada a jusante do Reservatório do Alto da Serra. Nessa caixa encontra-se uma válvula redutora de pressão, que permite regular a pressão de saída. Existe um *bypass* a essa válvula que permite carregar a conduta lentamente em caso de necessidade (por exemplo após reparação de rotura a jusante da caixa). Aqui são medidas as pressões a montante e jusante da válvula redutora de pressão e o caudal que passa na conduta. Foram consideradas essas 3 variáveis de entrada, embora por vezes se tenham utilizado apenas 1 ou 2 variáveis.

No segundo caso estudado considerou-se uma conduta de elevação por bombagem. Para tal utilizaram-se dados recolhidos na Estação Elevatória de Virtudes e no Ponto de Entrega de Casais da Lagoa, para onde a Estação Elevatória eleva. Existe um outro Ponto de Entrega no troço da conduta que não comunica com o Centro de Comando, pelo que não há registo de consumo dessa instalação.

Os dados retirados do servidor encontravam-se num conjunto de ficheiros no formato CSV (Comma Separated Values), cada um com os dados recolhidos para uma dada variável numa dada instalação. Assim existiam tantos ficheiros para cada instalação quantas as variáveis relevantes (pressão, caudal, nível, percentagem de abertura de válvulas…) existentes no SCADA para essa instalação.

O primeiro passo para que fosse possível trabalhar com os dados foi juntá-los num ficheiro comum por instalação. Todos os ficheiros representavam séries temporais, tendo-se verificado que por vezes havia duas leituras para a mesma data e hora (HH:MM). Assumiu-se que tal ocorreu devido a testes de gravação ou a erros de leitura. Desta forma foi necessário eliminar os elementos em excesso para cada minuto utilizando funções da biblioteca "*pandas*" (Pandas Devlopers, 2024). Deixou-se ficar o primeiro valor para cada série de valores duplicados. Utilizando a mesma biblioteca juntaram-se todas as variáveis relativas a cada instalação num só ficheiro. Criou-se um programa base em linguagem *Python* onde foi apenas necessário alterar a referência da instalação e o número de variáveis existentes de cada tipo (pressões, caudais…) para realizar as ações descritas anteriormente e gerar um ficheiro por instalação com todas as variáveis que lhe estão associadas.

Não se realizou no programa atrás referido qualquer normalização ou remoção de valores em falta, deixando-se essa fase para os algoritmos a desenvolver, de acordo com as necessidades que se encontrassem.

Uma vez que uma rotura pode durar várias horas e que existem registos de dados a cada minuto, é percetível que existe um conjunto alargado de dados relativos a cada rotura. O mesmo acontece para os eventos de alteração de regime após rotura, situações em que devido ao carregamento da conduta há alterações ao regime normal de funcionamento. Os caudais e as pressões relativos a situações de rotura e/ou carregamento de conduta podem assumir variações de grande amplitude e imprevisíveis. Quando comparadas com outras situações do mesmo tipo não existe um comportamento padrão.

Conforme já referido, uma das fases cruciais para que seja possível compreender a capacidade de classificação ou previsão de um algoritmo de aprendizagem supervisionada é a pré-classificação dos dados que vão ser utilizados para treino, teste e validação do modelo.

O conjunto de dados deve ser analisado amostra a amostra, para encontrar padrões que permitam definir se se trata de uma situação normal ou anómala. De acordo com o resultado dessa análise classifica-se a amostra.

No presente trabalho elaboraram-se gráficos para ajudar a classificar as amostras de dados. A título de exemplo, no caso da conduta de abastecimento gravítico existiam três variáveis de entrada, as pressões a montante e a jusante da válvula de regulação e o caudal, que se encontram representadas na Figura 12, onde se encontram já identificados alguns eventos que saltam à vista, nomeadamente um conjunto de roturas, que na figura se encontram rodeadas com retângulos e algumas alterações de regime, que se encontram rodeadas com elipses. No gráfico mais acima encontra-se representada a pressão a montante da válvula redutora de pressão cuja unidade é “bar”, no gráfico ao centro encontra-se representada a pressão a jusante da válvula redutora de pressão cuja unidade é “bar” e no gráfico de baixo encontra-se representado o caudal, cuja unidade é “$m^3/h$”.

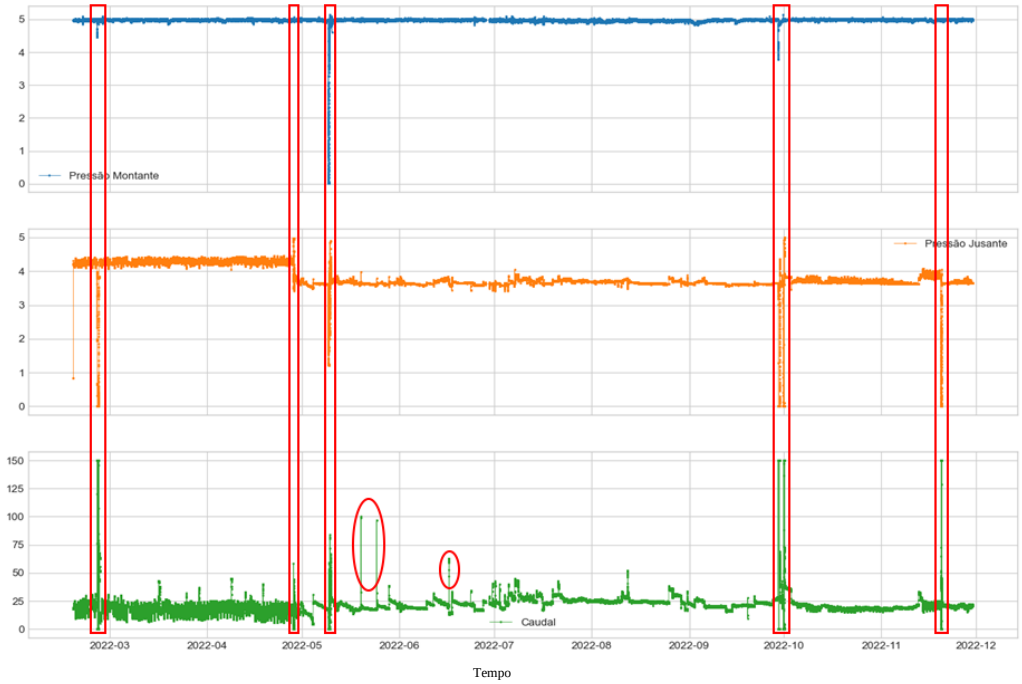


**Figura 12 – Conjunto de dados a analisar para a conduta de abastecimento gravítico (pressão em bar e caudal em $m^3/h$)**

Essas situações foram analisadas ponto a ponto para tentar compreender o comportamento das variáveis de entrada para cada uma delas. Confrontaram-se os gráficos com os registos de manutenção existentes para melhor identificar as anomalias detetadas. Esses registos encontram-se no software MAXIMO da IBM (International Business Machines), que é utilizado para gestão de ativos, trabalhos, inventários, serviços, contratos e compras. Foi realizada uma análise cuidada do gráfico para sinalizar dois tipos de defeito, que são as situações de rotura e as alterações de regime de funcionamento. Numa situação de rotura verifica-se a existência de um aumento brusco de caudal acompanhado de alterações nas pressões. Já as situações de alteração de regime de funcionamento comtemplam o carregamento da conduta após roturas e as alterações ao normal funcionamento do sistema, que representam alterações de pressão e de caudal menos bruscas. Caso as situações de alteração de regime se devam a carregamento de conduta, existe acompanhamento no local por parte de técnicos que controlam a velocidade de carregamento para que não apareçam novas roturas. Já as alterações de funcionamento devido a variações de caudal e pressão devem ser estudadas, pois podem ser causadas por defeito de funcionamento de equipamentos ou por roturas de pequena dimensão. A pré-classificação de dados realizada será abordada na secção 5.1.

Com o objetivo de prevenir situações de *overfitting*, em que o algoritmo fica demasiado familiarizado com os dados de treino para os quais obtém bons resultados de classificação, mas não consegue prever corretamente situações provenientes de novos conjuntos de dados, dividiu-se o conjunto de dados em subconjunto de treino (80%) e subconjunto de teste (20%) utilizando o "*train_test_split*" da biblioteca *Scikit Learn (Scikit-Learn Developers*, 2024-d*)*. Desta forma foi garantida a aleatoriedade na seleção dos dados para cada um dos subconjuntos.

## 4.2. Aprendizagem Supervisionada

Qualquer que seja o algoritmo de aprendizagem supervisionada utilizado, é de grande importância a seleção dos hiperparâmetros que garantam o melhor desempenho possível. Uma ferramenta que ajuda nessa tarefa é o *Optuna* (Optuna Contributors, 2018), que seleciona os melhores hiperparâmetros a partir de uma gama de valores pré-definida pelo utilizador. Há que definir uma função objetivo, onde são definidos possíveis valores a atribuir a cada um dos hiperparâmetros e o objetivo do estudo que, no caso específico deste trabalho, é maximizar a pontuação média resultante da validação cruzada do modelo. O utilizador define o número de combinações de hiperparâmetros a testar através do parâmetro "*n_trials*", na definição da otimização do modelo. O *Optuna* seleciona os hiperparâmetros que obtiverem o melhor resultado na satisfação da função objetivo.

O método de validação cruzada de dados divide os dados que lhe são apresentados em 'n' subconjuntos diferentes (*N-Fold Cross-Validation*). Em seguida realiza 'n' simulações em que todos os subconjuntos são utilizados uma vez como dados de teste, utilizando os restantes 'n-1' subconjuntos como dados de treino. Na Figura 13 encontra-se a divisão que é realizada para 5 subconjuntos, estando identificados a verde os subconjuntos de treino e a azul os subconjuntos de teste nas 5 divisões (*splits*) realizadas.

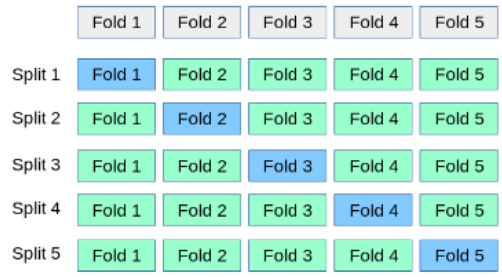


**Figura 13 – Princípio de funcionamento do método de validação cruzada. Adaptado de (*Scikit-Learn Developers*, 2024-a)**

O método *cross_val_score* fornece as pontuações obtidas para os processos de validação cruzada. No caso das funções objetivo criadas para este trabalho considerou-se o valor médio das 'n' simulações realizadas para cada combinação de hiperparâmetros.

Relativamente às métricas de erro utilizadas para avaliação dos modelos, apresentam-se os resultados obtidos para precisão, sensibilidade e *F1-Score* (Saias et al., 2018). Utilizou-se para avaliação dos modelos o *Macro Average F1-Score*, tendo em conta que este indicador é apurado através do valor médio entre as classes, não sendo afetado pelo desequilíbrio existente entre as mesmas (Saias et al., 2018).

Nos subcapítulos 4.2.1 a 4.2.4 serão apresentados os algoritmos utilizados no presente trabalho.

### 4.2.1. *Random Forest*

O algoritmo *Random Forest* deriva do algoritmo *Decision Tree* (ou Árvore de Decisão) que consiste na criação de um esquema em árvore que permite classificar eventos. Um exemplo simples de criação de uma árvore de decisão é dado por (Didática Tech, 2024), com um problema que consiste na questão "Vou para a praia?".

Para criar a árvore de decisão foram utilizados os dados da Tabela 3

**Tabela 3 – Tabela exemplificativa para criação de árvores de decisão, adaptado de (*Didática Tech*, 2024)**

| Dia | Sol? | Vento? | Vou para a praia? |
|---|---|---|---|
| 1 | Sim | Sim | Não |
| 2 | Sim | Sim | Não |
| 3 | Sim | Não | Sim |
| 4 | Não | Não | Não |
| 5 | Não | Sim | Não |
| 6 | Não | Sim | Não |

Analisando a tabela verifica-se que em dias sem sol a resposta à questão "Vou para a praia?" é sempre "não". Por outro lado, em dias de sol é sempre verificada a existência de vento para definir a resposta. Vendo a questão por outro prisma, nos dias em que há vento a resposta é sempre "não", mas nos dias em que não há vento a decisão depende de haver sol. Esta situação pode criar duas árvores de decisão, que se encontram na Figura 14.

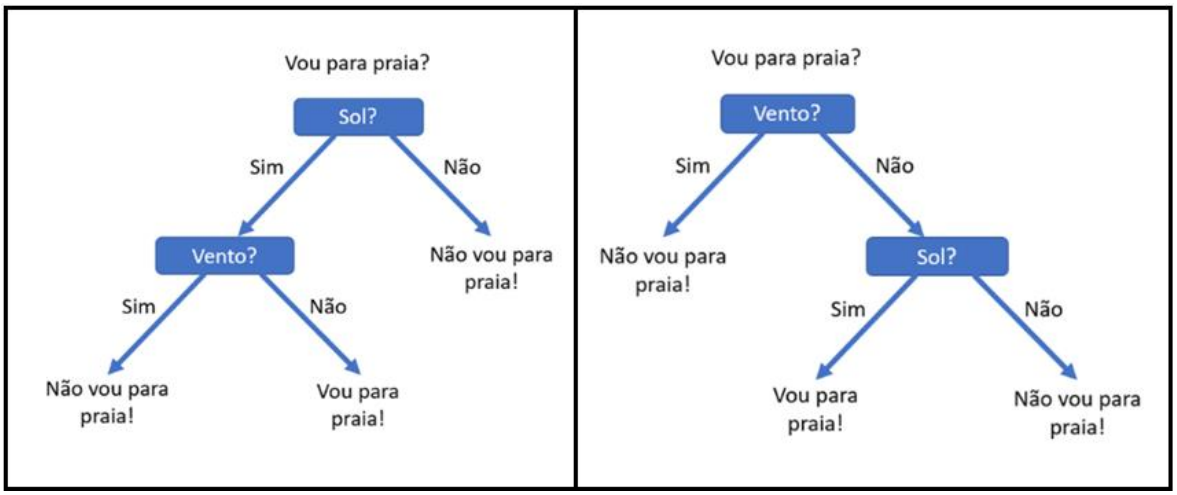


**Figura 14 – Árvores de decisão possíveis para o caso apresentado, retirado de (*Didática Tech*, 2024)**

Os campos que representam condições a verificar (“Sol?” e “Vento?”) são os nós da árvore, cuja construção se inicia pelo nó raiz que se encontra no seu topo. Cada resposta possível à questão colocada num dado nó dá acesso a um ramo, que representa o caminho a seguir se essa for a resposta escolhida. A árvore termina nas possíveis decisões a tomar, neste caso “Vou para a praia!” ou “Não vou para a praia!”, que representam as folhas da árvore.

Segundo (*Didática Tech*, 2024), embora possamos ter diferentes árvores de decisão os resultados podem ser os mesmos, o que é facilmente percetível neste exemplo.

Após criar esta árvore, é possível alguém aplicá-la diariamente para decidir se vai ou não à praia de acordo com as condições climatéricas. Isto mostra a aplicabilidade deste tipo de algoritmo à classificação de novos dados.

Conforme foi já referido, este é um exemplo simples com apenas duas variáveis de entrada, em que ambas têm influência na saída. A árvore gerada é bastante simples. Existem, no entanto, casos com muito mais variáveis de entrada em que nem sempre é linear a escolha das variáveis de entrada a utilizar nem onde colocar cada uma delas. A escolha da variável a colocar no nó raiz e de quais as variáveis que ficarão nos restantes nós, tem influência no resultado. Nas aplicações computacionais o algoritmo define as alocações das variáveis de entrada de acordo com a relação de cada uma delas com a saída, otimizando a árvore para obter os melhores resultados.

Compreendido o princípio de funcionamento do método *Decision Tree* é mais fácil compreender o funcionamento do *Random Forest*, que utiliza um conjunto de árvores para tomar uma decisão. No processo de criação de cada árvore são selecionadas aleatoriamente algumas amostras do conjunto de dados inicial utilizando o método de reamostragem *bootstrap*, que permite a existência de amostras repetidas no conjunto de dados a utilizar. Essas amostras são utilizadas na otimização dessa árvore.

A seleção dos nós de cada árvore acontece de forma aleatória, pelo que em determinadas situações pode haver variáveis de entrada que não sejam selecionadas para uma determinada árvore. É também usual a existência de árvores com desempenho inferior devido às variáveis de entrada selecionadas.

Após ser criado e otimizado um conjunto pré-definido de árvores, o algoritmo está pronto para analisar novos conjuntos de dados (entradas). Para tal, o conjunto de dados é apresentado a cada uma das árvores de decisão que o avalia. Caso se trate de um problema

de classificação, o resultado da avaliação é aquele que for apresentado por mais árvores. Se for um problema de regressão, o resultado é a média de todas as avaliações.

Neste trabalho foi utilizado o algoritmo *RandomForestClassifier* da biblioteca *Scikit Learn*. Conforme já foi referido, este algoritmo utiliza um conjunto de árvores de decisão, sendo imperativo definir a quantidade de árvores a utilizar através do parâmetro *n_estimators*. O intervalo utilizado na pesquisa foi [1, 101].

A profundidade máxima da árvore, ou *max_depth*, representa a distância máxima entre o nó raíz e cada uma das folhas e variou no intervalo [1, 121] ao longo da otimização. Também o número máximo de nós terminais ou folhas de cada árvore é parametrizável através do hiperparâmetro *max_leaf_nodes*, que foi otimizado no intervalo [2, 102]. Otimizou-se também o número mínimo de amostras por folha, *min_samples_leaf* e o número mínimo de amostras presentes num nó para que este se possa dividir, *min_samples_split*, ambos no intervalo [2, 38].

### 4.2.2. *XGBoost*

O método "*XGBoost"* ou *"eXtreme Gradient Boosting"* é, também ele, baseado em árvores de decisão. É bastante utilizado em conjuntos de dados tabulares e lida bem com relacionamentos não lineares entre dados de entrada e valores alvo (Masui, *T*, 2022).

Este método cria um primeiro modelo (árvore de decisão) e analisa o seu erro de classificação, que tentará otimizar ao longo do processo de treino. Para tal vai criando sequencialmente árvores de decisão baseadas nas anteriores com o objetivo de minimizar o erro de previsão a cada novo modelo, através da minimização de uma função de perda. Este método gera vários modelos, que poderão até ser considerados fracos, para criar um modelo robusto.

A seleção de instâncias para as novas árvores não é totalmente aleatória, pois são preferidas as instâncias consideradas mais difíceis de prever nas árvores anteriores. Para (Glander, 2018) este processo garante uma aprendizagem baseada nos erros cometidos anteriormente.

O erro de cada modelo contribui para a criação do modelo seguinte numa dada proporção que é definida através da taxa de aprendizagem (*learning rate*). Desta forma é possível

limitar a contribuição de cada uma das árvores criadas para o modelo final, que deverá ter uma margem de erro baixa (valores previstos bastante próximos dos reais no caso da regressão e assertividade na rotulagem dos dados de entrada no caso da classificação) graças à contribuição de todos os modelos criados ao longo do processo. Na Figura 15 encontra-se um gráfico que demonstra a descida da margem de erro ao longo das diversas iterações que consistem na adição de novas árvores de decisão ao modelo para o tornar mais eficiente, reduzindo o erro associado às etiquetagens de dados que realiza.

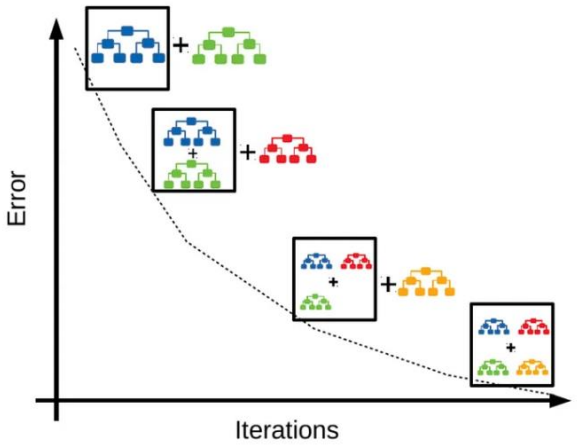


**Figura 15 – Variação do erro ao longo do processo de treino dos modelos baseados em *Gradient Boosting*, retirado de (Hemashreekilari, *2023*)**

Este classificador é bastante utilizado em problemas de classificação, especialmente na presença de grandes quantidades de dados. (Kazmi, 2024)

No presente trabalho foi utilizado o classificador *XGBClassifier* da biblioteca *XGBoost*. Sendo este classificador uma variação do *Random Forest*, os hiperparâmetros *n_estimators* e *max_depth* estão também presentes aqui e foram ambos otimizados no intervalo [1, 101]. Otimizou-se também o peso mínimo de cada novo nó, *min_child_weight*, no intervalo [1, 21]. O parâmetro de regularização, *gamma*, representa a redução mínima de perda necessária para criar uma nova divisão na árvore. Quanto maior o valor deste parâmetro mais conservador se torna o algoritmo. Este parâmetro foi otimizado para o intervalo [0, 100]. Também a fração de amostras e a proporção de subamostragem utilizadas na construção de cada árvore foram otimizadas utilizando os parâmetros *subsample* e *colsample_bytree*, respetivamente. Estes valores foram otimizados para o intervalo [0.1, 1].

Outro hiperparâmetro relevante é a taxa de aprendizagem do algoritmo, *learning_rate*, já referido anteriormente, que foi otimizado no intervalo [1e-5, 0.99991].

### 4.2.3. *Support Vector Machines*

Os modelos de aprendizagem baseados em máquinas de vetores de suporte (*Support Vector Machines* ou SVM) são bastante utilizados para classificação binária segundo (Boswell, 2002) e (Gandhi, 2018), e possuem boa capacidade de reconhecimento de padrões, podendo mesmo em alguns casos permitir obter resultados melhores que as redes neuronais artificiais (Lorena et al., 2007). As SVM têm implícito um processo estruturado de minimização de erro, contrariamente ao que acontece com as ANN, que têm implícito um processo empírico.

Estes modelos estimam a localização ótima de uma fronteira (hiperplano) que separa os pontos de um conjunto de dados em duas classes, definindo assim o limite de decisão entre elas. A otimização consiste na maximização da distância entre essa fronteira e os pontos de ambas as classes que lhe ficam mais próximos. Se existirem 2 variáveis de entrada, a fronteira será uma linha. Já na presença de 3 variáveis de entrada a fronteira será um plano (Gandhi, 2018), conforme a Figura 16.

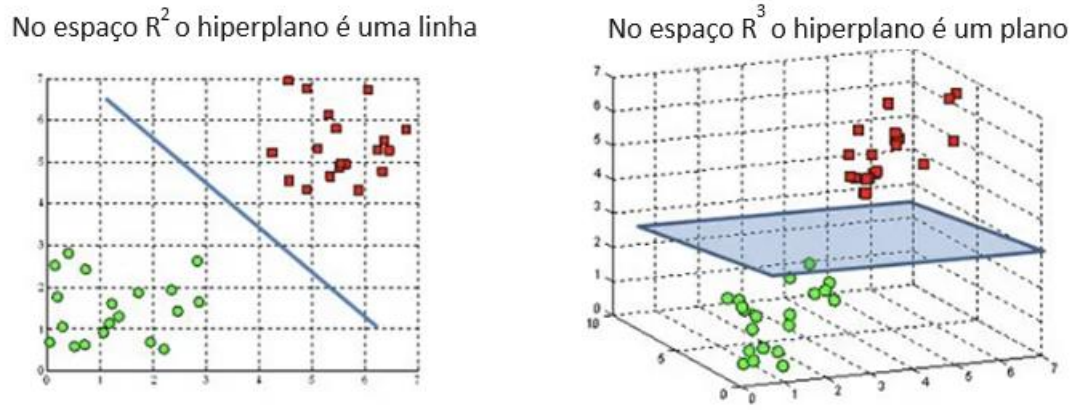


**Figura 16 – Representação de hiperplanos nos espaços R2 e R3, adaptado de (Gandhi, 2018)**

Os vetores de suporte assumem um papel muito importante na criação do hiperplano e são os pontos de cada classe que se encontram mais próximos deste. Conforme se pode constatar pela Figura 17, uma alteração no conjunto de dados que retire os vetores de suporte vai

obrigar à criação de um novo hiperplano, uma vez que as distâncias entre estes elementos serão diferentes.

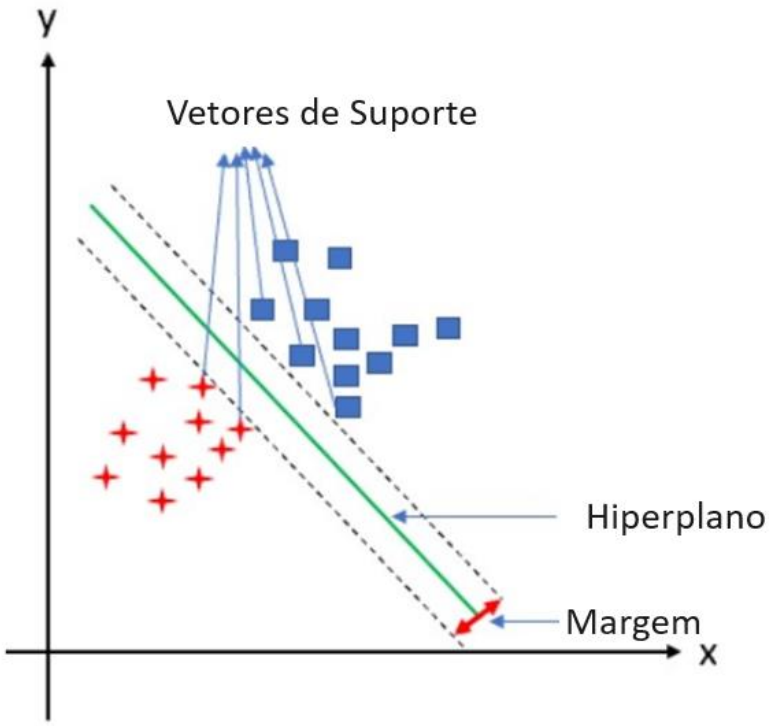


**Figura 17 – Representação gráfica de Vetores de Suporte e Hiperplano a que dão origem, adaptado de (Araújo, 2022)**

A distância entre os vetores de suporte e o hiperplano designa-se margem.

Ao serem apresentados novos dados para classificação ao algoritmo treinado, o hiperplano (limite de decisão) será utilizado para ajudar a definir qual a classe a que pertencem.

Neste estudo utilizou-se a implementação SVC da biblioteca *Scikit Learn*, para a qual se otimizaram os parâmetros C, *gamma*, *kernel* e *degree*.

O parâmetro C denomina-se parâmetro de regularização e é inversamente proporcional ao tamanho da margem, o que significa que quanto maior for C menor será a distância entre os vetores de suporte. Este parâmetro representa ainda o custo (penalização) que se pretende atribuir às más classificações. Se C assumir um valor baixo, a margem será muito grande e consequentemente será tolerado um maior número de más classificações. Já um valor elevado de C gera uma margem mais apertada que não permite tantas más classificações, embora aumente o risco de sobreajuste do modelo *(overfit)*. Para obter um valor para este parâmetro que não tornasse o modelo sobreajustado mas que também não levasse a um

número elevado de más classificações, recorreu-se à sua otimização. O intervalo utilizado foi [1, 81].

Já o parâmetro *gamma* define a distância entre os pontos considerados na escolha do limite de decisão. Quanto maior o valor de *gamma*, menor será a distância entre os pontos a considerar, conforme ilustrado na Figura 18. Um valor de *gamma* muito elevado (Figura 18(a)) leva a uma aproximação do limite de decisão aos pontos que lhe são mais próximos, havendo o risco de *overfit* do classificador. Já um valor mais baixo torna o classificador mais estável e diminui as possibilidades de *overfit* (Figura 18(b)).

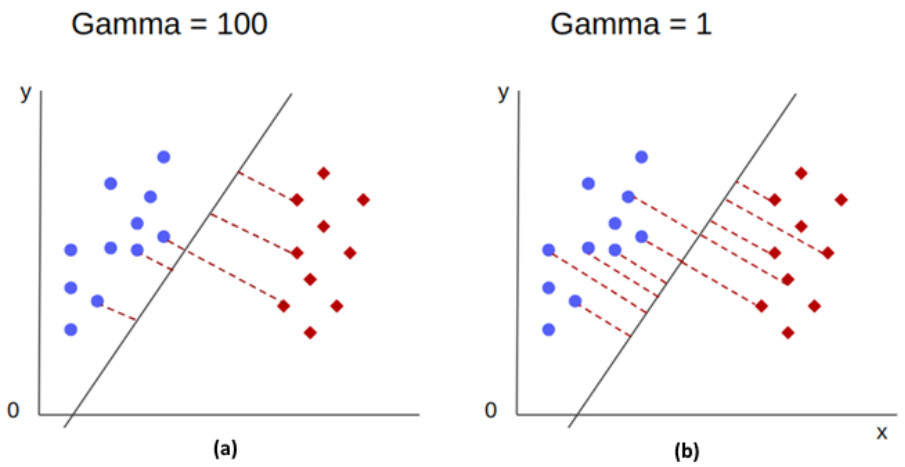


**Figura 18 – Influência do parâmetro *gamma* no limite de decisão. Retirado de (Sampaio, 2023)**

O intervalo de seleção utilizado para o parâmetro *gamma* foi [1, 47.8].

Ao utilizar funções matemáticas específicas, o algoritmo SVC necessita de uma ferramenta que transforme os dados de entrada num formato com que essas mesmas funções possam trabalhar. A ferramenta utilizada é a função *kernel*, que para este algoritmo específico pode traduzir-se numa função polinomial (‘*poly*’), sigmoide (‘*sigmoid*’), linear (‘*linear*’), função de base radial (‘*rbf*’) ou pré-definida pelo utilizador (‘*precomputed’*). Para este parâmetro foram utilizadas as funções polinomial, sigmoide, linear e base radial na otimização dos hiperparâmetros. Foi ainda otimizado o grau do polinómio, que apenas é relevante quando se utiliza a função polinomial. O intervalo de seleção definido foi [1, 4].

### 4.2.4. *Artificial Neural Networks*

Os modelos baseados em redes neuronais artificiais são geralmente utilizados para resolver problemas complexos, que tradicionalmente apenas eram resolvidos por humanos. Para (Santos, 2022), as redes neuronais têm em comum com o cérebro humano a capacidade de adquirir conhecimento através de aprendizagem e a capacidade de armazenar esse conhecimento nas ligações entre os neurónios, através dos pesos sinápticos.

Na Figura 19 encontra-se a configuração de uma rede neuronal simples, constituída por um único neurónio.

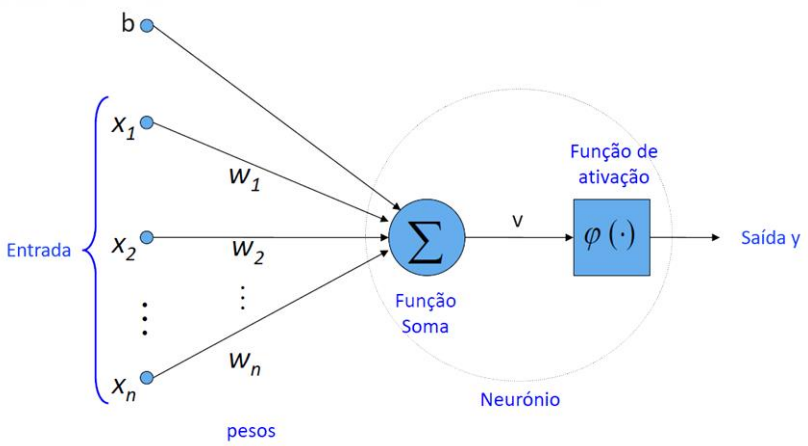


**Figura 19 – Representação de rede neuronal simples, adaptado de (Santos, 2022)**

À esquerda encontram-se as entradas do modelo, $x_1$ a $x_n$, que são multiplicadas pelos pesos sinápticos $w_1$ a $w_n$. Já no interior do neurónio dá-se a soma de todos os valores obtidos anteriormente com o viés (b), que é uma constante. O valor resultante da soma (v) é obtido através da seguinte fórmula:

$$v = \sum_{i=1}^{n} (x_n . w_n) + b$$

Ainda no neurónio este resultado é submetido a uma função de ativação que vai gerar a saída (y). A função de ativação pode ser linear ou não linear.

Segundo (Vulimiri & Stebner, 2020), as funções de ativação mais utilizadas em redes neuronais artificiais são as apresentadas na Figura 20.

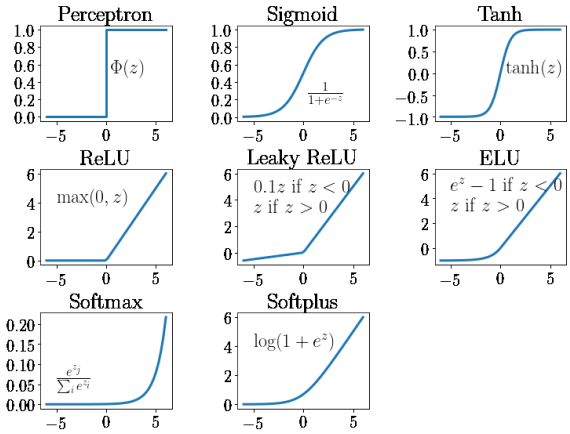


**Figura 20 - Funções de ativação mais utilizadas em ANN, retirado de (Vulimiri & Stebner, 2020)**

As redes neuronais artificiais podem ter múltiplos neurónios na sua constituição, que podem ser agrupados em várias camadas ocultas e a sua complexidade aumenta com o aumento do número de neurónios e de camadas ocultas que possuam. As saídas dos neurónios de uma camada contribuem para as entradas dos neurónios da camada seguinte, conforme se pode ver na Figura 21.

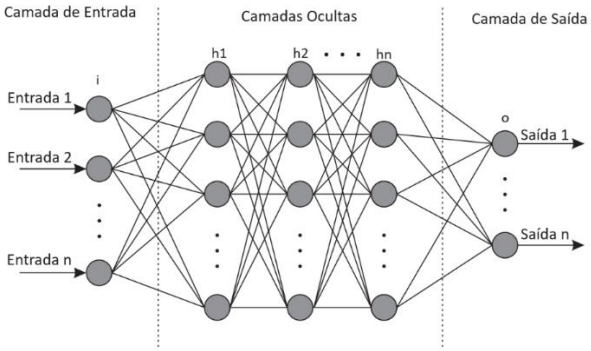


**Figura 21 – ANN com várias camadas ocultas, adaptado de (Bre et al., 2018)**

Da mesma forma que ocorre com a rede apresentada na Figura 19, existe associado à saída de cada neurónio um peso sináptico (w), que define a sua contribuição para a entrada dos neurónios da camada seguinte, conforme se pode ver na Figura 22.

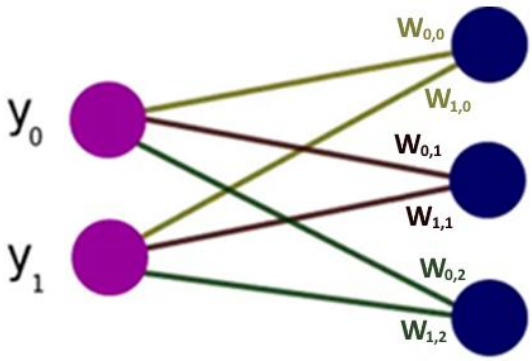


**Figura 22 – Pormenor de parte de uma ANN, adaptado de (Grübler, 2018)**

As redes neuronais são treinadas através de iterações, que permitem afinar os valores dos pesos sinápticos e do viés. Numa fase inicial são atribuídos a estes parâmetros valores aleatórios que são depois ajustados através da retropropagação do erro dos resultados obtidos relativamente ao valor real da saída, com o objetivo de minimizar esse erro e assim garantir um bom desempenho do modelo.

O classificador utilizado no estudo realizado foi o *Multilayer Perception Classifier (MLPClassifier)* da biblioteca *Scikit Learn*. Este classificador permite a utilização de múltiplas camadas ocultas com diferentes quantidades de neurónios por camada, utilizando a mesma função de ativação para todas elas.

As funções de ativação disponíveis para este classificador são as seguintes:

| Designação | Equação Matemática |
|---|---|
| **Identidade ('identity')** | $f(x) = x$ |
| **Sigmoide ('sigmoid')** | $f(x) = \frac{1}{1 + e^{-x}}$ |
| **Tangente hiperbólica ('tanh')** | $f(x) = \tanh(x)$ |
| **Unidade linear retificada ('relu')** | $f(x) = \max(0, x)$ |

Nos testes realizados, e para encontrar os hiperparâmetros que melhor se adequam às situações em estudo, utilizaram-se todas as funções de ativação.

Os '*solvers*' são algoritmos que guiam o processo de atualização dos parâmetros da rede (pesos sinápticos e vieses), otimizando o seu valor para garantir o melhor desempenho possível do modelo. O *MLPClassifier* pode utilizar os algoritmos '*lbfgs*', '*sgd*' e '*adam*' para realizar essa otimização. O primeiro pertence à família dos algoritmos quasi-Newton, o

segundo e o terceiro derivam do método do gradiente. Nos testes utilizaram-se todos os '*solvers*'. Utilizou-se também o parâmetro '*momentum*', que apenas tem utilidade quando se utiliza o solver '*sgd*' e, segundo (Bushaev, 2017), tem a função de tornar a convergência do '*solver*' mais rápida. Este parâmetro encontra-se normalmente no intervalo [0,1] (*Scikit-Learn Developers*, 2024-c).

Relativamente às camadas ocultas de neurónios, realizaram-se testes com um máximo de três camadas, tendo-se permitido ao algoritmo de otimização testar até 101 neurónios por camada oculta.

# 5. Caso I – Deteção de perdas de água numa conduta de abastecimento gravítico

Neste capítulo serão apresentados os resultados relativos aos ensaios realizados com os dados obtidos na caixa existente a jusante do Reservatório do Alto da Serra.

## 5.1. Análise prévia de dados

No presente subcapítulo apresenta-se a análise prévia dos dados, que foi realizada antes da aplicação de qualquer algoritmo de aprendizagem supervisionada, com o objetivo de criar familiarização com os dados e tentar perceber correlações, desvios e tendências.

Após criar o ficheiro com todas as variáveis obtidas na instalação, nomeadamente pressão a montante e a jusante da válvula redutora de pressão e caudal, verificou-se a correlação entre estas 3 variáveis. Constatou-se que esta instalação apenas teve todos os medidores operacionais a partir de 17 de fevereiro de 2022, pelo que os dados analisados são relativos ao período compreendido entre essa data e 29 de novembro de 2022. Os dados existentes antes do dia 17 de fevereiro foram eliminados dos dados submetidos a teste e treino para não influenciar os resultados.

Tendo-se verificado a existência de valores em falta na base de dados a submeter ao algoritmo de aprendizagem, foi necessário substituí-los pelo valor médio dos valores existentes para a variável em causa utilizando a biblioteca "*Scikit Learn*". Em seguida normalizaram-se os dados utilizando o comando "*MinMaxScaler* ()" da mesma biblioteca. A nova gama de valores para todas as variáveis foi [0, 1].

Foram efetuados testes de correlação entre as variáveis de entrada, utilizando para tal os coeficientes correlação de *Pearson* e *Spearman* (Ramzai, 2020). O coeficiente de *Pearson* é utilizado para medir a relação linear entre duas variáveis contínuas, podendo assumir valores no intervalo [-1, 1]. Um valor positivo indica que as variáveis variam linearmente (ambas crescem ou decrescem de igual forma no mesmo sentido), aumentando essa linearidade para valores próximos de 1. O valor 0 indica que não existe qualquer linearidade entre as variáveis. Um valor negativo indica que as variáveis variam de forma inversa (quando uma

cresce a outra diminui). Já o coeficiente de *Spearman* avalia relações monótonas entre variáveis, isto é, variações não lineares de variáveis na mesma direção ou em direções opostas.

Para a correlação entre as variáveis de entrada e a variável de saída (que pode assumir os valores 0, 1 e 2 que significam respetivamente "Regime Normal", "Rotura" e "Alteração de Regime") utilizou-se a correlação *point bisserial* (DATAtab Team, 2024), normalmente utilizada para situações em que se pretende encontrar relações entre variáveis contínuas e discretas. O coeficiente varia de forma idêntica ao de *Pearson* (DATAtab Team, 2024).

Os testes de correlação foram realizados numa primeira fase utilizando os coeficientes de *Spearman* e *Pearson* para Pressão Montante versus Pressão Jusante, Pressão Montante versus Caudal, Pressão Jusante versus Caudal. Já para Pressão Montante versus Saída, Pressão Jusante versus Saída e Caudal versus Saída foi utilizada a correlação *point bisserial* por se tratar de relações entre variáveis contínuas e discretas. Na Tabela 4 encontram-se os resultados obtidos dos testes de correlação de *Spearman* e *Pearson*.

**Tabela 4 - Correlação entre as entradas para os coeficientes de *Pearson* e *Spearman***

| | Coeficiente de Pearson | Coeficiente de Spearman |
|---|---|---|
| Pressão Montante/Pressão Jusante | 0,1977 | 0,0912 |
| Pressão Montante/Caudal | -0,034 | -0,3483 |
| Pressão Jusante/Caudal | -0,0761 | -0,0476 |

Verifica-se a existência de valores baixos de coeficientes de correlação entre as entradas, sendo a correlação mais expressiva entre a pressão a montante e o caudal para o coeficiente de *Spearman*.

Na Tabela 5 encontram-se os resultados obtidos dos testes de correlação *biserial*.

**Tabela 5 - Correlação *biserial* entre cada uma das entradas e a saída**

| | Coeficiente Biserial |
|---|---|
| Pressão Montante/Saída | -0,3229 |
| Pressão Jusante/Saída | -0,5365 |
| Caudal/Saída | 0,1227 |

Para a correlação entradas/saída, verifica-se que a melhor relação ocorre entre a pressão a jusante e a saída, havendo uma relação inversa entre as duas variáveis.

Em seguida realizaram-se testes de correlação idênticos, agora entre os valores atuais e alguns anteriores, até 9 leituras antes.

Para o coeficiente de *Pearson* os melhores valores obtidos são os mostrados na Tabela 6, para a correlação entre a pressão a jusante e a montante. Nesta situação a pressão a jusante é fixa e a pressão a montante varia entre a leitura anterior e a nona leitura anterior.

**Tabela 6 - Melhores valores obtidos para o coeficiente de *Pearson* a 9 leituras anteriores**

| | Coeficiente de Pearson |
|---|---|
| Pressão Jusante/Pressão Montante-1 | 0,1981 |
| Pressão Jusante/Pressão Montante-2 | 0,1985 |
| Pressão Jusante/Pressão Montante-3 | 0,1989 |
| Pressão Jusante/Pressão Montante-4 | 0,1993 |
| Pressão Jusante/Pressão Montante-5 | 0,1997 |
| Pressão Jusante/Pressão Montante-6 | 0,2000 |
| Pressão Jusante/Pressão Montante-7 | 0,2003 |
| Pressão Jusante/Pressão Montante-8 | 0,2007 |
| Pressão Jusante/Pressão Montante-9 | 0,2011 |

Para o coeficiente de *Spearman* os melhores valores obtidos são os mostrados na Tabela 7, para a correlação entre o caudal e a pressão a montante. Nesta situação o caudal é fixo e a pressão a montante varia entre a leitura anterior e a nona leitura anterior.

**Tabela 7 - Melhores valores obtidos para o coeficiente de *Spearman* a 9 leituras anteriores**

| | Coeficiente de Spearman |
|---|---|
| Caudal/Pressão Montante-1 | -0,3486 |
| Caudal/Pressão Montante-2 | -0,3486 |
| Caudal/Pressão Montante-3 | -0,3486 |
| Caudal/Pressão Montante-4 | -0,3487 |
| Caudal/Pressão Montante-5 | -0,3488 |
| Caudal/Pressão Montante-6 | -0,3488 |
| Caudal/Pressão Montante-7 | -0,3488 |
| Caudal/Pressão Montante-8 | -0,3489 |
| Caudal/Pressão Montante-9 | -0,3490 |

Desenharam-se em seguida gráficos para análise da auto-correlação em janelas temporais de 7 e 60 dias, tendo-se verificado um aumento de correlação para iguais períodos do dia no caso do caudal para o gráfico a 7 dias, conforme se pode ver na Figura 23.

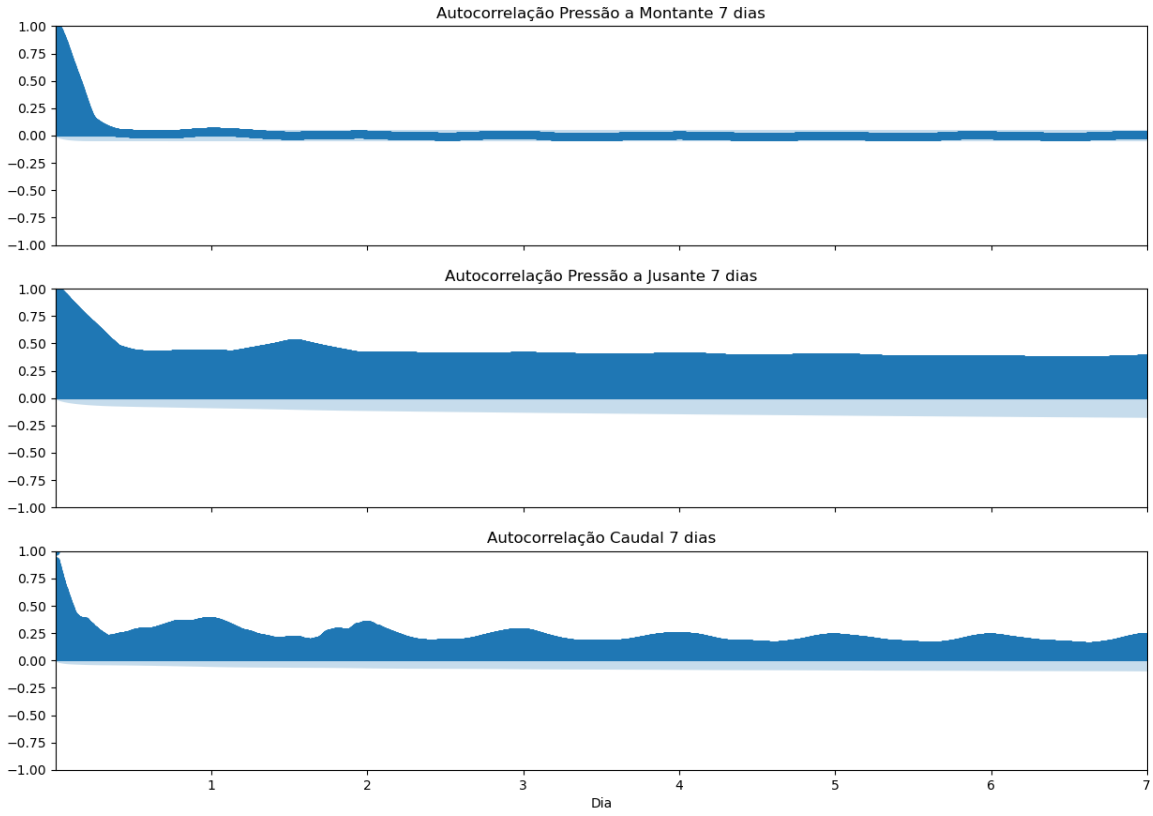


**Figura 23 - Gráfico de auto-correlação a 7 dias**

Já no gráfico a 60 dias é notório que a correlação do caudal vai baixando até ser praticamente nula nos últimos dias, conforme a Figura 24.

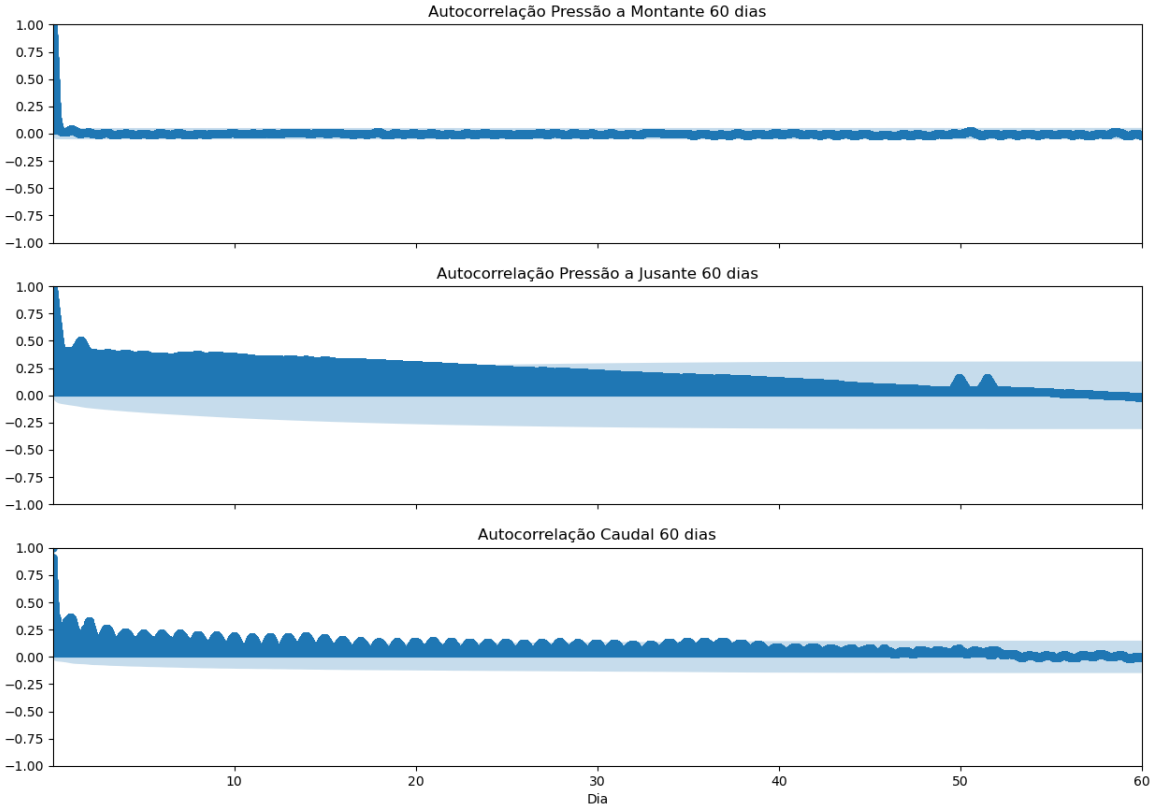


**Figura 24 – Gráfico de auto-correlação a 60 dias**

Antes de entrar nos algoritmos de aprendizagem automática produziram-se alguns gráficos para análise visual, nomeadamente de sobreposição de dados semanais e de confronto de dados reais com valores médios, este último para verificar a tendência de desvio em caso de anomalia.

Na Figura 25 encontra-se representada graficamente a sobreposição dos caudais médios diários de 4 semanas.

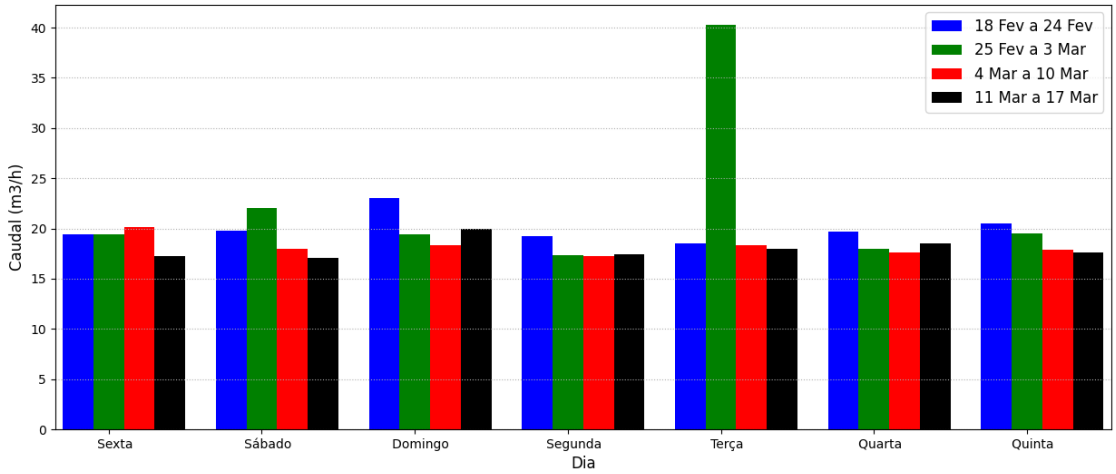


**Figura 25 - Sobreposição dos caudais médios diários de 4 semanas**

É visível que os valores médios diários de caudal num determinado dia da semana sofrem modificações de umas semanas para as outras, notando-se uma tendência de aumento de consumo antes e durante os fins-de-semana, que abranda no início de cada semana. Existe uma particularidade na terça-feira da segunda semana, dia em que se verificou uma rotura, o que levou a um aumento substancial do caudal médio diário.

Após analisar a questão da tendência semanal dos consumos, compararam-se os caudais médios diários com as médias desses caudais nos últimos 7 dias, tendo sido aplicadas médias móveis para evitar ruídos nos gráficos. A média do caudal nos últimos 7 dias pode ser considerada uma previsão de caudal para o dia atual (que apenas é visível 7 dias após o início dos dados reais). Na Figura 26 encontra-se o gráfico, onde é possível perceber que o caudal médio diário não é muito diferente do valor previsto quando não há picos consideráveis, que são minimizados na previsão. A verde é possível ver as roturas que existiram no período em análise (de acordo com os registos realizados pelas equipas técnicas que estiveram no local a proceder às respetivas reparações), verificando-se que nem todos os grandes picos se deveram a roturas. Os picos ocorridos durante o verão poderão estar relacionados com a elevação sazonal de caudal que se verifica normalmente nessa época do ano. Já o pico ocorrido em novembro deveu-se a um ajuste no caudal instantâneo máximo disponibilizado ao reservatório a jusante.

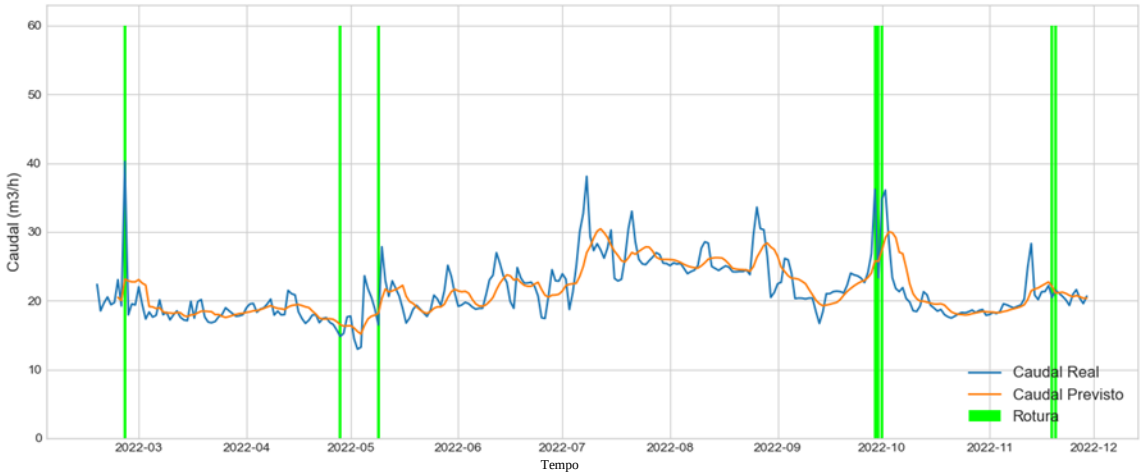


**Figura 26 – Gráfico de confrontação do caudal diário com a sua média dos últimos 7 dias**

Na Figura 27 encontra-se um gráfico idêntico ao anterior, mas agora com os valores de caudal instantâneo (intervalos de 1 segundo entre medições) e os valores médios de caudal por hora (que funcionam como previsão para o momento seguinte), sendo aqui visível que existem picos de caudal que não se refletem na média horária, pelo que seria fácil implementar a ativação um alarme caso o caudal se afastasse por exemplo de 20 a 30 $m^3/h$ do valor médio. Desta forma o algoritmo poderia detetar e informar o utilizador sobre picos de caudal, que merecem sempre análise para verificar a existência de roturas, erros de medição, aumentos de consumo ou roubo de água, por exemplo a partir de válvulas de descarga. Seriam apenas detetados eventos que causassem um aumento considerável de caudal, uma vez que uma pequena perda de água poderia ser facilmente considerada um normal aumento de consumo.

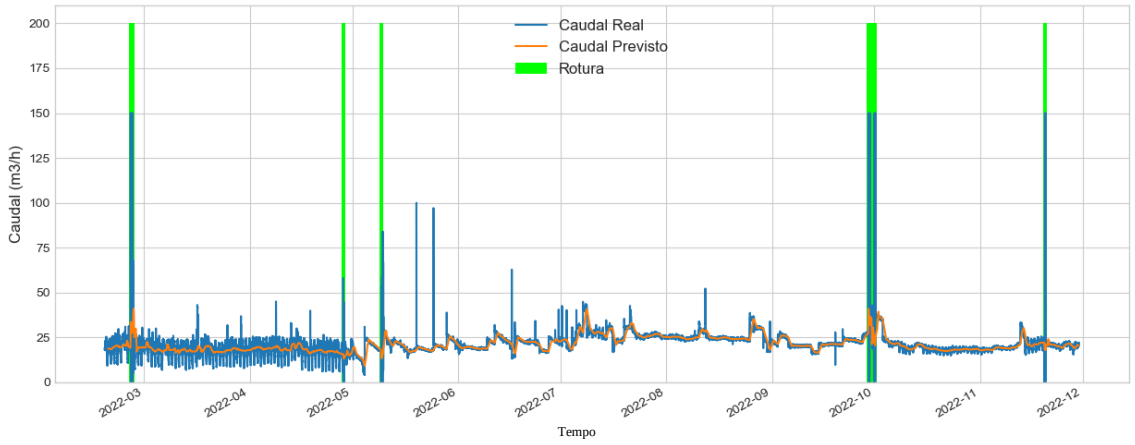


**Figura 27 - Gráfico de confrontação do caudal instantâneo com a média da última hora**

Pela análise realizada aos dados obtidos pode concluir-se que a variável com maior autocorrelação ao longo do tempo é o caudal. Verificou-se também que esta variável sofreu aumentos consideráveis nos dias em que se verificaram roturas.

## 5.2. Utilização de Algoritmos de Aprendizagem Supervisionada

Foi necessário classificar previamente os dados para poder utilizar algoritmos de aprendizagem supervisionada. Na Tabela 8 encontram-se registados os eventos em que o caudal na conduta foi a zero durante o período em estudo, sendo que para um dos casos não se encontraram registos relativos a roturas no software MAXIMO (software de gestão onde são criados registos relativos a avarias).

**Tabela 8 – Eventos verificados na conduta entre fevereiro e novembro de 2022**

| Data | Registo em MAXIMO |
|---|---|
| 25/02/2022 | Sim |
| 28/04/2022 | Não |
| 09/05/2022 | Sim |
| 29/09/2022 | Sim |
| 01/10/2022 | Sim |
| 19/11/2022 | Sim |

Na Figura 28 encontra-se a representação gráfica de pressões a montante e a jusante da válvula redutora de pressão e do caudal que passa na conduta. No gráfico mais acima encontra-se representada a pressão a montante da válvula redutora de pressão cuja unidade é “bar”, no gráfico ao centro encontra-se representada a pressão a jusante da válvula redutora de pressão cuja unidade é “bar” e no gráfico de baixo encontra-se representado o caudal, cuja unidade é “$m^3$/h”. Ao longo do presente Capítulo serão apresentados diferentes gráficos relativos às variáveis atrás descritas, cujas unidades são as que foram referidas.


**Figura 28 – Pressões e caudais na conduta entre fevereiro e novembro de 2022 (pressão em bar e caudal em $m^3/h$)**

Para além das interrupções de abastecimento já referidas, existem situações em que o caudal sobe acima dos 40$m^3$/h, que é o caudal máximo aproximado desta conduta nos períodos em que os consumos são mais elevados. A identificação de caudais acima do esperado é importante, pois estes podem derivar de problemas de afinação na válvula redutora de pressão ou roturas. Por vezes existem grandes oscilações que, embora não representem situações de rotura, são importantes do ponto de vista de operação/manutenção, uma vez que oscilações bruscas de caudal e pressão podem estar na origem de situações de desgaste das condutas que levam a roturas.

Posto isto, optou-se por criar duas classes, uma com a designação "Rotura" em que são identificadas todas as roturas e situações com comportamentos equiparados a tal e outra com a designação "Alteração de Regime", na qual se incluíram os caudais acima dos 40$m^3$/h (incluindo picos) que não representam roturas e as situações de carregamento de conduta, em que tanto as pressões como o caudal terão valores diferentes dos habituais.

Embora sejam dados provenientes de uma situação de rotura, os valores obtidos durante o carregamento de condutas não podem ser dados a conhecer ao algoritmo como situação de rotura, sob pena de haver más classificações em dados diferentes.

Na Figura 29 encontra-se uma vista geral da pré-classificação efetuada, constando no Anexo I o pormenor de cada uma das falhas.

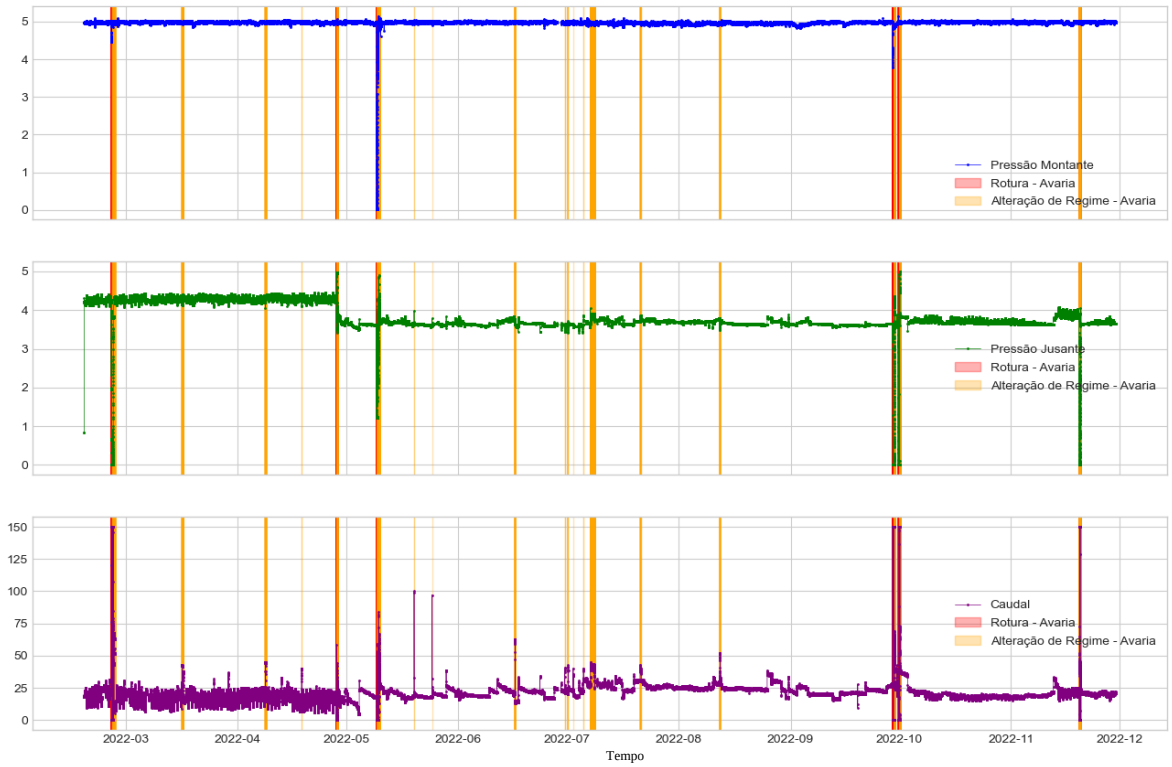


**Figura 29 – Pré-classificação de anomalias no troço entre fevereiro e novembro de 2022 (pressão em bar e caudal em $m^3/h$)**

Para obter uma uniformização das variáveis de entrada normalizaram-se os valores lidos para o intervalo [0,1].

Com o objetivo de dividir o conjunto de dados em subconjunto de treino (80%) e subconjunto de teste (20%) utilizou-se o "*train_test_split*" da biblioteca *Scikit Learn*. Desta forma foi garantida a aleatoriedade na seleção dos dados para cada um dos sub-conjuntos.

Em seguida procedeu-se à otimização dos hiperparâmetros de um conjunto de algoritmos de aprendizagem supervisionada, que foram depois treinados por forma a realizar a classificação de novos dados.

Nos sub-capítulos 5.2.1 a 5.2.4 encontram-se os resultados obtidos para cada um dos algoritmos.

### 5.2.1. Random Forest

Para o algoritmo Random Forest, após a otimização com o auxílio da ferramenta *Optuna* obtiveram-se os seguintes hiperparâmetros:

| *max_depth* | *max_leaf_nodes* | *min_samples_leaf* | *min_samples_split* | *n_estimators* |
|---|---|---|---|---|
| 81 | 102 | 2 | 30 | 73 |

Realizou-se então o treino do algoritmo, tendo-se obtido o valor 0,98454 para o parâmetro *Macro Average F1-Score*.

Na Tabela 9 encontra-se a matriz de confusão relativa a essa mesma classificação, sendo notória a classificação correta de quase todos os dados pré-classificados como regime normal (apenas um foi classificado como alteração de regime), de mais de 99% dos dados pré-classificados como rotura e de mais de 92% dos dados pré-classificados como alteração de regime. A maior parte dos casos de falha na classificação ocorreram em casos de alteração de regime que foram classificados como regime normal. É bastante notório o desequilíbrio de dados, com o número total de amostras de situações de anomalia (roturas + alterações de regime) a representar 1,6% do total de dados.

**Tabela 9 – Matriz de Confusão *Random Forest***

| | **Regime normal (Estimado)** | **Rotura (Estimado)** | **Alteração de Regime (Estimado)** |
|---|---|---|---|
| **Regime normal (Real)** | 79825 | 0 | 1 |
| **Rotura (Real)** | 2 | 650 | 2 |
| **Alteração de Regime (Real)** | 47 | 2 | 601 |

Na Tabela 10 encontra-se um conjunto de dados retirados do relatório de classificação obtido.

**Tabela 10 – Dados do Relatório de Classificação com *Random Forest***

| | Precisão (Precision) | Sensibilidade (Recall) | *F1-Score* | Suporte |
|---|---|---|---|---|
| **Regime normal** | 0,99939 | 0,99999 | 0,99969 | 79826 |
| **Rotura** | 0,99693 | 0,99388 | 0,99541 | 654 |
| **Alteração de Regime** | 0,99503 | 0,92462 | 0,95853 | 650 |
| **Exatidão (*Accuracy*)** | 0,99933 | 0,99933 | 0,99933 | 81076 |
| ***Macro Average*** | 0,99712 | 0,97283 | 0,98454 | 81130 |
| ***Weighted Average*** | 0,99933 | 0,99933 | 0,99932 | 81130 |

Na Figura 30 encontra-se o gráfico de deteção de eventos nos dados de teste.

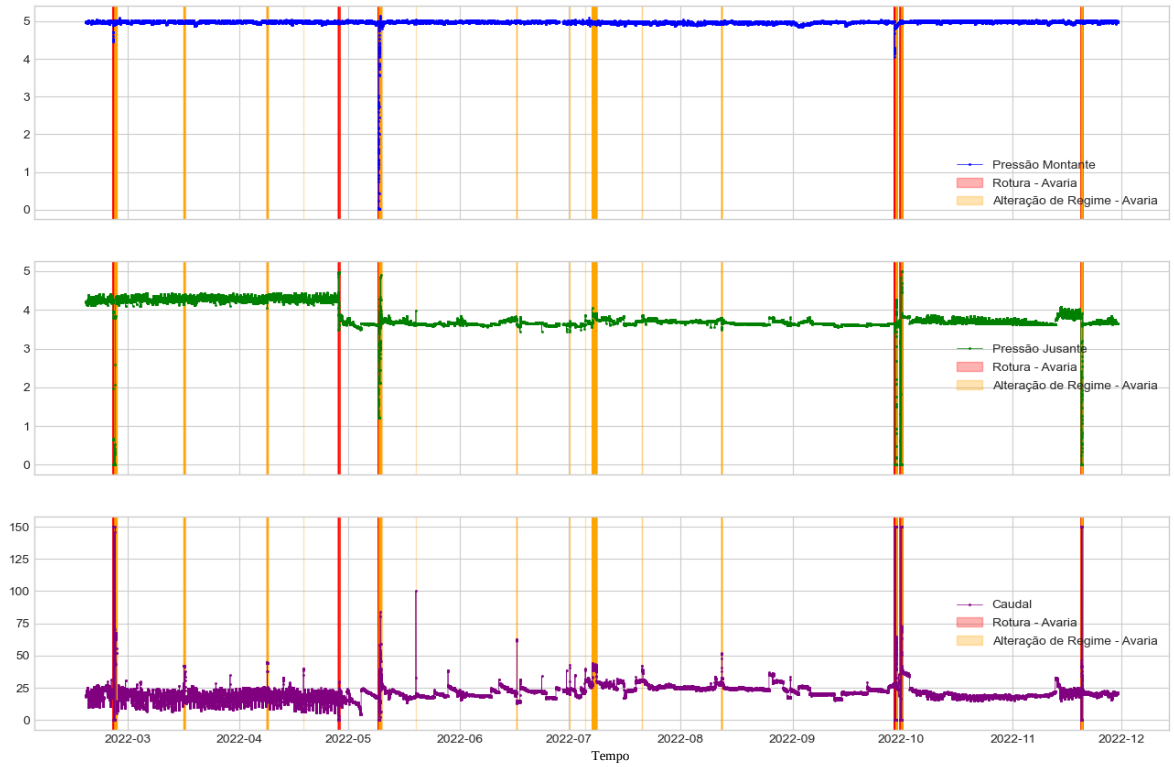


**Figura 30 – Representação gráfica da classificação dos dados de teste pelo algoritmo *Random Forest* (pressão em bar e caudal em $m^3/h$)**

O gráfico da Figura 30 tem apenas representados 20% dos dados totais presentes na pré-classificação da Figura 29 (devido à subamostragem de dados para formar o conjunto de treino), onde se encontram as classificações reais. É possível concluir que, apesar de o

número de amostras presentes na Figura 30 ser substancialmente inferior, estão representadas praticamente todas as anomalias pré-classificadas. Numa comparação rápida dos gráficos verifica-se que o algoritmo *Random Forest* detetou grande parte dos eventos que lhe foram apresentados, verificando-se algumas falhas na classificação de algumas alterações de regime ocorridas ao longo do mês de julho.

### 5.2.2. XGBoost

À semelhança do que foi descrito para o algoritmo Random Forest, utilizou-se o *Optuna* para selecionar os hiperparâmetros a utilizar no algoritmo baseado em *XGBoost*, tendo-se obtido o resultado abaixo.

| *min_child_weight* | *gamma* | *n_estimators* | *learning_rate* | *subsample* | *colsample_bytree* | *max_depth* |
|---|---|---|---|---|---|---|
| 6 | 0 | 21 | 0,66171 | 0,88846 | 0,82647 | 66 |

Neste caso, obteve-se um valor de 0,95808 para o parâmetro *Macro Average F1-Score* obtida na classificação dos valores de teste após o treino do algoritmo.

Os dados obtidos na matriz de confusão, que se encontra na Tabela 11, mostram uma pior deteção em todas as situações.

**Tabela 11 - Matriz de Confusão *XGBoost***

| | **Regime normal (Estimado)** | **Rotura (Estimado)** | **Alteração de Regime (Estimado)** |
|---|---|---|---|
| **Regime normal (Real)** | 79780 | 14 | 32 |
| **Rotura (Real)** | 39 | 612 | 3 |
| **Alteração de Regime (Real)** | 64 | 2 | 584 |

Na Tabela 12 encontra-se um conjunto de dados retirados do relatório de classificação obtido.

**Tabela 12 – Dados relevantes do Relatório de Classificação com *XGBoost***

| | Precisão (Precision) | Sensibilidade (Recall) | *F1-Score* | Suporte |
|---|---|---|---|---|
| **Regime normal** | 0,99871 | 0,99942 | 0,99907 | 79826 |
| **Rotura** | 0,97452 | 0,93578 | 0,95476 | 654 |
| **Alteração de Regime** | 0,94346 | 0,89846 | 0,92041 | 650 |
| **Exatidão (*Accuracy*)** | 0,99810 | 0,99810 | 0,99810 | 80976 |
| ***Macro Average*** | 0,97223 | 0,94456 | 0,95808 | 81130 |
| ***Weighted Average*** | 0,99807 | 0,99810 | 0,99808 | 81130 |

Na Figura 31 são confrontadas as classificações obtidas com os dados de teste que lhes deram origem.

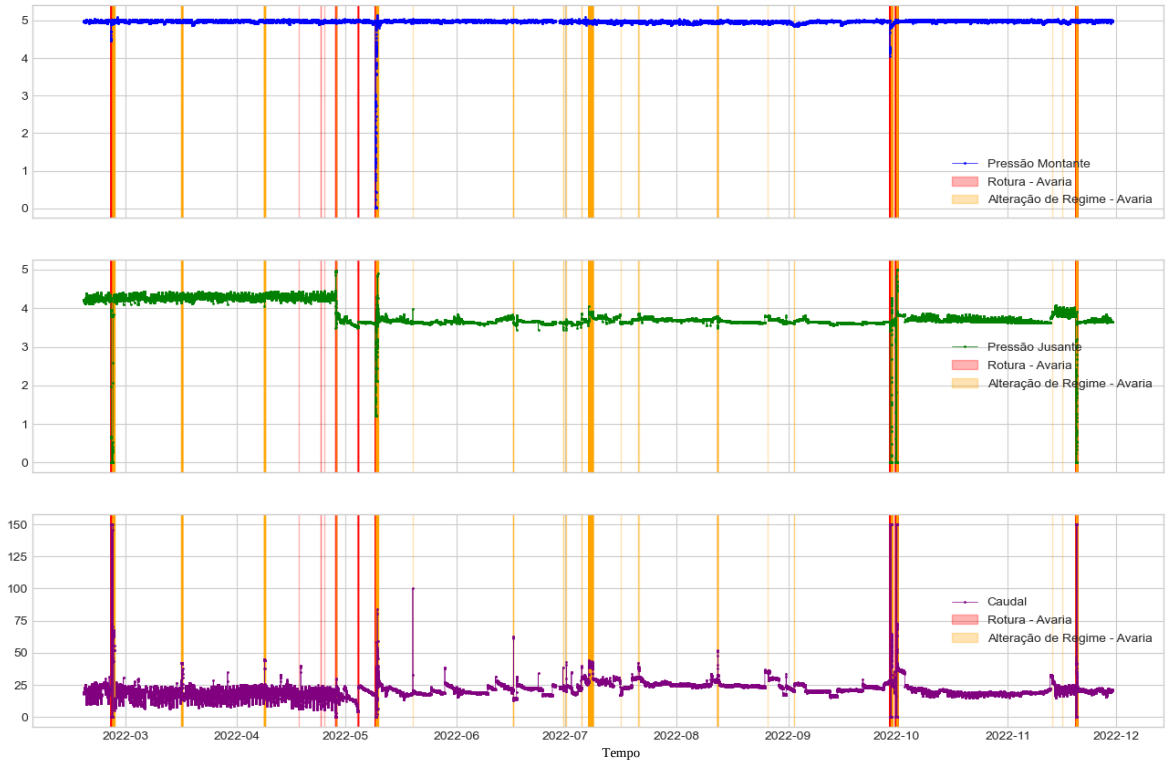


**Figura 31 - Representação gráfica da classificação dos dados de teste pelo algoritmo *XGBoost* (pressão em bar e caudal em $m^3/h$)**

Comparando os resultados agora obtidos com os obtidos pelo modelo *Random Forest* pode concluir-se que o modelo *XGBoost* obteve pior resultado na deteção de todas as classes. Pela comparação do gráfico da Figura 31 com a pré-classificação da Figura 29 é possível perceber que o algoritmo *XGBoost* detetou grande parte dos eventos que lhe foram apresentados,

tendo classificado erradamente algumas situações de funcionamento considerado normal como rotura. Verificam-se também algumas deteções de alterações de regime em situações pré-classificadas como funcionamento normal.

Na Figura 32 encontra-se representada a importância de cada uma das variáveis de entrada na determinação da saída, que o algoritmo *XGBoost* calculou no processo de treino. Por defeito, essa importância assume o valor do parâmetro '*gain'*, que calcula a melhoria média do valor da perda nas iterações em que a variável foi utilizada ao longo do processo de treino. (Filho, 2023; XGBoost Developers, 2022).

Verifica-se que o caudal é a variável mais influente, com um peso relativo expressivo de 70%.

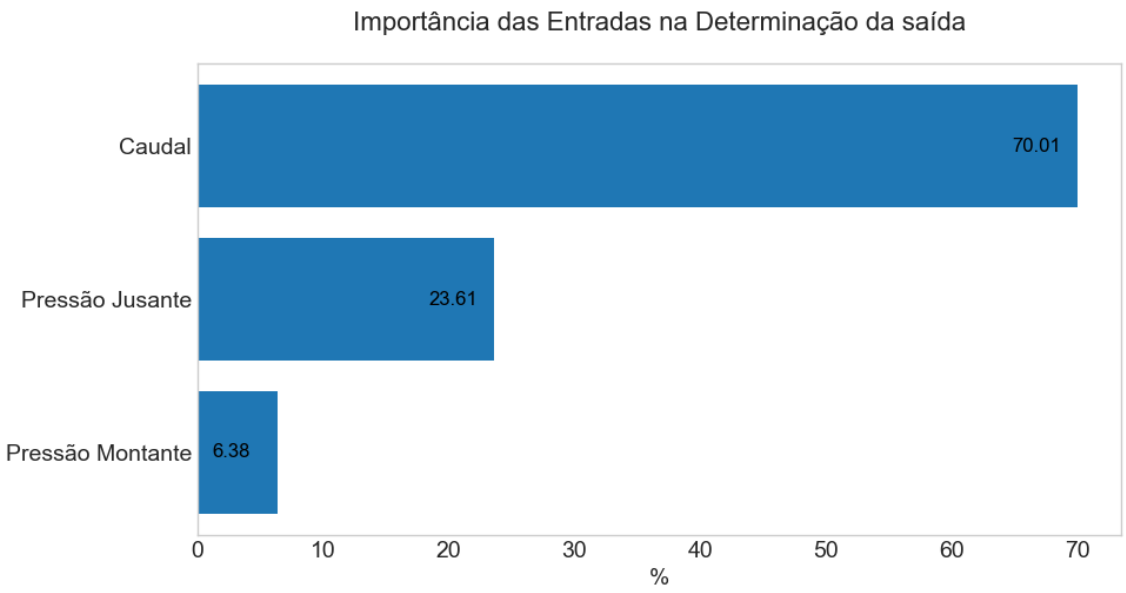


**Figura 32 – Importância de cada uma das entradas na determinação da saída no algoritmo *XGBoost***

### 5.2.3. *Support Vector Machines*

Realizou-se a otimização dos hiperparâmetros com o *Optuna*, tendo-se obtido os resultados abaixo.

| C | *degree* | *gamma* | *kernel* |
|---|---|---|---|
| 69 | 1 | 27 | *'rbf'* |

Após treinar o algoritmo obteve-se para o parâmetro *Macro Average F1-Score* o valor 0,97179 na classificação dos dados de teste.

Na Tabela 13 encontra-se a matriz de confusão obtida.

**Tabela 13 - Matriz de Confusão SVM**

| | **Regime normal (Estimado)** | **Rotura (Estimado)** | **Alteração de Regime (Estimado)** |
|---|---|---|---|
| **Regime normal (Real)** | 79817 | 1 | 8 |
| **Rotura (Real)** | 1 | 648 | 5 |
| **Alteração de Regime (Real)** | 80 | 2 | 568 |

Na Tabela 14 encontra-se um conjunto de dados retirados do relatório de classificação obtido.

**Tabela 14 – Dados relevantes do Relatório de Classificação com SVM**

| | **Precisão (Precision)** | **Sensibilidade (Recall)** | ***F1-Score*** | **Suporte** |
|---|---|---|---|---|
| **Regime normal** | 0,99899 | 0,99989 | 0,99944 | 79826 |
| **Rotura** | 0,99539 | 0,99083 | 0,99310 | 654 |
| **Alteração de Regime** | 0,97762 | 0,87385 | 0,92283 | 650 |
| **Exatidão (*Accuracy*)** | 0,99880 | 0,99880 | 0,99880 | 81033 |
| ***Macro Average*** | 0,99067 | 0,95485 | 0,97179 | 81130 |
| ***Weighted Average*** | 0,99879 | 0,99880 | 0,99877 | 81130 |

Na Figura 33 encontra-se a classificação realizada pelo algoritmo para os dados de teste.

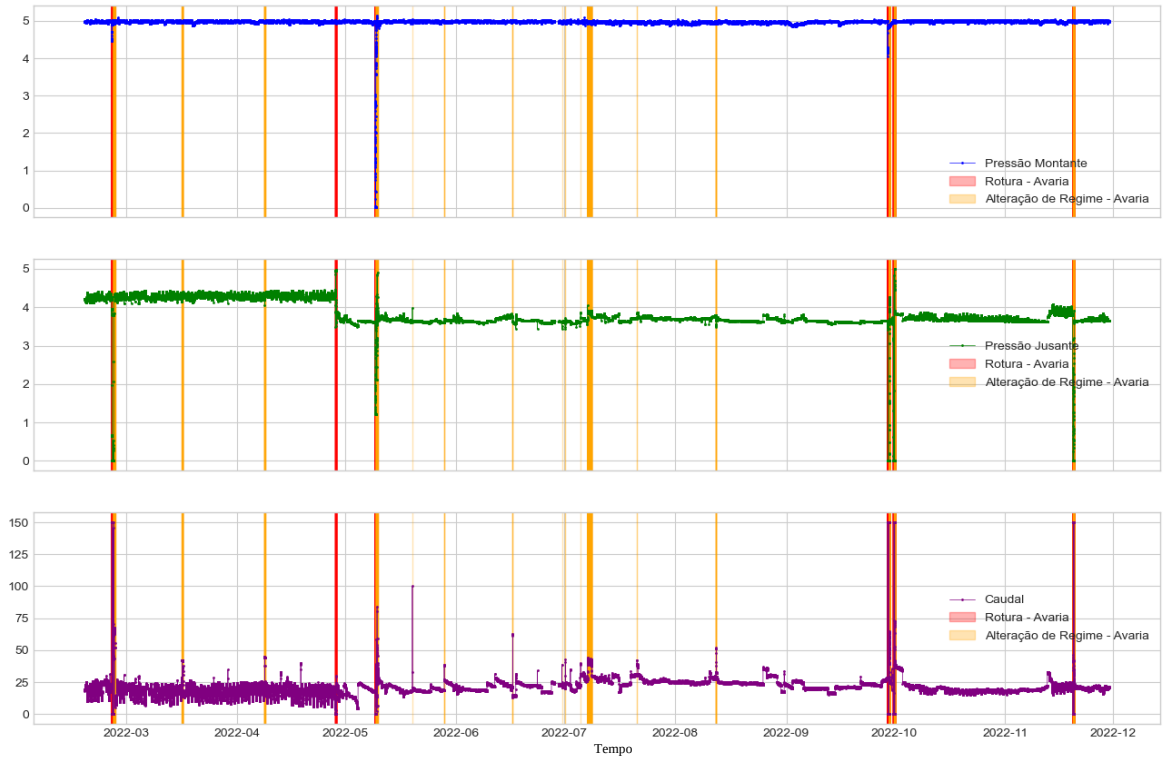


**Figura 33 - Representação gráfica da classificação dos dados de teste pelo algoritmo SVM (pressão em bar e caudal em $m^3/h$)**

Comparando os resultados agora obtidos com os obtidos pelos modelos *Random Forest* e *XGBoost*, verifica-se que este modelo obtém piores resultados que o *XGBoost* apenas para a deteção de alteração de regime, apresentando para os restantes casos resultados ligeiramente abaixo dos obtidos pelo modelo *Random Forest*. Pela comparação do gráfico da Figura 33 com a pré-classificação da Figura 29 é possível perceber que o algoritmo SVM detetou grande parte dos eventos que lhe foram apresentados, sendo visível que no caso da rotura ocorrida no dia 28 de abril de 2022 a fase de carregamento da conduta (alteração de regime) foi classificada como rotura. À semelhança do que ocorreu nas situações anteriores, algumas situações pré-classificadas como funcionamento normal foram classificadas como alteração de regime e vice-versa.

### 5.2.4. *Artificial Neural Networks*

À semelhança do que ocorreu com os restantes algoritmos, os hiperparâmetros utilizados foram otimizados com o *Optuna*, tendo-se obtido os resultados abaixo.

| *momentum* | *hidden_layer_sizes1* | *hidden_layer_sizes2* | *hidden_layer_sizes3* | *activation* | *solver* |
|---|---|---|---|---|---|
| 0,9935799 | 21 | 11 | 21 | *'relu'* | *'adam'* |

Após treinar o algoritmo obteve-se para o indicador *Average F1-Score* o valor 0,97215 na classificação dos dados de teste.

Na Tabela 15 encontra-se a matriz de confusão obtida.

**Tabela 15 – Matriz de Confusão ANN**

| | Regime normal (Estimado) | Rotura (Estimado) | Alteração de Regime (Estimado) |
|---|---|---|---|
| **Regime normal (Real)** | 79824 | 0 | 2 |
| **Rotura (Real)** | 5 | 645 | 4 |
| **Alteração de Regime (Real)** | 83 | 2 | 565 |

Na Tabela 16 encontram-se um conjunto de dados retirados do relatório de classificação obtido.

**Tabela 16 - Dados relevantes do Relatório de Classificação com ANN**

| | Precisão (Precision) | Sensibilidade (Recall) | *F1-Score* | Suporte |
|---|---|---|---|---|
| **Regime normal** | 0,99890 | 0,99997 | 0,99944 | 79826 |
| **Rotura** | 0,99691 | 0,98624 | 0,99154 | 654 |
| **Alteração de Regime** | 0,98949 | 0,86923 | 0,92547 | 650 |
| **Exatidão (*Accuracy*)** | 0,99882 | 0,99882 | 0,99882 | 81034 |
| ***Macro Average*** | 0,99510 | 0,95181 | 0,97215 | 81130 |
| ***Weighted Average*** | 0,99881 | 0,99882 | 0,99878 | 81130 |

Na Figura 34 encontra-se a classificação realizada pelo algoritmo para os dados de teste, verificando-se que todas as situações de rotura foram detetadas.

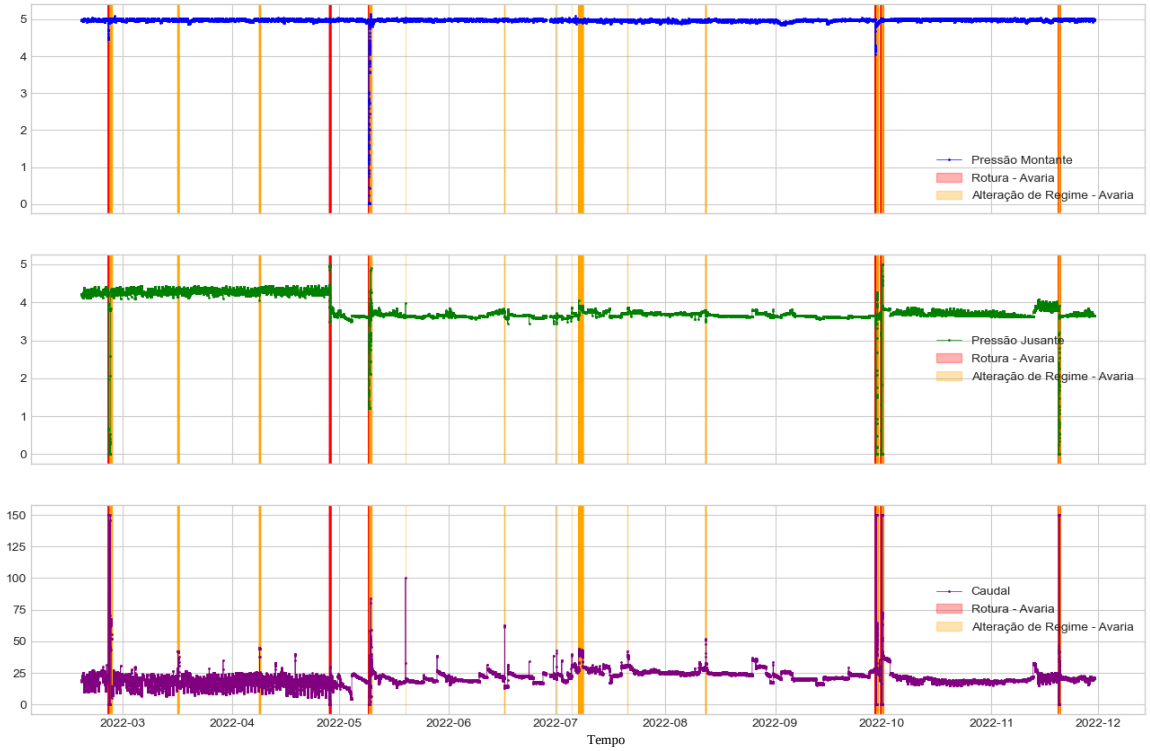


**Figura 34 - Representação gráfica da classificação dos dados de teste pelo algoritmo ANN (pressão em bar e caudal em $m^3/h$)**

Comparando os resultados agora obtidos com os obtidos pelos modelos anteriores, verifica-se que, à semelhança do que aconteceu com o modelo SVM, este modelo obtém resultados ligeiramente abaixo dos obtidos pelo modelo *Random Forest* para as situações de regime normal e rotura, piorando também para valores idênticos aos do SVM para situações de alteração de regime. Desta forma é possível concluir que o modelo com pior comportamento geral é o XGBoost, que obteve uma classificação mais baixa na generalidade. Verifica-se que a maior parte dos erros de classificação ocorreu nas situações de alteração de regime classificada como regime normal (situação de funcionamento normal). Os resultados presentes no gráfico da Figura 34 são idênticos aos obtidos pelo algoritmo SVM, o que é compatível com as semelhanças presentes nas matrizes de confusão obtidas para ambos os modelos.

### 5.2.5. Comparação das classificações obtidas e aplicação a um novo conjunto de dados

Da análise dos resultados de classificação verificou-se que o algoritmo que melhor classificou os dados de teste foi o *"Random Forest"*, conforme se pode verificar na Tabela 17.

**Tabela 17 – Comparação de resultados obtidos pelos algoritmos utilizados**

| | | Precisão (*Precision*) | *F1-score* |
|---|---|---|---|
| Regime normal | *Random Forest* | 0,99939 | 0,99969 |
| | *XGBoost* | 0,99871 | 0,99907 |
| | SVM | 0,99899 | 0,99944 |
| | ANN | 0,99890 | 0,99944 |
| Rotura | *Random Forest* | 0,99693 | 0,99541 |
| | *XGBoost* | 0,97452 | 0,95476 |
| | SVM | 0,99539 | 0,99310 |
| | ANN | 0,99691 | 0,99154 |
| Alteração de Regime | *Random Forest* | 0,99503 | 0,95853 |
| | *XGBoost* | 0,94346 | 0,92041 |
| | SVM | 0,97762 | 0,92283 |
| | ANN | 0,98949 | 0,92547 |
| Exatidão (*Accuracy*) | *Random Forest* | 0,99933 | 0,99933 |
| | *XGBoost* | 0,99810 | 0,99810 |
| | SVM | 0,99880 | 0,99880 |
| | ANN | 0,99882 | 0,99882 |
| *Macro Average* | *Random Forest* | 0,99712 | 0,98454 |
| | *XGBoost* | 0,97223 | 0,95808 |
| | SVM | 0,99067 | 0,97179 |
| | ANN | 0,9951 | 0,97215 |

**- Aplicação dos algoritmos a novos dados**

Numa fase mais avançada deste trabalho ficou disponível um novo conjunto de dados relativo ao período compreendido entre 1 de dezembro de 2022 e 16 de janeiro de 2024. Utilizou-se esse mesmo conjunto de dados para verificar a capacidade dos modelos para classificar situações de anomalia num período mais alargado, tendo-se obtido um decréscimo de performance. Na Figura 35 encontram-se as representações gráficas das classificações obtidas.

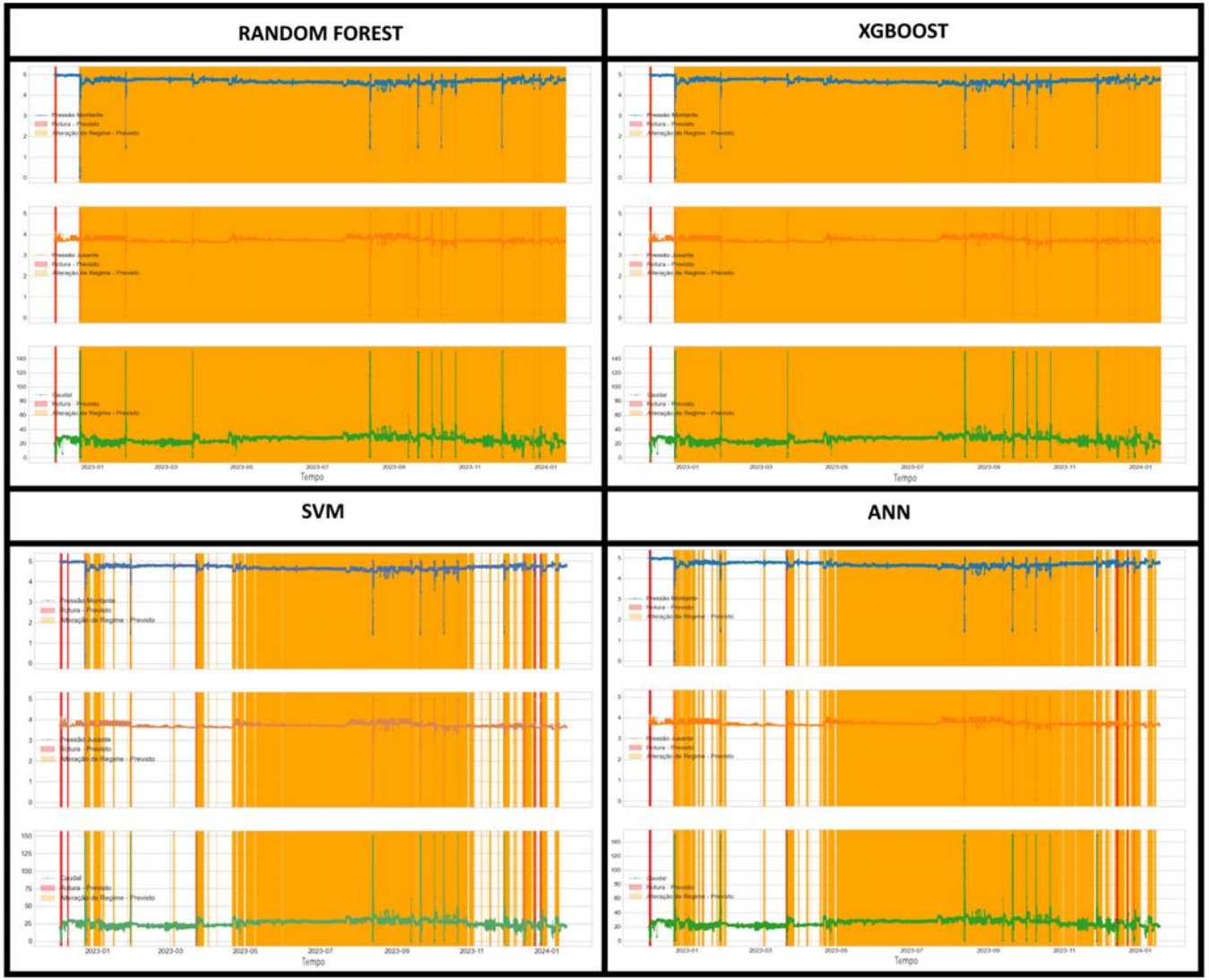


**Figura 35 – Classificações obtidas para novos resultados (pressão em bar e caudal em $m^3/h$**

É visível que todos os algoritmos detetam corretamente o primeiro evento (rotura que ocorreu no dia 1de dezembro de 2022 e cuja reparação terminou no dia 2 de dezembro). No entanto, após o evento ocorrido em 21 de dezembro de 2022 os algoritmos *Random Forest* e *XGBoost* passaram a detetar alteração de regime em contínuo. Esse evento não foi uma rotura mas sim uma substituição da válvula redutora de pressão, cujo diâmetro nominal passou de 250 para 65mm e por essa razão pode mesmo ser considerado uma alteração às condições de funcionamento da conduta. Desta situação pode concluir-se que os algoritmos SVM e ANN se adaptam melhor a novos dados, uma vez que apenas passaram a sinalizar alteração de regime em contínuo algum tempo depois dos restantes, após nova alteração de condições.

**- Retreino dos algoritmos (com os dados originais) e aplicação aos dados novos**

Tentou-se então verificar se uma diminuição do intervalo de seleção de cada um dos hiperparâmetros poderia melhorar a capacidade de classificação dos algoritmos em caso de alteração de condições (menor tendência para *overfitting*), o que poderia antever uma possibilidade de generalização dos algoritmos para condições idênticas em sistemas diferentes. Utilizou-se o conjunto de dados inicial para seleção dos hiperparâmetros (através do *Optuna*) e alteraram-se as dimensões dos conjuntos de dados de treino e teste[1]. No final tentou-se classificar o segundo conjunto de dados. Esta situação apenas foi considerada para tentar encontrar uma solução que pudesse ser aplicada de forma expedita noutras situações, ainda que perdendo alguma eficácia na deteção. Em seguida apresentam-se os melhores resultados obtidos.

No caso do algoritmo *Random Forest* obtiveram-se os seguintes hiperparâmetros:

| *max_depth* | *max_leaf_nodes* | *min_samples_leaf* | *min_samples_split* | *n_estimators* |
|---|---|---|---|---|
| 11 | 4 | 6 | 10 | 7 |

[1] Alterou-se a seleção de dados de treino e teste, para que fossem utilizados 50% dos dados para treino e outros 50% para teste, sem aleatoriedade na escolha.

No final, a classificação dos novos dados ficou conforme a Figura 36.

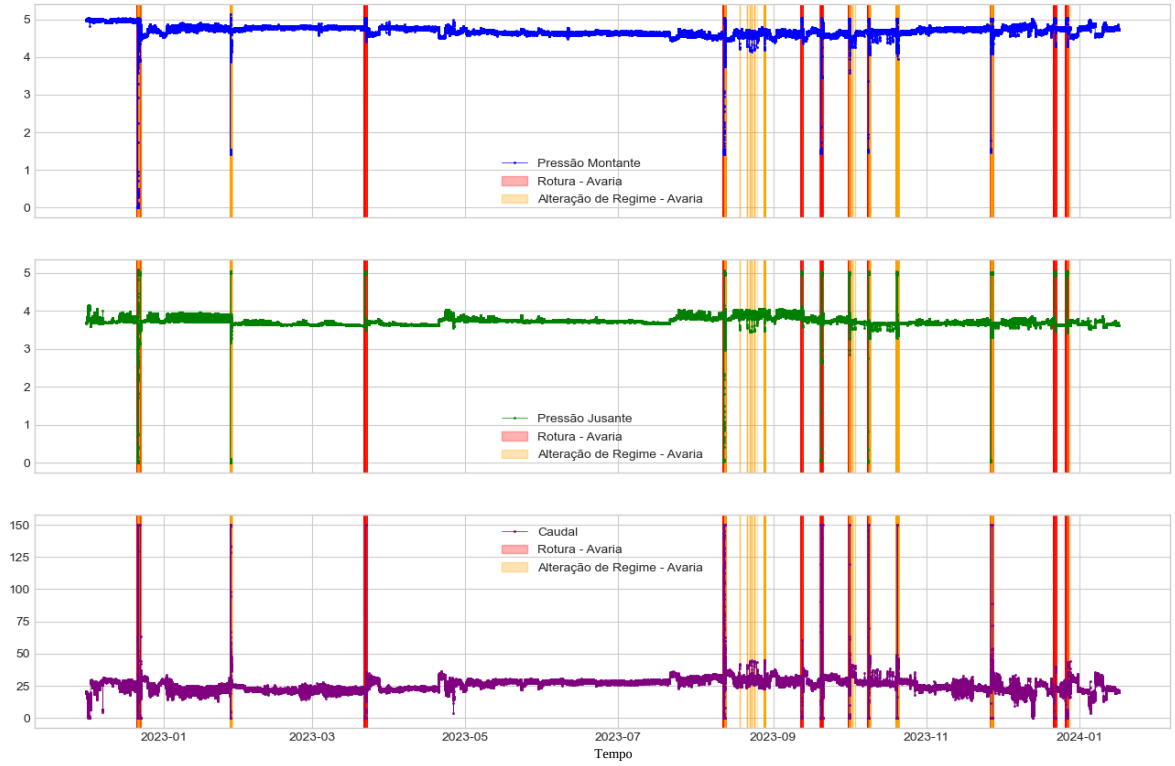


**Figura 36 – Representação gráfica da classificação dos novos dados pelo algoritmo *Random Forest* (pressão em bar e caudal em m$^3$/h)**

Da análise do gráfico sem uma pré-classificação de dados é possível concluir que uma boa parte das situações compatíveis com rotura (caudal a 0) foi detetada, embora seja visível que em alguns casos a fase de carregamento da conduta é considerada rotura. É ainda visível um conjunto de deteções de alterações de regime durante o mês de agosto.

Já para o algoritmo *XGBoost* obtiveram-se os seguintes hiperparâmetros:

| *min_child_weight* | *gamma* | *n_estimators* | *learning_rate* | *subsample* | *colsample_bytree* | *max_depth* |
|---|---|---|---|---|---|---|
| 1 | 5 | 4 | 0, 72461 | 0,469797 | 0,60987 | 2 |

Na Figura 37 são confrontadas as classificações com os dados que lhes deram origem.

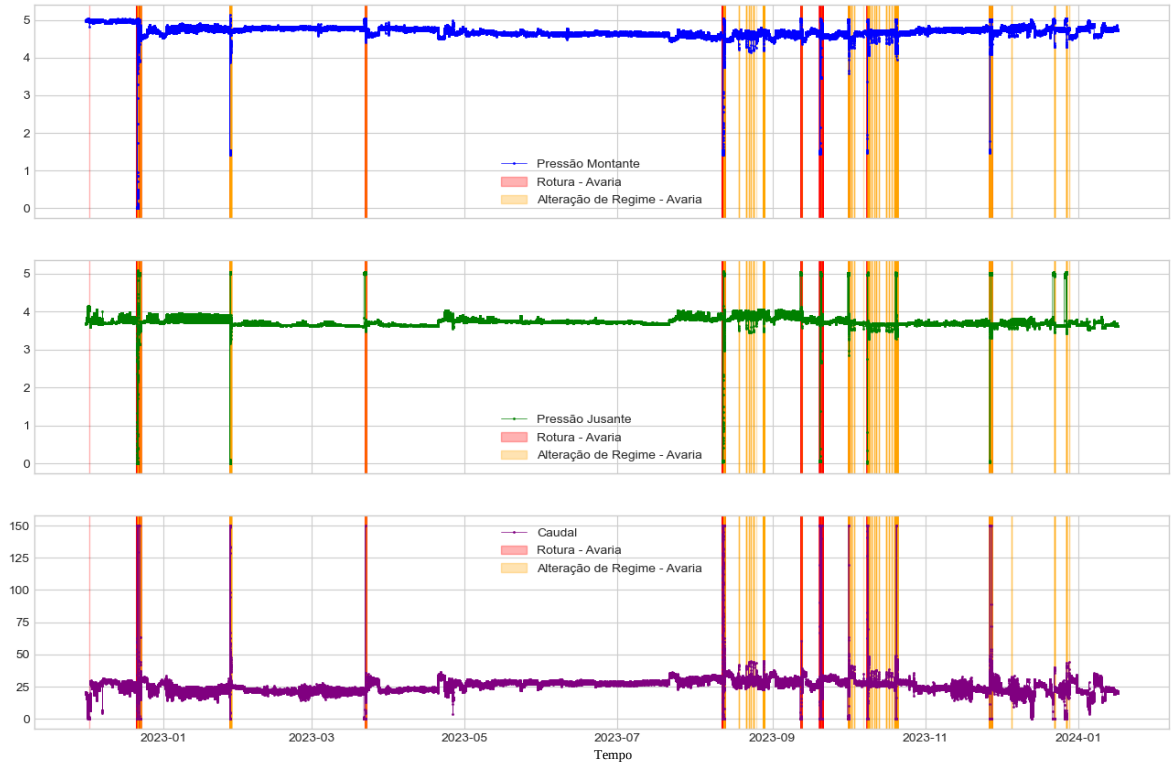


**Figura 37 - Representação gráfica da classificação dos novos dados pelo algoritmo *XGBoost* (pressão em bar e caudal em m³/h)**

Numa análise preliminar do gráfico é possível concluir que mais uma vez uma boa parte das situações compatíveis com rotura (caudal a 0) foi detetada, algumas delas tardiamente. Verifica-se que existem duas situações compatíveis com rotura em dezembro de 2023 que não foram detetadas, ao contrário do que aconteceu com o modelo *Random Forest*. À semelhança do que se verificou para o algoritmo *Random Forest*, é visível que em alguns casos a fase de carregamento da conduta é considerada rotura. É ainda visível um conjunto de deteções de alterações de regime durante os meses de agosto e outubro.

Para o algoritmo SVM obtiveram-se os seguintes hiperparâmetros:

| C | *degree* | *gamma* | *kernel* |
|---|---|---|---|
| 1 | 1 | *'auto'* | *'poly'* |

Após treinar o algoritmo obtiveram-se os resultados representados graficamente na Figura 38.

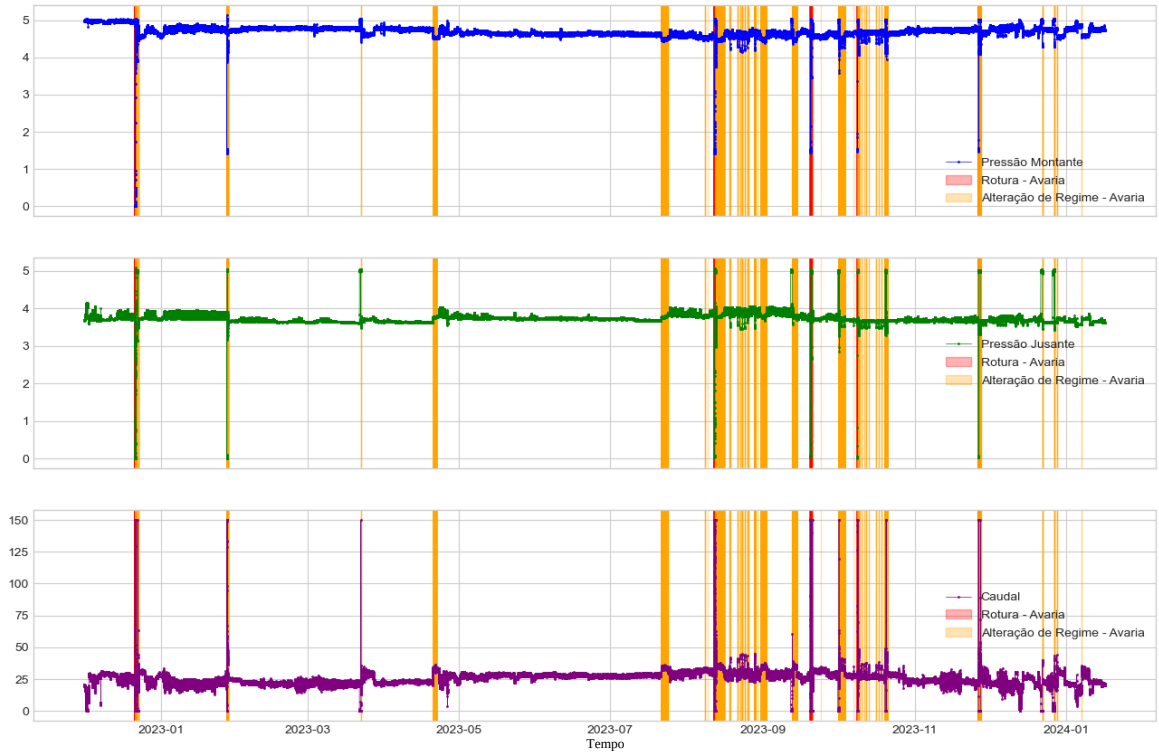


**Figura 38 - Representação gráfica da classificação dos novos dados pelo algoritmo *SVM* (pressão em bar e caudal em $m^3/h$)**

Analisando as deteções realizadas por este modelo é possível concluir que aumentaram as deteções de alterações de regime, especialmente entre agosto e novembro. Verifica-se também uma deterioração das deteções de situações compatíveis com rotura (caudal a 0), que são aqui maioritariamente detetadas como alteração de regime.

Já relativamente ao algoritmo ANN, o *Optuna* obteve os seguintes hiperparâmetros (apenas 1 camada oculta):

| *momentum* | *hidden_layer_sizes* | *activation* | *solver* |
|---|---|---|---|
| 0,9 | 2 | *'relu'* | *'adam'* |

Na Figura 39 encontra-se a classificação realizada pelo algoritmo para os dados novos.

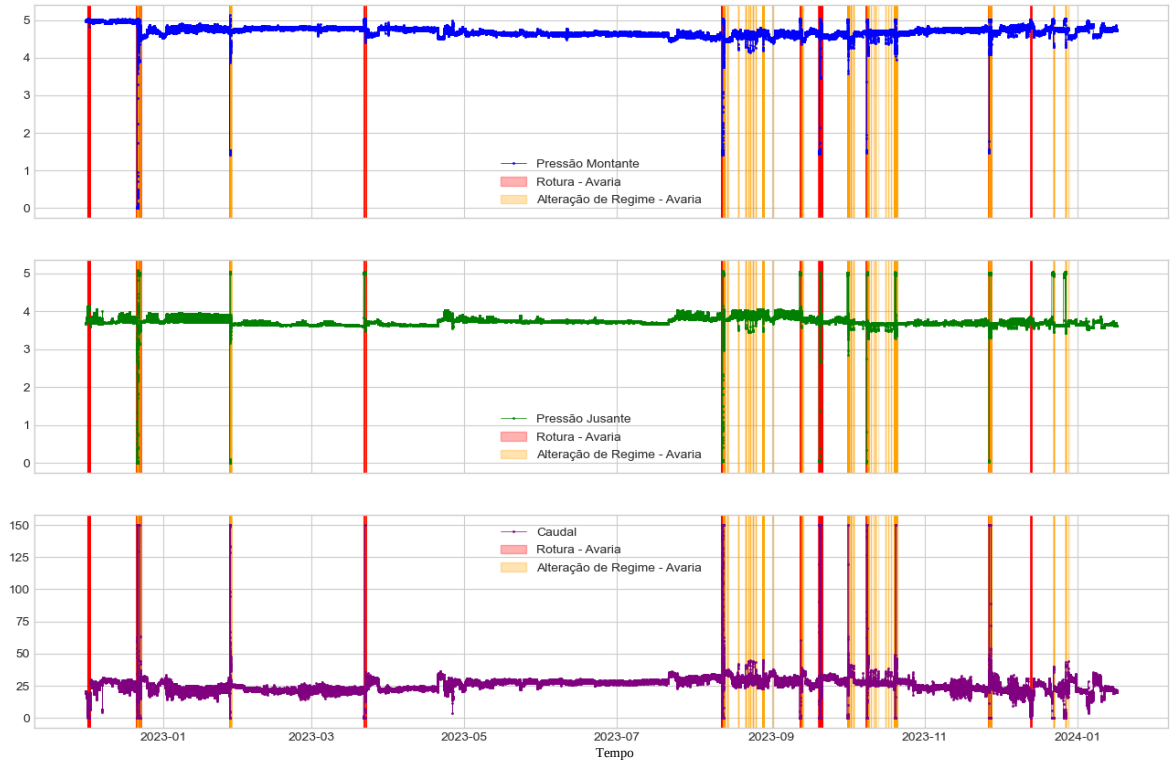


**Figura 39 - Representação gráfica da classificação dos novos dados pelo algoritmo ANN (pressão em bar e caudal em $m^3/h$)**

Analisando o gráfico da Figura 39 e comparando com os anteriores é possível perceber que o algoritmo ANN é o único que deteta o primeiro evento compatível com rotura existente nos dados. À semelhança do que ocorreu com o algoritmo SVM, entre agosto e novembro foram detetadas bastantes alterações de regime. Nesta situação foi detetado algo em todas as situações compatíveis com rotura, embora existam casos em que a deteção foi tardia e/ou o modelo classificou a deteção como alteração de regime.

Na Tabela 18 encontra-se a comparação de resultados, verificando-se uma degradação geral dos parâmetros de classificação, com maior ênfase para a situação de alteração de regime, em comparação com os dados apresentados na Tabela 17.

**Tabela 18 – Comparação de resultados obtidos pelos algoritmos utilizados**

| | | Precisão (*Precision*) | *F1-score* |
|---|---|---|---|
| Regime normal | *Random Forest* | 0,98007 | 0,98994 |
| | *XGBoost* | 0,96783 | 0,98357 |
| | SVM | 0,96572 | 0,97013 |
| | ANN | 0,97047 | 0,98496 |
| Rotura | *Random Forest* | 0,72605 | 0,71425 |
| | *XGBoost* | 0,95415 | 0,33467 |
| | SVM | 0,98414 | 0,21351 |
| | ANN | 0,96754 | 0,52177 |
| Alteração de Regime | *Random Forest* | 0,72398 | 0,05267 |
| | *XGBoost* | 0,70684 | 0,07044 |
| | SVM | 0,03303 | 0,03725 |
| | ANN | 0,70827 | 0,07352 |
| Exatidão (*Accuracy*) | *Random Forest* | 0,97579 | 0,97579 |
| | *XGBoost* | 0,96751 | 0,96751 |
| | SVM | 0,94188 | 0,94188 |
| | ANN | 0,97016 | 0,97016 |
| *Macro Average* | *Random Forest* | 0,81004 | 0,58562 |
| | *XGBoost* | 0,87627 | 0,4629 |
| | SVM | 0,66096 | 0,40696 |
| | ANN | 0,73245 | 0,75672 |

**- Retreino dos algoritmos (com os dados novos) e aplicação a um subconjunto de dados novos**

Não tendo obtido resultados positivos na reprogramação dos hiperparâmetros com utilização do primeiro conjunto de dados, utilizou-se o segundo conjunto de dados para seleção dos hiperparâmetros e treino dos algoritmos, sendo o mesmo dividido em 80% para treino e os restantes 20% para teste. Mais uma vez o algoritmo baseado em "Random Forest" obteve a melhor classificação geral, conforme a Tabela 19.

**Tabela 19 – Comparação de resultados obtidos pelos algoritmos utilizados após treino com os dados novos**

| | | Precisão (*Precision*) | *F1-score* |
|---|---|---|---|
| Regime normal | *Random Forest* | 0,99845 | 0,99910 |
| | *XGBoost* | 0,99844 | 0,99903 |
| | SVM | 0,99717 | 0,99853 |
| | ANN | 0,99835 | 0,99859 |
| Rotura | *Random Forest* | 0,99209 | 0,98955 |
| | *XGBoost* | 0,97303 | 0,97923 |
| | SVM | 0,99424 | 0,99281 |
| | ANN | 0,99125 | 0,99263 |
| Alteração de Regime | *Random Forest* | 0,87736 | 0,66192 |
| | *XGBoost* | 0,90187 | 0,64765 |
| | SVM | 0,90164 | 0,25581 |
| | ANN | 0,58725 | 0,53030 |
| Exatidão (*Accuracy*) | *Random Forest* | 0,99810 | 0,99810 |
| | *XGBoost* | 0,99774 | 0,99774 |
| | SVM | 0,99706 | 0,99706 |
| | ANN | 0,99716 | 0,99716 |
| *Macro Average* | *Random Forest* | 0,95597 | 0,88352 |
| | *XGBoost* | 0,95778 | 0,87531 |
| | SVM | 0,96435 | 0,74905 |
| | ANN | 0,85895 | 0,84051 |

Uma vez que foi o algoritmo baseado em "*Random Forest*" que obteve melhores resultados, serão apresentados apenas os dados obtidos com esse algoritmo.

Verifica-se alguma dificuldade na classificação correta de situações de alteração de regime, sendo uma parte dessas "más classificações" devida a um conjunto de eventos extra classificados como alteração de regime, conforme se pode ver na Figura 40, onde algumas dessas situações se encontram identificadas com uma elipse no gráfico relativo à pressão a montante da válvula redutora de pressão.

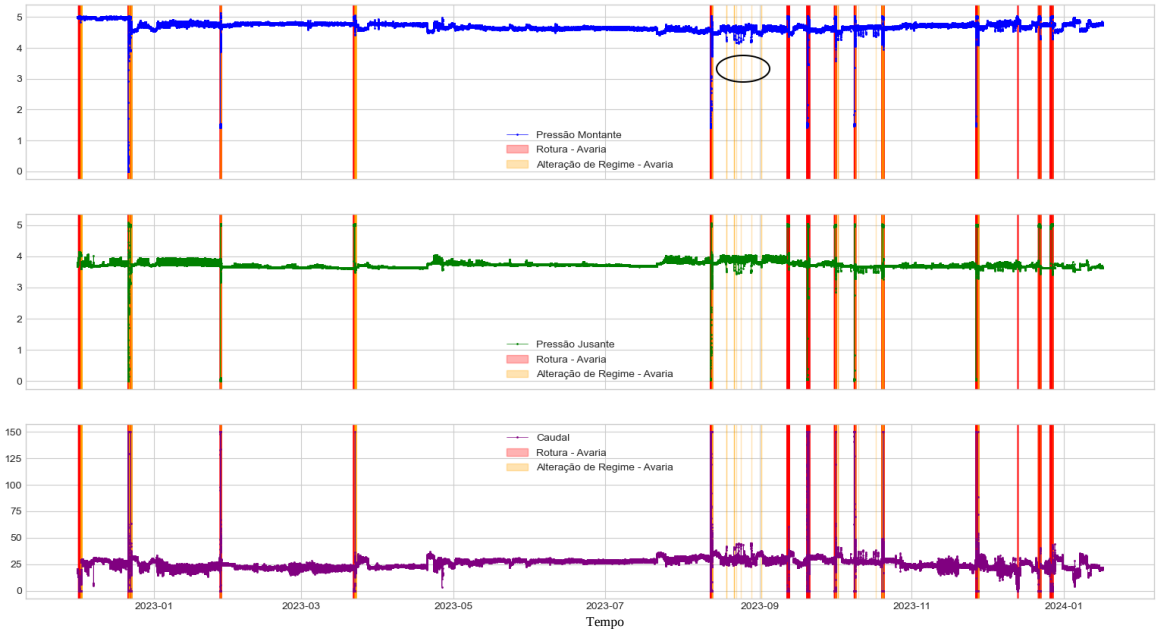


**Figura 40 – Classificação dos dados de teste com sinalização de deteções de alteração de regime não pré-classificadas (pressão em bar e caudal em $m^3/h$)**

Para esta situação foram pré-classificados 14 eventos de rotura ou eventos compatíveis. Na Tabela 20 encontram-se os eventos pré-classificados.

Para analisar o número de falsos positivos e negativos existentes nos eventos detetados obtiveram-se as matrizes de confusão da Figura 41. Tal como ocorreu nos casos anteriores, nas linhas encontram-se os valores reais e nas colunas os valores estimados pelo classificador. Note-se que o número de amostras correspondentes a Regime Normal (RN) verificadas em todas as matrizes de confusão (39) deve-se ao facto de se ter marcado um intervalo fixo de 20 amostras antes da rotura e 19 após o fim da alteração de regime.

**Tabela 20 - Deteção de eventos nos dados de validação**

| *Data* | *Evento* | *Ocorrências Rotura (número de amostras durante o evento)* | *Ocorrências Carregamento (número de amostras durante o evento)* | *Detetou?* | *Falsos Positivos/Negativos?* |
|---|---|---|---|---|---|
| 01-12-2022 | Rotura seguida de carregamento de conduta | 1445 | 337 | Sim | Sim |
| 21-12-2022 | Evento compatível com condições de rotura | 498 | 86 | Sim | Sim |
| 22-12-2022 | Rotura seguida de carregamento de conduta | 252 | 194 | Sim | Sim |
| 27-01-2023 | Rotura seguida de carregamento de conduta | 442 | 87 | Sim | Sim |
| 22-03-2023 | Rotura seguida de carregamento de conduta | 909 | 239 | Sim | Sim |
| 12-08-2023 | Rotura seguida de carregamento de conduta | 895 | 95 | Sim | Sim |
| 11-09-2023 | Rotura seguida de carregamento de conduta | 1022 | 126 | Sim | Sim |
| 20-09-2023 | Rotura seguida de carregamento de conduta | 1222 | 153 | Sim | Sim |
| 30-09-2023 | Rotura seguida de carregamento de conduta | 740 | 175 | Sim | Sim |
| 08-10-2023 | Rotura seguida de carregamento de conduta | 519 | 98 | Sim | Sim |
| 20-10-2023 | Rotura seguida de carregamento de conduta | 990 | 79 | Sim | Sim |
| 26-11-2023 | Rotura seguida de carregamento de conduta | 1164 | 89 | Sim | Sim |
| 21-12-2023 | Evento compatível com rotura e carregamento de conduta | 1029 | 65 | Sim | Sim |
| 26-12-2023 | Evento compatível com rotura e carregamento de conduta | 1331 | 10 | Sim | Sim |

**Evento de 01/12/2022**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 20 | 279 | 0 |
| **A.R.** | 5 | 0 | 63 |

**Evento de 21/12/2022**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 86 | 0 |
| **A.R.** | 2 | 0 | 13 |

**Evento de 22/12/2022**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 40 | 0 |
| **A.R.** | 18 | 0 | 20 |

**Evento de 27/01/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 88 | 0 |
| **A.R.** | 10 | 0 | 8 |

**Evento de 22/03/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 183 | 0 |
| **A.R.** | 7 | 0 | 41 |

**Evento de 12/08/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 169 | 0 |
| **A.R.** | 12 | 1 | 9 |

**Evento de 11/09/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 205 | 0 |
| **A.R.** | 20 | 0 | 0 |

**Evento de 20/09/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 243 | 0 |
| **A.R.** | 25 | 0 | 0 |

**Evento de 30/09/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 1 | 163 | 0 |
| **A.R.** | 20 | 0 | 7 |

**Evento de 08/10/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 117 | 0 |
| **A.R.** | 18 | 1 | 4 |

**Evento de 20/10/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 3 | 215 | 8 |
| **A.R.** | 4 | 1 | 6 |

**Evento de 26/11/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 263 | 1 |
| **A.R.** | 5 | 4 | 12 |

**Evento de 21/12/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 205 | 0 |
| **A.R.** | 9 | 1 | 3 |

**Evento de 26/12/2023**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 1 |
| **R.** | 0 | 253 | 0 |
| **A.R.** | 0 | 1 | 0 |

**Figura 41 – Matrizes de confusão das deteções nos dados de teste (R.N. – Regime Normal, R. – Rotura, A.R. – Alteração de Regime; Valores reais nas linhas e estimados nas colunas)**

Da análise às matrizes de confusão obtidas verifica-se que a situação com maior percentagem de classificações erradas foi a ocorrida em 22/12/2022, com cerca de 15,4% de más classificações (que representam 18 más classificações num universo de 117). Em seguida analisam-se os falsos positivos e falsos negativos obtidos em cada um dos eventos pré-classificados como anomalia.

No evento ocorrido no dia 01/12/2022, representado na Figura 42, verificou-se que os falsos negativos para a situação de rotura ocorreram logo no início, tendo a rotura sido detetada algum tempo após o momento em que se iniciou segundo a pré-classificação. Já os falsos negativos para alteração de regime ocorreram ao longo do carregamento, numa fase em que os valores de pressão a jusante se terão aproximado de valores que ocorrem em situações normais. Para além das classificações realizadas pelo algoritmo para os dados de teste é ainda possível ver no gráfico a pré-classificação dos dados relativos a rotura e alteração de regime que foram utilizados para treino do algoritmo.

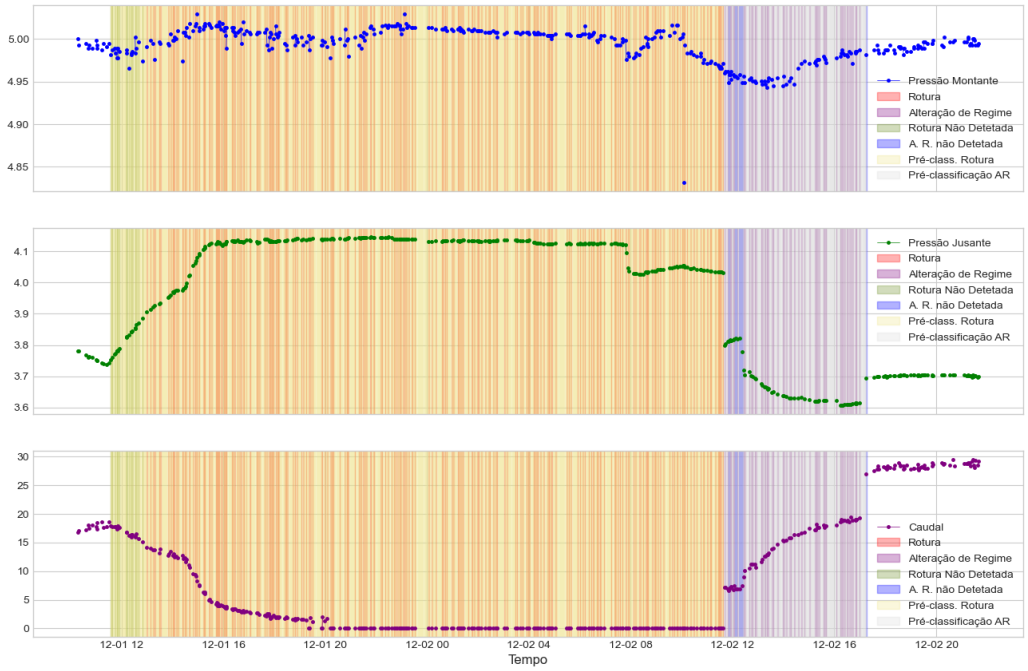


**Figura 42 – Deteções no evento ocorrido a 01/12/2022 (pressão em bar e caudal em $m^3/h$)**

Já os eventos ocorridos a 21/12/2022 (Figura 43), 22/12/2022 (Figura 44), 27/01/2023 (Figura 45), 22/03/2023 (Figura 46) e 12/08/2023 (Figura 47) foram todos detetados logo no primeiro registo anómalo, uma vez que o algoritmo classificou corretamente todos os dados

pré-classificados como rotura. Já relativamente aos dados relativos ao carregamento da conduta, em todas as situações ocorreram falsos negativos para alteração de regime, uma vez que durante essa fase ocorreram aproximações dos valores lidos aos normalmente atingidos em situações de funcionamento normal (A.R. não Detetada) e de rotura (A.R. Detetada como Rotura) para as três variáveis em análise.

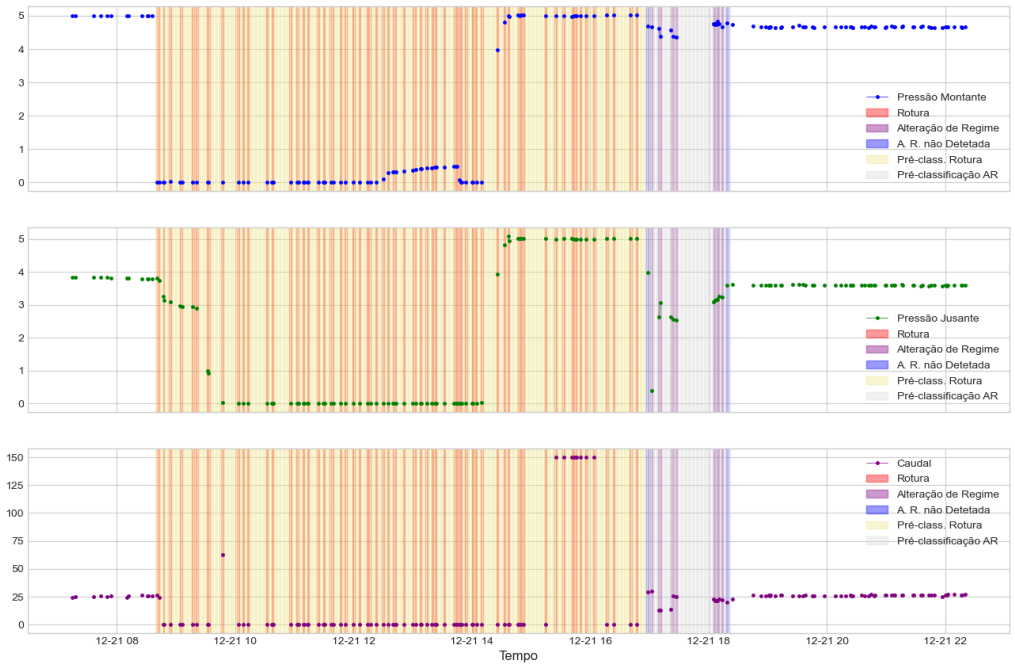


**Figura 43 – Deteções no evento ocorrido a 21/12/2022 (pressão em bar e caudal em $m^3/h$)**

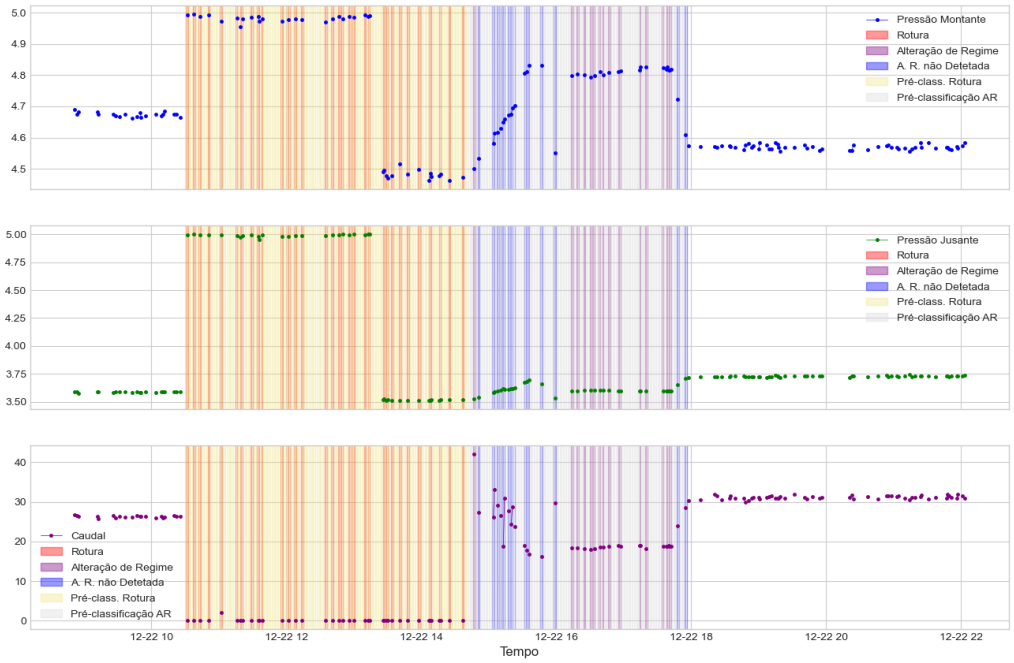


**Figura 44 – Deteções no evento ocorrido a 22/12/2022 (pressão em bar e caudal em $m^3/h$)**

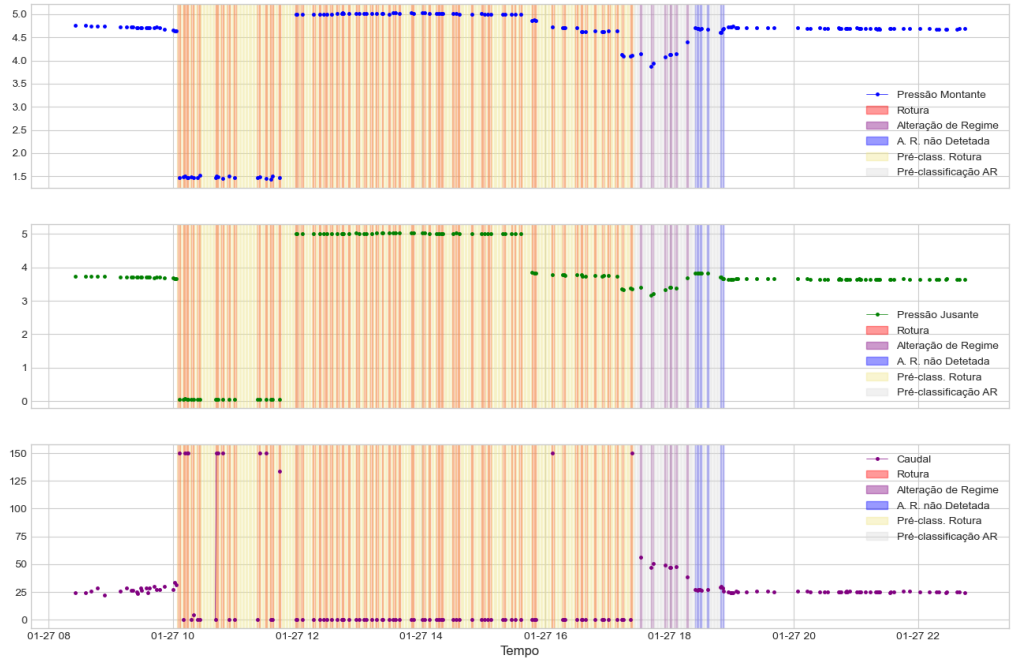


**Figura 45 – Deteções no evento ocorrido a 27/01/2023 (pressão em bar e caudal em $m^3/h$)**

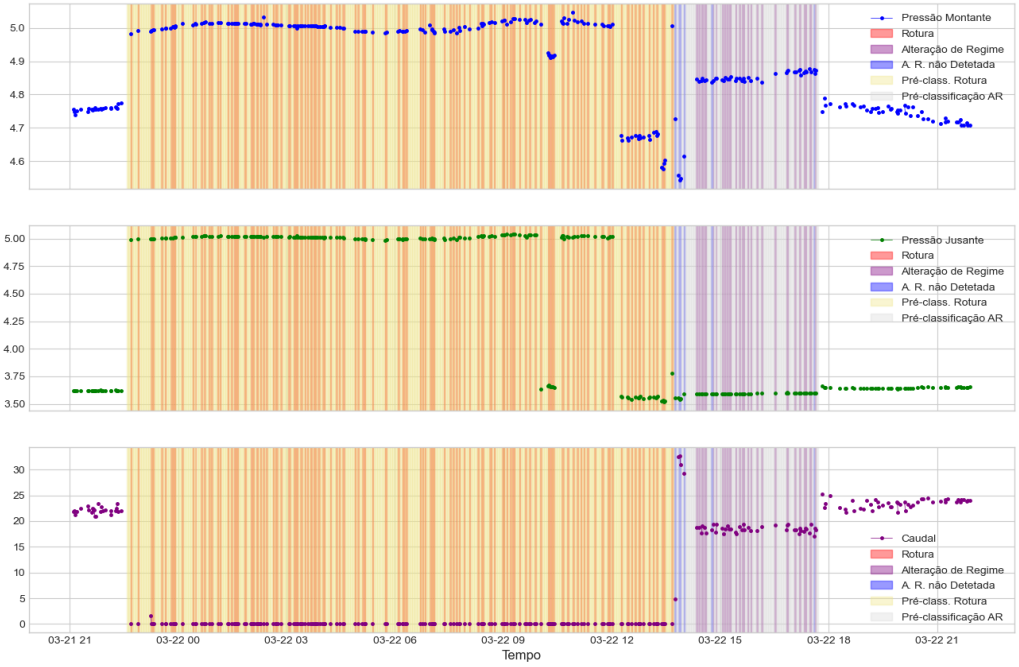


**Figura 46 – Deteções no evento ocorrido a 22/03/2023 (pressão em bar e caudal em $m^3/h$)**

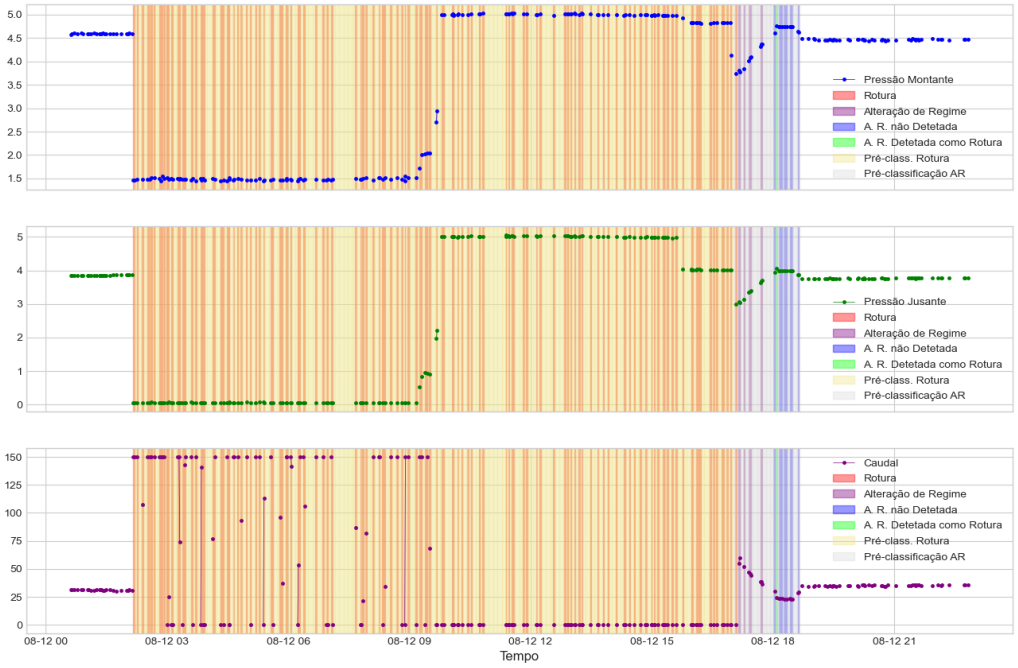


**Figura 47 – Deteções no evento ocorrido a 12/08/2023 (pressão em bar e caudal em $m^3/h$)**

No caso do evento ocorrido em 12/09/2023, representado na Figura 48 e do evento de 20/09/2023, representado na Figura 49, não foi detetada a alteração de regime devido ao carregamento da conduta. Esta situação poderá dever-se a uma proximidade entre os valores

de caudal e pressões ao longo do carregamento e dos valores desses parâmetros em determinadas condições de funcionamento normal.

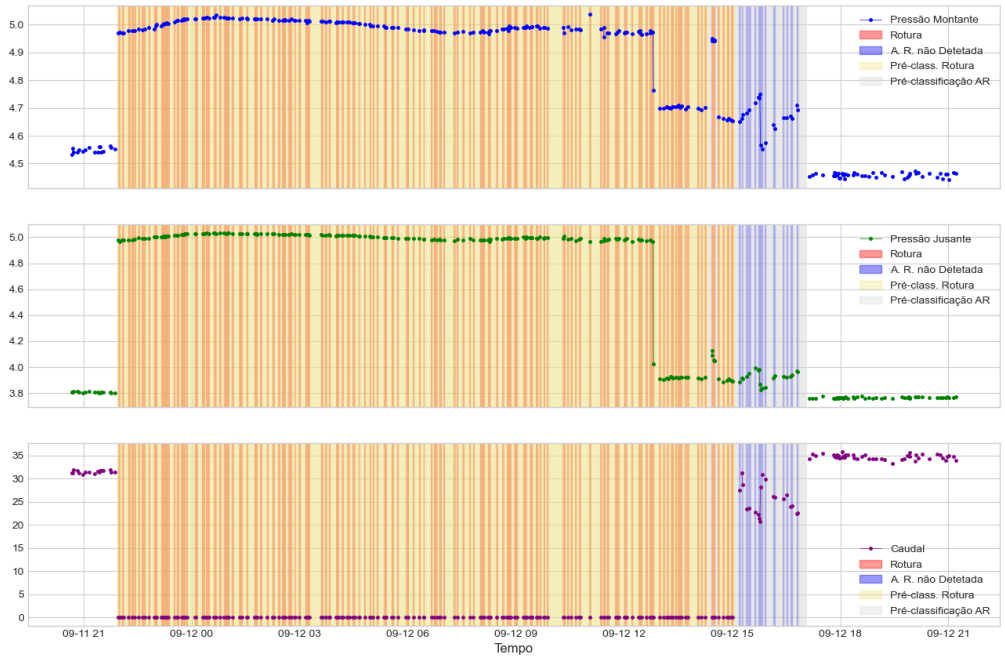


**Figura 48 – Deteções no evento ocorrido a 12/09/2023 (pressão em bar e caudal em $m^3/h$)**

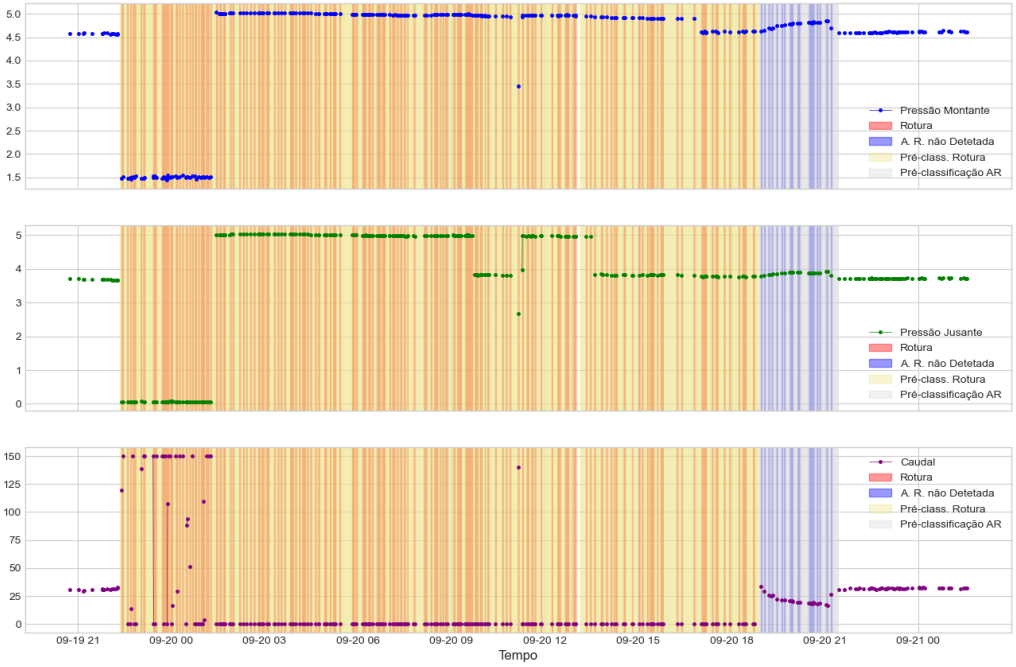


**Figura 49 – Deteções no evento ocorrido a 20/09/2023 (pressão em bar e caudal em $m^3/h$)**

Mais uma vez, nas roturas ocorridas em 30/09/2023 (Figura 50) e 08/10/2023 (Figura 51), ao longo do carregamento da conduta, foram detetadas algumas das situações de alteração de regime, havendo outras que não foram detetadas.

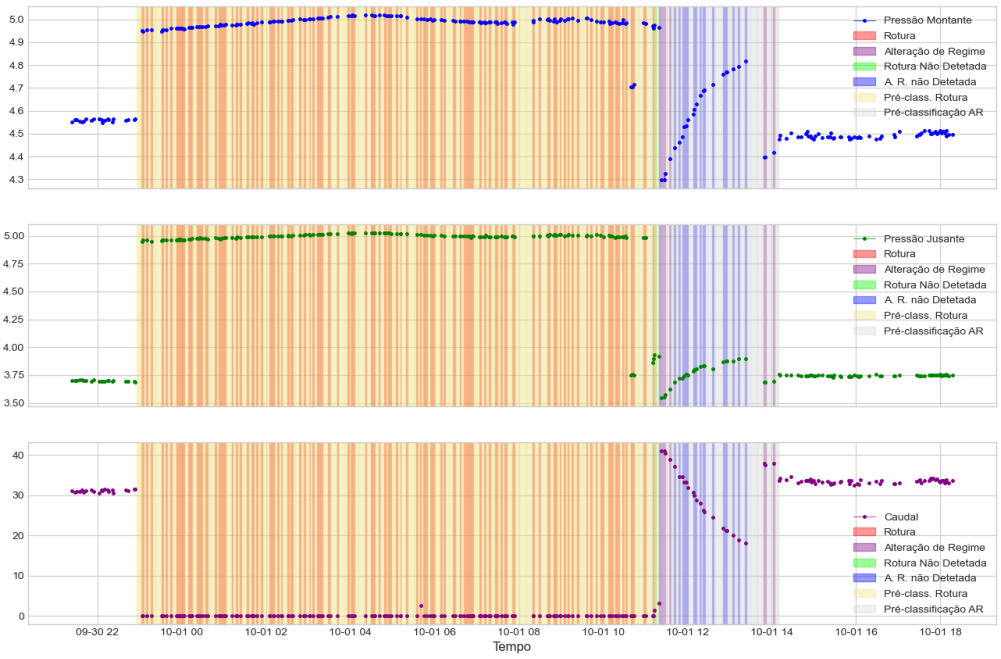


**Figura 50 – Deteções no evento ocorrido a 30/09/2023 (pressão em bar e caudal em $m^3/h$)**

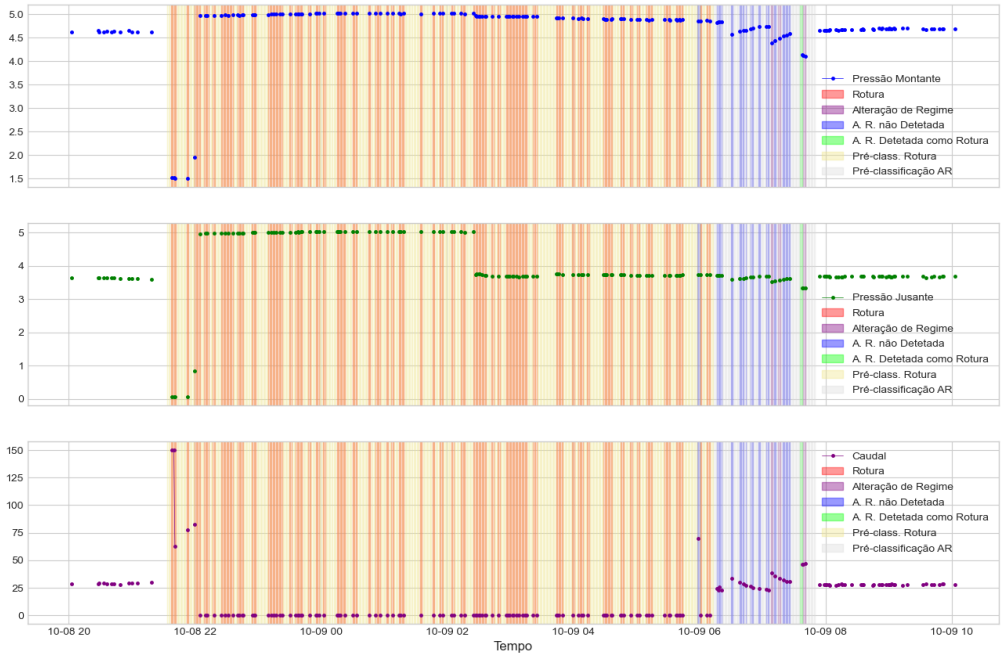


**Figura 51 – Deteções no evento ocorrido a 08/10/2023 (pressão em bar e caudal em $m^3/h$)**

Já a rotura ocorrida em 20/10/2023, representada na Figura 52, começou por ser detetada como alteração de regime numa fase inicial, tendo depois sido detetada como rotura. Também na fase de carregamento da conduta foram mal classificados alguns pontos, não tendo sido totalmente detetada a alteração de regime.

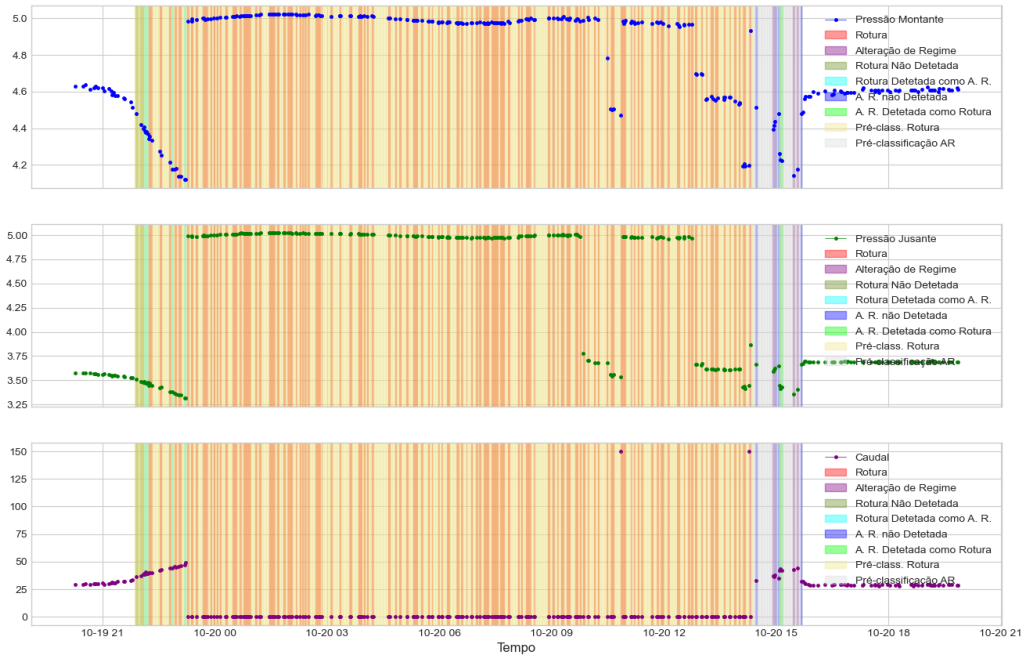


**Figura 52 – Deteções no evento ocorrido a 20/10/2023 (pressão em bar e caudal em $m^3/h$)**

O evento ocorrido em 26/11/2023 foi corretamente detetado no início, tendo apenas havido alguns erros de deteção na fase de carregamento da conduta, conforme se pode ver na Figura 53.

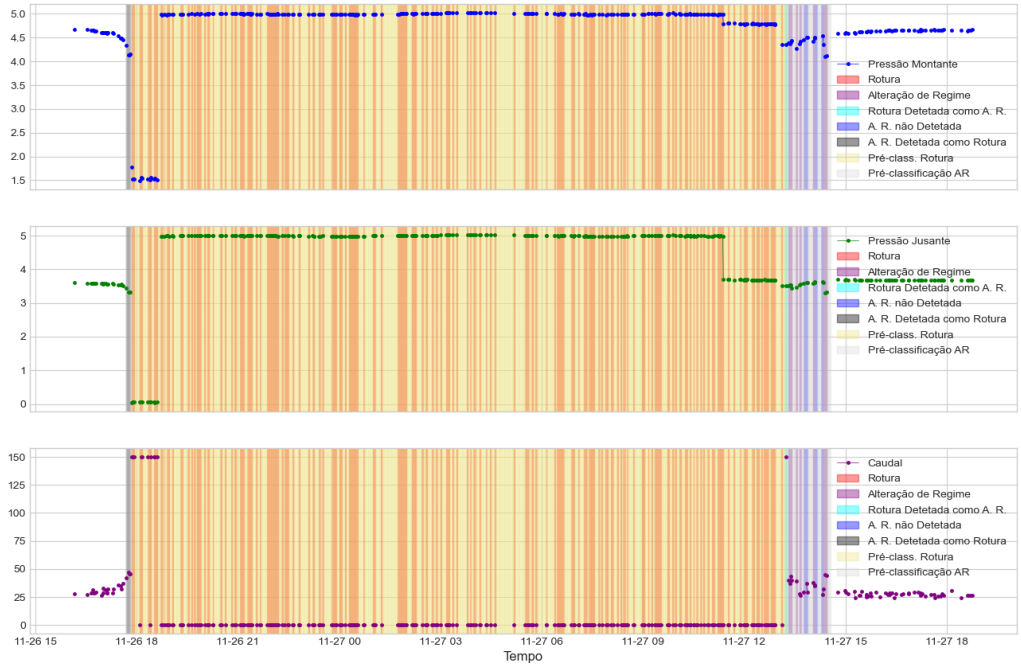


**Figura 53 – Deteções no evento ocorrido a 26/11/2023 (pressão em bar e caudal em $m^3/h$)**

Os eventos de 21/12/2023 (Figura 54) e de 26/12/2023 (Figura 55) deveram-se a operações de manutenção realizadas no reservatório de destino, tendo o troço sido colocado fora de serviço nessas datas. Na primeira situação o evento começou por ser detetado como alteração de regime, sendo depois detetado como rotura, uma vez que os valores de caudal e pressões são compatíveis com rotura. Já na segunda situação existem duas falsas deteções, uma alteração de regime detetada como rotura no início e uma falsa alteração de regime no final.

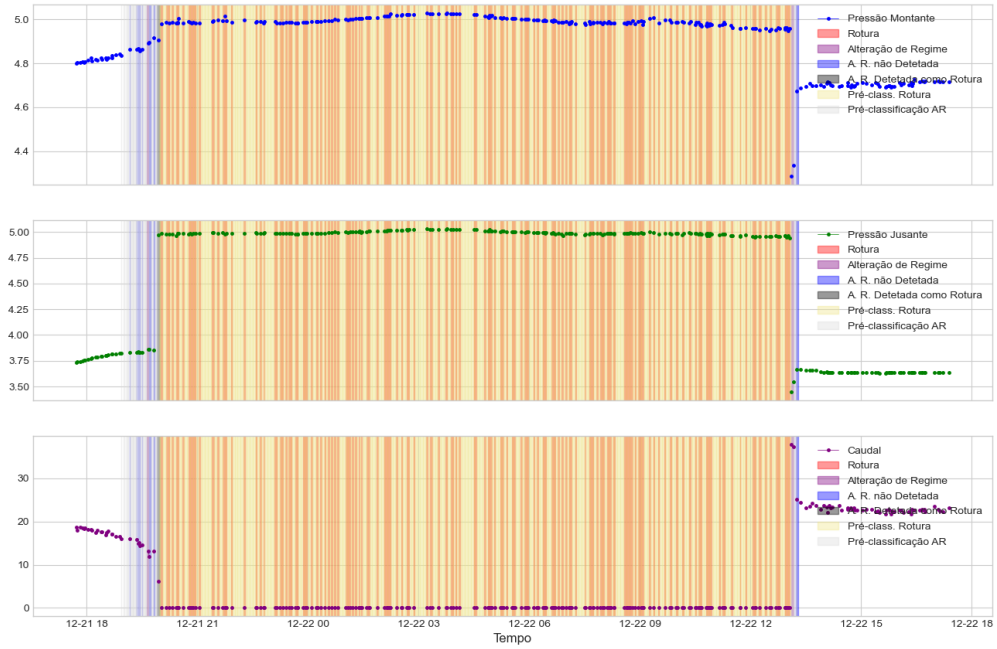


**Figura 54 – Deteções no evento ocorrido a 21/12/2023 (pressão em bar e caudal em m$^3$/h)**

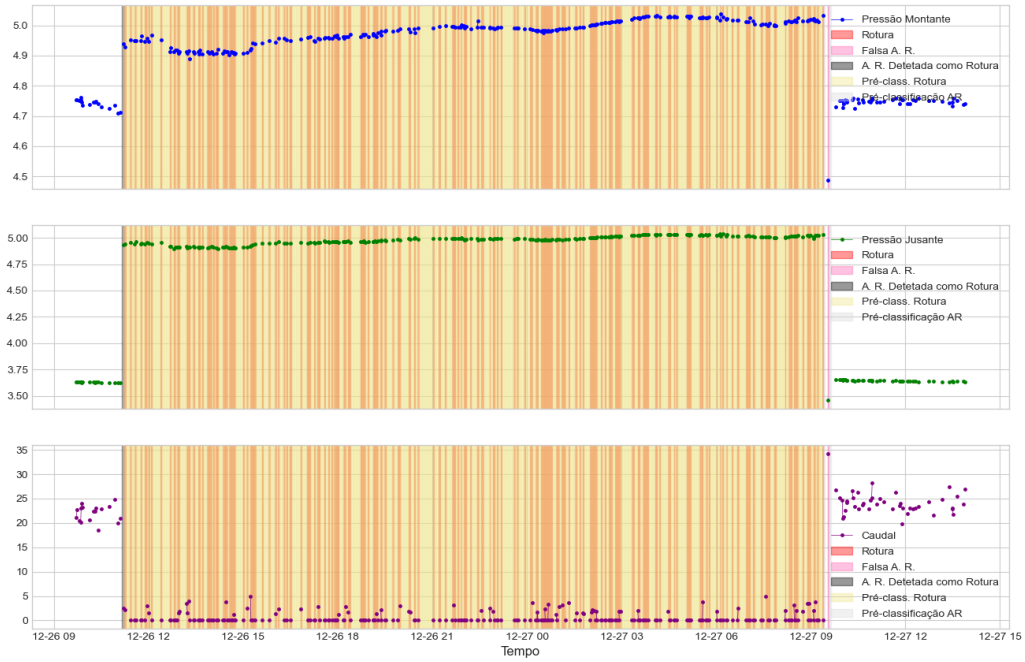


**Figura 55 – Deteções no evento ocorrido a 26/12/2023 (pressão em bar e caudal em m$^3$/h)**

# 6. Caso II – Deteção de perdas de água numa conduta de abastecimento elevatório

Após ter obtido resultados bastante satisfatórios para a conduta de abastecimento gravítico, aplicaram-se os mesmos algoritmos em dados provenientes de uma conduta de abastecimento elevatório. Utilizaram-se como variáveis de entrada para os algoritmos de deteção, a pressão na conduta à saída da estação de bombagem, o caudal à saída da estação de bombagem, o caudal à entrada do reservatório de destino e o nível do reservatório de destino. Verificou-se a existência de dados relativos a todas estas variáveis desde o ano de 2019 até ao ano de 2022.

À semelhança do que ocorreu anteriormente, classificaram-se manualmente todos os dados. Na Tabela 21 encontra-se o registo de eventos ocorridos na conduta ao longo dos anos em análise.

**Tabela 21 – Eventos verificados na conduta entre 28 de maio de 2019 e novembro de 2022**

| Data | Registo em MAXIMO |
|---|---|
| 30/05/2019 | Sim |
| 03/06/2019 | Sim |
| 12/06/2019 | Sim |
| 12/07/2019 | Sim |
| 16/10/2019 | Sim |
| 03/05/2020 | Sim |
| 09/05/2020 | Sim |
| 02/06/2020 | Sim |
| 17/09/2020 | Sim |
| 27/06/2021 | Sim |
| 14/07/2021 | Sim |
| 18/09/2021 | Sim |
| 22/07/2022 | Sim |

Conforme os dados apresentados na Tabela 22, entre 28 de maio e 31 de dezembro de 2019 ocorreram 5 roturas na conduta elevatória que transporta água desde a Estação Elevatória (EE) de Virtudes até ao Ponto de Entrega de Casais da Lagoa. Para além desses eventos, na

pré-classificação consideraram-se ainda algumas situações pontuais em que a pressão baixou abruptamente durante alguns instantes, atingindo valores apenas alcançados em situações de rotura. Classificaram-se esses valores como rotura, para que o sistema considerasse a ocorrência desses eventos como algo que não se enquadra numa situação normal nem numa alteração de regime. No total foram identificados 14 eventos, 5 dos quais representam roturas. Os restantes eventos representam alterações bruscas de pressão, operações de manutenção que obrigaram a descarregar o troço a montante da válvula de saída da Estação Elevatória, não havendo evidência de descarregamento da conduta a jusante da mesma. Existe ainda uma aparente avaria do transdutor de pressão, em que todos os dados disponíveis com exceção da pressão de saída da EE têm variações consideradas normais. Na Tabela 22 encontram-se registados os eventos pré-classificados, bem como o número de ocorrências de rotura e carregamento que cada um deles gerou. No Anexo II encontram-se os gráficos relativos aos 14 eventos identificados.

**Tabela 22 – Eventos classificados na conduta de abastecimento elevatório no período em análise**

| ***Data*** | ***Evento*** | ***Ocorrências Rotura (número de amostras durante o evento)*** | ***Ocorrências Carregamento (número de amostras durante o evento)*** |
|---|---|---|---|
| 30-05-2019 | Rotura seguida de carregamento de conduta | 232 | 104 |
| 03-06-2019 | Rotura seguida de carregamento de conduta, nova rotura e novo carregamento | 508 | 140 |
| 12-06-2019 | Rotura seguida de carregamento de conduta | 338 | 27 |
| 13-07-2019 | Rotura seguida de carregamento de conduta | 601 | 12 |
| 03-09-2019 | Manutenção Corretiva Hidráulica na EE | 208 | 0 |
| 04-09-2019 | Evento de perda de pressão compatível com condições de rotura | 3 | 0 |
| 05-09-2019 | Evento de perda de pressão compatível com condições de rotura | 3 | 0 |
| 06-09-2019 | Evento de perda de pressão compatível com condições de rotura | 2 | 0 |
| 10-09-2019 | Evento de perda de pressão compatível com condições de rotura | 13 | 0 |
| 24-09-2019 | Evento de perda de pressão compatível com condições de rotura | 8 | 0 |
| 02-10-2019 | Evento de perda de pressão compatível com condições de rotura | 2 | 0 |
| 03-10-2019 | Evento de perda de pressão compatível com condições de rotura | 5 | 0 |
| 04-10-2019 | Evento de perda de pressão compatível com condições de rotura | 260 | 0 |
| 16-10-2019 | Rotura em seguida de carregamento de conduta | 256 | 139 |

Recorreu-se mais uma vez ao "*train_test_split*" para selecionar aleatoriamente 80% dos dados para as fases de treino do algoritmo e 20% para a fase de teste.

Utilizou-se a ferramenta *Optuna* para selecionar os melhores hiperparâmetros para cada um dos algoritmos. No subcapítulo 6.1 apresentam-se os melhores resultados obtidos na fase de otimização dos algoritmos, para a qual foram utilizados os dados obtidos ao longo do ano de 2019. No subcapítulo 6.2 apresentam-se os resultados da deteção de situações de anomalia em novos dados, relativos aos anos de 2020, 2021 e 2022.

## 6.1. Dados relativos ao ano de 2019

Analisando os resultados relativos ao funcionamento de cada um dos algoritmos quando submetido aos dados de teste, que se encontram na

Tabela 23, verifica-se que o que obteve o melhor valor para o parâmetro *Macro Average F1-Score* foi o *Random Forest*, com 0,96880.

**Tabela 23 – Comparação de resultados obtidos pelos algoritmos utilizados**

| | | Precisão (*Precision*) | *F1-score* |
|---|---|---|---|
| Regime normal | *Random Forest* | 0,99945 | 0,99972 |
| | *XGBoost* | 0,99871 | 0,99917 |
| | SVM | 0,99939 | 0,99968 |
| | ANN | 0,99937 | 0,99967 |
| Rotura | *Random Forest* | 0,99348 | 0,97028 |
| | *XGBoost* | 0,94156 | 0,90437 |
| | SVM | 0,99784 | 0,97257 |
| | ANN | 0,99581 | 0,97236 |
| Alteração de Regime | *Random Forest* | 0,98780 | 0,93642 |
| | *XGBoost* | 0,93421 | 0,84024 |
| | SVM | 0,98701 | 0,92121 |
| | ANN | 0,97143 | 0,89474 |
| Exatidão (*Accuracy*) | *Random Forest* | 0,99939 | 0,99939 |
| | *XGBoost* | 0,99818 | 0,99818 |
| | SVM | 0,99936 | 0,99936 |
| | ANN | 0,99931 | 0,99931 |
| *Macro Average* | *Random Forest* | 0,99358 | 0,96880 |
| | *XGBoost* | 0,95816 | 0,91459 |
| | SVM | 0,99474 | 0,96449 |
| | ANN | 0,98889 | 0,95559 |

Constatou-se que o algoritmo *Random Forest* detetou 12 eventos de rotura ou compatíveis com situações de rotura entre maio e dezembro de 2019. Não foram detetadas alterações de regime para além daquelas ocorridas após situações de rotura, o que se deverá certamente à inexistência de alterações significativas no volume de caudal elevado durante os períodos de funcionamento dos grupos elevatórios.

Na Tabela 24 encontra-se alguma informação sobre a deteção dos eventos pré-classificados com o algoritmo *Random Forest*. Verifica-se que todos os eventos que estavam representados no conjunto de validação foram detetados, existindo algumas amostras que foram mal classificadas, originando falsos positivos/negativos em quatro deles.

**Tabela 24 – Deteção de eventos nos dados de validação**

| *Data* | *Evento* | *Detetou* | *Falsos positivos/negativos?* |
|---|---|---|---|
| | | | |
| 30-05-2019 | Rotura seguida de carregamento de conduta | Sim | Sim |
| 03-06-2019 | Rotura seguida de carregamento de conduta, nova rotura e novo carregamento | Sim | Sim |
| 12-06-2019 | Rotura seguida de carregamento de conduta | Sim | Não |
| 13-07-2019 | Rotura seguida de carregamento de conduta | Sim | Sim |
| 03-09-2019 | Manutenção Corretiva Hidráulica na EE | Sim | Não |
| 04-09-2019 | Evento de perda de pressão compatível com condições de rotura | Sim | Não |
| 05-09-2019 | Evento de perda de pressão compatível com condições de rotura | S/amostras presentes nos dados de teste | N/A |
| 06-09-2019 | Evento de perda de pressão compatível com condições de rotura | S/amostras presentes nos dados de teste | N/A |
| 10-09-2019 | Evento de perda de pressão compatível com condições de rotura | Sim | Não |
| 24-09-2019 | Evento de perda de pressão compatível com condições de rotura | Sim | Não |
| 02-10-2019 | Evento de perda de pressão compatível com condições de rotura | S/amostras presentes nos dados de teste | N/A |
| 03-10-2019 | Evento de perda de pressão compatível com condições de rotura | Sim | Não |
| 04-10-2019 | Evento de perda de pressão compatível com condições de rotura | Sim | Não |
| 16-10-2019 | Rotura em seguida de carregamento de conduta | Sim | Sim |

Numa análise rápida ao número de falsos positivos e negativos existentes nos eventos detetados obtiveram-se as matrizes de confusão da Figura 56, cuja análise demonstra uma

baixa taxa de falsas classificações, sendo o pior caso o do dia 13/7/2019 com cerca de 12% de falsas classificações, que representam 21 eventos pré-classificados como rotura que são classificados como regime normal (20) e como alteração de regime (1). Esta rotura é diferente de todas as outras ocorridas ao longo do ano de 2019 e as razões para tal são explicadas mais à frente neste documento, na análise às falsas classificações com base na Figura 59.

**Evento de 30/05/2019**

| | R.N. | R. | A.R. |
|---|---|---|---|
| R.N. | 39 | 0 | 0 |
| R. | 1 | 36 | 0 |
| A.R. | 3 | 2 | 21 |

**Evento de 03/06/2019**

| | R.N. | R. | A.R. |
|---|---|---|---|
| R.N. | 39 | 0 | 0 |
| R. | 3 | 96 | 0 |
| A.R. | 3 | 0 | 21 |

**Evento de 13/07/2019**

| | R.N. | R. | A.R. |
|---|---|---|---|
| R.N. | 39 | 0 | 0 |
| R. | 20 | 109 | 1 |
| A.R. | 0 | 0 | 1 |

**Evento de 16/10/2019**

| | R.N. | R. | A.R. |
|---|---|---|---|
| R.N. | 39 | 0 | 0 |
| R. | 0 | 44 | 0 |
| A.R. | 2 | 0 | 34 |

**Figura 56 – Matrizes de confusão das deteções de 2019 (R.N. – Regime Normal, R. – Rotura, A.R. – Alteração de Regime; Valores reais nas linhas e estimados nas colunas)**

Uma vez apresentados os dados relativos aos eventos ocorridos ao longo do ano de 2019, será apresentada em seguida cada uma das ocorrências em particular.

Olhando para a Figura 57, relativa ao evento ocorrido em 30/05/2019, constata-se que o falso negativo na deteção da rotura ocorre numa fase inicial desta. Tendo em conta que a instalação se encontrava parada, não foi encontrada nenhuma variação no caudal de saída e a pressão a jusante das bombas elevatórias ainda se encontrava em valores normais. Já as alterações de regime detetadas como roturas ocorreram após os arranques da bomba elevatória para carregar a conduta, sendo esses momentos caracterizados por variações de pressão e de caudal que o sistema avaliou erradamente como situações de rotura. Já na fase final de

carregamento o sistema considerou que a conduta já estaria carregada antes do momento definido na pré-classificação.

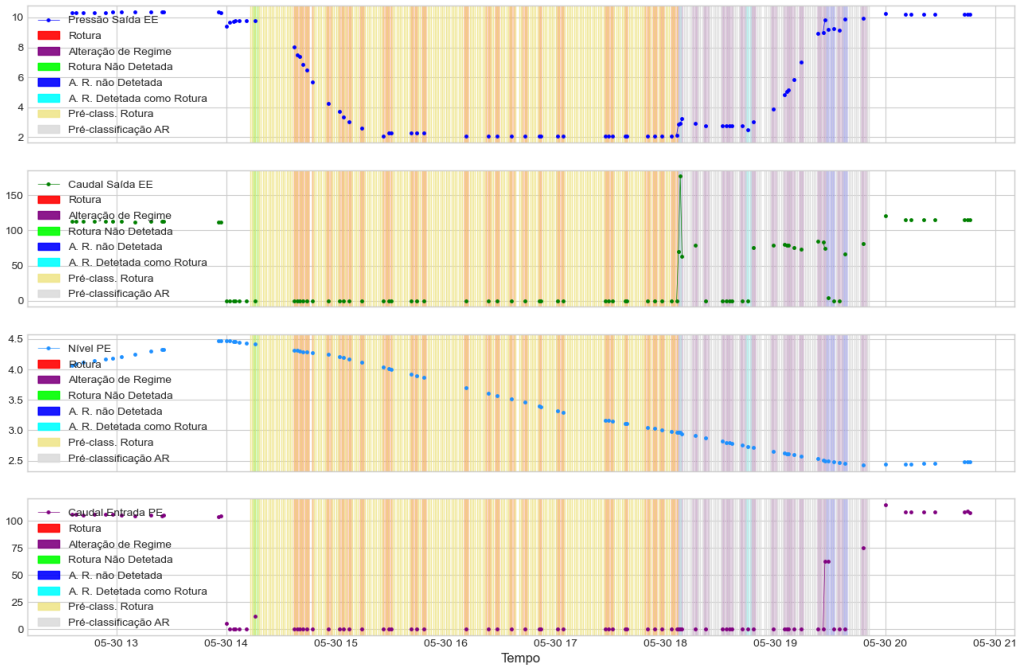


**Figura 57 – Deteções no evento ocorrido a 30/05/2019 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Na Figura 58, relativa ao evento ocorrido em 03/06/2019, verifica-se a existência de uma rotura que foi reparada e o aparecimento de outra, algum tempo após o carregamento da conduta. Mais uma vez ambas as situações não foram detetadas logo na primeira leitura considerada na pré-classificação como rotura. Possivelmente as razões não estarão muito longe das apresentadas para a situação de 30 de maio. Também na fase final de ambos os carregamentos da conduta o sistema atuou da mesma forma que anteriormente, deixando de detetar a alteração de regime momentos antes do definido na pré-classificação.

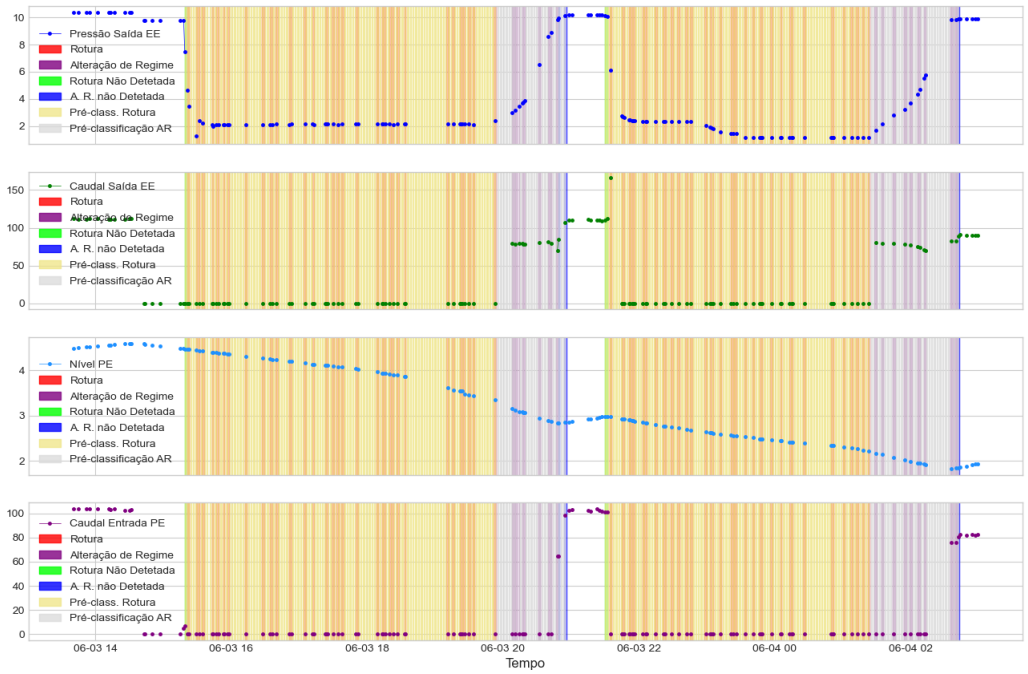


**Figura 58 – Deteções no evento ocorrido a 03/06/2019 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Quanto ao evento decorrido a 13 de julho de 2019, representado na Figura 59, a deteção foi bastante mais demorada que as anteriores, o que se traduz por um maior número de dados de rotura não detetados. Esta é uma situação especial, uma vez que a rotura foi junto ao reservatório de destino, já após o caudalímetro de chegada e sem alterações significativas nas leituras de caudal e pressão até à paragem do grupo elevatório. A única evidência de que algo não estaria bem até esse momento foi a descida constante do nível do reservatório de destino apesar de o grupo elevatório se encontrar em funcionamento. Esta situação é única nos dados relativos a 2019, mas o algoritmo conseguiu detetá-la, apesar de ter demorado mais algum tempo do que nos casos anteriores. Apesar de ter havido variações de pressão e caudal após a primeira paragem do grupo elevatório, foram ainda realizados dois arranques antes da deteção da rotura.

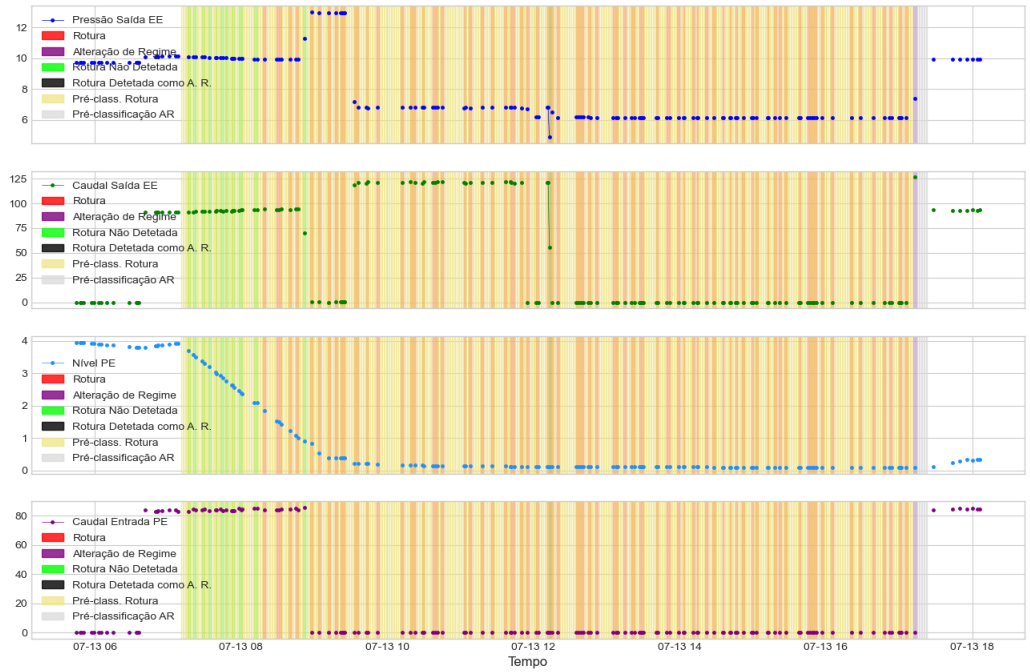


**Figura 59 – Deteções no evento ocorrido a 13/07/2019 (pressão em bar, nível em metros e caudal em $m^3/h$)**

A rotura de 16 de outubro apenas teve alterações de regime não detetadas na fase final do carregamento da conduta, conforme se verifica na Figura 60.

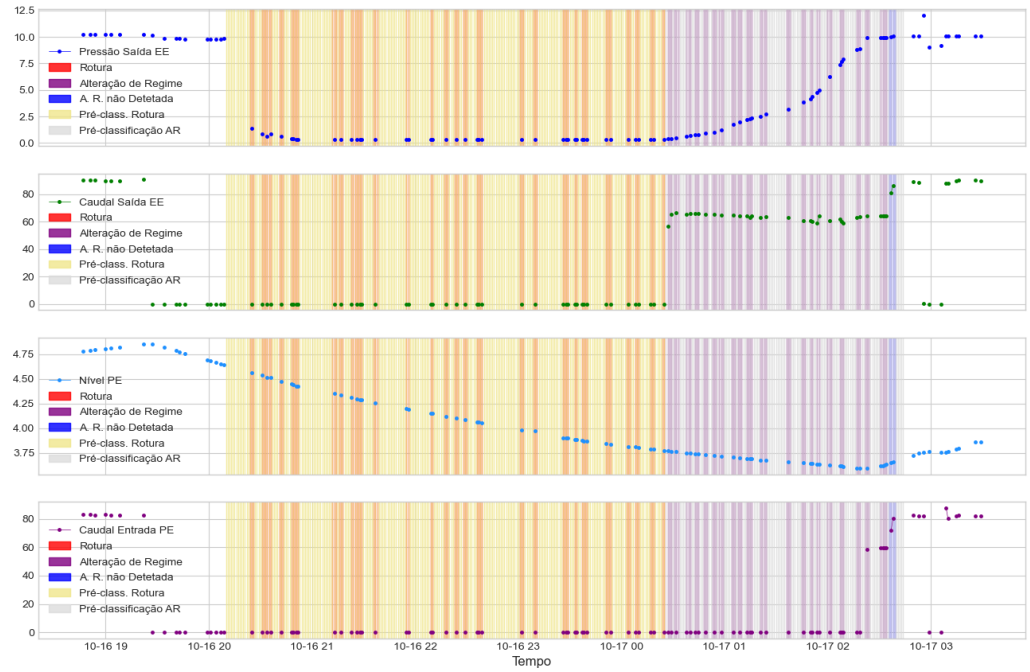


**Figura 60 – Deteções no evento ocorrido a 16/10/2019 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Foi ainda detetado um evento de alteração de pressão compatível com condições de rotura (falso positivo) no dia 3 de outubro às 01:00. O mesmo não tinha sido pré-classificado como possível rotura, no entanto é percetível a existência de um pico no momento da deteção, conforme se mostra na Figura 61.

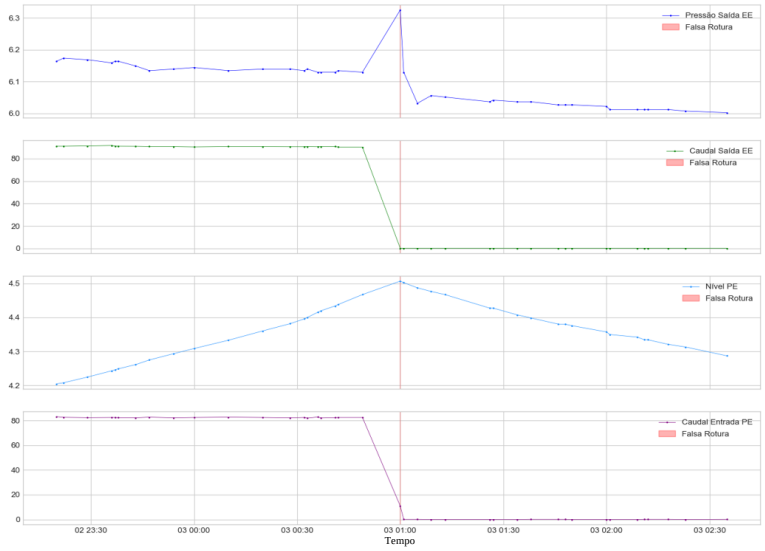


**Figura 61 – Pico de pressão não classificado que foi encontrado pelo algoritmo (pressão em bar, nível em metros e caudal em $m^3/h$)**

Dado que entre 4 de setembro e 4 de outubro de 2019 existiram múltiplas leituras de pressão compatíveis com condições de rotura e que não se voltou a verificar essa situação, assume-se que o transdutor de pressão tenha sido substituído no dia 4 de outubro.

## 6.2. Deteção em novos dados

Tendo-se obtido resultados bastante satisfatórios na deteção de anomalias nos dados de validação, classificaram-se novos dados relativos aos anos de 2020, 2021 e 2022 com o mesmo algoritmo. Tal como anteriormente, classificaram-se previamente os dados para poder obter uma classificação das deteções realizadas.

### - Ano de 2020

Ao longo do ano de 2020 contabilizaram-se quatro roturas. O algoritmo identificou oito eventos, quatro dos quais foram variações instantâneas de pressão apenas com uma amostra. Relativamente às roturas ocorridas na conduta ao longo desse ano, a primeira ocorreu no dia 3 de maio e as variações das variáveis relativas à conduta durante esse evento encontram-se na Figura 62. São visíveis algumas más classificações devido a alterações de regime detetadas como rotura e alterações de regime não detetadas ao longo do carregamento da conduta. As primeiras devem-se aos valores assumidos pela pressão e caudal à saída da instalação que em determinados pontos assumem valores compatíveis com situações de rotura e as segundas têm a ver com o ponto em que o algoritmo considera que a conduta se encontra carregada.

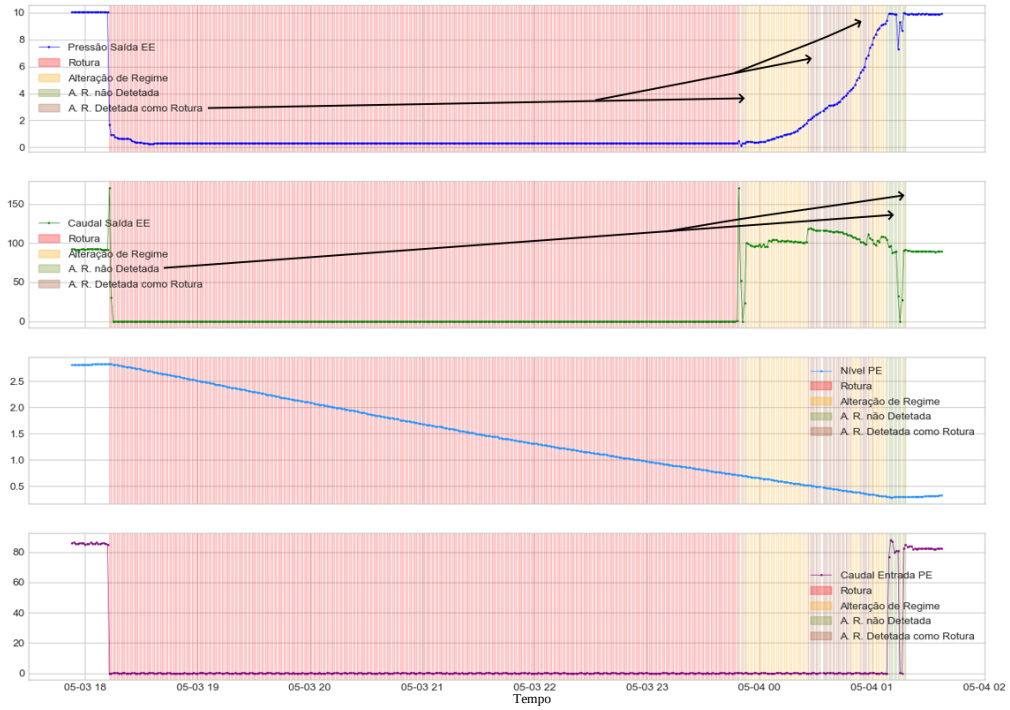


**Figura 62 – Representação gráfica dos parâmetros da rotura ocorrida em 3 de maio de 2020 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Também nas roturas que ocorreram a 9 de maio e 2 de junho, verificaram-se algumas alterações de regime não detetadas e alterações de regime detetadas como rotura, desta vez em menor quantidade, como se pode verificar na Figura 63 e na Figura 64.

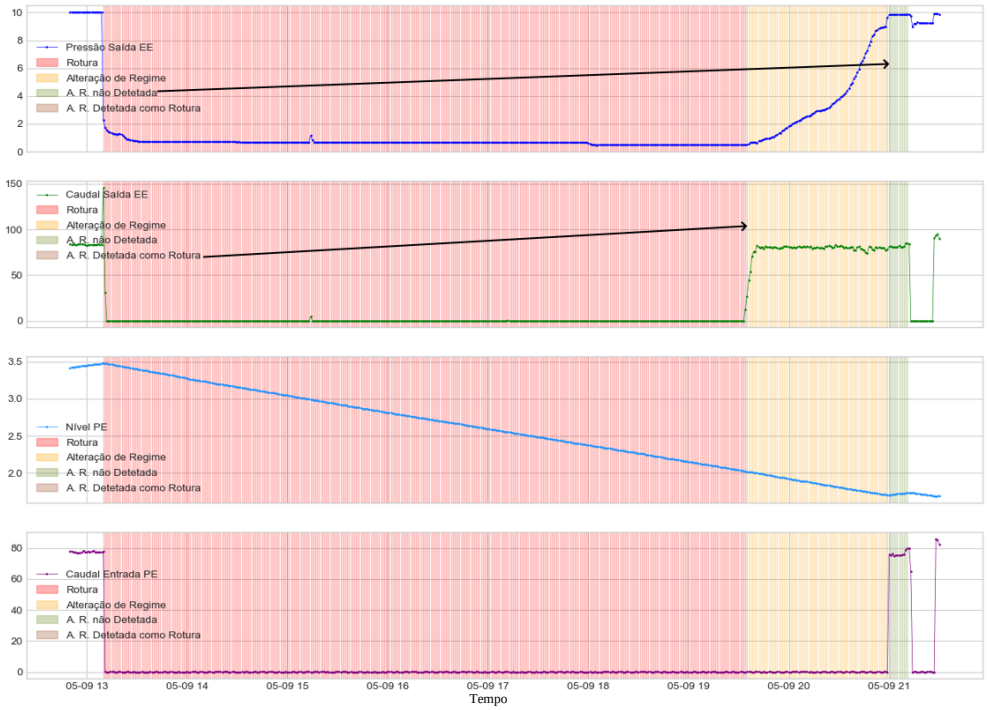


**Figura 63 – Representação gráfica dos parâmetros da rotura ocorrida em 9 de maio de 2020 (pressão em bar, nível em metros e caudal em $m^3/h$)**

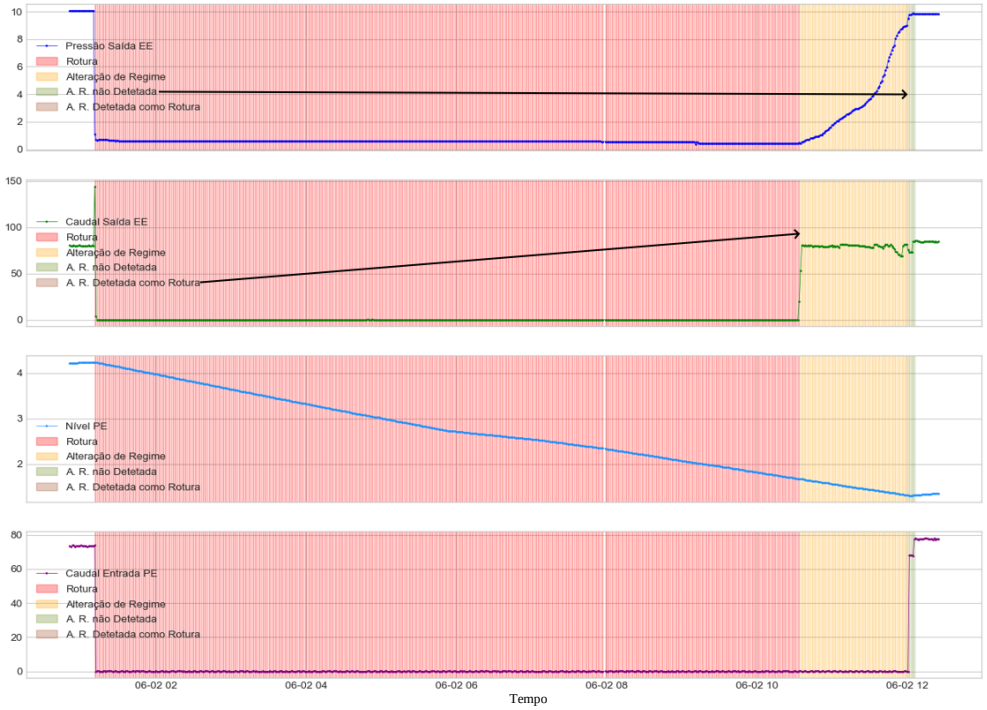


**Figura 64 – Representação gráfica dos parâmetros da rotura ocorrida em 2 de junho de 2020 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Contrariamente ao que ocorreu nas roturas anteriores, a de 17 de setembro de 2020 não foi detetada logo no primeiro ponto pré-classificado, o que originou falsos negativos para a situação de rotura. Foram ainda mal classificados alguns pontos de alteração de regime durante o carregamento da conduta, à semelhança do que já tinha ocorrido anteriormente. Os dados relativos a esse evento encontram-se na Figura 65.

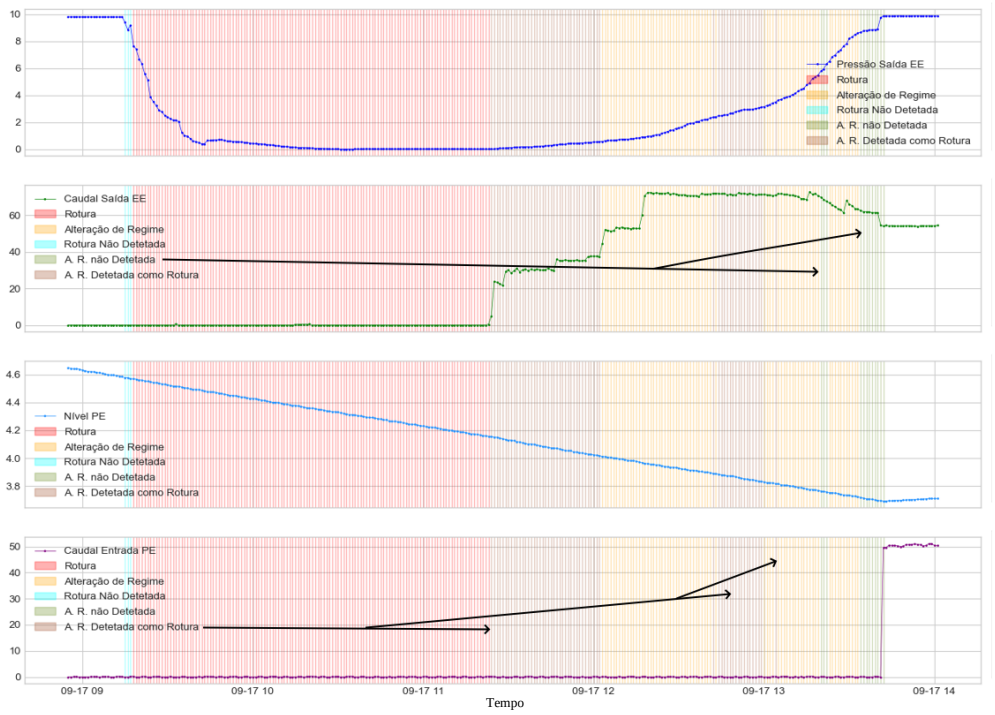


**Figura 65 – Representação gráfica dos parâmetros da rotura ocorrida em 17 de setembro de 2020 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Na Figura 66 encontram-se as matrizes de confusão relativas às roturas identificadas em 2020, todas elas detetadas pelo algoritmo. Verifica-se que a rotura de 17 de setembro foi aquela em que se verificou um maior número de falsas classificações, com 23% de erros de classificação para os casos de alteração de regime.

**Evento de 03/05/2020**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 336 | 0 |
| **A.R.** | 10 | 30 | 49 |

**Evento de 09/05/2020**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 384 | 0 |
| **A.R.** | 13 | 2 | 83 |

**Evento de 02/06/2020**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 562 | 0 |
| **A.R.** | 5 | 1 | 87 |

**Evento de 17/09/2020**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 3 | 126 | 0 |
| **A.R.** | 12 | 57 | 70 |

**Figura 66 – Matrizes de confusão das deteções de 2020 (R.N. – Regime Normal, R. – Rotura, A.R. – Alteração de Regime; Valores reais nas linhas e estimados nas colunas)**

### - Ano de 2021

Verificou-se a existência de duas roturas ao longo do ano de 2021, tendo o algoritmo detetado um total de trinta eventos.

A maior parte dos eventos detetados refere-se a picos de pressão no arranque do grupo elevatório, coincidentes com o que se encontra na Figura 67.

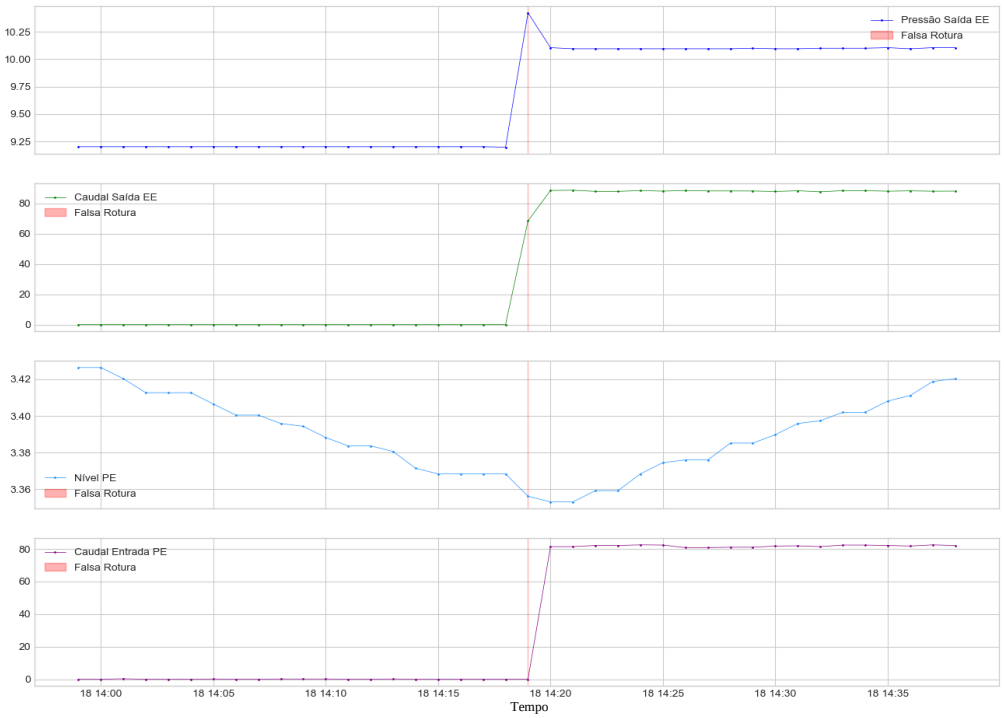


**Figura 67 – Pico de pressão não classificado que foi encontrado pelo algoritmo nos dados relativos a 2021 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Relativamente às verdadeiras roturas detetadas, a primeira ocorreu a 14 de julho de 2021 e na Figura 68 encontra-se a representação dos parâmetros relativos à conduta de transporte de água ao longo da mesma. Verifica-se que a rotura não foi logo identificada de início pelo algoritmo, o que deu origem a alguns falsos negativos. Pela análise do caudal é visível que a bomba arrancou durante a rotura, não tendo chegado caudal ao reservatório de destino e tendo o algoritmo entendido esse arranque como alteração de regime. Nos primeiros momentos do carregamento da conduta verificou-se ainda a deteção de alteração de regime como rotura e no final o algoritmo deixou de detetar o carregamento da conduta, considerando-a já carregada.

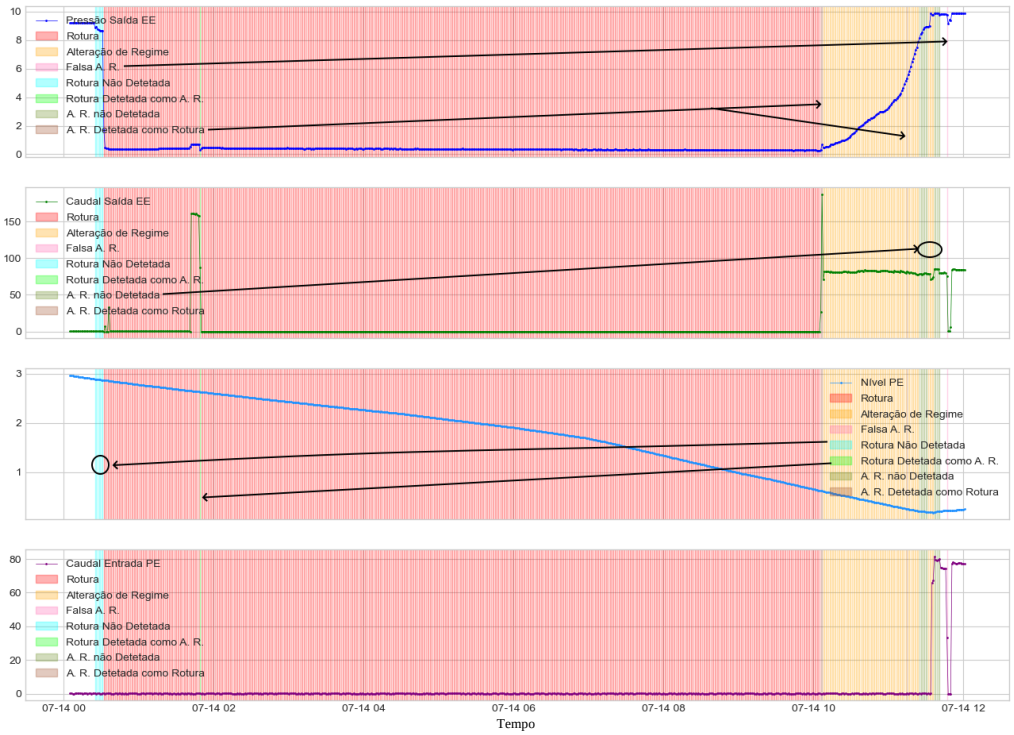


**Figura 68 – Representação gráfica dos parâmetros da rotura ocorrida em 14 de julho de 2021 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Na Figura 69 pode verificar-se que a rotura de 18 de setembro de 2021 foi bem classificada logo de início e que na fase de carregamento da conduta a pressão demorou bastante tempo a subir devido ao baixo caudal injetado na conduta. Dessa forma houve um conjunto de pontos classificados como rotura quando na realidade eram alteração de regime. No final a conduta foi considerada carregada antes do ponto considerado na pré-classificação.

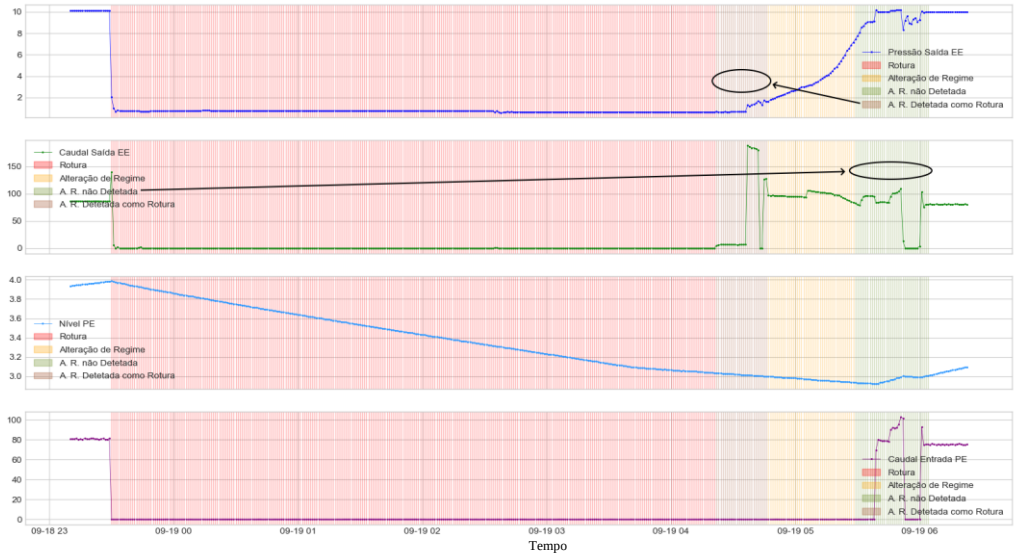


**Figura 69 – Representação gráfica dos parâmetros da rotura ocorrida em 18 de setembro de 2021 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Na Figura 70 é possível verificar que a maioria dos pontos correspondentes a rotura foi detetada, sendo evidente nos dados relativos ao carregamento da conduta após a rotura de 18 de setembro de 2021 uma grande dispersão dos pontos pré-classificados como alteração de regime, com 36 a serem classificados pelo algoritmo como regime normal e 25 como rotura. Tal como ocorreu nos casos anteriores, nas linhas encontram-se os valores reais e nas colunas os valores estimados pelo classificador.

**Evento de 14/07/2021**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 1 |
| **R.** | 7 | 573 | 1 |
| **A.R.** | 12 | 3 | 81 |

**Evento de 18/09/2021**

| | R.N. | R. | A.R. |
|---|---|---|---|
| **R.N.** | 39 | 0 | 0 |
| **R.** | 0 | 292 | 0 |
| **A.R.** | 36 | 25 | 42 |

**Figura 70 – Matrizes de confusão das deteções de 2021 (R.N. – Regime Normal, R. – Rotura, A.R. – Alteração de Regime; Valores reais nas linhas e estimados nas colunas)**

Verificou-se que no final do ano de 2021 o algoritmo identificou uma tipologia de evento que não havia ainda sido detetada de forma isolada. Em funcionamento normal era detetada

alteração de regime, que era obviamente falsa. Na Figura 71 encontra-se a representação gráfica de um desses eventos.

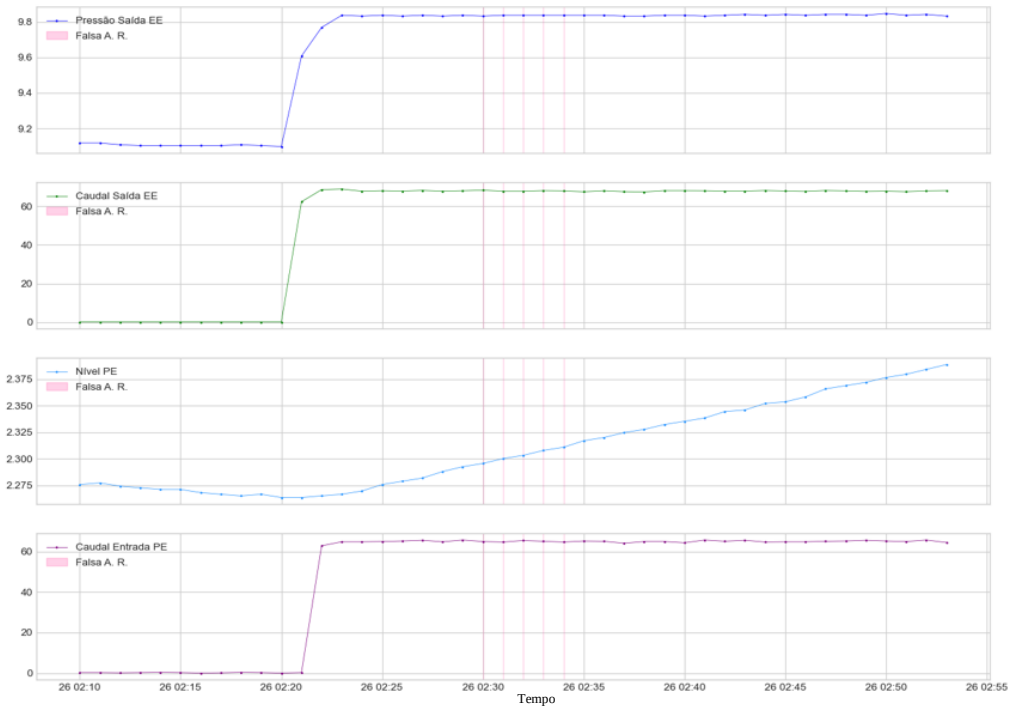


**Figura 71 – Representação gráfica de uma falsa alteração de regime (pressão em bar, nível em metros e caudal em $m^3/h$)**

Constatou-se então que em outubro de 2021 o caudal elevado por um dos grupos, que normalmente era 80 $m^3/h$, foi ajustado para cerca de 60 $m^3/h$ através de estrangulamento da válvula de compressão, tendo a pressão à saída da instalação descido também. Os novos valores de pressão e caudal estavam muito provavelmente no limite de valores que o algoritmo considera alteração de regime (carregamento de conduta), pelo que uma pequena variação das variáveis levaria à deteção desses eventos. Esta situação aliada à verificada no carregamento da conduta após a rotura de 18 de setembro de 2021 mostra que uma simples alteração de caudal, que representa uma alteração do modo de funcionamento da instalação em determinadas situações, pode tornar um modelo de deteção que funciona bem num modelo sem qualquer utilidade. À semelhança do que aconteceu na análise ao sistema de abastecimento gravítico, após a alteração ao regime de funcionamento, neste caso causada pela diminuição de caudal debitado por um dos grupos existentes na instalação, o modelo deveria ser novamente otimizado e treinado com dados que incluíssem o novo regime de funcionamento, para que lhe fosse possível reconhecê-lo.

### - Ano de 2022

Ao longo do ano de 2022 ocorreu apenas uma rotura na conduta e dois eventos de perda temporária do valor da pressão para realização de trabalhos na instalação, que foram consideradas anomalia para não interferir com os dados. O algoritmo detetou 15 eventos. Tal como ocorreu nos dados relativos aos anos anteriores, alguns desses eventos sinalizam variações de pressão e caudal decorrentes do arranque e paragem da elevação que são compatíveis com situações de rotura. Representam apenas uma ou duas amostras seguidas, voltando depois a pressão e o caudal para valores ditos normais. Verificou-se também a ocorrência das falsas alterações de regime, uma vez que um dos grupos elevatórios se manteve com velocidade reduzida.

Nas Figura 72 e Figura 73 encontram-se os gráficos relativos às perdas de pressão na compressão dos grupos para realização de trabalhos na instalação. É visível a correta deteção de rotura com a perda de pressão e o regresso ao "normal" quando os valores de pressão estão coincidentes com o que o algoritmo considera expectável em comparação com os restantes parâmetros.

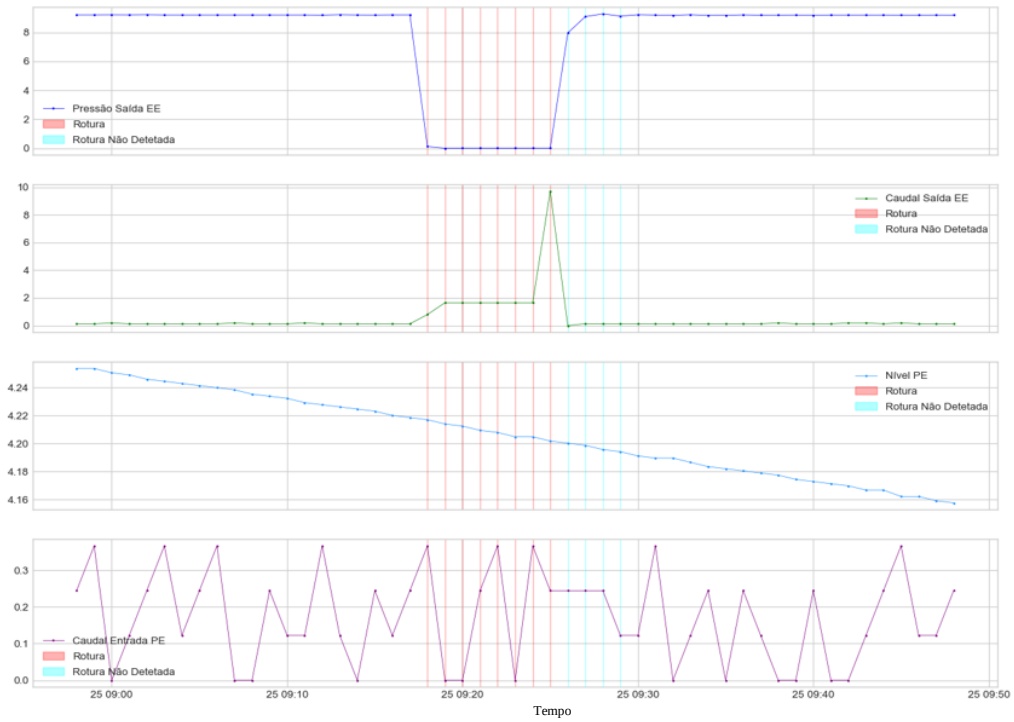


**Figura 72 – Perdas de pressão na compressão dos grupos devido à realização de trabalhos na instalação (pressão em bar, nível em metros e caudal em $m^3/h$)**


**Figura 73 – Perdas de pressão na compressão dos grupos devido à realização de trabalhos na instalação (pressão em bar, nível em metros e caudal em $m^3/h$)**

Relativamente à rotura ocorrida na conduta elevatória no dia 22 de julho de 2022, constata-se na Figura 74 que a mesma foi detetada conforme a pré-classificação, tendo por algumas vezes o algoritmo alterado erradamente a deteção para alteração de regime por estar na presença de caudal. No final verificaram-se as habituais alterações de regime detetadas como roturas e alterações de regime não detetadas.

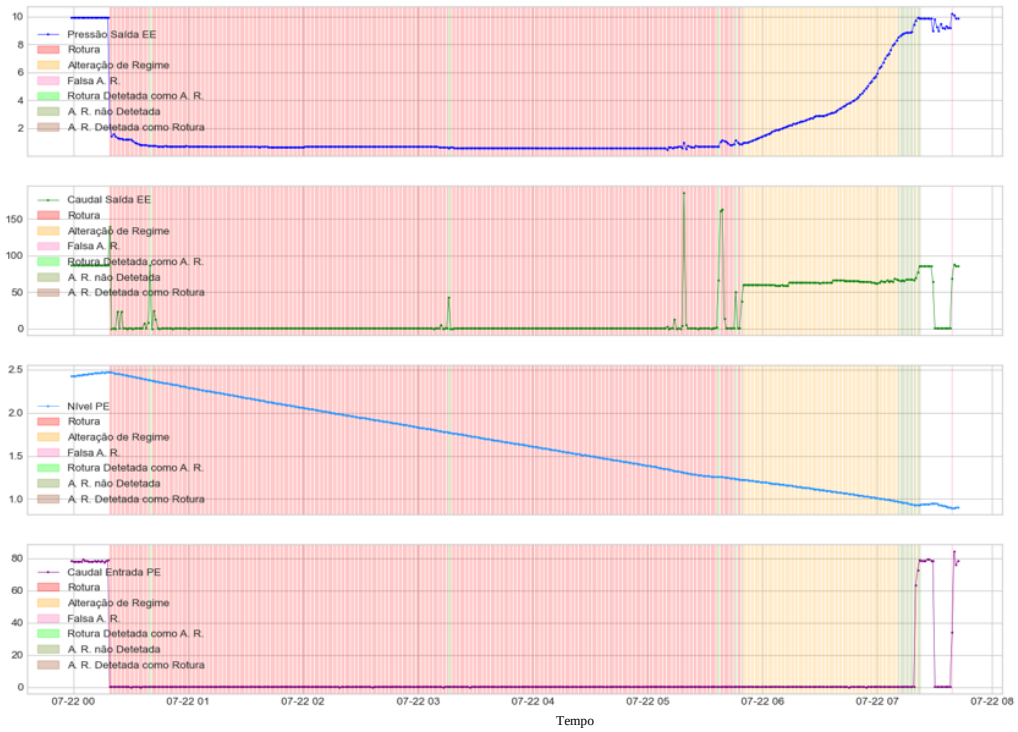


**Figura 74 – Representação gráfica dos parâmetros da rotura ocorrida em 22 de julho de 2022 (pressão em bar, nível em metros e caudal em $m^3/h$)**

Na Tabela 25 encontra-se a matriz de confusão relativa a esta rotura. Verifica-se uma taxa de erro de classificação na ordem dos 3,6%.

**Tabela 25 – Matriz de confusão rotura ocorrida em 22 de julho de 2022 (R.N. – Regime Normal, R. – Rotura, A.R. – Alteração de Regime; Valores reais nas linhas e estimados nas colunas)**

| | Evento de 22/07/2022 | | |
|---|---|---|---|
| | **R.N.** | **R.** | **A.R.** |
| **R.N.** | 39 | 0 | 1 |
| **R.** | 0 | 326 | 4 |
| **A.R.** | 11 | 1 | 82 |

# 7. Discussão de Resultados

Na realização do estudo apresentado no presente trabalho foram analisados dados provenientes de dois sistemas de abastecimento com naturezas diferentes.

O primeiro caso apresentado refere-se a um sistema de abastecimento gravítico, que é afetado por alterações de caudal e pressões devido a diversos fatores. Dois fatores importantes são o nível do reservatório onde se inicia o troço de abastecimento e a posição dos boiadores existentes nos reservatórios de destino. Quando o nível do reservatório de destino sobe, estes boiadores atuam sobre a válvula de entrada aumentando a pressão na conduta e diminuindo o caudal que chega ao reservatório.

Compararam-se os resultados de deteção obtidos por algoritmos baseados em *Random Forest*, ANN, *XGBoost* e SVM, tendo sido otimizados um conjunto de hiperparâmetros para cada um dos algoritmos. Essas otimizações foram efetuadas utilizando a ferramenta *Optuna*.

Para este primeiro caso foi necessário realizar duas otimizações de hiperparâmetros, uma para um conjunto de dados relativos ao ano de 2022 e outra para um segundo conjunto de dados relativo ao final do ano de 2022, ao ano de 2023 e a alguns dias de 2024. Esta situação deveu-se a uma alteração realizada na conduta no dia 21 de dezembro de 2022 que causou variações consideráveis nos dados em análise e inviabilizou a utilização do modelo otimizado com o primeiro conjunto de dados.

A métrica de avaliação utilizada foi o parâmetro *Macro Average F1-Score*, tendo o algoritmo baseado em *Random Forest* obtido os melhores resultados, conforme se pode observar na Tabela 26. Nessa mesma tabela é visível que a precisão apresenta sempre valores mais altos devido a não considerar as diferenças existentes entre o número de amostras pertencentes a cada classe. Na Tabela 27 apresentam-se os melhores hiperparâmetros para ambas as otimizações realizadas para o algoritmo baseado em *Random Forest*.

**Tabela 26 – Classificação global obtida para cada um dos algoritmos em cada conjunto de dados para o Caso I**

| | Primeiro conjunto de dados | | Segundo conjunto de dados | |
|---|---|---|---|---|
| | Precisão (*Precision*) | *Macro Average F1-score* | Precisão (*Precision*) | *Macro Average F1-score* |
| ***Random Forest*** | 0,99933 | 0,98454 | 0,9981 | 0,88352 |
| ***XGBoost*** | 0,9981 | 0,95808 | 0,99774 | 0,87531 |
| **SVM** | 0,9988 | 0,97179 | 0,99706 | 0,74905 |
| **ANN** | 0,99882 | 0,97215 | 0,99716 | 0,84051 |

**Tabela 27 – Hiperparâmetros resultantes da otimização do modelo *Random Forest* para as duas otimizações realizadas no Caso I**

| | Primeiro conjunto de dados | Segundo conjunto de dados |
|---|---|---|
| ***max_depth*** | 81 | 61 |
| ***max_leaf_nodes*** | 102 | 102 |
| ***min_samples_leaf*** | 2 | 2 |
| ***min_samples_split*** | 30 | 22 |
| ***n_estimators*** | 73 | 29 |

Para além de detetar todas as roturas, o algoritmo otimizado com os novos hiperparâmetros conseguiu ainda detetar dez casos em que houve picos de caudal e pressão. Apesar de não corresponderem a situações efetivas de rotura, são situações anómalas que poderão ter contribuído para uma degradação mais acentuada da conduta. Na sua maioria, essas alterações foram consideradas alterações de regime, tendo contribuído para o desempenho obtido na classificação de situações de alteração de regime. A maior ênfase dada a este segundo modelo em detrimento do primeiro, prende-se com o facto de ter sido treinado com dados mais atuais que permitem a sua aplicação na análise de novos dados. Já o primeiro modelo encontra-se desatualizado devido à alteração dos fatores que levaram à calibração dos hiperparâmetros.

Verificou-se também que uma parte dos erros de classificação dos dados de carregamento da conduta se devem à proximidade entre os valores das pressões e do caudal em determinados pontos de carregamento e em situações de funcionamento normal. Por outras palavras, valores que em determinadas situações representam um comportamento normal, noutras situações representam comportamentos anómalos. Provavelmente este tipo de erro não ocorreria se fossem analisadas as leituras de caudal de entrada nos reservatórios de

destino, já que que durante a fase de carregamento da conduta as somas destes valores devem ser discrepantes em relação ao caudal que passa no caudalímetro da caixa da válvula redutora de pressão. Uma vez que vão ser instalados caudalímetros no Reservatório de Baraçais e no PE Caniceira, provavelmente a capacidade de deteção do modelo irá melhorar.

Relativamente ao sistema de abastecimento elevatório (caso II), foram fornecidos dados recolhidos entre 2019 e 2022, que foram separados por ano civil. Os dados relativos ao ano de 2019 foram utilizados para treino dos algoritmos, tendo os restantes sido utilizados como dados novos para classificação. Mais uma vez foi o algoritmo *Random Forest* que obteve a melhor classificação global, conforme se pode constatar pela Tabela 28. Na Tabela 29 apresentam-se os melhores hiperparâmetros obtidos.

**Tabela 28 – Classificação global obtida para cada um dos algoritmos para o Caso II**

| | Dados 2019 | |
|---|---|---|
| | Precisão (*Precision*) | *Macro Average F1-score* |
| **Random Forest** | 0,99939 | 0,96880 |
| **XGBoost** | 0,99818 | 0,91459 |
| **SVM** | 0,99936 | 0,96449 |
| **ANN** | 0,99931 | 0,95559 |

**Tabela 29 – Hiperparâmetros resultantes da otimização do modelo *Random Forest* para as duas otimizações realizadas no Caso II**

| *max_depth* | *max_leaf_nodes* | *min_samples_leaf* | *min_samples_split* | *n_estimators* |
|---|---|---|---|---|
| 61 | 102 | 2 | 6 | 77 |

Verificou-se que todas as roturas ocorridas foram detetadas, bem como um conjunto de alterações de regime fora dos eventos de carregamento de conduta.

Verificou-se no final de 2021 uma alteração no regime de funcionamento de um dos grupos elevatórios, que passou a funcionar com velocidade reduzida, elevando assim menos caudal e causando um comportamento que por vezes o algoritmo identifica como alteração de regime de funcionamento. Neste tipo de situação há que analisar os efeitos da alteração de estado no sistema, podendo ser decidido manter a alteração de regime e otimizar o modelo para que a situação seja tida como normal ou reverter as modificações que levaram à alteração de regime por haver risco de causar danos a instalações e/ou equipamentos.

# 8. Conclusões e perspetivas de trabalho futuro

O objetivo do presente trabalho foi a exploração de métodos que permitam detetar situações de rotura e alterações de regime em condutas de abastecimento de água nos processos de transporte. Foram analisadas duas situações, a primeira das quais referente a uma conduta de abastecimento gravítico e a segunda referente a uma conduta elevatória. Recolheram-se de um servidor existente um conjunto de dados históricos reais relativos a níveis de reservatórios, pressões e caudais nas condutas. Realizou-se um pré-tratamento e uma análise prévia dos dados, tendo-se aplicado em seguida modelos de deteção baseados nos algoritmos *Random Forest, XGBoost, Support Vector Machines e Artificial Neural Networks* para deteção de situações de rotura e de alterações ao regime normal de funcionamento. Verificou-se que os modelos que obtiveram melhores resultados em ambas as situações foram os modelos baseados em *Random Forest*. Foi também analisado o desempenho dos modelos em situações de substituição de componentes hidráulicos e alteração do regime típico de funcionamento dos sistemas de distribuição.

Os resultados obtidos em ambos os casos estudados são bastante promissores, uma vez que para além da deteção de roturas os algoritmos conseguem também detetar alterações de funcionamento que podem pôr em causa a integridade do sistema de abastecimento. Ficou claro que alterações ao funcionamento típico da conduta alteram a capacidade de classificação do algoritmo, o que obriga a alterações aos hiperparâmetros dos modelos e a novos treinos sempre que essas alterações sejam para manter.

O sistema de abastecimento elevatório obteve melhores resultados na deteção devido à presença de informações relativas ao local de destino, nomeadamente caudal de entrada e nível do reservatório.

Também se conclui que em situações que causem alterações significativas ao regime típico de funcionamento de ambos os tipos de sistemas de abastecimento (gravítico e elevatório), o modelo referente ao caso I apresenta uma deterioração mais notória. Este modelo (relativo à conduta de abastecimento gravítico) deverá ser testado na classificação de novos dados para perceber como se comporta, uma vez que os dados mais atuais (após a substituição da válvula redutora de pressão) foram utilizados na totalidade para treino e teste. Este modelo

poderá ser muito mais assertivo nas classificações que realiza se contar com as leituras de caudal de chegada aos reservatórios de destino, uma vez que estas variáveis vão ajudar a classificar as situações em que o desempenho do algoritmo foi mais baixo, que são as alterações de regime.

Já o modelo relativo à conduta de abastecimento elevatório (caso II), otimizado e treinado com os dados relativos ao ano de 2019, apesar de ter aumentado o número de deteções de falsas alterações de regime devido à redução do caudal debitado por um dos grupos elevatórios (dados de 2021 e 2022), a degradação sentida não foi tão significativa. Possivelmente esta maior robustez deveu-se ao maior número de variáveis disponíveis.

O presente trabalho representa apenas os primeiros passos no desenvolvimento de uma ferramenta que pode ser melhorada e trabalhada, para que seja possível a sua aplicação a um sistema mais complexo, com um conjunto alargado de reservatórios de destino. Dessa perspetiva de desenvolvimento faz parte a exploração de algoritmos de aprendizagem não supervisionada, cujo sucesso permitiria uma versatilidade superior na adaptação a novas situações, tendo em conta que não seria necessária a pré-classificação de dados. Poderia definir um padrão de registos conformes e na presença de anomalias a este padrão típico de funcionamento existiria sinalização de situações potencialmente críticas. Também a mitigação do desequilíbrio de dados deverá ser revisitada, utilizando abordagens diferentes das referidas no Capítulo 4, numa tentativa de obter melhores resultados na classificação das classes minoritárias. Para tornar os algoritmos mais fiáveis relativamente aos picos de consumo que normalmente ocorrem nos meses mais quentes, será interessante incorporar nas bases de dados informação sobre o mês atual, a temperatura nas imediações da instalação onde se recolhem os dados e a época do ano. Também poderá ter bastante interesse a interpretação do funcionamento dos modelos *"Random Forest"* otimizados, para assim compreender melhor a importância dada pelo classificador a cada uma das variáveis de entrada.

# Referências Bibliográficas


Aditya Sharma. (2020, January). *Principal Component Analysis (PCA) in Python Tutorial*. Principal Component Analysis (PCA) in Python Tutorial. https://www.datacamp.com/tutorial/principal-component-analysis-in-python

AdVT. (2023). *Sub-Sistema de Abastecimento de Água e de Saneamento do Oeste*. Águas Do Vale Do Tejo. Retrieved December 30, 2023, from https://www.advt.pt/index.php/pt/menu/atividade/abastecimento-de-agua/subsistemas/

Alves Coelho, J., Glória, A., Sebastião, P. (2020). Precise Water Leak Detection Using Machine Learning and Real-Time Sensor Data. *Internet of Things*, *1*(2), 474–493. https://doi.org/10.3390/iot1020026

André de Oliveira. (2020). *Estudo e análise preditiva sobre as fugas na rede de distribuição de água de Vila Nova de Gaia*. Retrieved December 10, 2023, from https://hdl.handle.net/10216/131462

Fontes, J. (2016, October). *Detecção de Fugas em Redes de Distribuição de Água*. Retrieved December 10, 2023, from http://users.isr.ist.utl.pt/~jsm/SPARSIS/2016_Joao_Fontes_MSc.pdf

Anselmo, F. (2019, December 22). *Modelos de Machine Learning - DBScan*. https://pt.linkedin.com/pulse/modelos-de-machine-learning-dbscan-fernando-anselmo

Anselmo, Lourenço, Maria, Filipe, Paço, Rosa, Florentino, Serrinha, Inácio, Simões, Pereira. (2010). *EPAL e os Municípios* (1st ed.). Edição EPAL

Araújo, R. (2022, March 11). *Introdução ao SVM como Classificador*. https://pt.linkedin.com/pulse/introdu%C3%A7%C3%A3o-ao-svm-como-classificador-rodrigo-araujo-

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Blázquez-García, A., Conde, A., Mori, U., Lozano, J. A. (2021). Water leak detection using self-supervised time series classification. *Information Sciences*, *574*, 528–541. https://doi.org/10.1016/j.ins.2021.06.015

Boswell, D. (2002, August 6). *Introduction to Support Vector Machines*, February 1, 2024, from https://home.work.caltech.edu/~boswell/IntroToSVM.pdf

Bre, F., Gimenez, J. M., Fachinotti, V. D. (2018). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, *158*, 1429–1441. https://doi.org/10.1016/j.enbuild.2017.11.045

Brownlee, J. (2021, January 5). *Random Oversampling and Undersampling for Imbalanced Classification - MachineLearningMastery.com*. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

Bushaev, V. (2017, December 4). *Stochastic Gradient Descent with momentum*. Towards Data Science. https://towardsdatascience.com/stochastic-gradient-descent-with-momentum-a84097641a5d

Camila Waltrick. (2020, May 7). *Machine Learning — O que é, tipos de aprendizagem de máquina, algoritmos e aplicações*. https://medium.com/camilawaltrick/introducao-machine-learning-o-que-e-tipos-de-aprendizado-de-maquina-445dcfb708f0

Campos, D. (2011). *Reabilitação de sistemas de abastecimento de água*. Retrieved February 28, 2023, from http://hdl.handle.net/10400.26/22294

Caputo, A. C., Pelagagge, P. M. (2003). Using neural networks to monitor piping systems. *Process Safety Progress*, *22*(2), 119–127. https://doi.org/10.1002/prs.680220208

Chen, T., He, T. (2024). *xgboost: eXtreme Gradient Boosting*. Retrieved February 12, 2024, from https://rdrr.io/cran/xgboost/f/inst/doc/xgboost.pdf

Costa, A. M. P. da. (2020). *Unsupervised Anomaly Detection in Water System Networks using Recurrent Neural Networks*. Retrieved February 12, 2023, from https://fenix.tecnico.ulisboa.pt/cursos/meic-a/dissertacao/846778572212579

DATAtab Team. (2024). *Point Biserial Correlation.* https://datatab.net/tutorial/point-biserial-correlation

Deng, J., Lin, Y. (2022). The Benefits and Challenges of ChatGPT: An Overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/FCIS.V2I2.4465

Departamento de Sustentabilidade Empresarial da EPAL. (2023)-a. *Relatório de Governo Societário 2022*. https://epal.pt/EPAL/docs/default-source/epal/relat%C3%B3rio-do-governo-societ%C3%A1rio/rgs-2022.pdf?sfvrsn=4

Departamento de Sustentabilidade Empresarial da EPAL. (2023)-b. *Relatório de Sustentabilidade 2022*. Retrieved November 23, 2023, from https://www.epal.pt/EPAL/docs/default-source/epal/relatorio-sustentabilidade/2022.pdf?sfvrsn=8

Diário de Notícias. (2023, March 22). *Água não faturada representou perdas de 347 milhões de euros em 2021*. Diário de Notícias. https://www.dn.pt/sociedade/agua-nao-faturada-representou-perdas-de-347-milhoes-de-euros-em-2021-16049312.html

Didatica Tech. (2024). *Como funciona o algoritmo de Árvore de Decisão (Decision Tree)*. Retrieved February 7, 2024, from https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/

Education Ecossistems. (2018, September 12). *Understanding K-means Clustering in Machine Learning*. Towards Data Science. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

El-Zahab, S., Zayed, T. (2019, June 11). *Leak detection in water distribution networks: an introductory overview*. https://doi.org/10.1186/s40713-019-0017-x

Engenharia do Departamento de Operações e Abastecimento da EPAL. (2019). *ETA de Vale da Pedra*. Retrieved December 28, 2023, from https://www.epal.pt/EPAL/docs/default-source/agua/eta-vale-da-pedra.pdf

Engenharia do Departamento de Manutenção (Adução) da EPAL. (2012). *Identificação de oportunidades de melhoria da qualidade da água em PMX*. Edição EPAL

Engenharia do Departamento de Operações e Abastecimento da EPAL. (2007). Caracterização do Sistema de Abastecimento da EPAL. In *Plano de Segurança da Água para Consumo Humano no Sistema de Abastecimento da EPAL, S.A.* Edição EPAL

Engenharia do Departamento de Operações e Abastecimento da EPAL. (2011). Caracterização do Sistema de Abastecimento da EPAL, S.A. In *Caracterização do Sistema de Abastecimento da EPAL, S.A.* (1st ed.). Edição EPAL

ERSAR. (2022). *O Setor da Água e do Saneamento de Águas Residuais*. Retrieved November 16, 2023, from https://www.ersar.pt/pt/setor/caracterizacao

ERSAR. (2023). *Decisão sobre a definição dos valores de ANF a considerar no cálculo da TRH, para efeitos de repercussão no utilizador final*. Retrieved November 23, 2023, from https://www.ersar.pt/pt/site-comunicacao/site-noticias/documents/decis%C3%A3o%20sobre%20a%20defini%C3%A7%C3%A3o%20dos%20valores%20de%20%C3%A1gua%20n%C3%A3o%20faturada%202023.pdf#:~:text=Assim%2C%20o%20Conselho%20de%20Administra%C3%A7%C3%A3o%20da%20ERSAR%20decide,de%2011%20de%20junho%2C%20na%20reda%C3%A7%C3%A3o%20em%20vigor

Fan, X., Zhang, X., Yu, X. ( B. (2021). Machine learning model and strategy for fast and accurate detection of leaks in water supply network. *Journal of Infrastructure Preservation and Resilience*, *2*(1). https://doi.org/10.1186/s43065-021-00021-6

Fang, S., Sun, W., Huang, L. (2019). Anomaly Detection for Water Supply Data using Machine Learning Technique. *Journal of Physics: Conference Series*, *1345*(2). https://doi.org/10.1088/1742-6596/1345/2/022054

Fares, A., Tijani, I. A., Rui, Z., Zayed, T. (2023). Leak detection in real water distribution networks based on acoustic emission and machine learning. *Environmental Technology (United Kingdom)*, *44*(25), 3850–3866. https://doi.org/10.1080/09593330.2022.2074320

Ferreira, A. (2017). *Deteção e Avaliação de Fugas e Perdas em Sistemas de Abastecimento de Água*. https://hdl.handle.net/10316/83191

Filho, M. (2023). *How to Get Feature Importance in XGBoost in Python*. https://forecastegy.com/posts/xgboost-feature-importance-python/

Gandhi, R. (2018, June 7). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Glander, S. (2018, November 29). *Machine Learning Basics - Gradient Boosting & XGBoost*. https://shirinsplayground.netlify.app/2018/11/ml_basics_gbm/

Grübler, M. (2018, June 11). *Entendendo o funcionamento de uma Rede Neural Artificial*. Medium. https://medium.com/brasil-ai/entendendo-o-funcionamento-de-uma-rede-neural-artificial-4463fcf44dd0

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. In *Current Reviews in Musculoskeletal Medicine* (Vol. 13, Issue 1, pp. 69–76). Springer. https://doi.org/10.1007/s12178-020-09600-8

Hemashreekilari. (2023, September 5). *Understanding Gradient Boosting*. Medium. https://medium.com/@hemashreekilari9/understanding-gradient-boosting-632939b98764

Hoare, J. (2023). *What is Gradient Boosting? - Gradient Boosting Explained*. DisplayR. Retrieved February 7, 2024, from https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/

Hu, Z., Tan, D., Chen, B., Chen, W., Shen, D. (2021). Review of model-based and data-driven approaches for leak detection and location in water distribution systems. In *Water Supply* (Vol. 21, Issue 7, pp. 3282–3306). IWA Publishing. https://doi.org/10.2166/ws.2021.101

Huang, J., Li, Y. F., Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology*, *67*, 108–127. https://doi.org/10.1016/j.infsof.2015.07.004

Institute of Electrical and Electronics Engineers, IEEE Robotics and Automation Society, IEEE Engineering in Medicine and Biology Society, & IEEE Systems, M. (2018). *Unsupervised Learning Based On Artificial Neural Network: A Review*. https://doi.org/10.1109/CBS.2018.8612259

Kammoun, M., Kammoun, A., Abid, M. (2023). LSTM-AE-WLDL: Unsupervised LSTM Auto-Encoders for Leak Detection and Location in Water Distribution Networks. *Water Resources Management*, *37*(2), 731–746. https://doi.org/10.1007/s11269-022-03397-6

Kazmi, H. (2024). *Classification using XGBoost in Python*. https://www.educative.io/answers/classification-using-xgboost-in-python

Leland McInnes, John Healy, & Steve Astels. (2016). *Predicting clusters for new points*. Predicting Clusters for New Points. https://hdbscan.readthedocs.io/en/latest/prediction_tutorial.html

Liu, Y., Ma, X., Li, Y., Tie, Y., Zhang, Y., Gao, J. (2019). Water pipeline leakage detection based on machine learning and wireless sensor networks. *Sensors (Switzerland)*, *19*(23). https://doi.org/10.3390/s19235086

Lorena, A. C., De Carvalho, A. C. P. L. F. (2007). *Uma Introdução às Support Vector Machines*. https://doi.org/10.22456/2175-2745.5690

Marmolejo-Ramos, F., Tejo, M., Brabec, M., Kuzilek, J., Joksimovic, S., Kovanovic, V., González, J., Kneib, T., Bühlmann, P., Kook, L., Briseño-Sánchez, G., Ospina, R. (2023). Distributional regression modeling via generalized additive models for location, scale, and shape: An overview through a data set from learning analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(1). https://doi.org/10.1002/WIDM.1479

Mashhadi, N., Shahrour, I., Attoue, N., El Khattabi, J., Aljer, A. (2021). Use of machine learning for leak detection and localization in water distribution systems. *Smart Cities*, *4*(4), 1293–1315. https://doi.org/10.3390/smartcities4040069

Masui, T. (2022a, January 20). *All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression*. Towards Data Science. https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Masui, T. (2022b, February 7). *All You Need to Know about Gradient Boosting Algorithm − Part 2. Classification*. Towards Data Science. https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-2-classification-d3ed8f56541e

McCue, C. (2007). Predictive Analytics. *Data Mining and Predictive Analysis*, 117–141. https://doi.org/10.1016/B978-075067796-7/50029-5

Muharemi, F., Logofătu, D., Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294–307. https://doi.org/10.1080/24751839.2019.1565653

Nascimento, W. (2021). *Detecção de vazamentos em dados de fluxo de água com seleção e otimização de modelos*. Retrieved January 30, 2023, from http://repositorio.utfpr.edu.br/jspui/handle/1/26307

Oliveira, A., Fonseca, A., Ramalhão, A. (2007). *Auditoria Energética EPAL*. EPAL Technical Editions.

Optuna Contributors. (2018). *Optuna: A hyperparameter optimization framework — Optuna 3.6.0 documentation*. https://optuna.readthedocs.io/en/stable/

Pandas Devlopers. (2024). *Cookbook — pandas 2.2.1 documentation*. https://pandas.pydata.org/docs/user_guide/cookbook.html#cookbook

Pereira, C., Nunes, D., Saraiva, H., Hilaco, S. (2022). *Manual do Sistema de Responsabilidade Empresarial*. Edição EPAL

Pinheiro, N. (2021). *Pré-processamento de dados com Python*. Retrieved December 21, 2023, from https://medium.com/data-hackers/pr%C3%A9-processamento-de-dados-com-python-53b95bcf5ff4

Ramzai, J. (2020). *Clearly explained: Pearson V/S Spearman Correlation Coefficient*. https://towardsdatascience.com/clearly-explained-pearson-v-s-spearman-correlation-coefficient-ada2f473b8

Rodrigues, P. (2021). *Estudo das Perdas Aparentes por Erros de Medição na Rede de Distribuição de Água de Beja*. http://hdl.handle.net/10400.1/19795

Şahin, E., Yüce, H. (2023). Prediction of Water Leakage in Pipeline Networks Using Graph Convolutional Network Method. *Applied Sciences (Switzerland)*, *13*(13). https://doi.org/10.3390/app13137427

Saias, J., Maia, M., Rato, L., Gonçalves, T. (2018). *Machine Learning: um estudo sobre conceitos, tarefas e algoritmos relacionados com predição e recomendação*. http://hdl.handle.net/10174/30174

Sampaio, C. (2023, July 2). *Understanding SVM Hyperparameters*. https://stackabuse.com/understanding-svm-hyperparameters/

Santos, J. M. (2022). *REDES NEURONAIS Conceitos*. Retrieved February 29, 2024, from https://www.isep.ipp.pt/files/Redes%20Neuronais%20-%20Conceitos%20-%20JMS%20(vers%C3%A3o_final).pdf

Sardinha, J., Serranito, F., Donnelly, A., Marmelo, V., Saraiva, P., Dias, N., Guimarães, R., Morais, D., Rocha, V. (2017). *Controlo Ativo de Perdas de Água* (2nd ed.). EPAL Technical Editions. Retrieved February 24, 2023 https://www.studocu.com/pt/document/instituto-universitario-de-lisboa/computer-engineering/controlo-ativo-de-perdas-de-agua/40026829

Sarvandani, M. (2023). *Top 10 applications of regression in machine learning*. https://medium.com/@mohamadhasan.sarvandani/top-applications-of-regression-in-machine-learning-2b9599da8090

Scikit-Learn Developers. (2024-a). *3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.4.1 documentation*. Retrieved February 29, 2024, from https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

Scikit-Learn Developers. (2024-b). *sklearn.ensemble.RandomForestClassifier — scikit-learn 1.4.0 documentation*. Retrieved February 7, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit-Learn Developers. (2024-c). *sklearn.neural_network.MLPClassifier*. Retrieved February 25, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Scikit-Learn Developers. (2024-d). *User guide: contents — scikit-learn 1.4.1 documentation*. Retrieved March 24, 2024, from https://scikit-learn.org/stable/user_guide.html

SciPy v1.12.0 Manual. (2024-a). *scipy.stats.pearsonr — SciPy v1.12.0 Manual*. Retrieved March 24, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

SciPy v1.12.0 Manual. (2024-b). *scipy.stats.pointbiserialr — SciPy v1.12.0 Manual*. Retrieved March 24, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pointbiserialr.html

SciPy v1.12.0 Manual. (2024-c). *scipy.stats.spearmanr — SciPy v1.12.0 Manual*. Retrieved March 24, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Seo, Y. S., Bae, D. H. (2013). On the value of outlier elimination on software effort estimation research. *Empirical Software Engineering*, *18*(4), 659–698. https://doi.org/10.1007/S10664-012-9207-Y

Shakeel, S. et al. (2021). *COVID-19 prediction models: a systematic literature review*. https://doi.org/10.24171/j.phrp.2021.0100

Sharma, A. (2024, February 8). *How Does DBSCAN Clustering Work? Understanding the Basics*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

Sharma, S., Yadav, R. (2023). Chat GPT – A Technological Remedy or Challenge for Education System. *Global Journal of Enterprise Information System*, *14*(4), 46–51. https://doi.org/10.18311/gjeis/2022

Sousa, J. (2022). *Aprendizagem Automática UC de Otimização e Aprendizagem Automática 1º ano do curso de Mestrado em Engenharia Eletrotécnica ESTG / IPLEIRIA*.

Stewart, G., Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences (Switzerland)*, *12*(5). https://doi.org/10.3390/app12052405

Twala, B., Cartwright, M. (2010). Ensemble missing data techniques for software effort prediction. *Intelligent Data Analysis*, *14*(3), 299–331. https://doi.org/10.3233/IDA-2010-0423

XGBoost Developers. (2022). *Python API Reference — xgboost 2.1.0-dev documentation*. https://xgboost.readthedocs.io/en/latest/python/python_api.html

Valero-Carreras, D., Alcaraz, J., Landete, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix. *Computers and Operations Research*, *152*. https://doi.org/10.1016/j.cor.2022.106131

Varone, M., Mayer, D., Melegari, A. (2020). *What is Machine Learning? A definition*. Expert System. https://www.expert.ai/blog/machine-learning-definition/#:~:text=A%20definition,-Expert%20System%20Team&text=Machine%20learning%20is%20an%20application,use%20it%20learn%20for%20themselves

Vivas, E., Leite, P., Pinto, S. (2021). *Utilização de Ferramentas de Inteligência Artificial no Apoio à Deteção de Fugas em Sistemas de Abastecimento de água*. Retrieved December 3, 2022, https://www.aprh.pt/congressoagua2021/docs/15ca_91.pdf

Vladimiro González Zelaya, C. (2019). *Towards Explaining the Effects of Data Preprocessing on Machine Learning*. https://doi.org/10.1109/ICDE.2019.00245

Vulimiri, P., Stebner, A. (2020). *Machine Learning for Materials Developments in Metals Additive Manufacturing*. https://www.researchgate.net/publication/341310767

Wittel. (2022). *O que é machine learning e quais são as suas aplicações no mercado?* Retrieved January 3, 2024, from https://blog.wittel.com/o-que-e-machine-learning/

Yıldırım, S. (2021). *15 Must-Know Machine Learning Algorithms | by Soner Yıldırım | Towards Data Science*. Retrieved December 3, 2023, from https://towardsdatascience.com/15-must-know-machine-learning-algorithms-44faf6bc758e

Zhang, S., Zhang, C., Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180

# Anexos

# Anexo I – Pormenor da Pré-Classificação de Eventos no caso da Conduta de Abastecimento Gravítico, dados de relativos ao período fevereiro-novembro de 2022

No presente Anexo apresenta-se a representação gráfica da pré-classificação de cada um dos eventos de rotura ou alteração de regime para o Caso I no período compreendido entre fevereiro e novembro de 2022, cujo panorama geral foi apresentado na secção 5.2 (Figura 29).

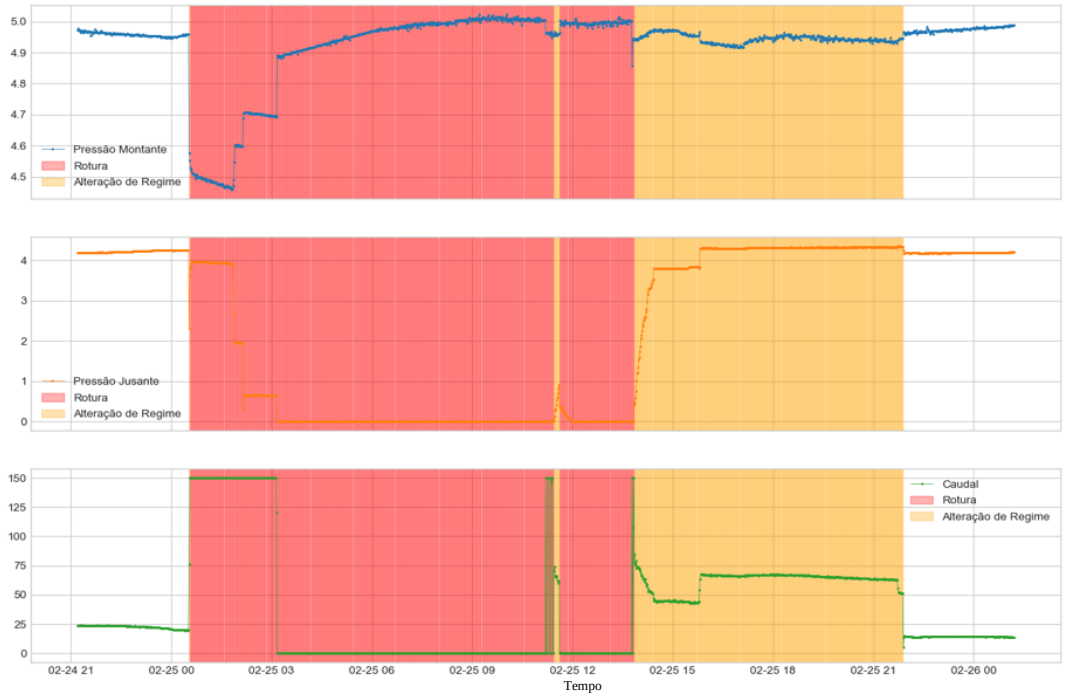


**Figura 75 - Rotura seguida de carregamento da conduta em 25-02-2022 (pressão em bar e caudal em $m^3/h$)**

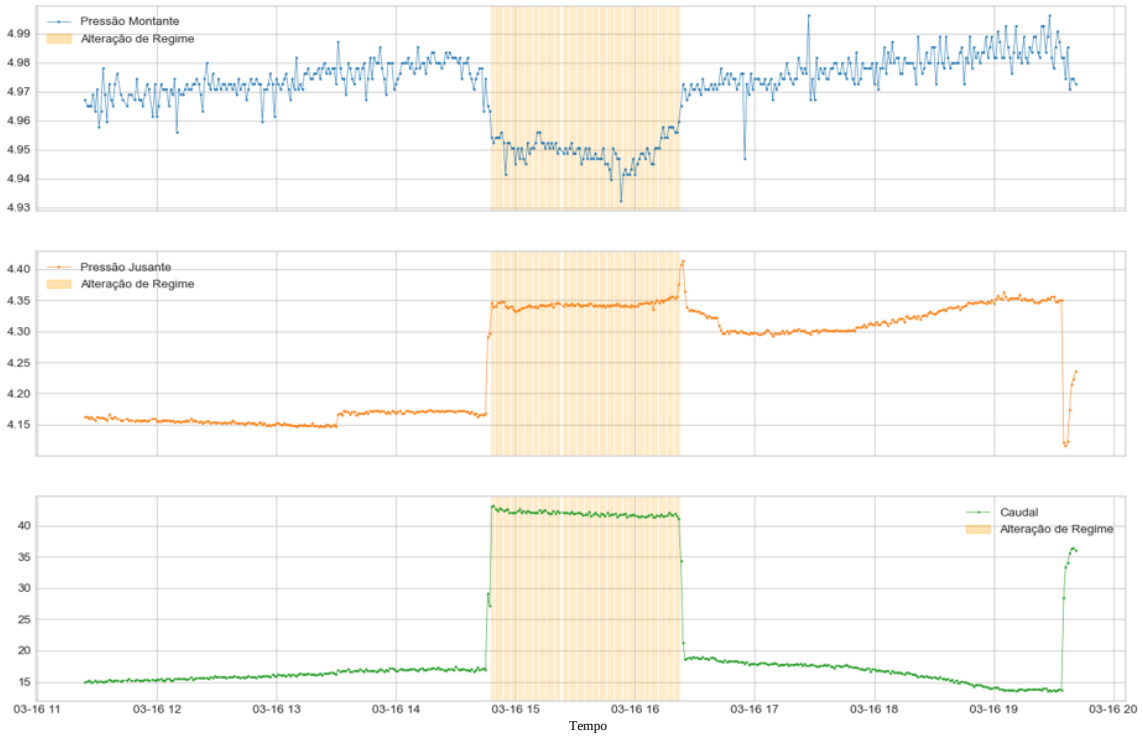


**Figura 76 - Alteração de regime em 16-03-2022 (pressão em bar e caudal em $m^3/h$)**

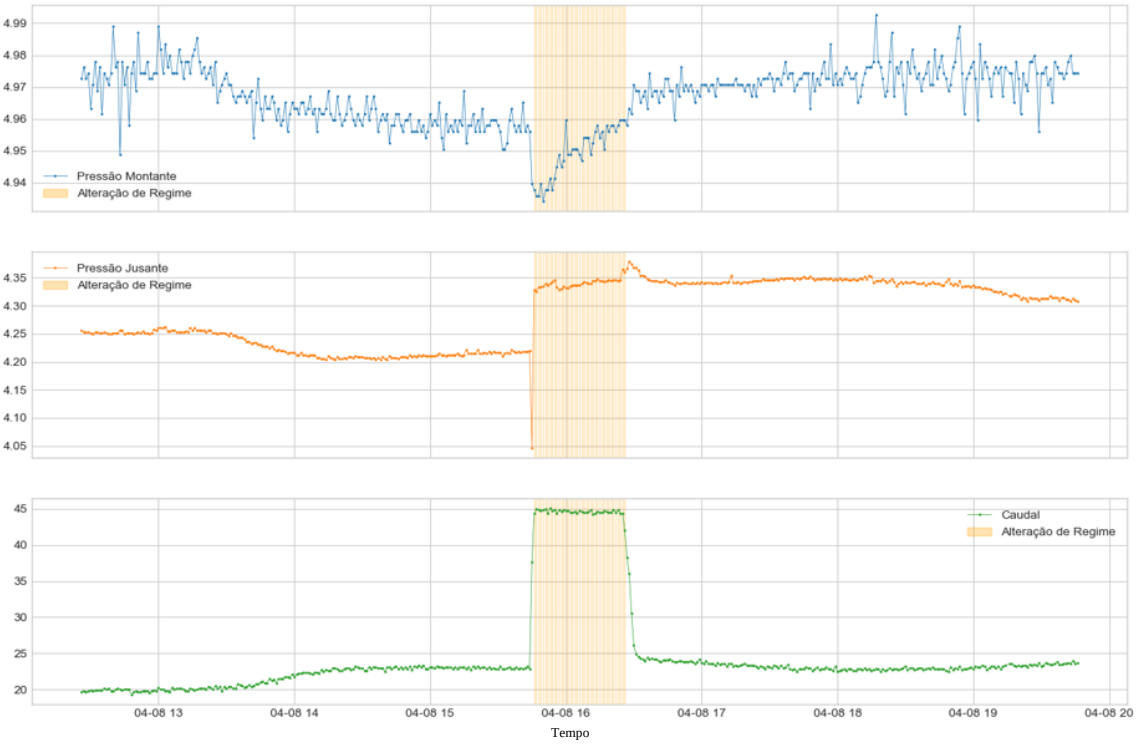


**Figura 77 - Alteração de regime em 08-04-2022 (pressão em bar e caudal em $m^3/h$)**

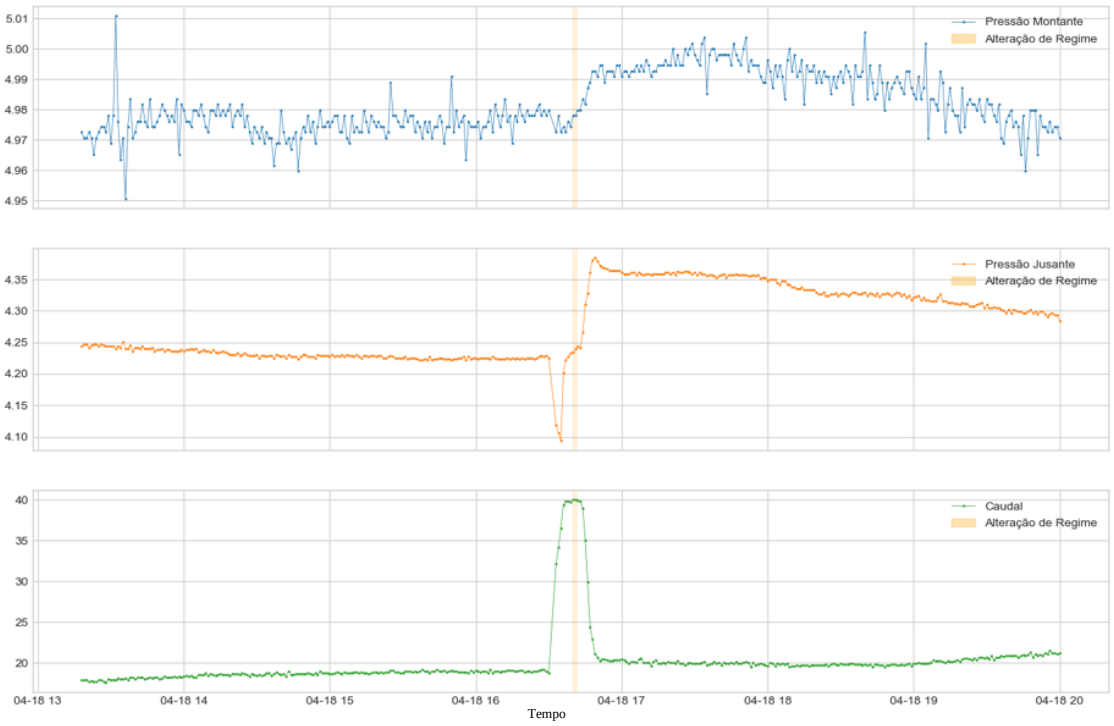


**Figura 78 - Alteração de regime em 18-04-2022 (pressão em bar e caudal em $m^3/h$)**

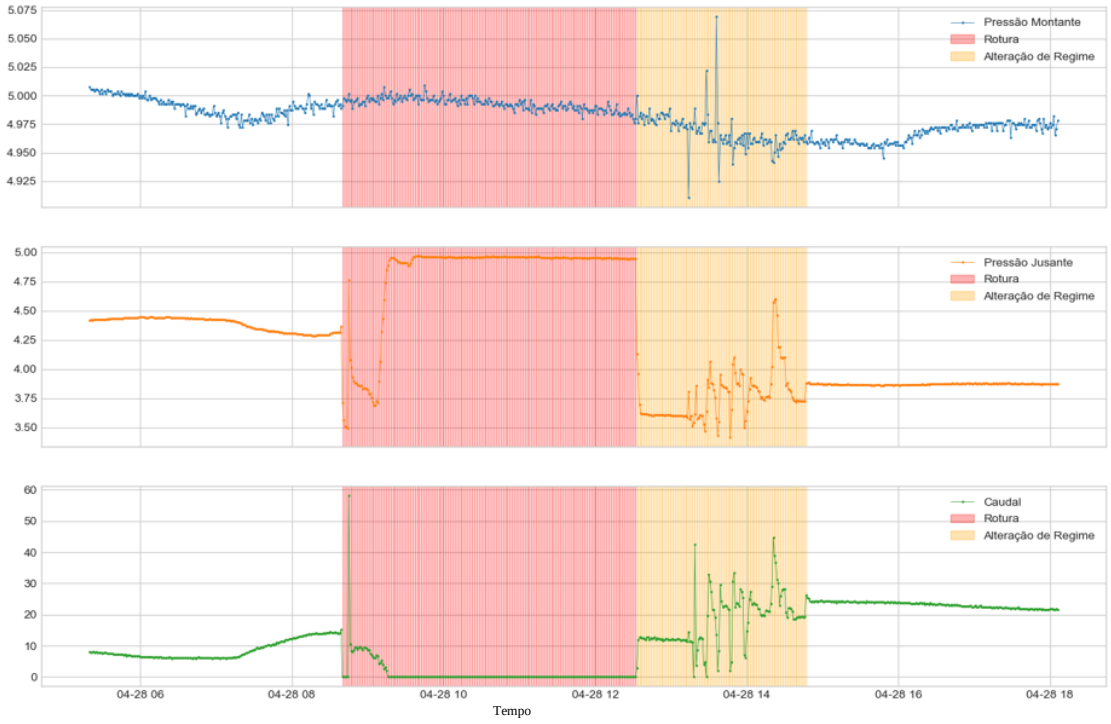


**Figura 79 - Rotura seguida de carregamento da conduta em 28-04-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 80 - Rotura seguida de carregamento da conduta em 09-05-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 81 - Alteração de regime em 19-05-2022 (pressão em bar e caudal em $m^3/h$)**

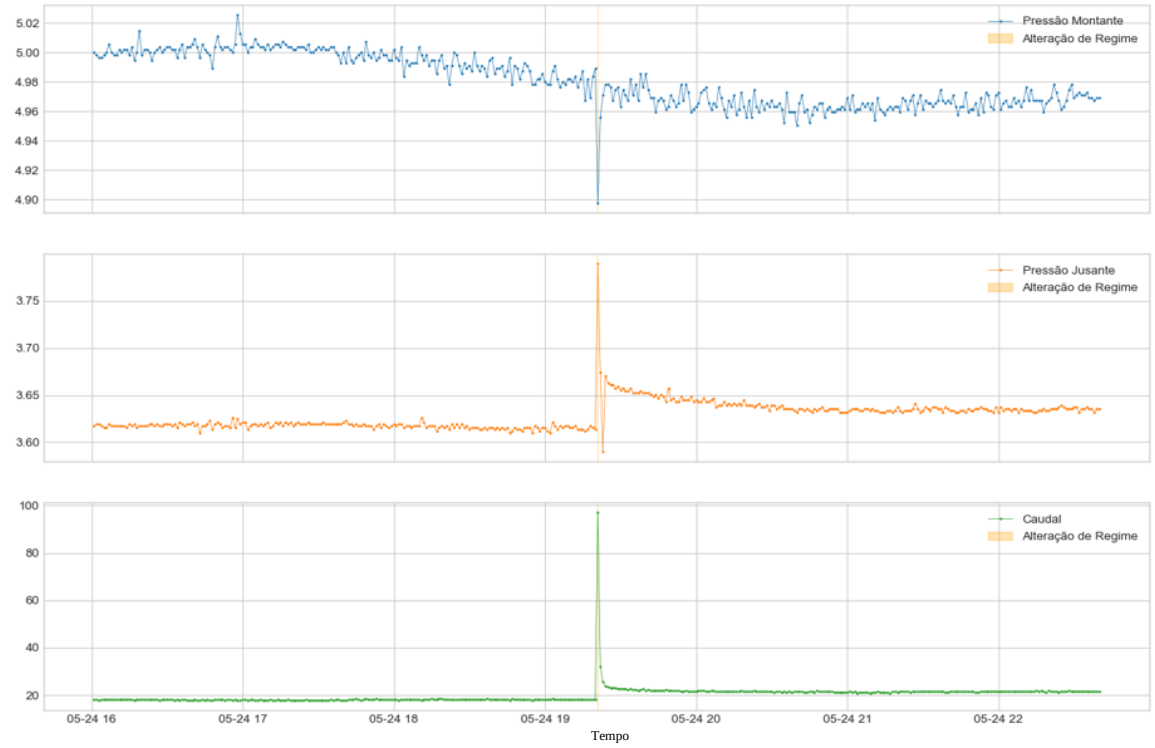


**Figura 82 - Alteração de regime em 24-05-2022 (pressão em bar e caudal em $m^3/h$)**

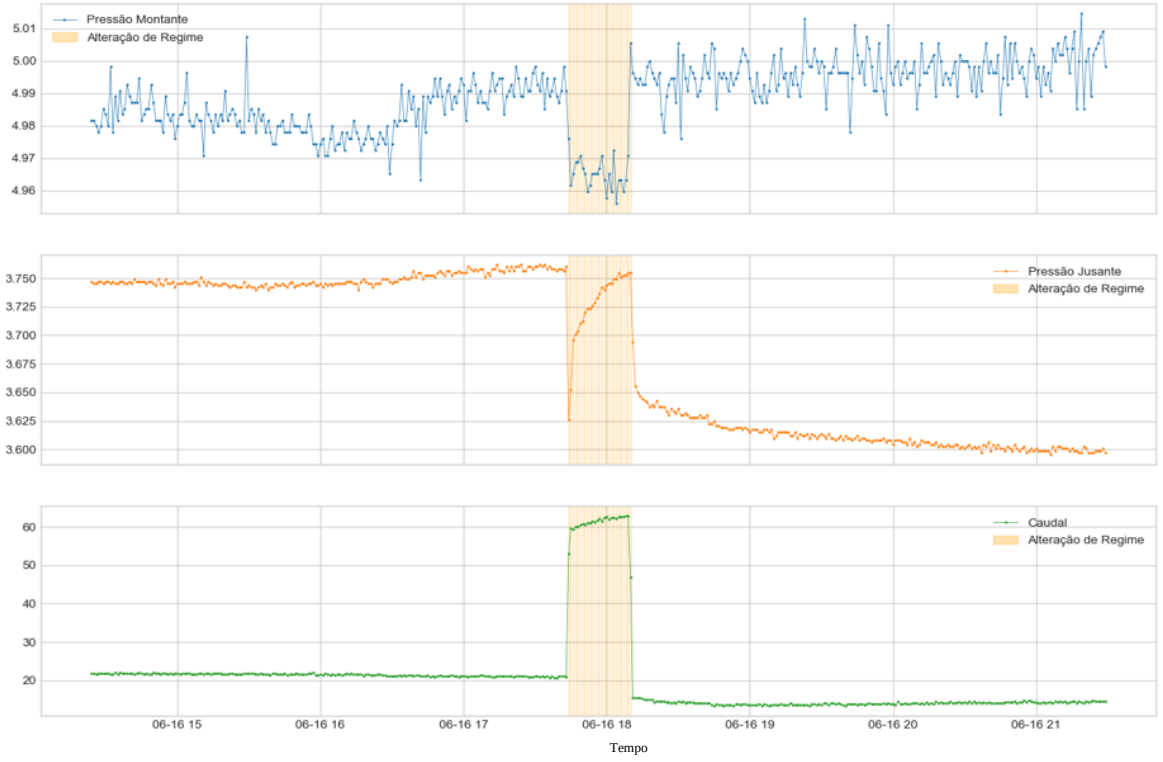


**Figura 83 - Alteração de regime em 16-06-2022 (pressão em bar e caudal em $m^3/h$)**

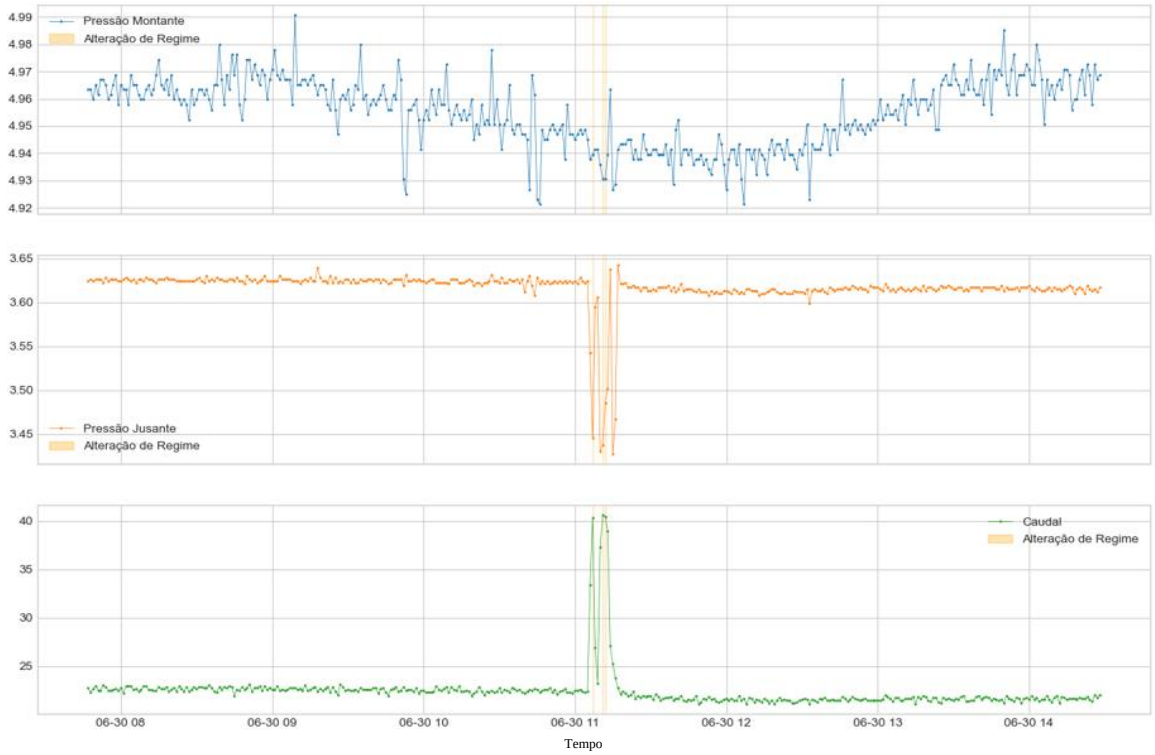


**Figura 84 - Alteração de regime em 30-06-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 85 - Alteração de regime em 01-07-2022 (pressão em bar e caudal em $m^3/h$)**

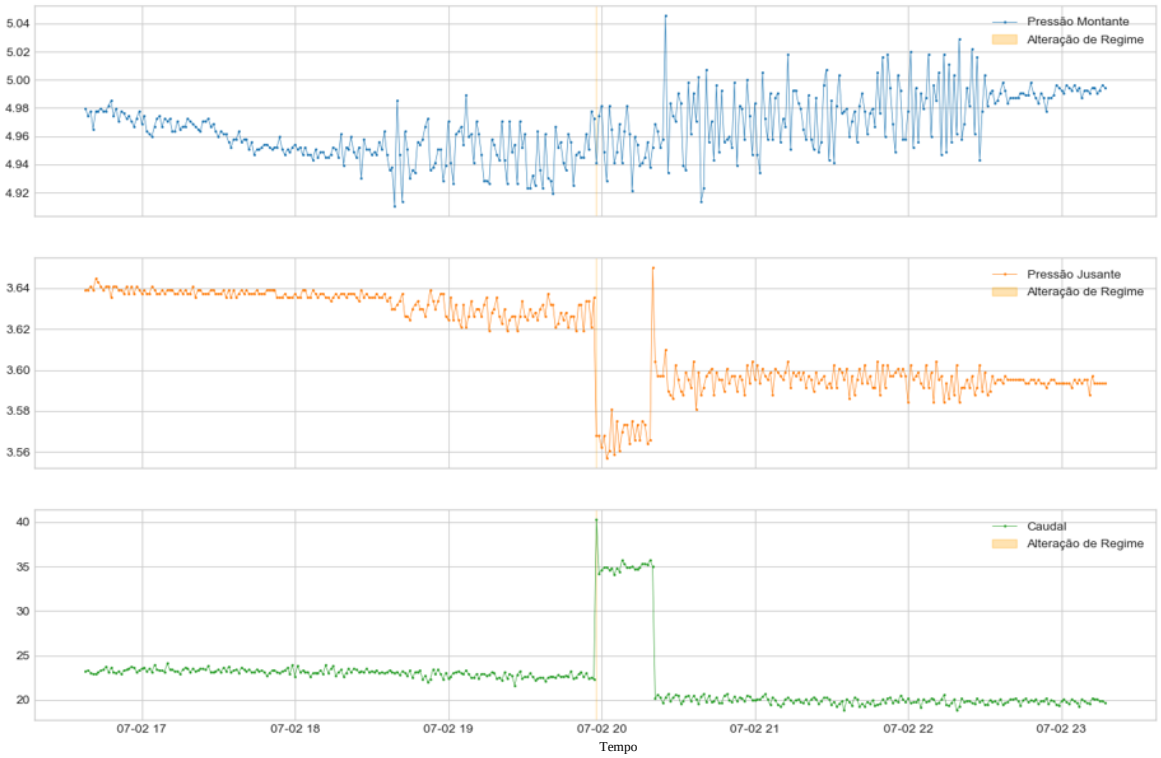


**Figura 86 - Alteração de regime em 02-07-2022 (pressão em bar e caudal em $m^3/h$)**

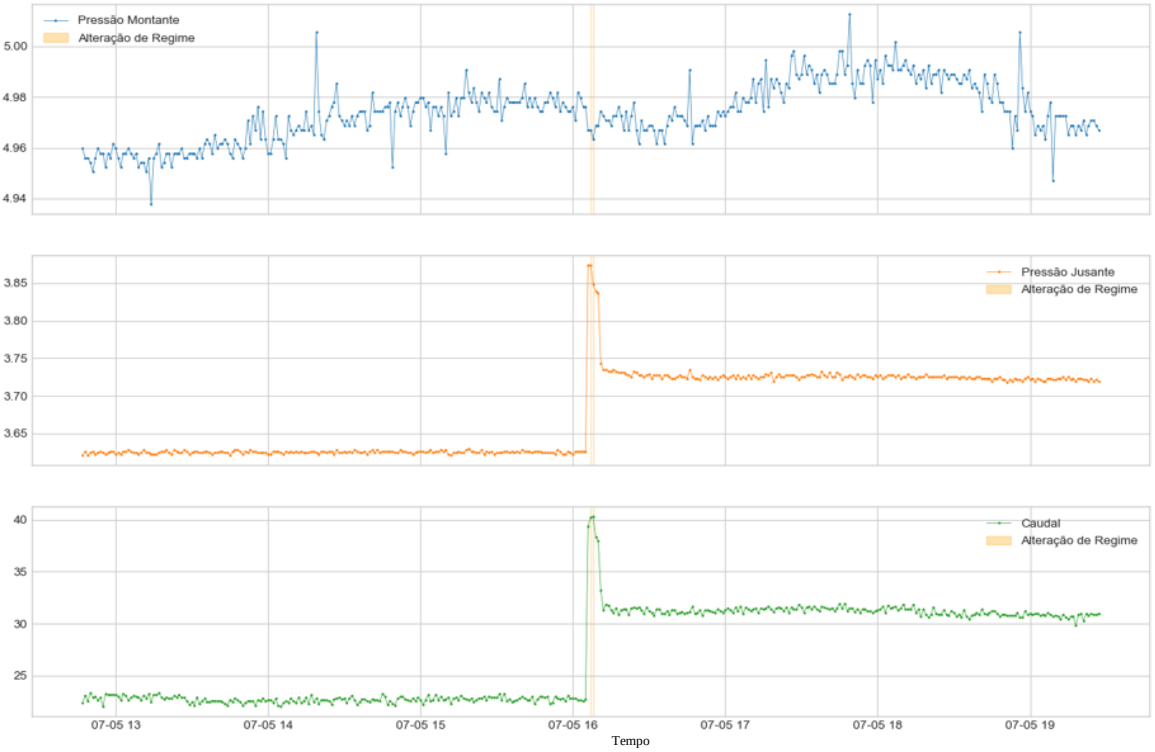


**Figura 87 - Alteração de regime em 05-07-2022 (pressão em bar e caudal em $m^3/h$)**

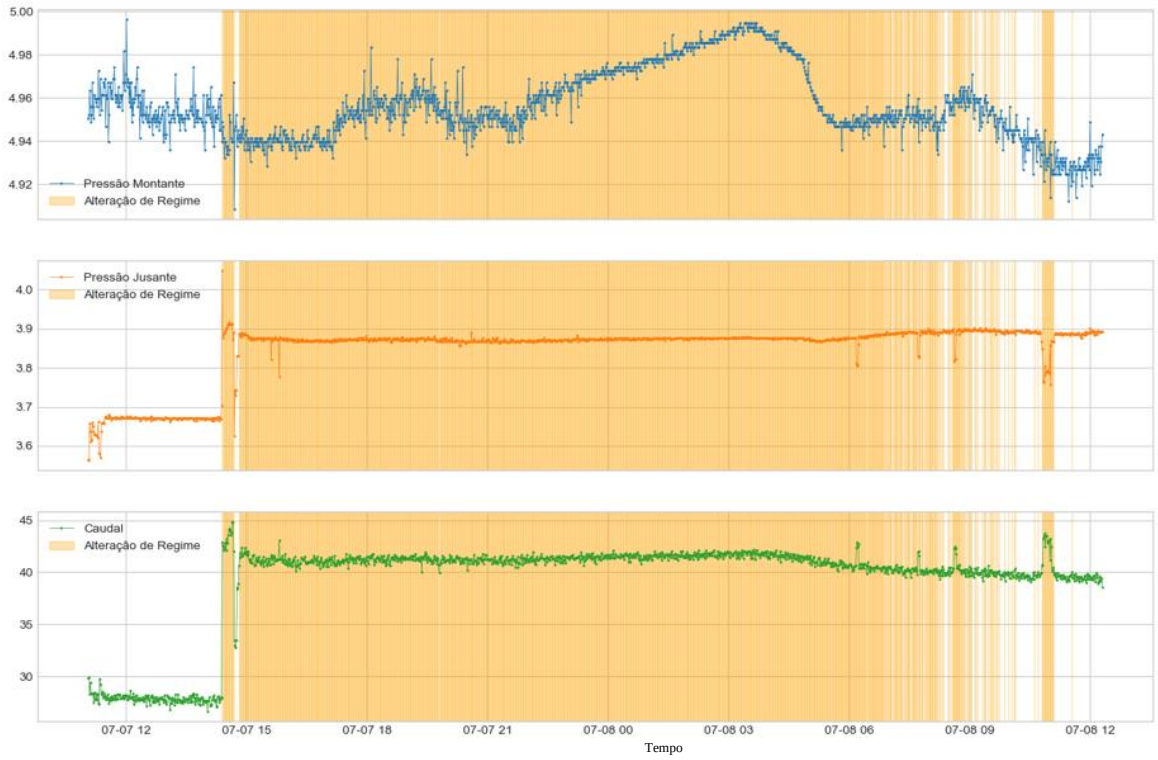


**Figura 88 - Alteração de regime entre 07-07-2022 e 08-07-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 89 - Alteração de regime em 08-07-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 90 - Alterações de regime em 21-07-2022 (pressão em bar e caudal em $m^3/h$)**

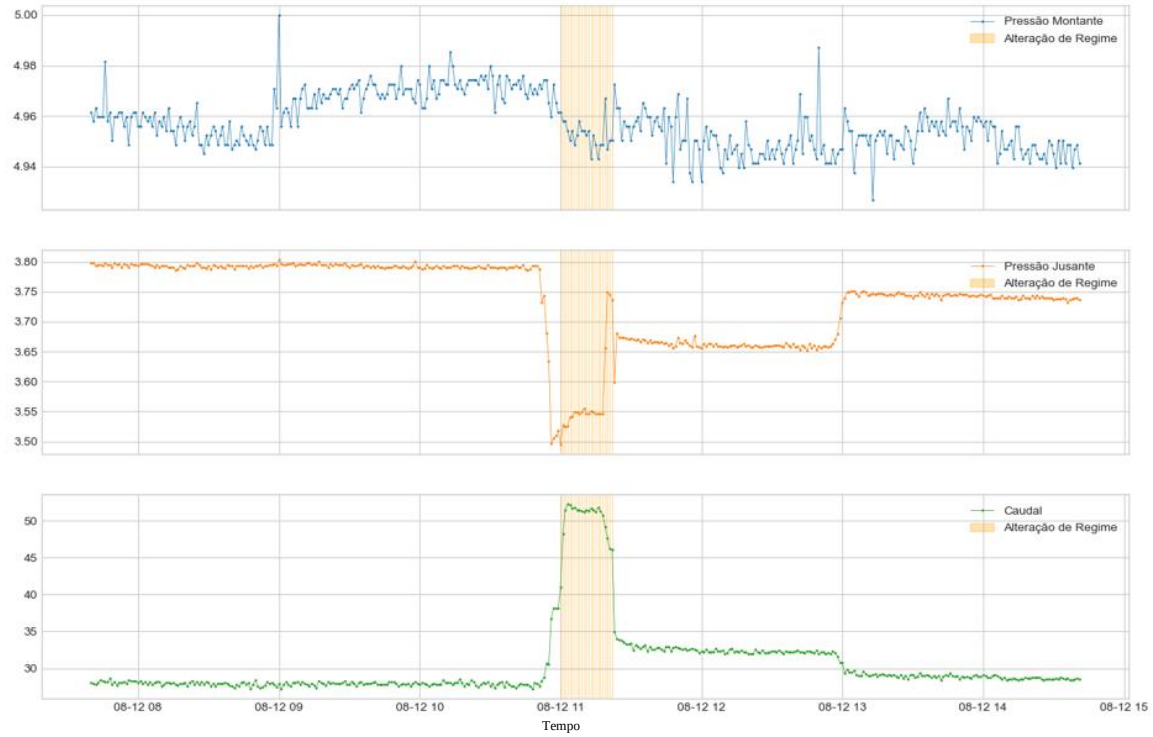


**Figura 91 - Alterações de regime em 12-08-2022 (pressão em bar e caudal em $m^3/h$)**

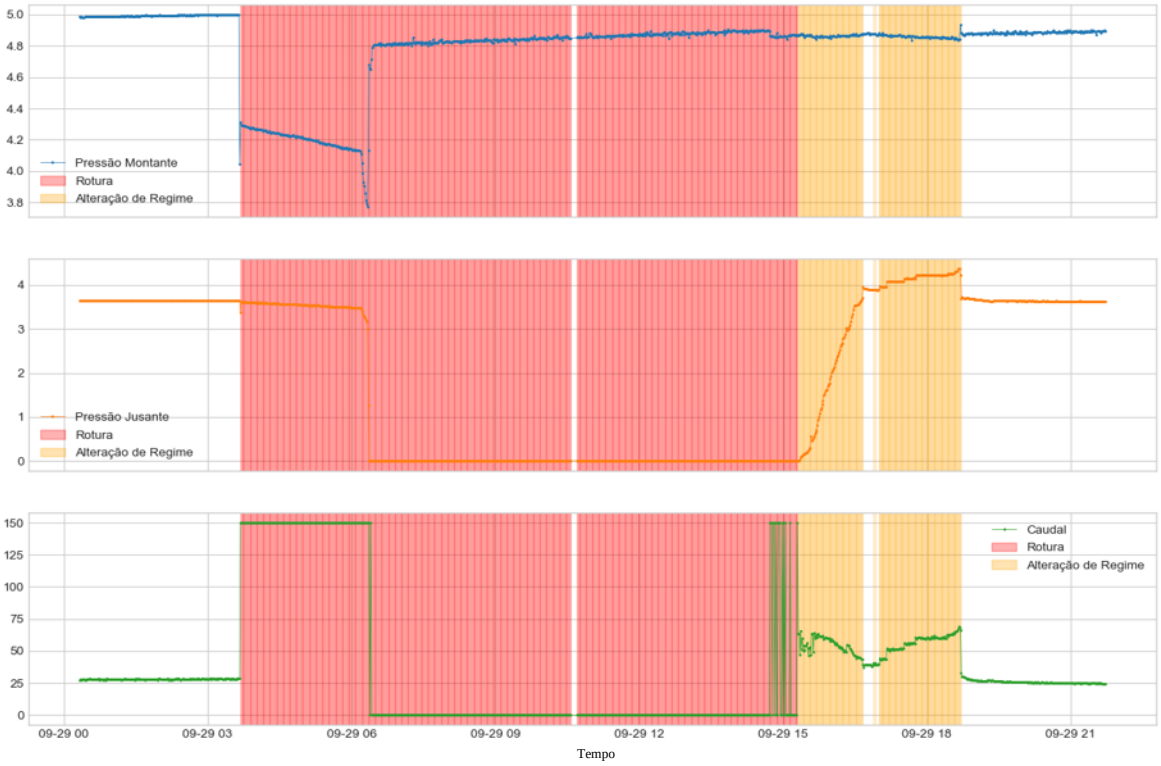


**Figura 92 - Rotura seguida de carregamento da conduta em 29-09-2022 (pressão em bar e caudal em $m^3/h$)**

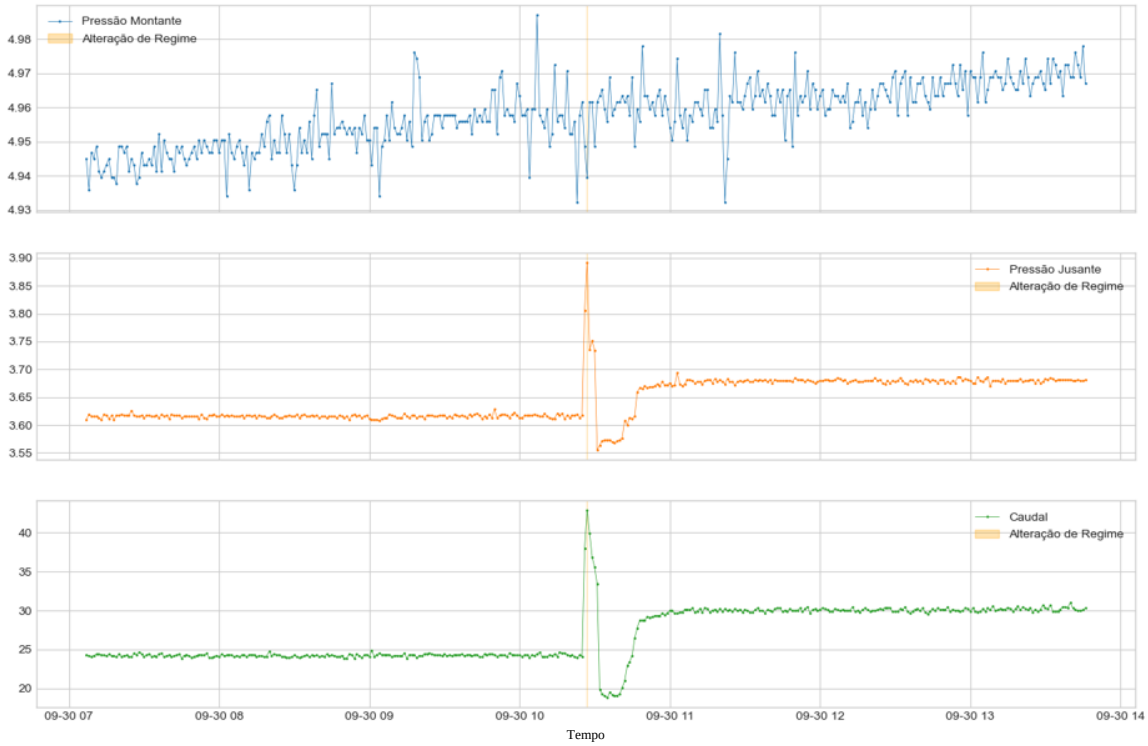


**Figura 93 - Alterações de regime em 30-09-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 94 - Rotura seguida de carregamento da conduta em 30-09-2022 (pressão em bar e caudal em $m^3/h$)**


**Figura 95 - Rotura em 19-11-2022 seguida de carregamento de conduta, nova rotura seguida e carregamento da conduta (pressão em bar e caudal em $m^3/h$)**

# Anexo II – Pormenor da Pré-Classificação de Eventos no caso da Conduta de Abastecimento Elevatório, dados de relativos ao ano de 2019

No presente Anexo apresenta-se a representação gráfica da pré-classificação de cada um dos eventos de rotura ou alteração de regime para o Caso II ocorridos em 2019, apresentado no Capítulo 6.

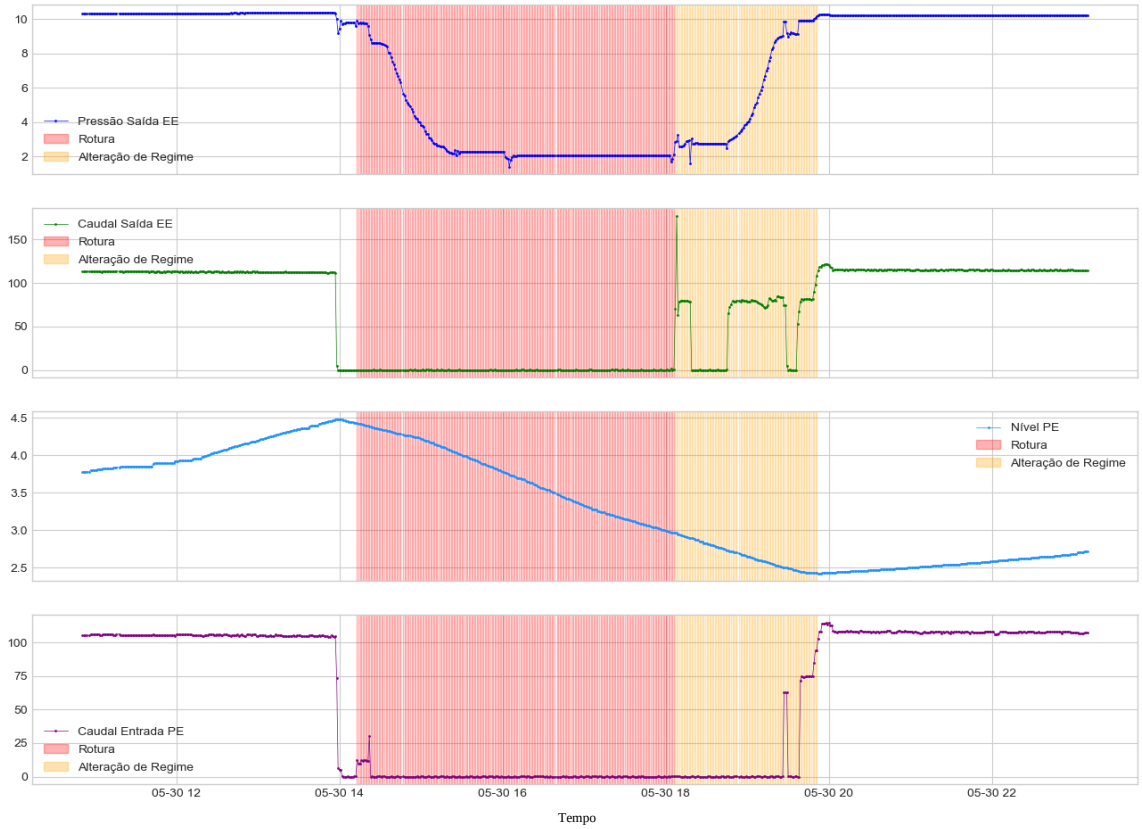


**Figura 96 - Rotura em 30-05-2019 seguida de carregamento de conduta (pressão em bar, nível em metros e caudal em $m^3/h$)**

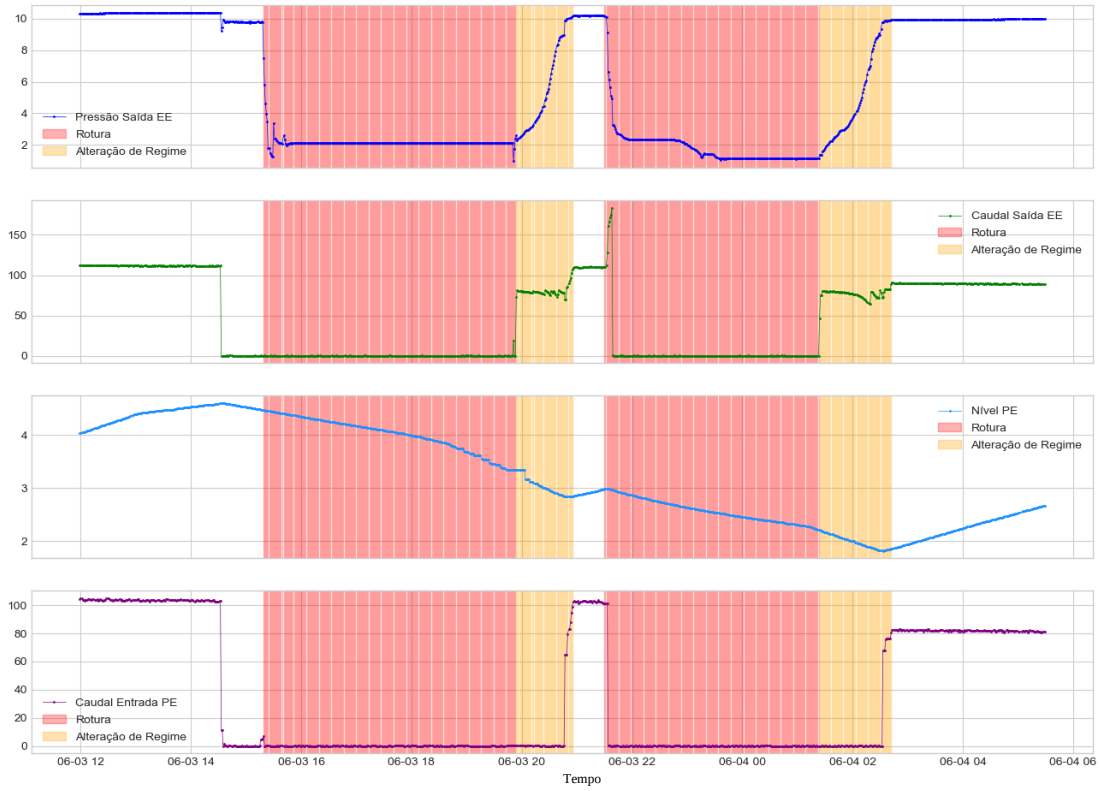


**Figura 97 - Rotura em 03-06-2019 seguida de carregamento de conduta, nova rotura e novo carregamento (pressão em bar, nível em metros e caudal em $m^3/h$)**

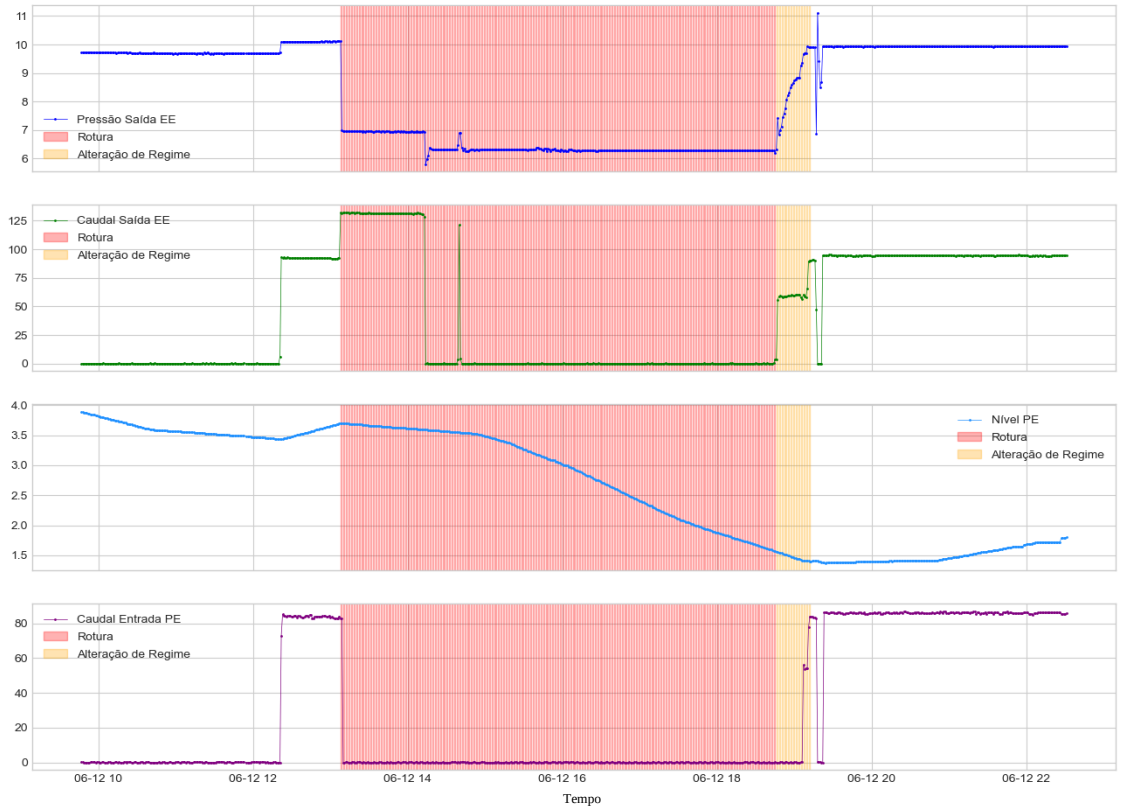


**Figura 98 - Rotura em 12-06-2019 seguida de carregamento de conduta (pressão em bar, nível em metros e caudal em $m^3/h$)**


**Figura 99 - Rotura em 13-07-2019 seguida de carregamento de conduta (pressão em bar, nível em metros e caudal em $m^3/h$)**

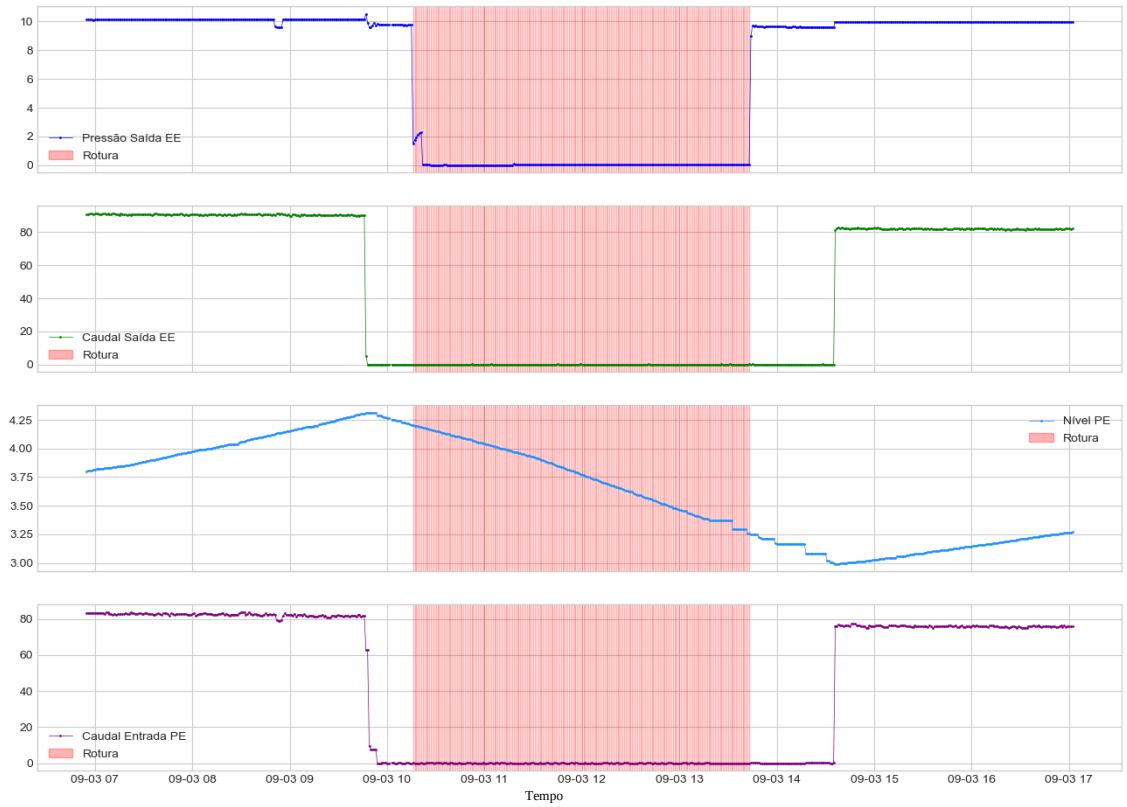


**Figura 100 - Manutenção hidráulica na EE em 03-09-2019 (pressão em bar, nível em metros e caudal em $m^3/h$)**

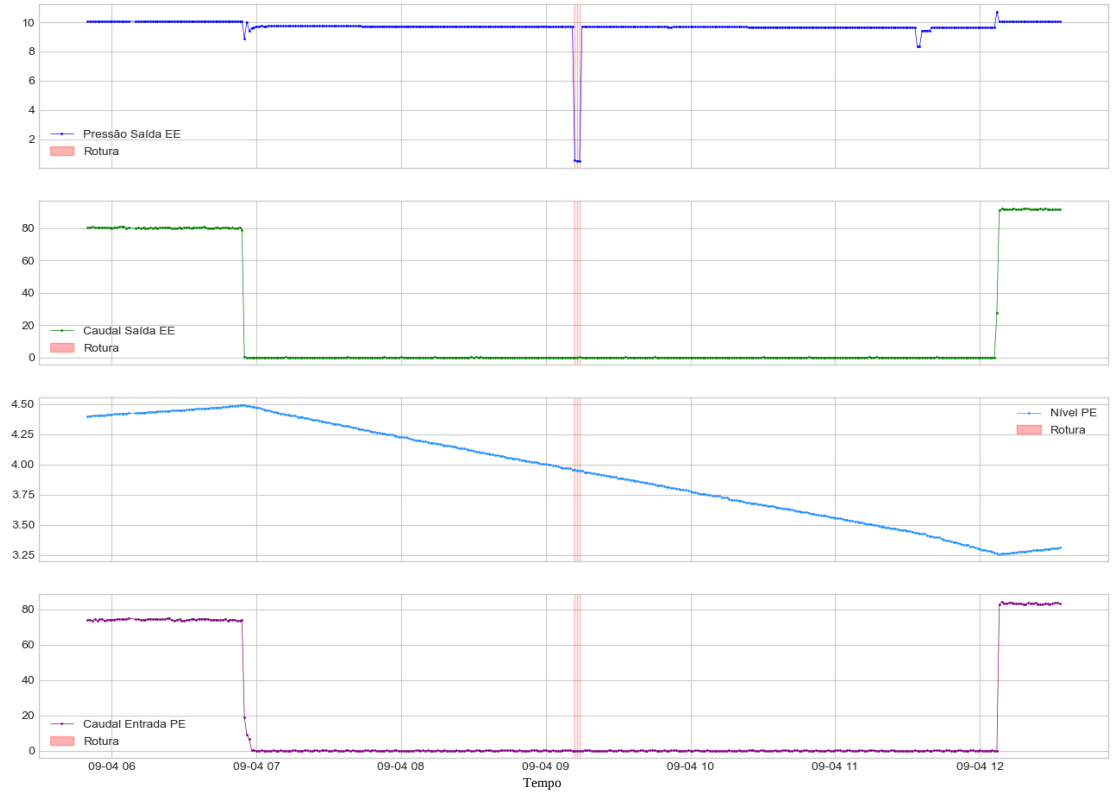


**Figura 101 - Evento de perda de pressão em 04-09-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

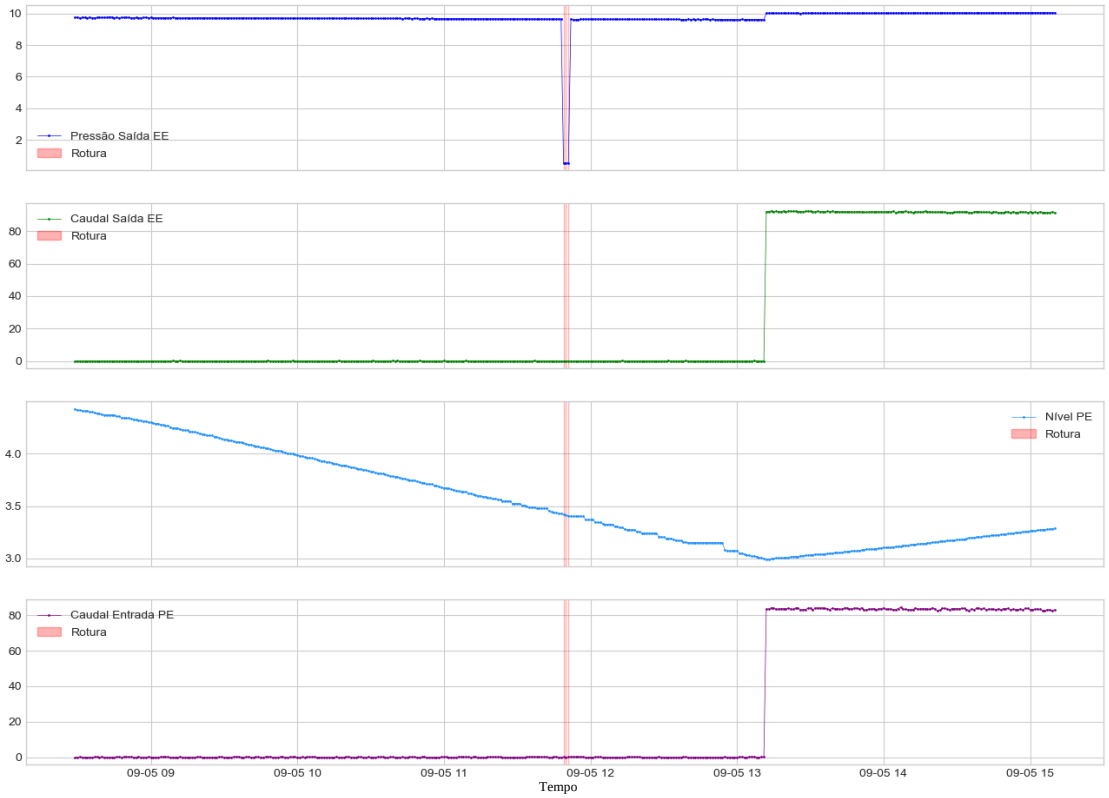


**Figura 102 - Evento de perda de pressão em 05-09-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

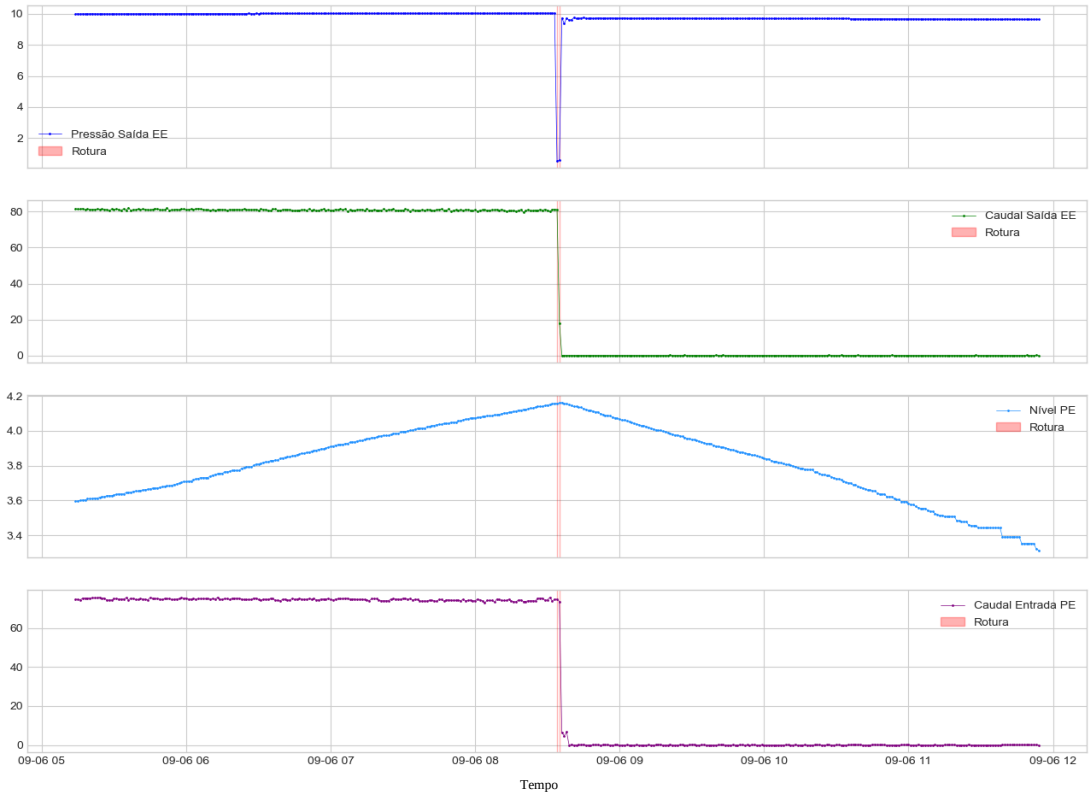


**Figura 103 - Evento de perda de pressão em 06-09-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

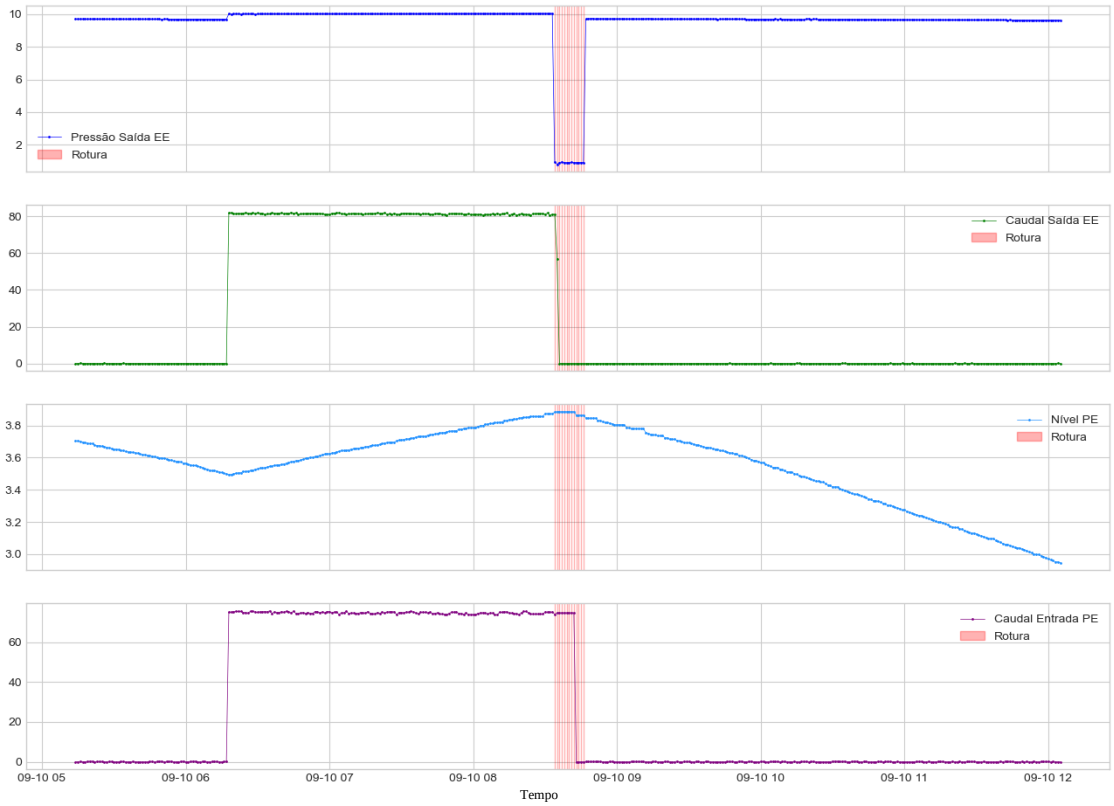


**Figura 104 - Evento de perda de pressão em 10-09-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

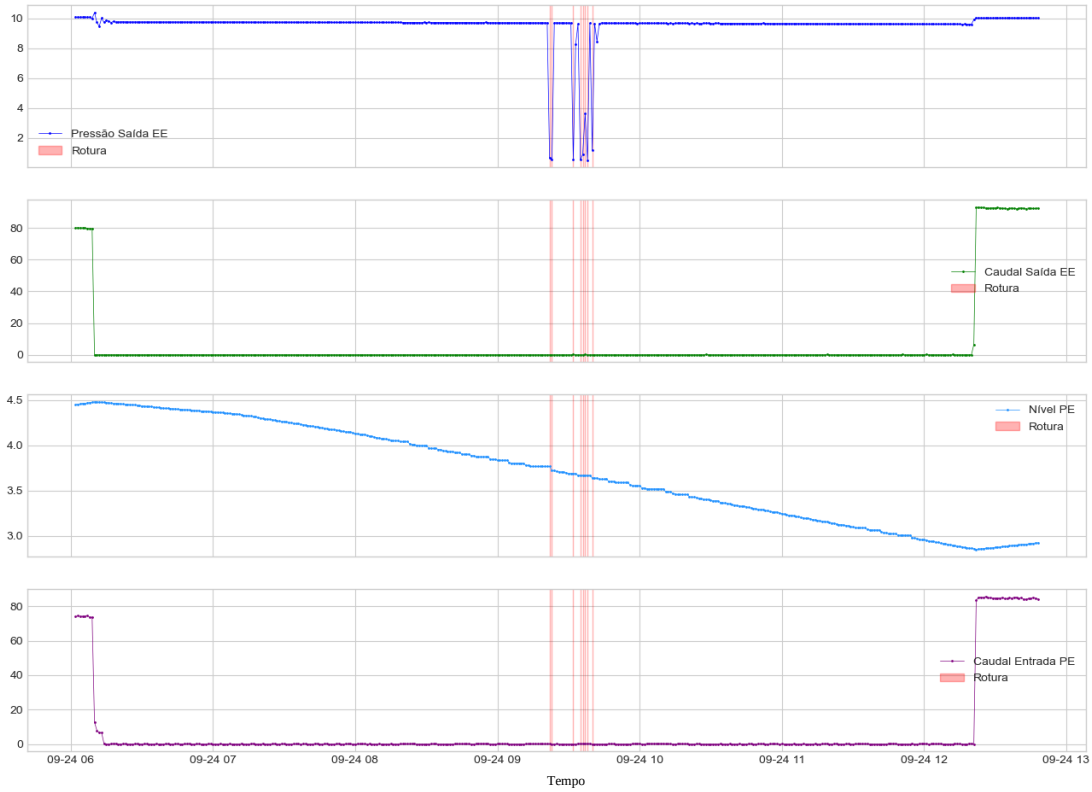


**Figura 105 - Evento de perda de pressão em 24-09-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

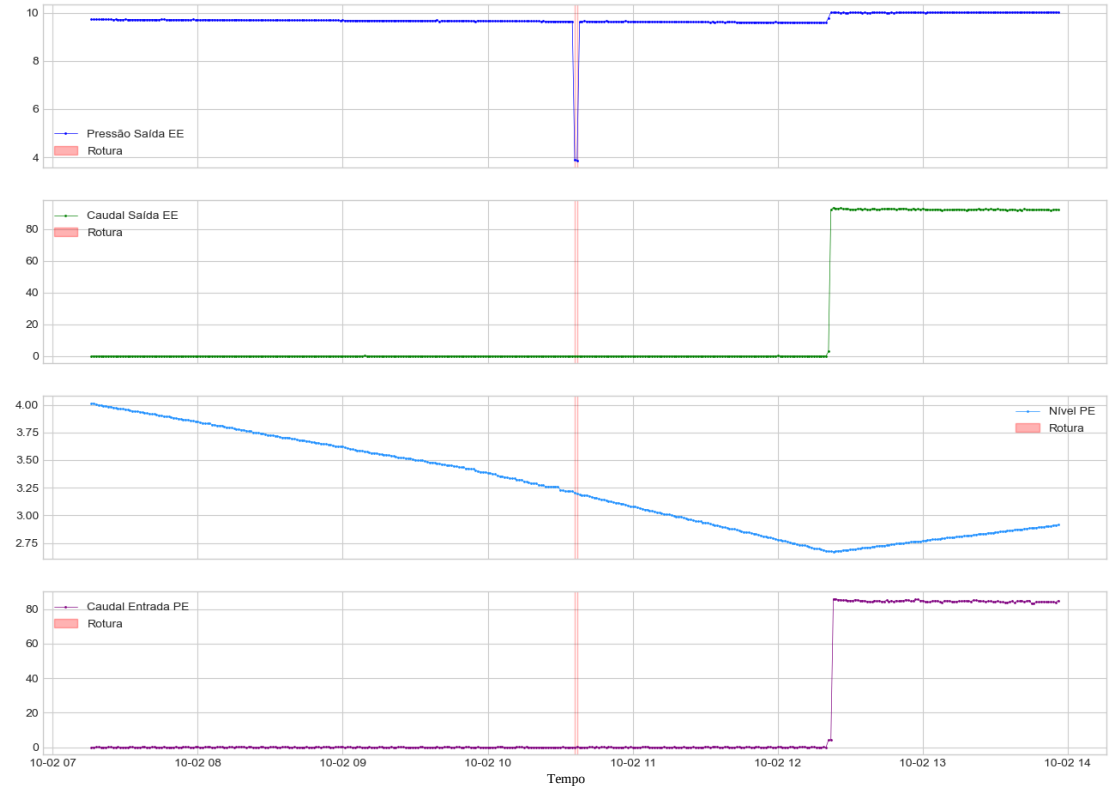


**Figura 106 - Evento de perda de pressão em 02-10-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

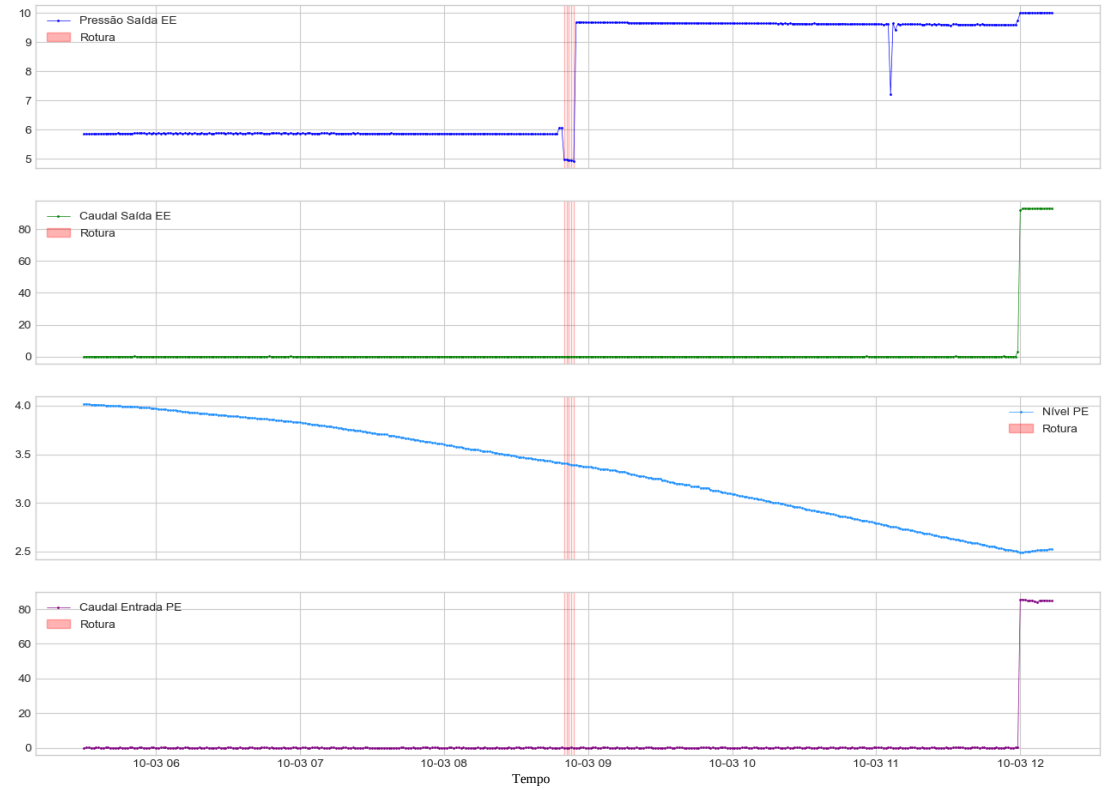


**Figura 107 - Evento de perda de pressão em 03-10-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

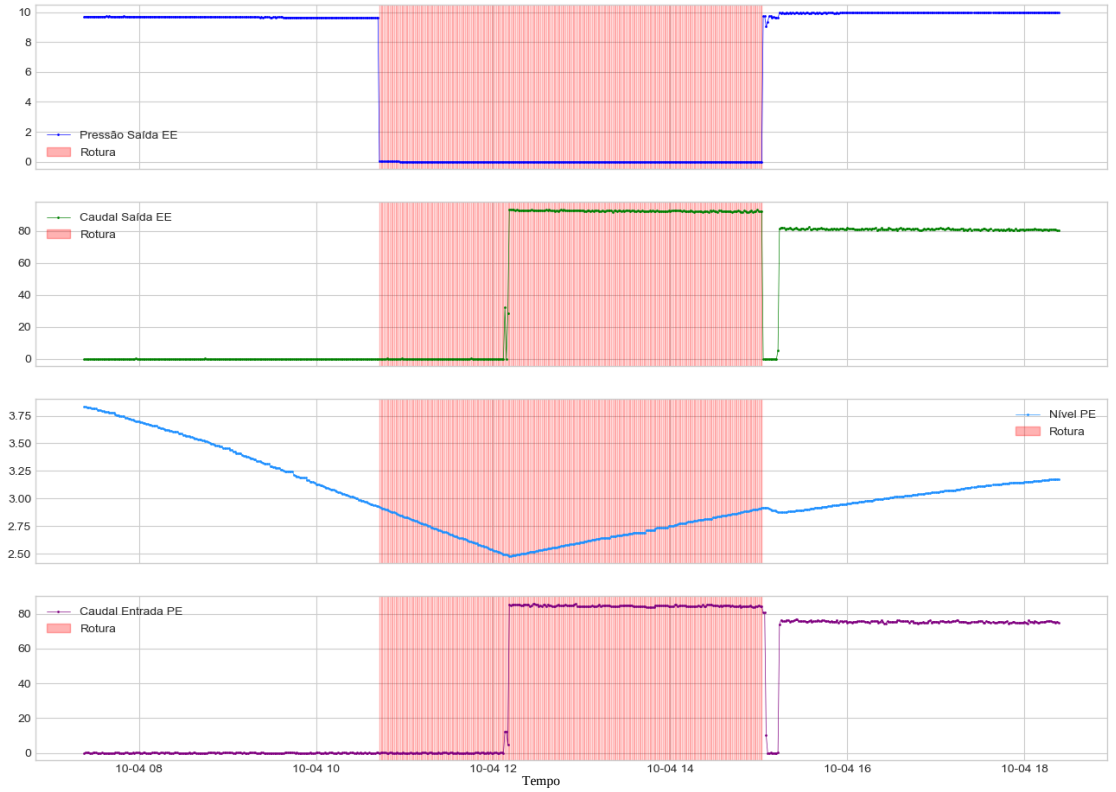


**Figura 108 - Evento de perda de pressão em 04-10-2019 compatível com condições de rotura (pressão em bar, nível em metros e caudal em $m^3/h$)**

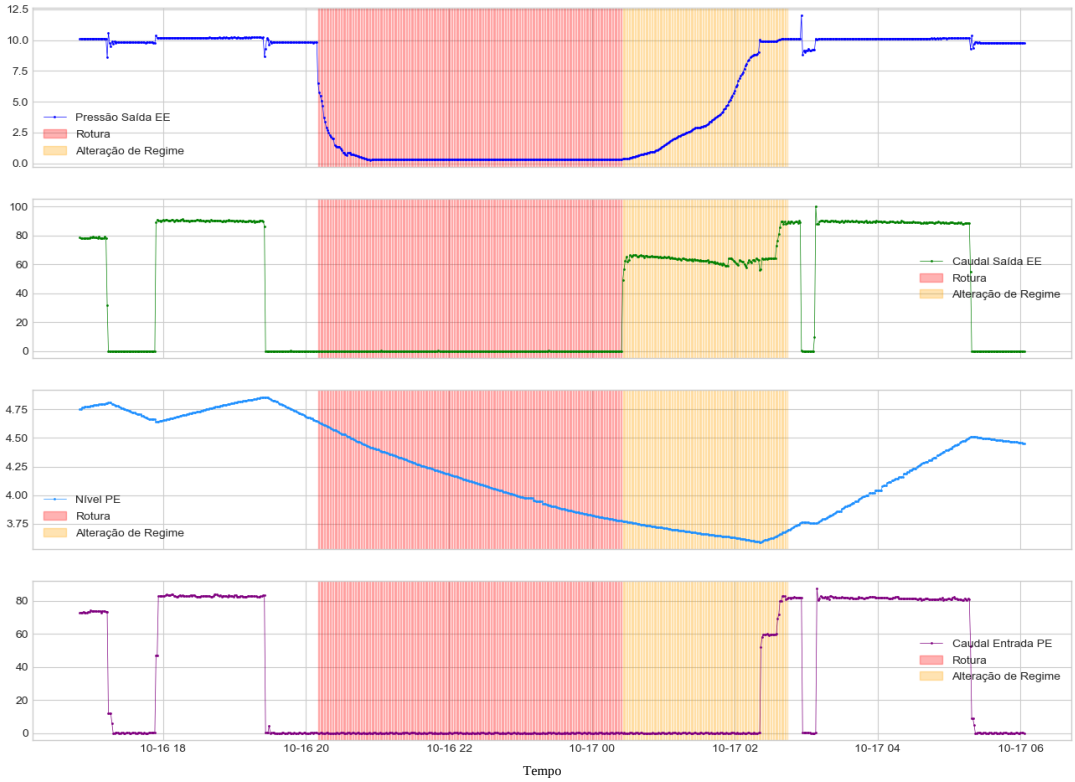


**Figura 109 - Rotura em 16-10-2019 seguida de carregamento de conduta (pressão em bar, nível em metros e caudal em $m^3/h$)**

# Anexo III – Apresentação do Sistema de Abastecimento da EPAL

No presente Anexo apresenta-se o sistema de abastecimento da EPAL e Oeste, com referência às principais Captações e Estações de Tratamento de Água da EPAL e aos principais adutores, que permitem que a água chegue aos vários sistemas municipais (rede de abastecimento em "baixa") e à cidade de Lisboa. Esta informação complementa a secção 3.1.

## O Subsistema de Castelo do Bode

A água captada na Barragem de Castelo do Bode através de uma Torre de Toma de Água aí instalada é bombeada através de duas estações elevatórias, Estação Elevatória 1 de Castelo do Bode (EECB1) e Estação Elevatória 2 de Castelo do Bode (EECB2), localizadas junto à Barragem, seguindo para a Estação de Tratamento de Água da Asseiceira através de duas condutas com comprimento aproximado de 8,7 quilómetros e dimensionadas para transportar em conjunto 1 000 000 $m^3$/dia (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007). Esta instalação possui duas linhas de tratamento com capacidade de tratamento para 625 000 m3/dia (linha 1 – 500 000 m3/dia; linha 2 – 125 000 m3/dia) (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007;Anselmo et al., 2010).

Segundo (Anselmo et al., 2010), estas instalações entraram em funcionamento em 1987, contando apenas com a EECB1 e com uma linha de tratamento na ETA da Asseiceira, que garantia uma capacidade de produção de 375 000 m3/dia. Face ao aumento do consumo de água foi necessária uma ampliação da linha nos anos 90 para garantir um aumento da capacidade de produção para os 500 000 m3/dia, que se mantém no presente. Mais tarde foi criada a EECB2 e a segunda linha de tratamento na ETA da Asseiceira, tendo sido também criada uma conduta que liga a nova estação elevatória à ETA. Esta alteração garantiu um aumento de produção de 125 000 m3/dia, aumentando assim a capacidade de produção para os atuais 625 000 m3/dia.

Segundo (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007, 2011), a ETA da Asseiceira abastece o Subsistema do Médio Tejo, que se divide em Médio

Tejo Norte e Médio Tejo Sul, e também a conduta adutora (Adutor) de Castelo do Bode, que tem um comprimento total de aproximadamente 80 quilómetros, compreendido entre a ETA e a Estação Elevatória de Vila Franca de Xira. Nesse trajeto encontra-se o Reservatório de Alcanhões, que induz uma perda de carga no sistema e tem capacidade de armazenamento para 18 000 $m^3$ de água. Existe também em Alcanhões a ligação a um Reservatório que alimenta uma grande parte do Sistema Oeste, que será abordado mais à frente. O Adutor segue então em direção a Várzea das Chaminés, de onde segue até à Estação Elevatória de Vila Franca de Xira. O Adutor de Castelo de Bode pode receber água vinda da ETA de Vale da Pedra e das captações de Valada I, II e III (que neste momento se encontram fora de serviço e por isso não são abordadas neste trabalho).

## O Subsistema Tejo

Na Captação de água de Valada-Tejo a água proveniente do Rio Tejo é captada de duas formas diferentes, de acordo com o nível do rio Tejo. Sempre que o nível da água do rio está acima do nível de soleira da captação gravítica, a água entra nas câmaras de aspiração da Estação Elevatória II. Quando o nível do rio não permite a entrada na captação gravítica, a Estação Elevatória I faz a captação de água através de um sistema de mastros oscilantes com 4 bombas submersíveis, colocando a água nas caleiras de captação que levam às câmaras de aspiração da Estação Elevatória II. A água é então elevada através da Estação Elevatória II que possui 7 grupos eletrobomba para a ETA de Vale da Pedra, através de uma conduta com cerca de 8 quilómetros (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007; Anselmo et al., 2010).

A ETA pode disponibilizar até 240 000 $m^3$/dia de água a partir das duas linhas de tratamento aí existentes (120 000 m3/dia por linha) (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007).

Segundo (Anselmo et al., 2010), as instalações atrás referidas entraram em funcionamento na década de 60 do século passado, tendo a ETA apenas uma linha de tratamento com capacidade de tratamento diário de 80 000 $m^3$, posteriormente aumentada para 120 000 $m^3$. Em 1976 foram realizadas obras de duplicação das instalações para conseguir o aumento da capacidade de produção para 220 000 m3/dia, que embora não corresponda ao dobro da capacidade anterior representa um aumento na ordem dos 80%.

A água tratada é bombeada através da Estação Elevatória da ETA até ao início do Adutor do Tejo (aproximadamente 9,9 quilómetros), na Várzea das Chaminés. O Adutor do Tejo pode, em caso de necessidade, receber água do Adutor de Castelo do Bode, dos furos de Alenquer e das Lezírias e abastece instalações nos concelhos de Azambuja, Vila Franca de Xira e Loures, podendo ainda colocar água no aqueduto do Alviela em Alhandra e na Verdelha (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007, 2011).

Para que a ETA de Vale da Pedra possa funcionar na sua capacidade máxima de produção sem recorrer aos restantes adutores, a velocidade da água que passa pelo adutor Tejo sofre um aumento ao passar pela Estação de Sobrelevação da Azambuja, garantindo assim um aumento do caudal que passa no adutor (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007).

O Adutor do Tejo termina em Lisboa, no Reservatório dos Olivais, e tem uma extensão aproximada de 42 quilómetros (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007, 2011).

## O Subsistema do Alviela

O Adutor do Alviela entrou ao serviço em 1890 e liga a captação dos Olhos de Água (nascentes do Rio Alviela) ao Reservatório dos Barbadinhos, transportando a água graviticamente ao longo de todo o seu percurso. Este aqueduto tem 114 quilómetros de extensão e a sua função é a distribuição de água em "alta", porque nos últimos anos a captação dos Olhos de Água tem estado fora de serviço, bem como parte da conduta. O Adutor recebe agora água do Adutor de Castelo do Bode e dos furos da Ota, de Alenquer e das Lezírias. Pode ainda receber água a partir das estações elevatórias da Pimenta, Alhandra e Verdelha I. Destaca-se a sua importância no abastecimento aos Subsistemas de Torres Vedras-Mafra e de Arruda-Sobral, ambos pertencentes à Zona Oeste. Abastece ainda diversas instalações nos concelhos de Alcanena, Santarém, Azambuja, Alenquer, Vila Franca de Xira e Loures (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007).

## Captações Subterrâneas

Quanto às captações subterrâneas, a água captada é tratada através de cloragem.

Segundo (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007), no caso das principais captações subterrâneas, Alenquer e Ota entregam água aos adutores do Alviela e do Tejo e ao sistema Oeste em Alenquer IV (que alimenta o sistema Torres Vedras-Mafra), Casal Machado e no ponto AL8. Já as captações das Lezírias entregam água ao adutor das Lezírias e aos adutores do Alviela e do Tejo.

## Adutores Vila Franca-Telheiras e Circunvalação

O Adutor Vila Franca de Xira-Telheiras inicia-se na Estação Elevatória 1 de Vila Franca de Xira e termina no Reservatório de Telheiras, com uma extensão total de 34.488 metros (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007). Este adutor tem uma derivação que permite abastecer o reservatório de Camarate.

O Adutor de Circunvalação liga a Estação Elevatória 2 de Vila Franca de Xira ao Reservatório de Vila Fria, numa extensão total de 46 240 metros (Engenharia do Departamento de Operações e Abastecimento da EPAL, 2007). A água é bombeada para o Reservatório de A-dos-Bispos, de onde segue depois de forma gravítica até ao reservatório de Vila Fria. Pelo caminho é distribuída água para outros reservatórios. Este adutor permitiu melhorar o abastecimento aos concelhos de Sintra, Cascais, Oeiras e Amadora devido ao aumento de capacidade de transporte a jusante de Vila Franca de Xira.

Estes adutores encontram-se interligados por forma a garantir a troca de caudais entre si.

Na Figura 110 encontra-se um diagrama que, embora tenha já alguns anos permite ter uma ideia do que foi exposto relativamente ao Sistema de Abastecimento da EPAL.

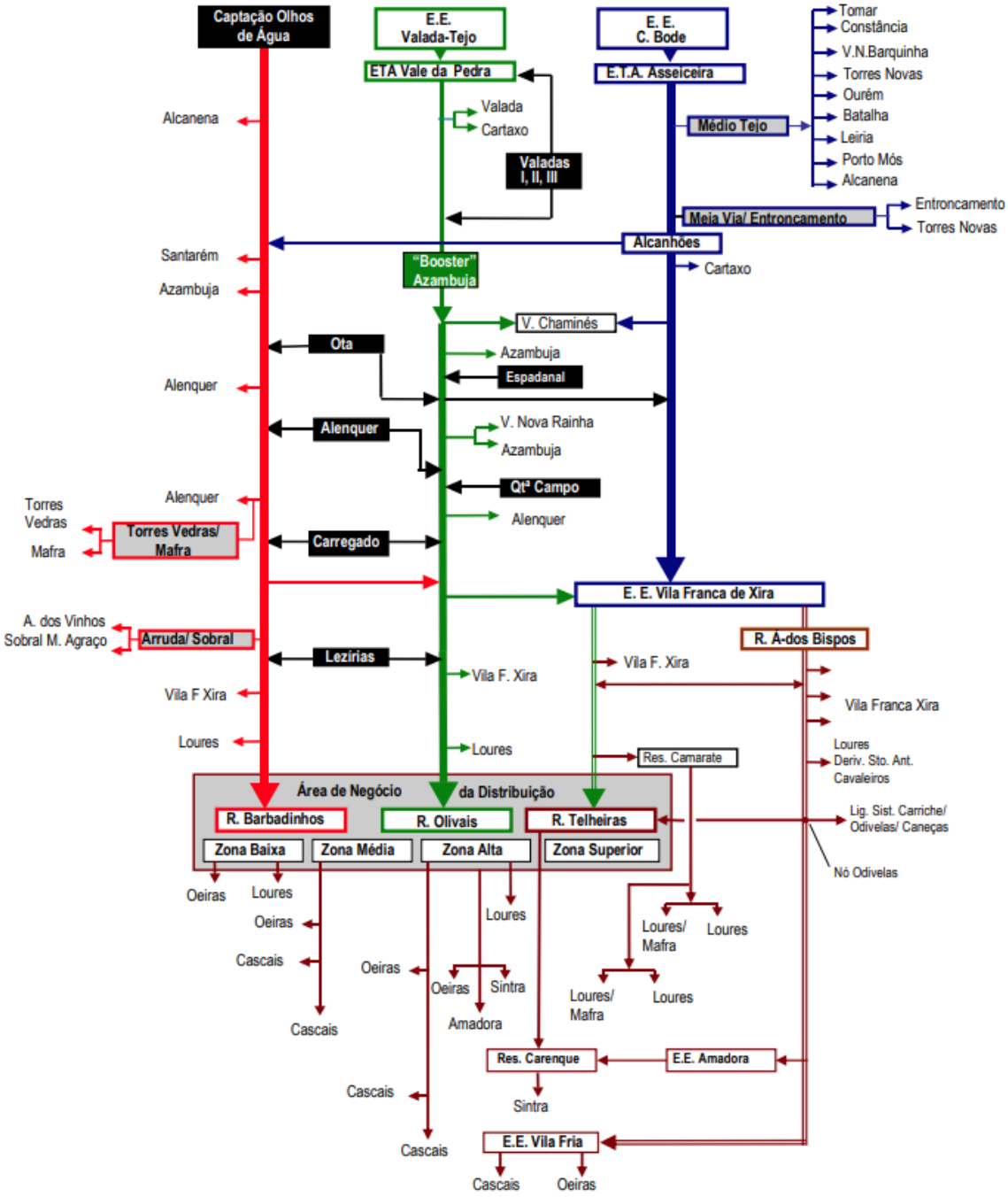


**Figura 110 - Diagrama geral do Sistema de Produção e Transporte. Retirado de (Oliveira et al., 2007)**

Estando apresentados os principais sistemas de produção e transporte de água que permitem a chegada de água aos municípios abastecidos pela EPAL, ao Sistema Oeste e a Lisboa apresentar-se-á em seguida o Sistema Oeste, ao qual pertencem as duas condutas de transporte de água estudadas para a realização do presente trabalho.

## O Sistema de Abastecimento do Oeste

O Sistema de Abastecimento de água à zona Oeste abastece em "alta" total ou parcialmente os concelhos de Alcobaça, Alenquer, Arruda dos Vinhos, Azambuja, Bombarral, Cadaval, Caldas da Rainha, Lourinhã, Mafra, Nazaré, Óbidos, Peniche, Rio Maior, Sobral de Monte Agraço e Torres Vedras (Departamento de Sustentabilidade Empresarial da EPAL, 2023-a), conforme se pode ver na Figura 111.

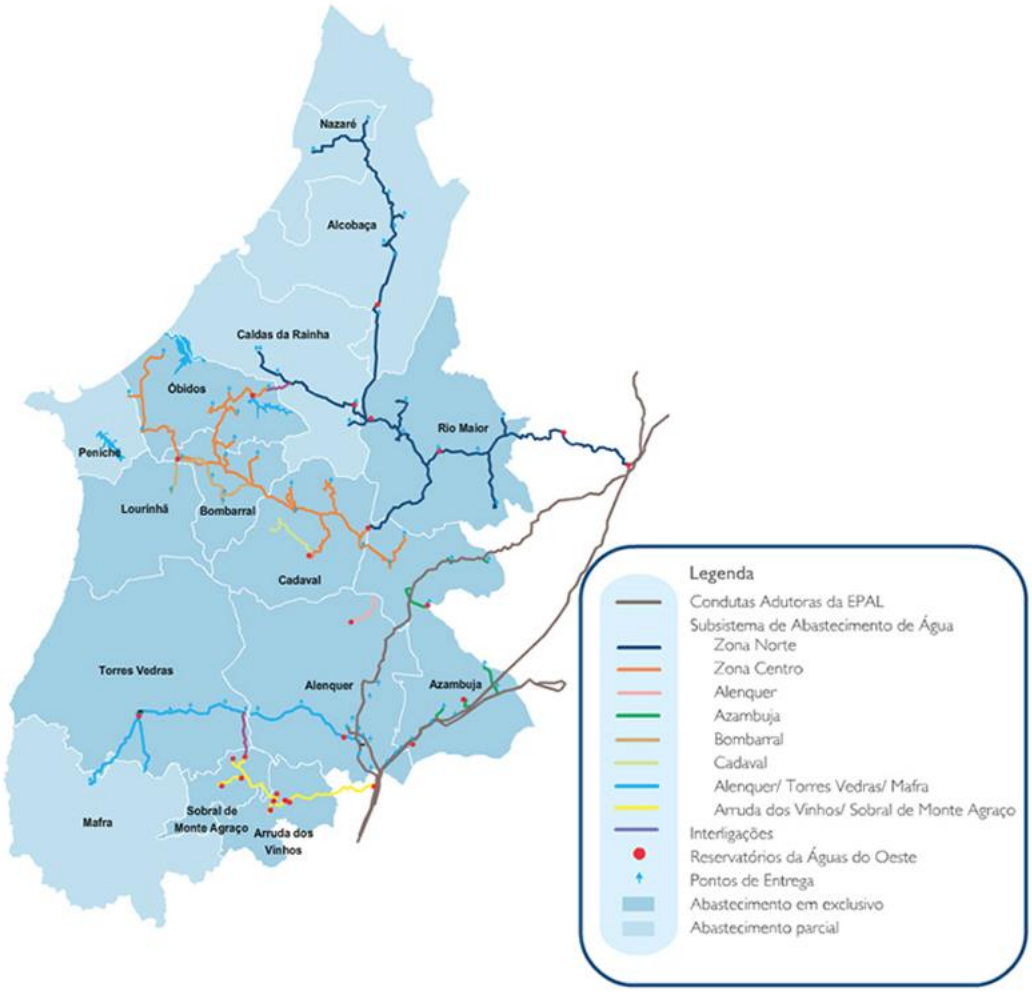


**Figura 111 - Área de intervenção do Sistema de Abastecimento da Zona Oeste. Retirado de (AdVT, 2023)**

Este sistema de abastecimento recebe água dos Adutores atrás apresentados. O Sistema Norte-Centro recebe água do Adutor de Castelo do Bode em Alcanhões. Já o Sistema Sul é composto pelos subsistemas Arruda-Sobral e Torres Vedras-Mafra, ambos abastecidos a partir do Aqueduto do Alviela.

Os Sistemas Autónomos são compostos por sistemas independentes com captações próprias e sistemas apenas com uma estação elevatória alimentada a partir de um dos adutores principais que eleva para um reservatório municipal.

Nos subcapítulos que se seguem apresenta-se cada um dos sistemas acima abordados. Na Figura 112 encontra-se a legenda relativa às imagens que serão apresentadas.

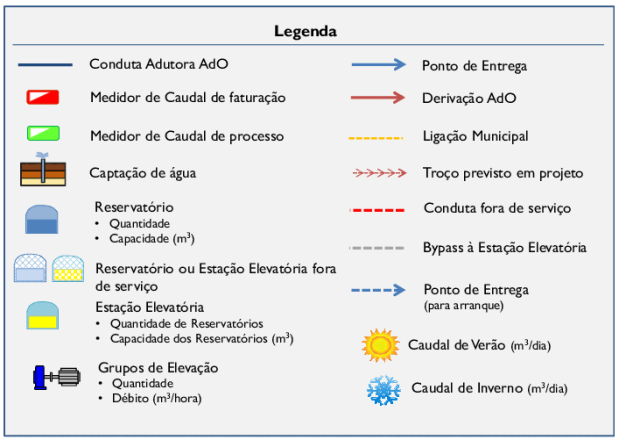


**Figura 112 – Legenda das Figuras que serão apresentadas nos subcapítulos seguintes, baseado num esquema geral da Zona Oeste.**

## O Sistema Norte-Centro

O Sistema Norte-Centro encontra-se esquematizado na Figura 113.

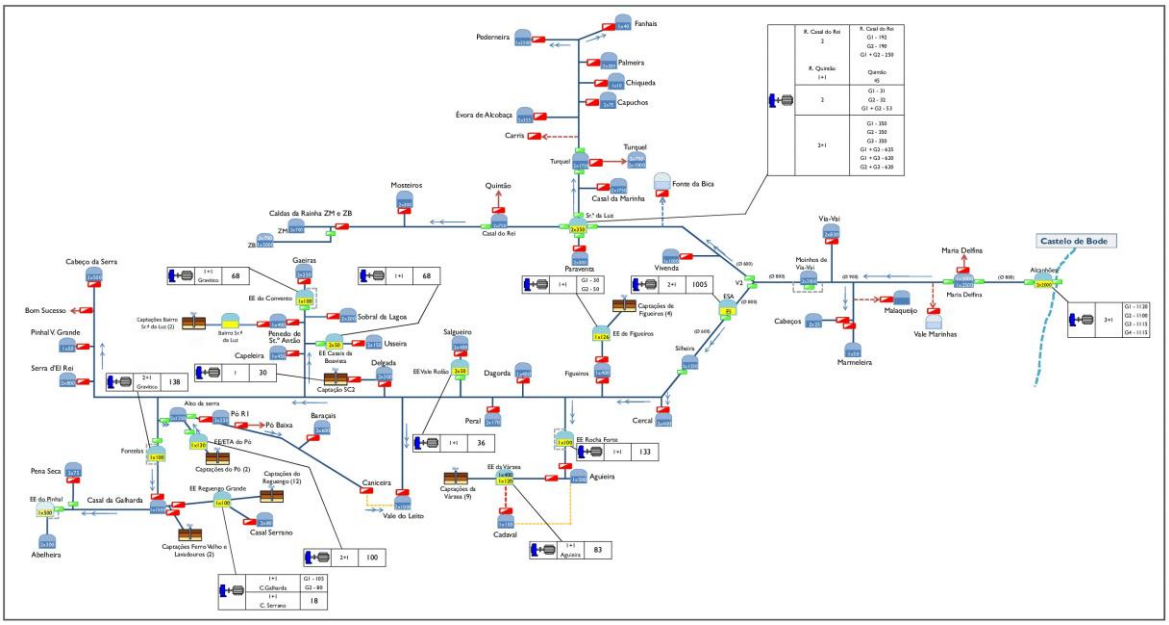


**Figura 113 – Esquema do Sistema Norte-Centro do Sistema de Abastecimento da Zona Oeste, baseado num esquema geral da Zona Oeste.**

Este sistema tem início na Estação Elevatória Alcanhões, que recebe em duas cubas de aspiração água proveniente do Adutor de Castelo do Bode, saída do Reservatório de Alcanhões. Essa água é depois bombeada até ao Reservatório de Maria Delfina, de onde segue graviticamente até alguns Reservatórios Municipais que abastecem as populações locais e até ao Reservatório de Moinhos de Viavai, de onde vai sair graviticamente até à Caixa de Derivação V2. É neste ponto que se iniciam o Sistema Norte e o Sistema Centro, que serão apresentados nos próximos subcapítulos. Na Figura 114 encontra-se a parte do esquema do Sistema Norte-Centro apresentada acima em pormenor.

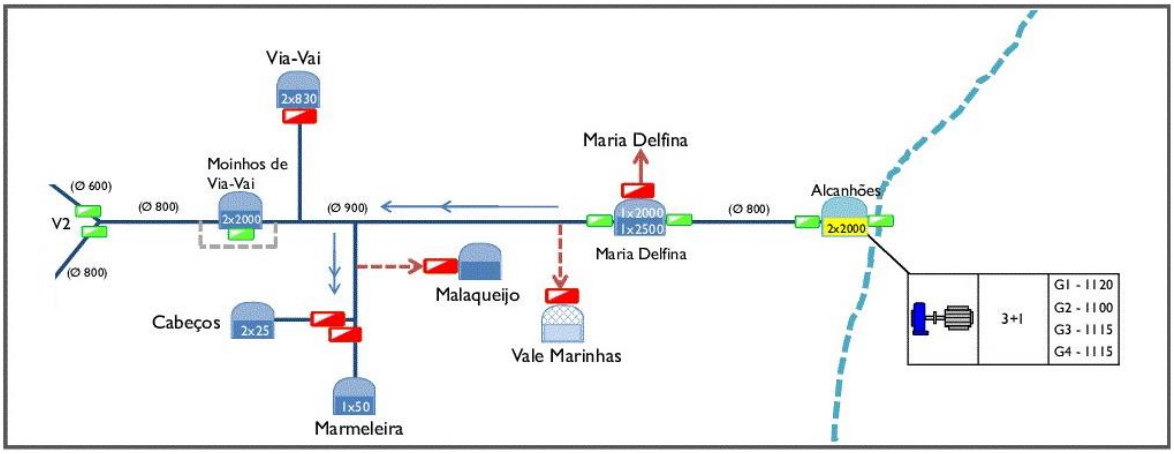


**Figura 114 – Pormenor da área do Sistema Norte-Centro a montante da Caixa de Derivação V2, baseado num esquema geral da Zona Oeste.**

## O Sistema Norte

O Sistema Norte abastece parcial ou totalmente os concelhos de Rio Maior, Caldas da Rainha, Alcobaça e Nazaré, tendo início na Caixa de Derivação V2 onde a água chega de forma gravítica a partir do Reservatório de Moinhos de Viavai (AdVT, 2023). Na Figura 115 encontra-se esta área em pormenor.

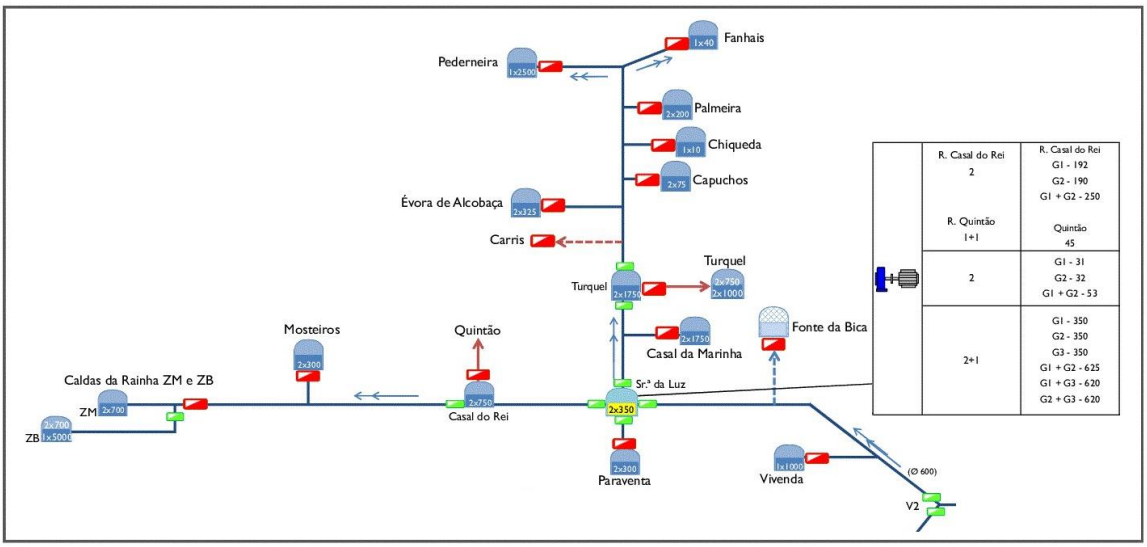


**Figura 115 – Pormenor da área Norte do Sistema Norte-Centro a jusante da Caixa de Derivação V2, baseado num esquema geral da Zona Oeste.**

A montante da Estação Elevatória da Senhora da Luz é abastecido apenas o Ponto de Entrega da Vivenda, uma vez que o Ponto de Entrega da Fonte da Bica não se encontra ativo. A EE Sr.ª da Luz abastece 4 reservatórios através de bombagem. São eles Casal do Rei, Turquel, Casal da Marinha e Paraventa. Os dois últimos são municipais e abastecem as populações locais em "baixa". Já Casal do Rei abastece graviticamente quatro reservatórios municipais, dois dos quais na cidade de Caldas da Rainha.

Relativamente ao Reservatório de Turquel, abastece 6 reservatórios municipais no concelho de Alcobaça (Turquel, Carris, Évora de Alcobaça, Capuchos, Chiqueda e Palmeira) e dois no concelho da nazaré (Pederneira e Fanhais). A conduta de abastecimento Turquel-Pederneira é aquela que no Sistema Oeste costuma ter pressões mais elevadas, que rondam normalmente os 20 bar no PE Chiqueda.

## O Sistema Centro

O Sistema Centro abastece parcial ou totalmente os concelhos Azambuja, Bombarral, Cadaval, Lourinhã, Óbidos e Peniche (AdVT, 2023), tendo início na Caixa de Derivação V2 onde a água chega de forma gravítica a partir do Reservatório de Moinhos de Viavai. Na Figura 116 encontra-se esta área em pormenor.

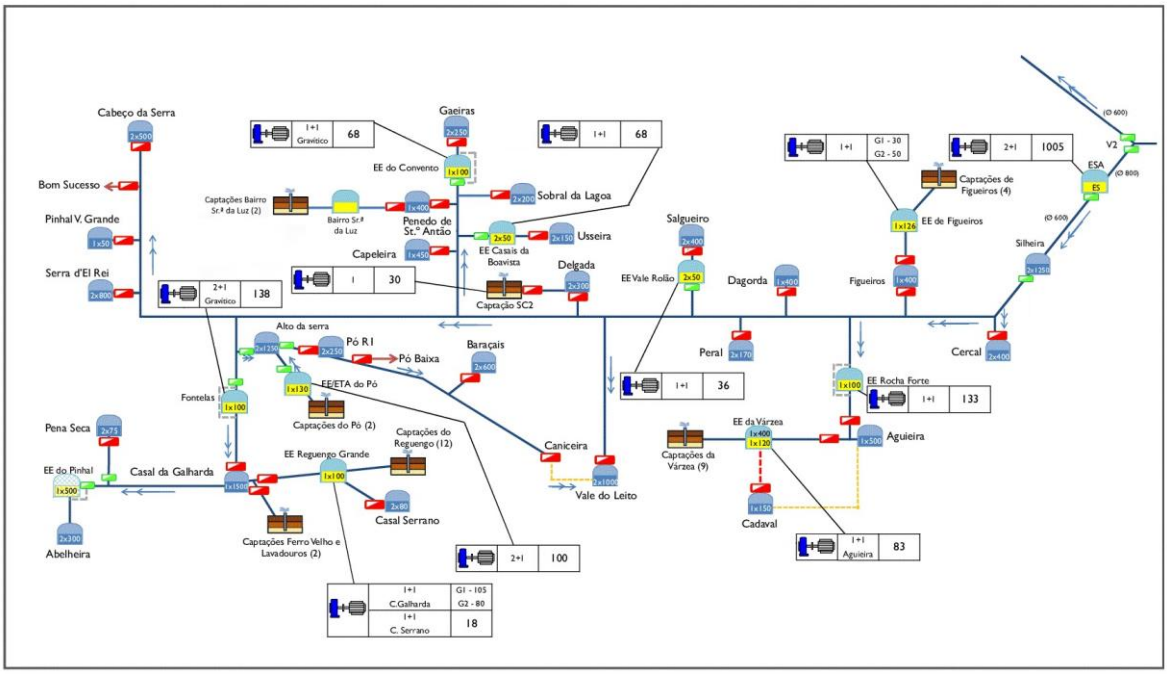


**Figura 116 – Pormenor da área do Sistema Centro a jusante da Caixa de Derivação V2, baseado num esquema geral da Zona Oeste.**

A água proveniente da Caixa de Derivação V2 é bombeada na Estação Elevatória da Amieira sem recurso a reservatório/cuba de aspiração, funcionando os grupos elevatórios ali existentes como hidropressores com o objetivo de fazer chegar a água ao Reservatório da Silheira que se encontra a uma cota elevada e vai permitir abastecer graviticamente uma área bastante significativa. Neste sistema existe um conjunto de captações que reforçam os reservatórios municipais da zona em que se encontram. Esse reforço é de grande importância no verão, quando os consumos são superiores aos habituais devido ao aumento de população em determinadas áreas durante os períodos de férias.

Este sistema abastece graviticamente 17 reservatórios municipais e 5 estações elevatórias, que bombeiam a água para um conjunto de reservatórios municipais mais elevados, onde a água poderia não chegar devido à cota a que se encontram ou chegar em quantidades insuficientes, como é o caso do Reservatório da Galharda que pode ser abastecido graviticamente com um caudal bastante mais reduzido do que o caudal que se obtém com o sistema de bombagem existente em Fontelas. Relativamente ao esquema apresentado, há a referir que a Estação Elevatória da Várzea se encontra como Reserva Operacional no abastecimento ao reservatório da Aguieira e como inoperacional no abastecimento ao

reservatório do Cadaval. Também a captação SC2 da Delgada e algumas das captações de Figueiros e do Reguengo Grande se encontra como Reserva Operacional.

## O Sistema Sul

O Sistema Sul garante o abastecimento a instalações nos municípios de Alenquer, Torres Vedras, Mafra, Arruda dos Vinhos e Sobral de Monte Agraço. Na Figura 117 encontra-se esta área em pormenor.

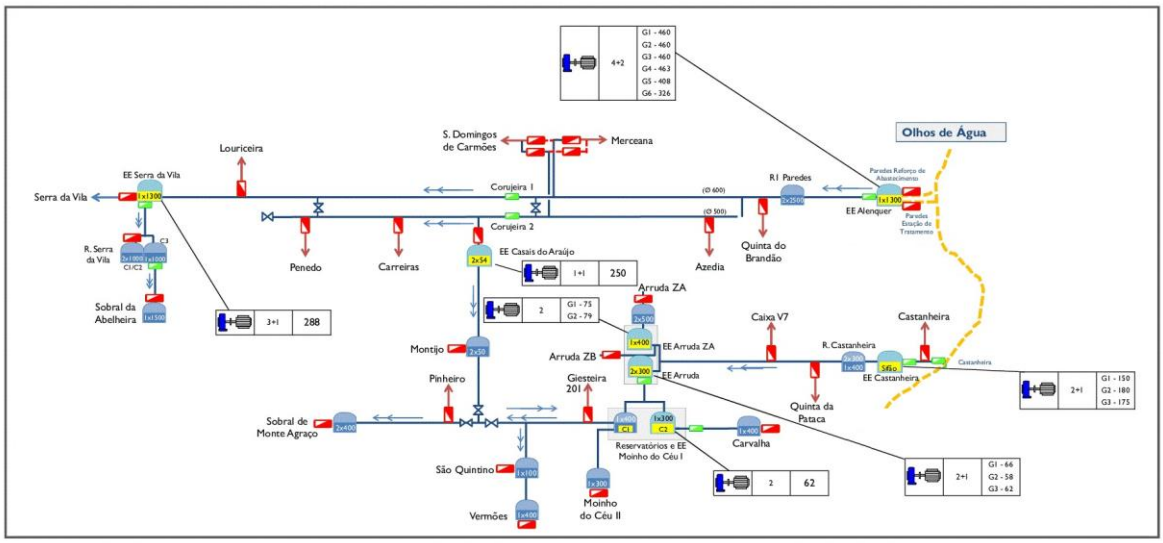


**Figura 117 – Pormenor da área do Sistema Sul, baseado num esquema geral da Zona Oeste.**

A Estação Elevatória de Alenquer IV bombeia água para o Reservatório de Paredes que a faz chegar graviticamente à estação elevatória de Serra da Vila, junto a Torres Vedras, e a um conjunto de pontos de entrega, bem como à estação elevatória de Casais do Araújo que funciona como sobrepressora para que a água chegue ao reservatório do Montijo.

O sistema Alenquer-Torres Vedras-Mafra permite abastecer pontos do concelho de Alenquer a partir da derivação da Merceana. Já o concelho de Torres Vedras é abastecido através das derivações de Carreiras, Penedo e Louriceira, pelo reservatório de Montijo, pelo reservatório da EE Serra da Vila e pelo Reservatório Municipal de Serra da Vila. A parte norte do município de Mafra é abastecida graviticamente através do Reservatório Municipal de Serra da Vila (AdVT, 2023). Caso exista impossibilidade temporária de fazer chegar água aos reservatórios de Arruda dos Vinhos é possível colocar água nos reservatórios de São Quintino, Vermões e EE Moinho do Céu 1 a partir do Reservatório do Montijo.

O sistema Arruda-Sobral inicia-se na Estação Elevatória da Castanheira, de onde a água vinda do Adutor do Alviela é bombeada para o Reservatório da Castanheira, de onde segue graviticamente para a Estação Elevatória de Arruda dos Vinhos.

Na Estação Elevatória de Arruda a água segue graviticamente para o reservatório Arruda Zona Baixa e é bombeada para o reservatório Arruda Zona Alta e para a Estação Elevatória de Moinho do Céu 1.

Conforme se pode ver na Figura 117, a água sai graviticamente da EE Moinho do Céu 1 para todos os reservatórios a jusante com exceção do Reservatório da Carvalha, para o qual é bombeada.

## Sistemas Autónomos

Conforme já foi referido, os sistemas considerados autónomos no Sistema Oeste são abastecidos apenas por captações locais ou possuem uma estação elevatória que recebe água de um dos adutores do Sistema EPAL e a eleva para reservatórios municipais. Na Figura 118 encontram-se essas instalações.

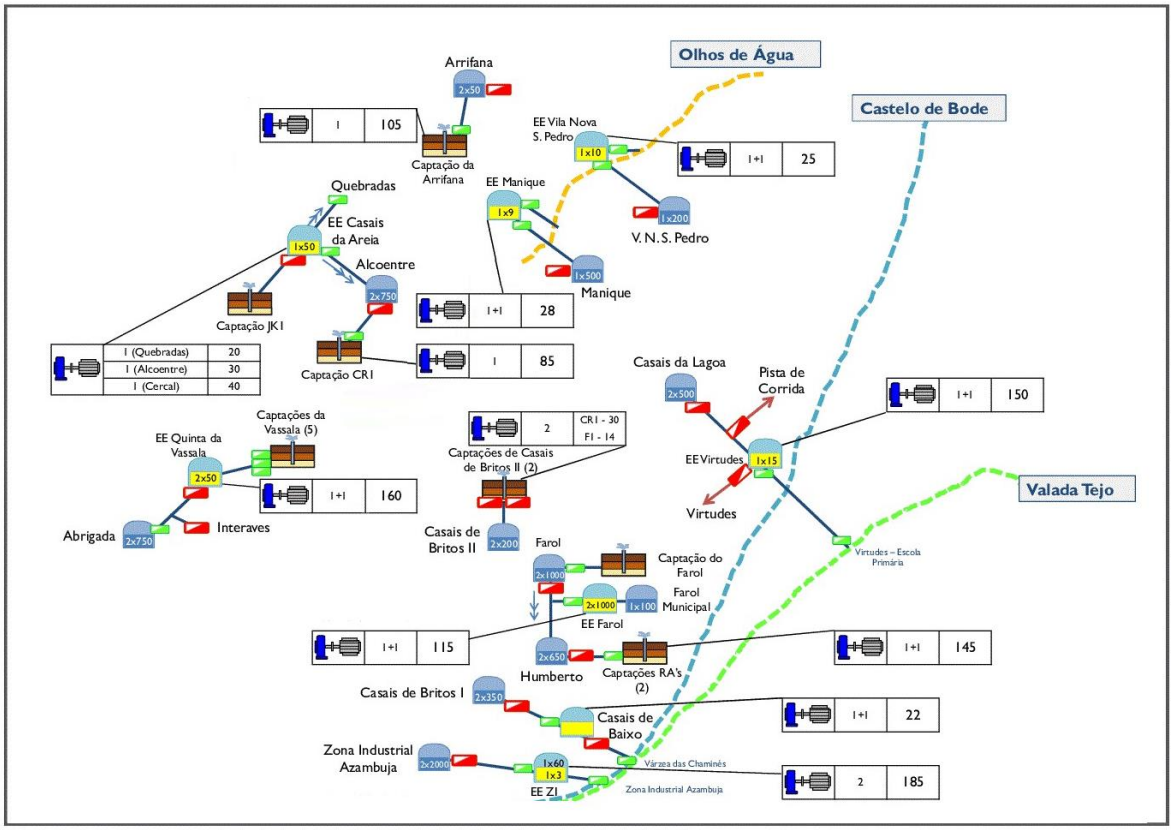


**Figura 118 – Pormenor dos Sistemas Autónomos, baseado num esquema geral da Zona Oeste.**

Conforme se pode perceber pela Figura 118, a EE Virtudes é abastecida pelo adutor de Valada (antes da chegada à Várzea das Chaminés e de se chamar Adutor Tejo), a EE Manique e a EE Vila Nova de São Pedro são abastecidas pelo Adutor do Alviela. A EE Zona Indusatrial (ZI) e a EE Casais de Baixo são abastecidas pelo Adutor Tejo, embora essa situação seja impercetível na Figura.