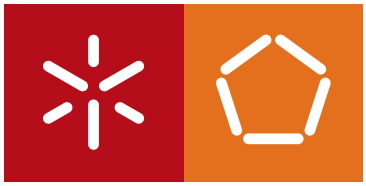
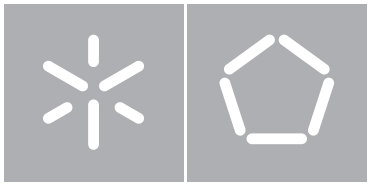


Universidade do Minho
Escola de Engenharia

Daniel José Silva Ribeiro

Support Vector Machines na Previsão do Comportamento de uma ETAR

Dezembro de 2012



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Daniel José Silva Ribeiro

Support Vector Machines na Previsão do Comportamento de uma ETAR

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de
Professor Orlando Manuel de Oliveira Belo

Dezembro de 2012

Aos meus Pais.

Agradecimentos

Aos meus Pais, por tudo que me ensinaram, e pelo suporte que sempre me concederam ao longo de toda a minha formação. A toda a minha família, em especial ao meu Irmão e à Carla, por todo o apoio prestado durante a realização deste trabalho.

Agradeço aos meus colegas e amigos, pela amizade, companheirismo e entre ajuda, fundamentais ao longo do meu percurso académico.

Um agradecimento aos intervenientes deste projeto pertencentes à ETAR em estudo, pela informação e dados fornecidos.

Um agradecimento muito especial para o meu orientador, Professor Doutor Orlando Belo, pelo seu profissionalismo, e pelo o incentivo e acompanhamento permanentes durante este projeto. Agradeço ainda os seus conselhos e disponibilidade, que foram determinantes na realização desta dissertação.

Resumo

Support Vector Machines na Previsão do Comportamento de uma ETAR

O *Data Mining* é um processo de exploração de grandes quantidades de dados, com um potencial enorme para ajudar as empresas na extração de conhecimento que está oculto nos mais diversos sistemas de dados. Esta tecnologia é utilizada pelas empresas nos mais variados domínios, com o intuito de as ajudar em atividades de tomada de decisões. Entre os diversos campos de aplicações encontramos o domínio da Biologia e do Ambiente, em particular, as questões relacionadas com as *Estações de Tratamento de Águas Residuais* (ETAR). As ETAR são infraestruturas essenciais para manter o equilíbrio do meio-ambiente, sendo caracterizadas por terem várias fases de tratamento, nas quais são removidas impurezas como sólidos, matéria orgânica e nutrientes. Todo este processo dinâmico e complexo deve ser processado de forma eficiente, permitindo que o efluente final que nelas é tratado tenha a melhor qualidade possível. A previsão da qualidade da água tratada, com base nos vários fluxos que dão entrada nas ETAR, permite medir a eficácia do tratamento e, assim, obter alguma informação útil para um melhor controle de toda a infraestrutura. A ETAR em estudo neste trabalho de dissertação, localiza-se no Norte de Portugal e serve uma população de cerca de 45 mil habitantes. Os dados fornecidos para alimentação dos processos de interação levados a cabo são referentes a tratamentos realizados nessa ETAR durante o período de um ano. Este estudo pretendeu explorar técnicas de *Data Mining* preditivas, nomeadamente modelos de regressão, por forma a prever com eficácia os valores dos parâmetros de qualidade da ETAR. As medidas de qualidade do tratamento analisadas neste estudo,

basearam-se nos parâmetros de previsão *Carência Bioquímica de Oxigénio* (CBO) e *Sólidos Suspensos Totais* (SST). Por sua vez, as técnicas de regressão adotadas neste trabalho são baseadas em *Support Vector Machines*, mais concretamente nos algoritmos *Support Vector Regression* e numas das suas variantes: *Sequential Minimal Optimization*. Este conjunto de técnicas tem sido aplicadas com sucesso em diferentes áreas, inclusive em alguns trabalhos relacionados com as ETAR. Pretendeu-se assim, à custa da utilização destas técnicas de previsão, definir um modelo comportamental para a ETAR em questão, por forma a analisar a sua capacidade preditiva neste tipo de sistemas complexos. Neste problema, as fases de análise e preparação dos dados mostraram-se determinantes na obtenção dos resultados alcançados. Analisaram-se ainda as diversas tarefas de modelação desenvolvidas neste estudo. Os modelos desenvolvidos demonstraram uma boa capacidade preditiva, especialmente na previsão do parâmetro do efluente final CBO. As técnicas de previsão utilizadas, para além da capacidade de modelação preditiva não linear, permitem ainda uma análise aos atributos mais influentes à qualidade dos parâmetros de previsão.

Palavras-chave: Data Mining, Regressão, Support Vector Machines, Support Vector Regression, Sequential Minimal Optimization, Estações de Tratamento de Águas Residuais, Carência Bioquímica de Oxigénio, Sólidos Suspensos Totais e Previsão de Comportamento de ETARs.

Abstract

Support Vector Machines on the Prediction of a WWTP Performance

Data mining is a process of exploration of large data sets with a huge potential to assist companies in the extraction of knowledge that is hidden in their data systems. Companies in various fields use this technology, in order to assist them in decision-making. Among the various fields are included the Biology and Environment domains, in particular, issues related to the *Wastewater Treatment Plants* (WWTP). WWTP are essential infrastructures to maintain the environmental balance. Treatment plants are characterized by having several treatment stages in which is done the removal of solids, organic matter and nutrients. All of this dynamic and complex process must be handled efficiently to ensure a good quality effluent. The prediction of the treated wastewater quality, based on the measured inflow parameters, allow the treatment performance evaluation and yet to obtain some useful information for a better control of the entire infrastructure. The data used in this study were collected from a WWTP located in northern Portugal that serves a population of about 45,000 inhabitants, whose data was provided regarding the treatments performed in this WWTP during one year. This study aimed to explore data mining techniques for prediction, namely regression models, in order to successfully predict the concentrations of the quality parameters like *Biochemical Oxygen Demand* (BOD) and *Total Suspended Solids* (TSS), which are actually the selected outflow parameters to be predicted in this work. The regression techniques used herein are based on *Support Vector Machines* (SVM), more particularly *Support Vector Regression* and in one of its variants: *Sequential Minimal Optimization*. This set of

techniques has been successfully applied in different areas, including some WWTP related work, thus we intended to explore the SVM and analyze their predictive ability in this type of complex systems. The stages of data preparation and data analysis were shown to be crucial to obtain the results achieved. Several regression models for both predictive parameters were analyzed and compared, where the results show that accurate estimates can be achieved especially on the concentrations of BOD. The SVM, beyond the capability of non-linear predictive modeling, yet allow the analysis of the features that are most related to the quality of the prediction parameters.

Keywords: Data Mining, Regression, Wastewater Treatment Plants, Support Vector Machines, Support Vector Regression, Sequential Minimal Optimization, Biochemical Oxygen Demand, Total Suspended Solids and Prediction of WWTPs Performance.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

Vivemos numa era em que estamos inundados de dados. O volume de informação electrónica aumentou bastante nos últimos anos. Atualmente, este crescimento continua a progredir rapidamente, não se prevendo o fim desta evolução num futuro próximo (Witten et al., 2011). A presença de grandes volumes de informação nas empresas fez com que estas passassem a olhar de forma diferente para os seus dados e, assim, tentarem obterem mais elementos sobre os seus negócios. Como sabemos, esta tendência contribuiu para a emergência da área *Business Intelligence* (BI) (Connoly & Berg, 2010). O BI proporciona um conjunto de soluções de software aos agentes de tomada de decisões das empresas, que os ajuda a identificar e perceber os factores chave do seu negócio, permitindo assim efetuar as melhores de decisões para as suas empresas. Os sistemas baseados em *Data Warehouses* (DW) têm sido uma das principais ferramentas do BI, devido em grande medida às vantagens que lhes são inerentes. A adesão das grandes e médias empresas a este novo tipo de *Sistemas de Suporte à Decisão* (SSD) foi rápida. No entanto, foram surgindo novas necessidades nas empresas que as obrigou sistematicamente a optimizar as suas decisões bem como os seus processos inerentes (Golfareli & Rizzi, 2009).

O simples armazenamento da informação num DW não traz todos os benefícios que as empresas procuram, sendo necessário extrair o conhecimento que possa estar oculto nesses repositórios para que se consiga atingir o real valor dos DW. Em resposta a esta necessidade, surgiu uma nova geração de técnicas computacionais e ferramentas que permitem a extração de conhecimento útil a partir de grandes *Base de Dados* (BD). Este conjunto de técnicas e ferramentas são hoje parte integrante de domínios científicos e técnicos de grandes potencialidades: *Data Mining* (DM) *e Knowledge Discovery in Databases* (KDD) (Fayyad et al., 1996a).

Estas ferramentas de extração de conhecimento, doravante descritas simplesmente por DM, são também parte integrante do domínio da biologia e do ambiente. Vários estudos científicos são feitos nestas áreas, nas quais se inclui o ramo de investigação do tratamento de águas residuais. As *Estações de Tratamento de Águas Residuais* (ETAR) são as infraestruturas responsáveis por efetuar este tipo de tratamentos. Como sabemos, estas são de suma importância para a preservação da saúde pública e do meio-ambiente que nos rodeia. Todavia, as ETAR pressupõem um processo de tratamento complexo, cujas dificuldades advêm principalmente da variedade de constituintes nocivos presentes nas águas residuais. Com base nisto, por forma a remover eficazmente estas impurezas, consegue-se perceber que existem muitas dificuldades no tratamento de águas residuais. Assim, procurando colmatar algumas dessas dificuldades, podem ser aplicadas técnicas de DM como uma tecnologia de auxílio às ETAR. Isto, lembrando que só com um processo de tratamento das águas residuais eficaz, é que as ETAR garantem descargas com qualidade para o meio-ambiente.

Das várias técnicas que se podem enquadrar em DM, destacam-se as técnicas de previsão. Estas permitem modelar o comportamento das ETAR. Deste modo, a previsão da qualidade da água tratada permite medir a eficácia do tratamento e, portanto, obter informação útil para um melhor controle de toda a infraestrutura. Contudo, os diferentes tipos de tratamentos físicos, biológicos e químicos, fazem das ETAR sistemas dinâmicos e complexos. Este facto sugere que sejam utilizadas técnicas avançadas de DM, especialmente modelos de previsão não lineares, por forma a modelar com sucesso o comportamento das ETAR. Neste âmbito, podemos utilizar as técnicas de previsão baseadas em *Support Vector Machines* (SVM), conhecidas pelo seu poder preditivo e pela sua não linearidade.

1.2 Motivação e Objetivos

O impacto que as ETAR têm nos nossos municípios, e na preservação do nosso ambiente, é de grande importância. O aperfeiçoamento do processo de tratamento de águas residuais, através da integração técnicas de previsão, com base em ferramentas de DM, é uma temática bastante interessante e um desafio muito aliciante. Cada vez mais existe maior preocupação com as questões ambientais do nosso planeta, facto este que incentiva a investigação de todo o tipo de métodos que possibilitem a redução do impacte ambiental, que, como sabemos, é provocado principalmente pela poluição causada pelo homem. Por outro lado, temos a motivação de explorar novas tecnologias, visto que estas podem trazer melhorias significativas nas mais variadas áreas. Isto aplica-se não só em empresas, cujo o principal objetivo passa por gerar lucro, mas também nas entidades que têm objectivos de ordem humanitária e de preservação ambiental. Dentro deste grupo estão inseridas as ETAR, uma vez que o seu principal objetivo é proteger a saúde pública e ao mesmo tempo evitar problemas ambientais, sociais, económicos e políticos (Metcafl & Eddy, 2003). Adicionalmente, como estamos perante um trabalho de cariz académico, procura-se aprofundar conhecimentos na área de SSD, em particular do processo de desenvolvimento e aplicação de técnicas de DM num caso real de uma ETAR. Por conseguinte, têm-se o desafio de, segundo dados reais, familiarizar-se com esses dados e estudar os melhores procedimentos a aplicar nas várias fases de desenvolvimento de uma aplicação de DM.

Esta dissertação desenvolveu-se em torno do estudo de uma ETAR do norte de Portugal e na aplicação de técnicas de previsão sobre o comportamento dos seus sistemas, e cujo objectivo principal passou por modelar esses comportamentos. Portanto, tendo em conta os vários factores que poderão levar a um aperfeiçoamento e uma maior precisão dos resultados, foi necessário obter-se conhecimento do processo de DM, em particular das técnicas relacionadas com as SVM. Com este conhecimento técnico, conjugado com uma compreensão dos dados e respectivos tratamentos da ETAR, pretendeu-se alcançar um desempenho preditivo eficaz, bem como apresentar informação potencialmente útil aos agentes decisores de uma ETAR.

1.3 Estrutura do Documento

Depois deste capítulo introdutório, este documento está estruturado em mais seis capítulos. Os capítulos 2 e 3 consistem, basicamente, em fundamentações teóricas das ETAR e da tecnologia de DM. Os capítulos seguintes acompanham os passos da metodologia de DM adoptada para estudo, sendo, no capítulo 4, abordadas as questões da análise e da preparação de dados, e no capítulo 5, a parte relativa à modelação dos processos. Nos capítulos finais são apresentadas as experiências realizadas e respectivos resultados (Capítulo 6). Por fim, são tiradas conclusões sobre o trabalho realizado (Capítulo 7). Complementa-se esta descrição geral, apresentando em seguida uma descrição mais detalhada dos vários capítulos:

- **Capítulo 2 - Estações de Tratamento de Águas Residuais.**
  Neste capítulo são caracterizadas as águas residuais e o processo de tratamento das ETAR. São ainda expostas as principais motivações para a aplicação de técnicas de DM nestas infraestruturas, e apresentados alguns trabalhos científicos relacionados com o problema em estudo.

- **Capítulo 3 - *Data Mining* e *Support Vector Machines*.**
  Introduz-se a tecnologia de DM, descrevendo o seu processo de desenvolvimento e algumas das suas técnicas mais usuais. Além disso, demostra-se a fundamentação teórica que motivou a origem das SVM, bem como as suas principais características.

- **Capítulo 4 - Análise e Preparação de Dados.**
  Neste capítulo, é inicialmente apresentada a análise de negócio e respectivos objetivos, seguindo-se a análise de dados, que inclui a caracterização dos dados relativos à ETAR em estudo. É demonstrado também todo o processo que envolveu o tratamento e preparação do conjunto de dados de trabalho.

- **Capítulo 5 - Modelação da ETAR.**
  Neste capítulo, descreve-se o processo de modelação adoptado para realizar as várias tarefas de previsão da ETAR em estudo. São portanto apresentadas as técnicas e estratégias adoptadas nesta fase decisiva do processo de DM.

- **Capítulo 6 - Experiências e Resultados.**
  Uma vez descrito o processo de modelação, são aqui apresentados os resultados obtidos referentes às diferentes tarefas de modelação executadas.

- **Capítulo 7 – Conclusões e Trabalho Futuro.**
  Por último, são apresentadas as conclusões, algumas considerações finais do estudo realizado, e um esboço de algumas ideias para eventuais trabalhos futuros.

Capítulo 2

Estações de Tratamento de Águas Residuais

2.1 A Criticidade da Água

O planeta Terra nos seus primórdios continha uma quantidade de água finita. Hoje em dia, a quantidade de água de que dispomos não é maior nem menor que essa quantidade inicial. Com isto em mente, para manter a natureza da mesma forma que a conhecemos hoje, é crucial preservar e proteger a água. Com o passar dos tempos o valor da água potável é cada vez maior, valor este que se poderá tornar bem mais alto do que possamos imaginar. No futuro poderemos ter água potável a ser vendida ao preço, por exemplo, do petróleo, ou ainda mais cara (Spellman, 2003).

Embora exista uma grande abundância de água no planeta, no entanto cerca de 97,5% da água é salgada, ou seja, imprópria para consumo. Os restantes 2,5% de água doce de que dispomos encontram-se em águas subterrâneas, calotes polares e, apenas uma pequena parte (0,3%), nos lagos e rios. Convém ainda destacar que no total de água doce existente na Terra está incluída a

água poluída. Segundo as estatísticas da UN-Water[i], o ritmo de consumo de água no Planeta está a aumentar para mais do dobro do crescimento da população. Existem atualmente locais em que a quantidade de água que é retirada para utilização (i.e. consumo doméstico, regas e indústria), é muito superior ao que a natureza consegue repor. Para agravar, prevê-se que até 2025 exista um aumento da utilização da água de 50% nos países em vias de desenvolvimento e de 18% nos países desenvolvidos, que como sabemos indicia graves problemas de sustentabilidade dos recursos hídricos (UN-Water, n.d.).

Para além dos problemas de escassez de recursos hídricos, existem ainda preocupações relativas à qualidade desses recursos. O relatório da WHO & UNICEF (2012) no âmbito do *Joint Monitoring Programme*[ii] (JMP) refere que a percentagem de população mundial com acesso a água própria para consumo em 2015 será de 92%, superando assim o patamar projetado pelos *Millennium Development Goals*[iii] (MDG). Ao passo que, por sua vez, as metas definidas no MDG acerca da população com acesso a saneamento básico (75%) não serão alcançadas na projeção feita para 2015 (67%). Apesar do aumento da taxa da população com acesso a saneamento, que passou de 49% em 1990 para 63% em 2010, é de notar que mesmo assim existe uma grande parte de população mundial sem acesso a estes sistemas. As zonas do globo onde o acesso à água própria para consumo é escassa, geralmente coincidem com as zonas que carecem de saneamento básico. É nos países em vias de desenvolvimento que o risco para a saúde pública devido à poluição da água é mais acentuado. Consequentemente, estas zonas têm uma elevada taxa de mortalidade devido em grande parte a doenças provenientes de águas poluídas, que são um efeito da ausência de sistemas de saneamento apropriados. Os factores já referidos, demonstram bem a importância

[i] UN-Water é o mecanismo que reforça a coordenação entres as diferentes entidades das Nações Unidas que trabalham nos problemas relacionados com a água e saneamento.

[ii] WHO/UNICEF JMP é o mecanismo oficial das Nações Unidas com a tarefa de monitorização do progresso das metas definidos na MDG relacionadas com a água para consumo e saneamento.

[iii] Millennium Development Goals são um conjunto objectivos de desenvolvimento internacional que se pretendem alcançar até ao ano 2015. Estes foram declarados em 2000, na cimeira do milénio.

da água e a criticidade do tratamento de águas, sendo por isso necessário tratar, purificar e até reutilizar a água que utilizámos. Deste modo, por forma a preservar a saúde pública e o meio-ambiente, torna-se fundamental a presença de infraestruturas como as ETAR no seio das populações. Só com uma rede de saneamento básico nos centros urbanos, assim como um processo de tratamento de águas eficaz e evoluído tecnologicamente nas ETAR, é possível obter um meio-ambiente limpo e garantir a sustentabilidade da água no nosso Planeta.

2.2 Tratamento de Águas Residuais

A água, devido às suas propriedades como solvente e à sua capacidade de transportar partículas, contém várias impurezas que definem portanto a sua qualidade. Essa qualidade é essencialmente resultado de duas causas: os fenómenos naturais e a interferência do homem. A primeira causa acontece, por exemplo, nas cheias e enxurradas que arrastam as impurezas dos solos para os rios. A segunda causa é bem conhecida como o principal factor de poluição da água, devido, por exemplo, às descargas de águas residuais e também à poluição dos solos com pesticidas e fertilizantes (Sperling, 2007). Nas nossas comunidades existe uma rede de condutas por onde passa a água de abastecimento que é usada nas diversas tarefas domésticas ou industriais. Depois de usada essa água fica imprópria para consumo, entrando assim num novo circuito que encaminha as então águas residuais para as ETAR. As águas residuais podem ser assim definidas como águas que contêm e transportam os desperdícios provenientes das habitações, do comércio, da agricultura e da indústria. É bastante óbvio que é importante não despejar estas águas contaminadas no meio-ambiente sem nenhum tipo de tratamento. As ETAR têm aqui a sua razão de ser. Estas estão hoje presentes nos nossos municípios e, embora não nos apercebamos da presença destas infraestruturas, elas são de uma importância vital para a obtenção de uma boa qualidade da água. De uma forma simples, pode-se dizer que as ETAR são infraestruturas que tratam as águas residuais, cujo o principal objectivo é proteger a saúde pública e ao mesmo tempo evitar problemas ambientais, sociais, económicos e políticos.

2.2.1 Caracterização das Águas Residuais

A acumulação de águas residuais sem tratamento implica o desenvolvimento de condições sépticas, que são bastante incomodativas, inclusive devido aos maus odores que produzem. Adicionalmente, as águas residuais contêm microrganismos patogénicos, que permanecem no sistema digestivo do homem e que podem causar doenças como a *Cólera*, a *Amebíase*, a *Hepatite* A, entre outras. Estas contêm ainda nas suas composições nutrientes, que podem estimular o crescimento de plantas aquáticas indesejáveis, e ainda componentes *tóxicas*, *mutagénicas* e *carcinogénicas* (Metcafl & Eddy, 2003). É por estes factores que a poluição dos recursos hídricos onde são realizadas descargas sem tratamento prévio, provoca a destruição da vida aquática, o ambiente em geral é afetado e as paisagens são alteradas. A utilização destes recursos naturais fica portanto limitada devido à poluição, não sendo possível usufruir deles de forma livre e sem precauções. Isto realça que é crucial remover as impurezas das águas residuais. Assim, o tratamento de águas residuais tem como objetivo identificar esses constituintes e tratar a água por forma a remove-los. Na tabela 2.1, descreve-se algumas das componentes das águas residuais que são geralmente removidas nas ETAR, algumas das quais já mencionadas anteriormente.

Dentre os diferentes grupos de constituintes apresentados na tabela 2.1, consideram-se vários parâmetros que caracterizam as águas residuais. Esses parâmetros são medidos durante o tratamento com o objetivo de proporcionar aos operadores das ETAR informação sobre as características das águas residuais em tratamento. Desta forma, os operadores adquirem conhecimento sobre o nível de poluição presente nas águas, o que é importantíssimo para executar tratamentos eficientes. Os parâmetros ou características das águas residuais podem ser divididos em três categorias, os parâmetros físicos, químicos e biológicos. Na tabela 2.2 são listados alguns dos parâmetros mais comuns e suas respectivas categorias. Note-se que os parâmetros químicos e biológicos correspondem, de forma geral, a constituintes que devem ser removidos (Tabela 2.1).

Tabela 2.1 - Principais componentes das águas residuais a serem tratados – tabela adaptada de (Metcafl & Eddy, 2003)

Constituinte	**Importância do Tratamento**
Sólidos Suspensos	Conduzem ao crescimento de depósitos de lamas, criando condições anaeróbicas quando as águas não tratadas são despejadas no meio hídrico.
Organismos Biodegradáveis	Compostos por proteínas, hidratos de carbono, gorduras, organismos biodegradáveis, são normalmente medidos como CBO (carência bioquímica de oxigénio) e CQO (carência química de oxigénio). Se descarregados sem tratamento, podem levar à redução do oxigénio nos recursos naturais e ao desenvolvimento de condições sépticas.
Patogénicos	As doenças podem ser transmitidas por organismos patogénicos presentes nas águas residuais.
Nutrientes	Os nutrientes, i.e. fósforo e carbono, são essenciais para o crescimento, portanto, podem levar ao crescimento de vida aquática indesejável. Quando descarregados em grandes quantidades, podem ainda poluir as águas subterrâneas.
Poluentes Prioritários	Componentes orgânicos e inorgânicos selecionados conforme as suspeitas de mutagenicidade, de carcinogenicidade ou de uma grande toxicidade.
Organismos Refratários	São organismos que tendem a resistir aos tratamentos efectuados pelas ETAR. Alguns exemplos típicos incluem os tensioativos, os fenóis e pesticidas agrícolas.
Metais Pesados	Normalmente são provenientes do comércio e da indústria, estes devem ser removidos principalmente se as águas residuais forem reutilizadas.
Inorgânicos Dissolvidos	Componentes como o cálcio, sódio e sulfato, que devem ser removidos caso as águas residuais forem reutilizáveis.

Tabela 2.2 - Parâmetros físicos, químicos e biológicos que caracterizam as águas residuais - tabela adaptada de (Sperling, 2007)

Categoria	Parâmetro	Descrição
Físicos	Temperatura	Varia com as estações do ano e influencia a atividade microbial, a solubilidade dos gases e a viscosidade dos líquidos.
	Cor	Diferentes cores indicam diferentes tipos de águas, i.e. água mais escura indica fortes condições sépticas, água colorida indica que provém dos desperdícios industriais.
	Odor	Águas residuais sépticas têm odores mais fortes e desagradáveis. Resíduos industriais têm odores característicos.
	Turbidez	Causada pela grande variedade de sólidos suspensos.
Químicos	Sólidos Suspensos Totais (SST)	Parte de sólidos orgânicos e inorgânicos que não são filtráveis. Compostos minerais não oxidáveis pelo calor e inertes.
	Sólidos Suspensos Voláteis (SSV)	Parte de sólidos orgânicos e inorgânicos que não são filtráveis. Compostos minerais oxidáveis pelo calor.
	Carência Bioquímica de oxigénio (CBO)	Medição do oxigénio consumido pelos microrganismos, após 5 dias na estabilização bioquímica da matéria orgânica.
	Carência Química de Oxigénio (CQO)	Representa a quantidade de oxigénio que é necessária para estabilizar quimicamente a matéria orgânica carbonosa.
	Nitrogénio Total	Inclui nitrogénio orgânico, amônio, nitrito e nitrato. É um nutriente necessário ao crescimento dos microrganismos.
	Fósforo Total	Existe em ambas as formas, orgânica e inorgânica. É um nutriente essencial para o tratamento biológico das águas residuais.
	pH	Indicador da acidez ou alcalinidade das águas residuais.
Biológicos	Bactérias	Organismos unicelulares, presentes em várias formas e tamanhos. Algumas bactérias são patogénicas, causam principalmente doenças intestinais.
	Protozoários	Essenciais no tratamento biológico para manter o equilíbrio dos vários grupos. Alguns são patogénicos.
	Vírus	Organismos parasitas, patogénicos e de difícil remoção no tratamento.

Os parâmetros físicos correspondem a medições ou observações das características físicas das águas, ou seja, características que são perceptíveis de forma sensorial, isto é, através da visão, do toque e do cheiro, daí a estarem inseridos nesta categoria a cor, a turbidez, a temperatura e o odor (Spellman, 2003). Na categoria de parâmetros químicos, encontra-se grande parte dos parâmetros ditos quantitativos, visto que são medidas as quantidades das concentrações desses parâmetros. Já na categoria de parâmetros biológicos, constam essencialmente microrganismos referidos como qualitativos, pois é registada a presença ou não desses microrganismos biológicos — por vezes estes parâmetros são também designados por microfauna.

Os parâmetros a partir dos quais são caracterizadas as águas residuais são medidos durante o tratamento através de análises das águas ou por equipamentos de monitorização (sensores). Como as quantidades de constituintes nocivos presentes nas águas residuais pode variar bastante devido a diversos fatores, é crucial dotar as ETAR de sistemas de monitorização precisos e fiáveis, só desta forma se obtém uma representação fiel dos fluxos de tratamento. Adicionalmente, é necessário obter os valores das concentrações dos parâmetros regulamentados por lei. Cada vez mais existem preocupações e, consequentemente, maiores exigências sobre as questões ambientais, logo é imprescindível uma regulamentação rigorosa sobre os tratamentos de águas residuais. Esta regulação é bastante importante, pois impõe às ETAR a obrigação emitir efluentes com boa qualidade reduzindo assim o impacte ambiental.

Em Portugal, a lei que regula a qualidade das águas residuais após tratamento consta do Decreto de Lei n.º 152/97 de 19 de Junho. Neste decreto de lei são regulamentados os parâmetros químicos como o CBO, CQO e SST, com limites nas concentrações de 25, 125 e 35 mg/l, respectivamente. Para as zonas consideradas sensíveis, existe ainda a obrigatoriedade de redução de nutrientes como o fósforo total e o nitrogénio total. É de realçar que com o passar do tempo e com a evolução tecnológica, são descobertos novos compostos e novas doenças. Com base nisso, a legislação deverá ser atualizada, assim como as ETAR deverão adaptar-se a novas realidades.

2.2.2 Processo de Tratamento das ETAR

O tratamento das impurezas presentes nas águas residuais é realizada em diferentes fases, cada uma delas orientada a determinados tipos de constituintes. É desta forma que se purifica a água das diferentes substâncias nocivas. Sabendo que estas têm diferentes características, consequentemente são necessários diferentes processos de tratamento. Na tabela 2.3 são descritas as fases de tratamento e respetivas impurezas tratadas, demonstrando assim a relação entre o tratamento realizado e os constituintes removidos.

Tabela 2.3 - Fases de tratamento das ETAR - adaptado de (Metcafl & Eddy, 2003)

Fase de Tratamento	Descrição
Preliminar	Remoção de partículas de maiores dimensões, como flutuáveis, sólidos grosseiros, areias e gorduras, visto que estas podem causar problemas operacionais nos tratamentos a jusante.
Primário	Remoção de parte de sólidos suspensos (SS) e alguma matéria orgânica.
Secundário	Essencialmente é removida a matéria orgânica biodegradável, embora sejam também reduzidas as cargas de SS.
Terciário	É realizado o tratamento de nutrientes e organismos patogénicos, usualmente através processos de desinfeção.

A título representativo, são ainda demonstradas as fases e os tipos de tratamento utilizados pela ETAR que será o alvo de estudo desta dissertação. Na figura 2.1, podemos observar um diagrama abstrato do processo de tratamento da ETAR em estudo. Note-se que este exemplo ajuda a ilustrar os tipos de tratamento executados na generalidade das ETAR.

O processo de tratamento inicia-se com a fase de tratamento preliminar. Depois do afluente bruto dar entrada na ETAR, são removidos os sólidos de maiores dimensões, areias, óleos e gorduras. Habitualmente a água bruta é filtrada através de sistemas de gradagem, com o objetivo de remover os sólidos mais grosseiros. Para além desta técnica podem ser utilizados crivos e

tamisadores, que têm capacidade de remover sólidos de menores dimensões. Ainda no tratamento preliminar, o tratamento prossegue para os desarenadores onde são removidas as areias das águas. A terminar esta fase é executado ainda o tratamento dos óleos e gorduras através de desoleadores e desengorduradores. Para uma remoção mais eficiente usam-se sistemas de flotação, que consistem na injeção de ar ascendente por forma a arrastar assim as impurezas para o topo, sendo estas posteriormente removidas.

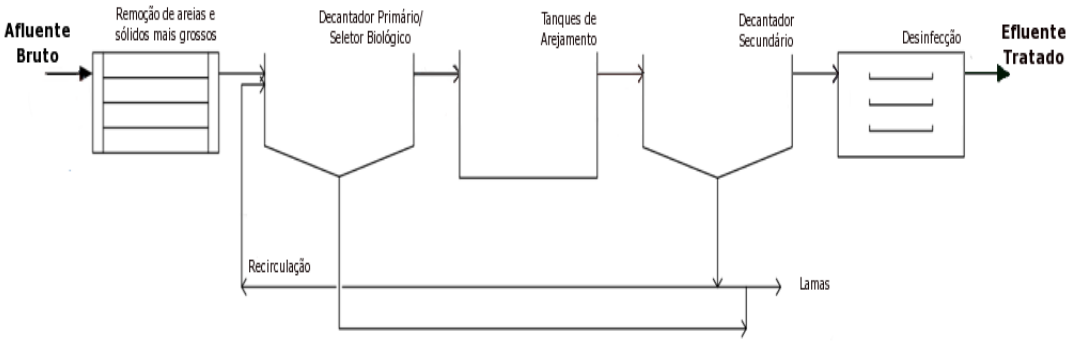


Figura 2.1 - Ilustração do processo de tratamento da ETAR em estudo

Passando ao tratamento primário, neste processo são reduzidas essencialmente as cargas de sólidos suspensos (SS) e de matéria orgânica (i.e. CBO e CQO). Para este efeito, são normalmente utilizados decantadores primários que, através da sedimentação, reduzem em cerca de 60% e 30% as cargas de SS e CBO/CQO respectivamente (Levy, 2000). De grosso modo, o processo de decantação consiste na remoção das partículas suspensas que se depositam na parte inferior do tanque e nas espumas que se acumulam no topo. Adicionalmente, a utilização de flotadores na fase primária, permite também uma maior eficácia de remoção dos resíduos. Ainda a apontar que no caso particular da ETAR em estudo (Figura 2.1), por vezes o decantador primário é utilizado como seletor biológico.

Após a fase primária segue-se a fase de tratamento secundário, na qual são realizados essencialmente tratamentos biológicos que procuram remover a carga orgânica (CBO e CQO).

Nesta fase existem várias tecnologias de tratamentos diferentes, como os sistemas de leitos percolares, discos biológicos, lagunagem e lamas ativadas (Azevedo, n.d.). O sistema de tratamento secundário demonstrado no diagrama, logo, utilizado pela ETAR em estudo, consiste num tratamento através de lamas ativadas. Depois do tratamento do decantador primário, o fluxo dá entrada nos tanques de arejamento, onde ocorre um processo de oxidação da matéria orgânica. O licor misto resultante da retenção nos tanques de arejamento, segue depois para o decantador secundário. De realçar, que em paralelo a este processo realizam-se os tratamentos das lamas, correspondentes à fase sólida do tratamento. Uma vez que os microrganismos podem crescer em demasia, é necessário proceder ocasionalmente à remoção das lamas. Deste modo, as lamas resultantes do decantador secundário seguem para o respetivo tratamento, no qual uma parte é misturada com o fluxo em tratamento de lamas, e a outra volta a entrar na linha de tratamento líquida, nomeadamente nos tanques de arejamento através de um processo de recirculação permanente. No fim do tratamento secundário, a água está pronta para ser descarregada no meio-ambiente já com um nível de qualidade aceitável.

Muitas ETAR apenas dispõem de processos de tratamento até à fase secundária. Todavia, em função da qualidade exigida e do meio receptor, que pode ser considerado uma zona crítica, é necessário por vezes uma maior qualidade do efluente final. O tratamento terciário permite este aumento de qualidade, onde, em alguns locais, as águas residuais chegam mesmo a ser reutilizadas. Na fase terciária são utilizadas técnicas de remoção de nutrientes, como o fósforo e o nitrogénio, ou ainda técnicas de desinfeção de organismos patogénicos, com recurso ao cloro, ao ozono, ou a raios ultravioleta. Em Portugal ainda não se atingiram níveis que permitam a reutilização das águas residuais para consumo direto da população, no entanto a reutilização da água tratada pode servir para diversos fins, como por exemplo para a rega de campos agrícolas. De realçar que existe neste momento a sabedoria e a tecnologia necessárias para aplicar um tratamento de águas mais eficaz a todo o território nacional, basta que para isso que existam verbas e vontade política capazes de realizar estas melhorias (Azevedo, n.d.).

Esta descrição de algumas das técnicas de tratamento mais comuns que são adoptadas pelas ETAR, dá para ter uma ideia que o tratamento de águas residuais é um processo bastante complexo e sensível. Por isso percebe-se que é bastante difícil manter a infraestrutura operacional

e em condições ótimas, por forma a assegurar-se que o efluente esteja nas condições ideais para ser despejado no meio-ambiente.

2.3 Aplicação de Técnicas de Data Mining nas ETAR

2.3.1 Motivação para a Previsão do Comportamento das ETAR

As técnicas de análise das águas residuais sofreram grandes progressos, em parte graças aos avanços tecnológicos ocorridos nos últimos tempos. Com a evolução da tecnologia de monitorização nas ETAR, nomeadamente, foram desenvolvidos instrumentos de medição mais sofisticados, hoje é possível obter processos de tratamento mais avançados e eficientes. Apesar dos sensores serem muitas vezes vistos como o elo mais fraco do sistema de tratamento, devido principalmente a problemas de fiabilidade, os progressos obtidos nos últimos tempos aumentaram bastante a performance e fiabilidade destas componentes. Ainda assim, existem entraves à adoção destes sensores mais avançados por parte das ETAR. O maior obstáculo deve-se ao facto que as ETAR não foram projetadas para possuírem, por exemplo, equipamentos de medições on-line ou que proporcionem um controle em tempo real. Outro obstáculo passa pelos custos elevados de alguns sensores. Todavia, com regulamentações sobre a qualidade do efluente cada vez mais exigentes, fazem com que a evolução das ETAR seja cada vez mais uma necessidade (Vanrolleghem & Lee, 2003). Com base nisto, é natural que surjam com maior frequência ETAR dotadas de sistemas de tratamento avançados e mais complexos, especialmente através da presença de sistemas avançados de monitorização, medição e consequentemente com melhores sensores. Com recurso a estes sistemas, as ETAR passam assim a disponibilizar aos seus operadores um maior e mais preciso leque de informação.

Nas últimas décadas, a tendência para o aumento da complexidade e quantidade de dados disponíveis verificou-se nos mais variados domínios, inclusive nas ETAR. Hoje em dia, existe já algum desenvolvimento de sistemas de informação que auxiliam a monitorização e controle das ETAR. Como é o caso dos sistemas *Supervisory Control and Data Acquisition* (SCADA) e o

TELEMAC. O sistema SCADA permite melhorar o controle de várias plataformas industriais, nas quais se incluem as ETAR. Algumas das funcionalidades do SCADA são a monitorização gráfica de toda a planta, registo de dados, alarmes, funções de diagnóstico etc. (Sosik, n.d.). Por sua vez, o TELEMAC trata-se de um sistema de monitorização remota, projetado para prestar serviços de análise e otimização do funcionamento das ETAR de algumas empresas (Lardon et al., 2002). Ambos os sistemas armazenam por tempo indefinido todo o histórico dos dados medidos durante o tratamento. De acrescentar que vários estudos de previsão realizados no domínio das ETAR recorrem aos dados provenientes destas ferramentas. Contudo, embora haja um aumento de dados guardados, grande parte desses dados não são utilizados. Por exemplo, para os operadores das ETAR, as ferramentas tradicionais de análise e visualização de dados não capturam toda a informação relativa aos tratamentos. Já para o analista de uma ETAR, os dados recolhidos nunca são suficientes, devido a vários fatores como o custo de fazer análises suplementares ou de realizar algumas medições que por vezes não são exequíveis. Assim sendo, a aplicação de métodos modernos de análise e exploração de dados, particularmente técnicas de DM, pode ajudar tanto os operadores como os analistas das ETAR. Se por um lado possibilita a extração de informação oculta dentre os dados, que pode ajudar os operadores a otimizar a eficácia do tratamento. Por outro lado, não só maximiza o uso dos dados que estão disponíveis, como pode possibilitar a obtenção de mais dados de forma menos dispendiosa, realizando assim um conjunto de medições mais completo (Dürrenmatt, 2011).

Além da motivação para o uso de técnicas de DM devido ao grande volume de informação, e à dificuldade inerente de tirar todo o partido dos dados, existem outras dificuldades no controlo do processo de tratamento. Aqui, a aplicação de técnicas de DM também pode auxiliar as ETAR a contornar algumas dessas dificuldades. Segundo Belanche et al. (1999a), as principais dificuldades do controlo operacional de uma ETAR passam pelo seguinte:

- a própria natureza biológica do processo de tratamento que é bastante diversificada;
- a variabilidade e complexidade da composição das águas brutas que dão entrada nas ETAR;
- a falta de sensores disponíveis e os atrasos de algumas análises feitas em laboratórios;
- todo o dinamismo que envolve o processo de tratamento de águas residuais.

Perante estas dificuldades, várias tarefas de DM poderão ser aplicadas nas ETAR com o objetivo comum de melhorar a eficácia do processo de tratamento. Em tarefas de análise da atividade da planta, poderão usar-se técnicas de DM denominadas descritivas (i.e. segmentação). Através deste tipo de técnicas podem, por exemplo, ser agrupados registos que correspondam a um determinado estado da planta de tratamento — estados como dia chuvoso, problemas nas lamas etc. Esses estados poderão ser atribuídos pelo analista da ETAR, assim que este observe as características dos agrupamentos descobertos. Desta feita, é facilitada a compreensão do complexo processo de tratamento, tornando assim a visualização e interpretação dos dados mais intuitiva.

Contudo, muitas das aplicações de DM nas ETAR consistem em técnicas de modelação preditiva (i.e. classificação e regressão). Estas técnicas variam essencialmente no tipo de dados do atributo de saída (previsão), sendo este um valor nominal ou numérico. Um exemplo típico de classificação é o da previsão das condições graves que podem ocorrer durante o processo de tratamento das ETAR. Isto, através de um modelo previamente desenvolvido, onde perante novos dados é classificado o respectivo estado da ETAR. Esta tarefa é bastante útil, por exemplo, na análise das variadas águas que dão entrada nas ETAR e dos efeitos estas poderão reproduzir no tratamento. Mas não só, mesmo em fases de tratamento posteriores poderão ser detectadas falhas no tratamento.

A modelação preditiva através de técnicas de regressão, é a tarefa de DM mais frequente em trabalhos deste domínio. Este factor pode ser justificado pelo facto que as medições realizadas durante as várias fases tratamento, inclusive os parâmetros de qualidade finais, tratarem-se valores quantitativos. Torna-se assim interessante efetuar uma modelação que contemple a complexidade das ETAR e a traduza numa previsão da concentração dos parâmetros. Parâmetros esses que, no fundo, refletem a eficácia do processo de tratamento efetuado, e que podem ser referentes a fases de tratamento intermediárias, ou então referentes ao efluente final. Começando pelo caso dos parâmetros intermédios, cujas medições provêm de sensores, surge uma das tendências de previsão do comportamento das ETAR que consiste no desenvolvimento dos chamados "*sensores por software*". Segundo Dürrenmatt (2011), entende-se sensores por software como, simplesmente, uma parte de software que dá como saída um sinal quantitativo, em função de um modelo com vários sinais de entrada provenientes de outras medições

realizadas. Basicamente é efetuada uma tarefa de regressão. Esta é uma abordagem muito interessante, pois existem sensores muito dispendiosos que poderão assim ser parcialmente substituídos por este tipo de sistemas. Caso esta aplicação seja bem sucedida, poderá proporcionar às ETAR mais informação em tempo útil (sem atrasos), auxiliando assim os analistas das ETAR em recolher mais dados e atenuando algumas dificuldades acima descritas.

De forma similar, podem ser aplicadas técnicas de regressão na previsão do parâmetros relativos ao efluente final. Na verdade, esta é a tarefa de previsão mais comum nos trabalhos realizados neste domínio. Como se trata da previsão do resultado final do tratamento e, sendo esse resultado a medida de qualidade das águas que serão descarregadas, é interessante prever essa qualidade de forma precisa. Desta forma, é possível medir a eficácia dos tratamentos efectuados e assim obter informação útil para um melhor controle da planta de tratamento. Adicionalmente, sabendo que alguns parâmetros são regulamentados por lei, a previsão das suas concentrações permite ainda descobrir se o tratamento está em conformidade com a legislação em vigor ou não.

Em jeito de conclusão, a aplicação de técnicas de DM poderá ser bastante útil às ETAR, mesmo naquelas que não disponham de sistemas de informação avançados como o SCADA ou o TELEMAC. Também nessas ETAR as técnicas de DM poderão trazer vantagens ou até motivar para um investimento em sistemas de informação adequados. Este é o caso da ETAR em estudo, cujos registos constam de simples folhas de cálculo com os registos mensais dos tratamentos efectuados. Em todo o caso, fica aqui retratada a ideia que as técnicas de DM podem ser uma mais valia para estas infraestruturas, auxiliando-as em vários aspectos, sempre com o objectivo de melhorar e facilitar o controle do tratamento das águas residuais.

2.3.2 Trabalhos de Previsão Realizados no Domínio

Segundo Dürrenmatt (2011), com uma pesquisa na BD da *ISI WEB Knowledge* sobre os artigos que estão relacionados com águas residuais, podemos retirar algumas conclusões interessantes. De referir que foi efectuada uma pesquisa por "*Data Mining*" ou "*Data Driven Modelling*", ambos associados ao termo águas residuais "*Wastewater"*, no título, resumo ou palavras-chave dos

artigos. Verificou-se que estes artigos são apenas 0,05% do total de artigos sobre águas residuais. A mesma pesquisa, mas para os últimos 5 anos, foi realizada, onde se observou que o rácio passou a ser de 0,1%. Embora se esteja a pôr de lado alguns artigos que mencionem técnicas de DM em específico, não referindo portanto os termos gerais, é interessante reparar que a percentagem de trabalho realizado é bastante pequena. Porém, é igualmente interessante verificar que o trabalho realizado nos últimos anos tem vindo a aumentar.

As publicações científicas que abordam a temática de DM ou de previsão do comportamento das ETAR, são, de uma maneira geral, trabalhos realizados ao longo da última década. Procurando não ser-se demasiado exaustivo, serão abaixo descritos alguns trabalhos acerca da aplicação de técnicas de DM nas ETAR. Estes trabalhos têm como objetivo investigar e aplicar algumas das tarefas descritas na secção anterior, isto variando o tipo de métodos e de algoritmos adotados nos vários estudos.

Em tarefas de modelação descritiva são usadas principalmente técnicas de segmentação (do Inglês, *Clustering*), muitas vezes aplicadas como um processo que antecede as tarefas de classificação. No trabalho de Comas et al. (2001), a segmentação é usada para que posteriormente os especialistas das ETAR identifiquem os grupos que resultam da aplicação desta técnica. Deste modo, são atribuídos pelos especialistas estados acerca de condições físicas e biológicas da planta de tratamento. Exemplo de algumas condições físicas são o estado do tempo, o tipo e quantidade de afluente bruto de entrada, a estação do ano etc. Quanto às condições biológicas, são exemplos de estados a presença abundante ou falta de determinados microrganismos, assim como os efeitos que estes reproduzem nos processos de tratamento — este exemplo envolve a utilização de dados qualitativos. Ainda neste trabalho, após serem identificados os agrupamentos, são realizadas tarefas de classificação através de árvores e regras. As regras são do tipo condicional (*Se c1 E c2 Então r1*), e neste caso geradas através do algoritmo CN2. Na classificação com árvores de decisão a tarefa é similar às regras de classificação, com a vantagem de se obter uma visualização gráfica e mais intuitiva das regras. De sublinhar que, embora estas tarefas sejam de previsão, também proporcionam a descrição dos dados. Daí que por vezes sejam considerados modelos mistos, ou seja, existem modelos de previsão que são descritivos e vice-versa. Parte do estudo de Atanasova

& Kompare (2002a) também aplica este procedimento, segmentação seguida de classificação, no entanto são utilizadas apenas árvores de decisão na tarefa de classificação.

Passando aos trabalhos de investigação que aplicam somente algoritmos de classificação, no estudo de Cărbureanu (2010) são utilizadas árvores de decisão com o objetivo de determinar e analisar a poluição de um rio onde é descarregado o efluente final após o tratamento da ETAR local. Para além de classificação com árvores de decisão, são também utilizados algoritmos como *K-Nearest-Neighbor (k-NN)*, Redes Neuronais (do Inglês, *Neural Networks* - NN) e SVM. Belanche et al. (1999b) utiliza no seu estudo métodos de NN e k-NN, explorando a influência dos dados qualitativos no desempenho de uma ETAR. Enquanto que Hong et al. (2008) investiga métodos SVM na classificação do estado da ETAR, com a particularidade de serem também exploradas técnicas de seleção de atributos.

Os trabalhos de investigação, cujos objectivos passam pela previsão da concentração dos parâmetros de qualidade, são predominantes nesta área. Normalmente são previstos os parâmetros referentes ao efluente final como o CBO, CQO, SST e alguns nutrientes. Nestas tarefas de regressão também podem ser utilizados métodos baseados em árvores, nomeadamente árvores de regressão. Nos estudos de Atanasova & Kompare (2002a, 2002b) a previsão de vários parâmetros é feita usando árvores de regressão, embora no primeiro trabalho sejam também apresentados modelos de classificação e de segmentação, tal como já foi mencionado. Devido à complexidade do processo de tratamento das ETAR, é comum usarem-se modelos de previsão não lineares. Em alguns trabalhos são até comparados os modelos lineares com não lineares, em que é mostrada a superioridade dos modelos mais avançados como as NN (Gallop et al. 2004; Dixon et al. 2007). Acrescentar que estes dois trabalhos têm como objectivo auxiliar o sistema de deteção de falhas do TELEMAC e incluem ainda tarefas de DM descritivas. No trabalho de Hamed et al. (2004) também são utilizadas as NN para regressão, porém, neste caso, são previstas as quantidades de CBO e SST. O estudo de Luo et al. (2009) compara a eficácia das NN testando diferentes abordagens de seleção de atributos, isto na previsão de CQO e mais dois nutrientes.

É perceptível que grande parte dos trabalhos de previsão utilizam técnicas de regressão baseadas em NN. Contudo surgem mais recentemente alguns trabalhos que exploram técnicas baseadas em SVM, como são os casos de: Wang & Chen (2008), que utilizam SVM na previsão de CQO e SST;

Yang et al. (2011), que aplicam a variante de SVM denominada *Least Square Support Vector Machine* na previsão de CQO e onde os resultados obtidos são comparados com as NN; e ainda de Huang et al. (2009), que também utilizam o método LS-SVM na previsão de vários parâmetros de qualidade, mas neste caso são efetuadas duas modelações distintas, a primeira para o estado de tempo de sol e a segunda para tempo chuvoso. Também a tese de Yang (2005) centra-se na aplicação de técnicas SVM de regressão, porém, neste caso, são previstas as concentrações de nitrogénio com recurso a dados simulados, e é também comparada a eficácia de previsão das SVM com as NN.

Para terminar esta passagem pela literatura relacionada com esta dissertação, de mencionar ainda dois estudos um pouco distintos daqueles que foram já apresentados. O primeiro é a tese de doutoramento de Dürrenmatt (2011), que apresenta um estudo abrangente sobre aplicação de técnicas de DM nas ETAR, investigando quatro tarefas distintas: a implementação de sensores por software; a identificação da entrada de descargas industriais na planta de tratamento; a estimativa das distribuições das descargas nos pontos de separação de caudais; e a modelação automática dos reatores hidráulicos. O segundo é o trabalho de Poch et al. (2000), que consiste no desenvolvimento e implementação de um sistema integrado de supervisão e monitorização de uma ETAR. Contrariamente à generalidade dos trabalhos que consistem em estudos teóricos sobre a aplicação de técnicas de DM nas ETAR, este trabalho centra-se no estudo dos problemas práticos inerentes à implementação e integração destas técnicas num caso real. São portanto apresentados diagramas com as arquiteturas de software necessárias à integração deste software na planta de tratamento, bem como fluxogramas dos modelos desenvolvidos etc.

Olhando para o conjunto de trabalhos aqui apresentados, constata-se que diferentes abordagens poderão ser adotadas, porém todas elas com o intuito de obter-se mais conhecimento sobre a plataforma e seus processos de tratamento. Fica também aqui descrito o estado da arte desta temática, uma vez que estas investigações ilustram bem os procedimentos utilizados na aplicação de técnicas de DM nas ETAR.

2.4 Sumário

Com o passar do tempo os recursos hídricos estão a tornar-se cada vez mais escassos, devido principalmente ao aumento do seu consumo por parte da população. Adicionalmente, alguns locais que não dispõem de sistemas de saneamento básico, têm uma elevada taxa de mortalidade devido a doenças relacionadas com a água. Com base neste factos, é bem perceptível a necessidade da presença sistemas de tratamento de águas residuais para purificar as águas, permitindo até reutilizá-las. A variedade de constituintes presentes nas águas residuais, faz com que diferentes fases de tratamento sejam processadas, cada uma vocacionada para tratar certos constituintes. Ao longo destas fases são realizadas medições de parâmetros, quer através de análises ou de sensores de medição. Os parâmetros representam assim os constituintes das águas residuais, mas também podem indicar outras características, como a temperatura, turbidez, pH, etc.

Com a evolução dos instrumentos de medições, é hoje possível dotar as ETAR de sensores mais avançados. No entanto, esta não é uma realidade da maioria das ETAR devido à dificuldade de implantação e aquisição destas componentes. De qualquer forma, com leis cada vez mais restritas para a obtenção de um tratamento eficaz, as ETAR deverão modernizar-se. Daí, a inclusão de sistemas de informação integrados na planta de tratamento que, entre as várias vantagens que oferecem às ETAR, permitem o armazenamento dos registos referentes aos tratamentos realizados. Surge assim uma motivação para a aplicação de técnicas de DM, devido à dificuldade de se utilizar toda a informação registada. Porém, é de realçar, que estas técnicas não só permitem maximizar o uso dos dados, como poderão ainda acrescentar informação útil para as ETAR. Isto, mesmo no caso das ETAR que não disponham destes sistemas de informação. Nesta situação, as técnicas de DM poderão auxiliar as ETAR a contornar as suas dificuldades, como a complexidade do tratamento, a variedade das águas de entrada e os atrasos nas análises. Enquanto os modelos descritivos proporcionam uma melhor visualização e tornam a análise dos dados mais intuitiva, os modelos preditivos, perante dados novos, possibilitam a previsão do respectivo estado da ETAR ou da concentração dos parâmetros de qualidade. Encontra-se na literatura relacionada com este trabalho, várias investigações que abrangem estes dois tipos de modelos, isto, através de um variado leque de técnicas de DM.

Capítulo 3

Data Mining e Support Vector Machines

3.1 Data Mining

Atualmente, é consensual dizermos que vivemos na era da informação. Observamos hoje em dia, uma omnipresença dos computadores em nosso redor, que permitem facilmente registar coisas que antes eram desperdiçadas. A este facto está ainda associado o baixo custo de armazenamento, que justifica o aumento contínuo do volume de informação que se armazena dia após dia. De grosso modo, todos os "passos" que damos, seja no supermercado, nos bancos ou na *World Wide Web* (WWW) são registados. Nota-se claramente que está a aumentar o fosso entre a geração dos dados e a capacidade de os compreender. Não surpreende, pois, o surgimento da necessidade de encontrar padrões nos dados por forma a compreende-los. Na verdade, desde sempre que se procuram padrões, por exemplo, os caçadores que procuram padrões nas migrações das aves ou os agricultores que procuram padrões no crescimento da suas colheitas. No entanto o DM realiza esta procura de forma automática sobre dados electrónicos. Mas até aqui não temos nada de novo, pois vários estatísticos, economistas e engenheiros de comunicações desenvolveram muito trabalho na pesquisa automática de padrões nos dados. O que passou a ser novidade foi sim o aumento vertiginoso de oportunidades de procura de padrões nos dados. Este aumento descontrolado do crescimento das BD, proporcionou o emergir da tecnologia de DM. Deste modo, o DM pode ser definido como o processo de extração de informação implícita,

previamente desconhecida e potencialmente útil dos dados. A ideia passa pela construção de programas computacionais, que executem este processo de procura e descoberta de padrões de forma automática, sobre uma ou mais bases de dados (Witten et al., 2011). Olhando para esta definição, compreende-se as vantagens que esta tecnologia pode trazer na obtenção de conhecimento, particularmente perante grandes volumes de dados. O DM trata-se de uma tecnologia que é aplicada em diversos domínios, como o marketing, vendas, reconhecimento de imagens, economia etc. Entre os diversos campos de aplicações está incluído o domínio da biologia ou do ambiente, em particular as ETAR, que são o foco deste estudo.

3.1.1 Metodologias

A aplicação e integração de técnicas de previsão num caso real, como as ETAR, pressupõe um processo de desenvolvimento iterativo. O processo de descoberta de conhecimento em base de dados (do Inglês, *Knowledge Discovery in Databases* - KDD) foi definido por Fayyad et al. (1996a) como um processo não trivial de identificar padrões nos dados válidos, novos, potencialmente úteis, e que sejam em última instância perceptíveis. Existe alguma controvérsia entre os termos KDD e DM. Na verdade, o termo DM é utilizado pela generalidade da comunidade, estatísticos, investigadores de BD e até pelas empresas que desenvolvem software de gestão de informação. Contudo, Fayyad et al. (1996a) refere que o DM consiste numa das fases do processo KDD.

O processo KDD é composto por 5 fases: seleção de dados, pré-processamento, transformação, mineração/prospecção de dados (DM) e interpretação/avaliação. A fase inicial consiste na seleção de dados, que inclui a aprendizagem do domínio da aplicação e a escolha do conjunto de dados de trabalho. A fase seguinte é o pré-processamento e a limpeza de dados. O terceiro passo, o da transformação de dados, consiste em fazer reduções ou projeções de dados em função dos objectivos da tarefa de DM. A fase denominada "*Data Mining*" no processo KDD, termo que gera alguma discussão no seio da comunidade, consta da aplicação de algoritmos específicos para a extração de padrões dos dados. Este passo consiste na escolha da função de DM, como classificação, regressão ou segmentação, com base no propósito do modelo a ser derivado pelo algoritmo. De sublinhar que a maioria dos trabalhos realizados na investigação de DM focam-se

nesta fase. No entanto, as outras fases são igualmente importantes para se ter sucesso na aplicação destas técnicas. A última fase do processo KDD é a avaliação e interpretação dos padrões descobertos. Possivelmente poderá surgir a necessidade de voltar a passos anteriores, isto, em função dos resultados obtidos (Fayyad et al., 1996a).

Além do processo KDD existem outras metodologias, como a *Cross Industry Standard Process for Data Mining* (CRISP-DM) e a SEMMA (*Sample, Explore, Modify, Model, Assess*) da empresa SAS. Estas duas metodologias são de certo modo similares ao processo KDD e seguem portanto a proposta inicial de Fayyad et al. (1996a). Aqui, apenas será referida, em maior detalhe, a metodologia CRISP-DM (CRISP-DM, 2006), uma vez que foi a selecionada para "conduzir" este projeto. De acrescentar que esta é a metodologia mais utilizada pela comunidade de DM, isto segundo a última votação realizada em (KDnuggets, 2007). Assim, nos capítulos seguintes, todos os passos que foram executados no processo de desenvolvimento deste projeto são descritos seguindo a metodologia CRISP-DM.

O CRISP-DM foi inicialmente desenvolvido em 1996 por três "veteranos" do mercado de DM, na altura ainda uma tecnologia recente e imatura. Surgiu então a ideia desenvolver uma metodologia de desenvolvimento padrão. Esta ideia rapidamente despertou o interesse por parte da comunidade de DM, e, por conseguinte, houve uma partilha de ideias que contribuiu para o refinamento do CRISP-DM. No ano 2000 surgiu a primeira versão do CRISP-DM, resultado da união de esforços das empresas NCR Systems Engineering Copenhagen (EUA e Dinamarca), DaimlerChrysler AG (Alemanha), SPSS Inc. (EUA) e OHRA Verzekeringen en Bank Groep B.V. (Holanda). Refere-se ainda que o CRISP-DM não foi desenvolvido com base teórica, nem segundo processos académicos, mas sim como resultado da experiência obtida através de um conjunto de práticas que foram aplicadas com sucesso em casos reais (CRISP-DM, 2000). As fases que integram o ciclo de vida do CRIPS_DM são:

- **Análise de negócio –** Esta fase centra-se em compreender os objectivos do projeto do ponto de vista de negócio, ou seja, compreender a área ou domínio da aplicação e as características do caso prático onde consiste o problema. Este conhecimento deve ser traduzido depois num plano inicial de projeto e numa definição do problema de DM.

- **Análise de dados –** Nesta fase é feita a recolha inicial de dados e executado um conjunto de procedimentos de exploração de dados, por forma a adquirir conhecimento e familiaridade sobre os dados.

- **Preparação de dados –** Esta fase cobre a aplicação de várias técnicas que são essenciais para obter-se um conjunto de dados final, já pronto para a fase seguinte de modelação. São portanto utilizadas técnicas de tratamento de dados, seleção de atributos entre outras.

- **Modelação –** Nesta fase são aplicadas as técnicas de modelação e, como existem diversas técnicas ou algoritmos de DM, é nesta fase que são selecionadas quais as técnicas a utilizar. Neste passo, são ainda calibrados os parâmetros dos algoritmos para os valores ótimos. De acrescentar que por vezes são utilizadas várias técnicas de modelação, logo, como certas técnicas têm requisitos próprios, pode ser necessário voltar à fase anterior de preparação de dados.

- **Avaliação –** Neste ponto os modelos já foram construídos. Assim, antes de se passar para a fase de implementação, é importante rever os passos que foram dados e assegurar-se que os modelos cumprem os objetivos de negócio propostos.

- **Implementação –** A última fase do CRISP-DM é a de implementação. Já com os modelos desenvolvidos, é importante organizar o conhecimento por forma a apresenta-lo ao utilizador final. Esta fase, dependendo dos requisitos de negócio, pode passar simplesmente pela elaboração de um relatório ou pelo desenvolvimento de uma interface integrada num sistema de suporte à decisão. Por vezes esta última fase é executada pelo cliente mas, em todo o caso, o analista de dados deverá ter um papel importante na implementação. Até porque, é crucial que se realizem as ações necessárias para tirar todo o partido dos modelos criados.

Na figura 3.1 pode-se observar as fases do processo CRISP-DM, segundo um diagrama que mostra o ciclo de vida dos projetos de DM. Olhando para a figura 3.1, verifica-se ainda que este processo de desenvolvimento pressupõe a possibilidade de recuar a fases anteriores. Isto acontece em particular nas fases de análise de dados, modelação e avaliação. Adicionalmente, no manual de referência com a linhas orientadoras desta metodologia (CRISP-DM, 2000), não se encontram apenas as fases do processo e respectivas descrições. Dentro de cada fase, são mencionadas várias tarefas a desempenhar, as documentações que devem ser elaboradas e algumas recomendações que devem ser seguidas durante todo o processo de desenvolvimento.

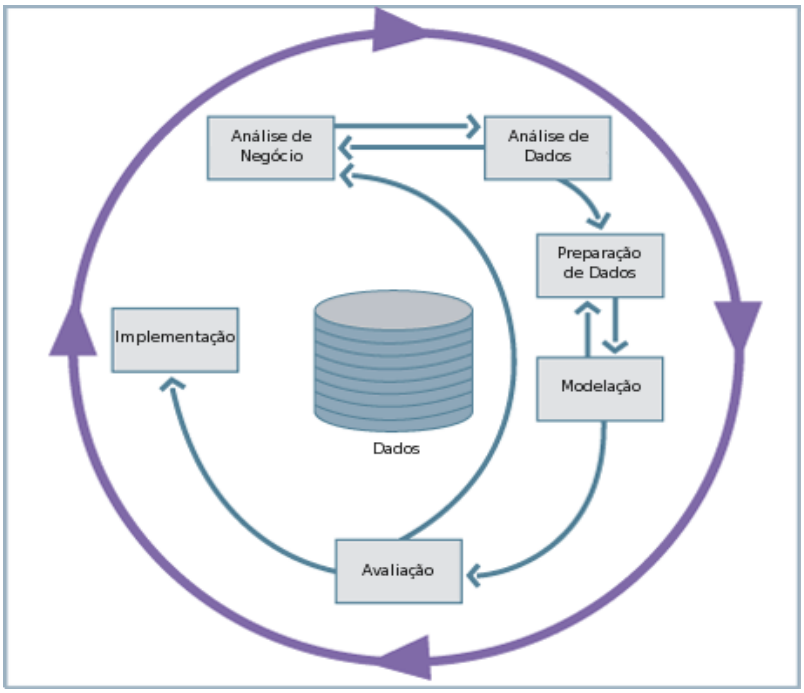


Figura 3.1 - Fases da metodologia CRISP-DM - adaptado de (CRISP-DM, 2000)

3.1.2 Técnicas de Data Mining

Como vimos no processo de desenvolvimento, a fase descrita como modelação consiste na aplicação de métodos ou algoritmos usualmente referidos como técnicas de DM. Os dois principais objetivos de DM são, na prática, a previsão e a descrição. Daí, tal como já foi referido, podemos separar as técnicas de DM em dois tipos de modelações, a descritiva e a preditiva. A modelação descritiva foca-se essencialmente em descobrir padrões descritivos dos dados que sejam interpretáveis pelo homem. Por sua vez, a modelação preditiva consiste na utilização de variáveis presentes nas BD, com o objetivo de prever um valor futuro, ou desconhecido, de outras variáveis de interesse. Contudo, esta separação não é assim tão rígida. Alguns modelos preditivos podem ser descritivos, na medida em que é possível interpretá-los e vice-versa. De qualquer forma torna-

se importante realizar esta distinção, por forma a esclarecer qual o objetivo geral da tarefa de DM a realizar (Fayyad et al., 1996b).

Dentro dos dois tipos de modelação encontramos diferentes tarefas de DM, tal como se pode observar na seguinte listagem onde são exemplificadas algumas das tarefas de DM mais comuns:

- **Modelos Descritivos**
    - Segmentação (*Clustering*)
    - Sumarização
- **Modelos Preditivos**
    - Classificação
    - Regressão

Segmentação

A segmentação, também conhecida como *clustering*, é uma tarefa descritiva bastante comum, em que procura-se encontrar um número finito de agrupamentos naturais, ou categorias, que descrevam os dados. A cada exemplo ou instância é atribuído um agrupamento onde, segundo métricas de similaridade, é escolhido o agrupamento mais "próximo" da instância a segmentar. Adicionalmente, alguns algoritmos de segmentação permitem que uma instância pertença simultaneamente a mais do que um grupo, produzindo assim diagramas de agrupamentos sobrepostos. Este é o caso representado na figura 3.2, onde podemos visualizar três agrupamentos de clientes de um banco. Outros algoritmos são probabilísticos, atribuindo a cada instância a probabilidade de esta pertencer a cada um dos agrupamentos. Existe ainda um outro tipo de segmentação que produz uma hierárquica de agrupamentos (dendrograma), cujos agrupamentos que estão juntos nos níveis mais baixos têm relacionamentos mais fortes (Witten et al., 2011). Frequentemente são utilizadas técnicas de classificação juntamente com a segmentação. Procura-se então classificar novos exemplos, por forma a descobrir a que agrupamento pertencem esses mesmos exemplos. Tal como foi mencionado nos trabalhos de DM nas ETAR, em que esta combinação de técnicas é adotada por vários autores.

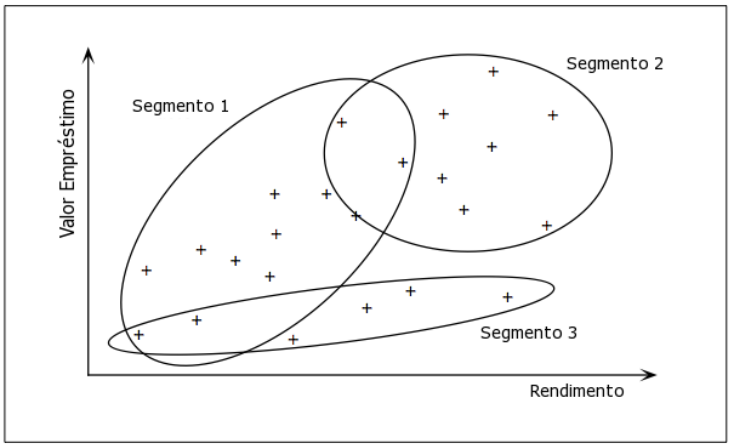


Figura 3.2 - Exemplo de segmentação - adaptado de (Fayyad et al., 1996b)

Sumarização

As técnicas de sumarização, tal como o nome indica, consistem em métodos que procuram resumir ou compactar a descrição de um subconjunto de dados. Alguns exemplos de sumarização são as tabelas com as médias e desvios padrão dos atributos ou as técnicas de visualização de vários atributos. Estas técnicas são muitas vezes utilizadas nas fases de análise e exploração de dados. Ainda dentro da sumarização encontram-se técnicas que derivam regras de associação. As regras de associação, ao contrário das regras de classificação, não procuram classificar apenas um atributo em especifico, mas têm sim a liberdade de prever qualquer um. Como podemos imaginar, bastantes regras de associação diferentes podem ser geradas, onde para cada regra existem valores de suporte e de confiança da regra que indicam as associações mais significativas. Esta é uma técnica típica de modelação descritiva, que é bastante usada em aplicações de DM.

Para terminar esta caracterização dos modelos descritivos, de mencionar ainda mais dois tipos de técnicas descritivas que, segundo Fayyad et al. (1996b), estão fora do grupo de técnicas de sumarização. As primeiras são as técnicas de modelação de dependências, que descrevem os relacionamentos entre variáveis, quer seja através de gráficos ou através de valores numéricos que indiquem a força de relacionamento (dependência). Por fim, as técnicas de deteção de alterações e

desvios nos dados, que consistem em encontrar alterações significativas dos valores dos dados medidos anteriormente.

Classificação

A classificação consiste na aprendizagem de uma função que usualmente faz o mapeamento de uma instância para uma das várias classes pré-definidas. Na figura 3.3, pode-se observar um exemplo de classificação, no qual existem 23 exemplos de pessoas que pediram empréstimos à banca. Neste conjunto de dados, os valores da variável de saída, também conhecidos como classes, são cliente bom pagador ("cliente seguro" - O) e clientes em dívida com o banco ("cliente de risco"- X). De notar que este tipo de dados é categórico ou nominal. Os algoritmos de classificação realizam assim uma separação dos dados, procurando classificar corretamente o maior número de dados possível. Neste caso, na região a cinza os exemplos são classificados como "cliente de risco", não sendo portanto concedido o empréstimo. Por outro lado, na região complementar, os exemplos são classificados como "cliente seguro", indicando que o banco deve conceder empréstimo a estes clientes. Curiosamente verifica-se que usando uma separação linear, como neste caso, não é possível classificar corretamente todos os exemplos, pois existem dois erros de classificação.

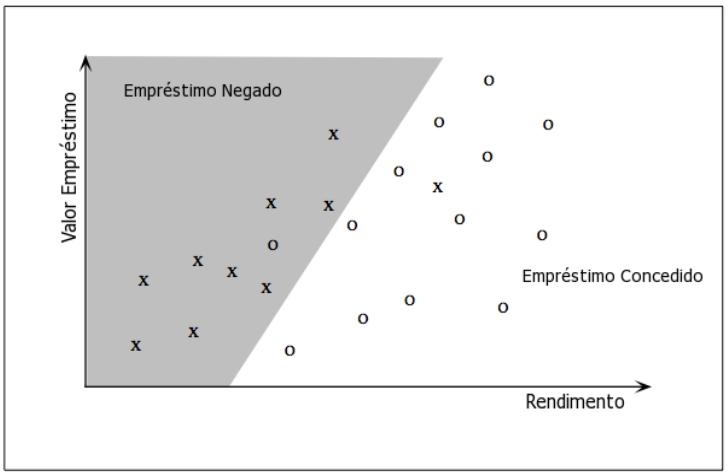


Figura 3.3 - Exemplo de classificação - adaptado de (Fayyad et al., 1996b)

As técnicas de classificação envolvem a aprendizagem de uma função, ou seja, os algoritmos de classificação realizam uma aprendizagem (treino) sobre um conjunto de dados. Com base no valor da variável de saída de cada exemplo são, assim, aprendidos padrões ou regras que permitem desenvolver um modelo de previsão. Desta forma, aplicando este modelo em exemplos novos, está-se a classificar (prever) qual a classe que corresponde ao exemplo novo, isto com base nas características do exemplo a classificar. Tomando o exemplo da figura 3.3, o algoritmo de classificação que foi aplicado "aprendeu" que os clientes com empréstimos de valor mais alto e rendimentos mais baixos são clientes de risco. Caso um novo cliente, cujos valores das variáveis "Valor Empréstimo" e "Rendimento" sejam respectivamente alto e baixo venha a ser classificado pelo modelo desenvolvido, a sua classe será "cliente de risco", logo não será concedido empréstimo. Este processo de aprendizagem é análogo nos problemas de regressão, sendo vulgarmente conhecido como aprendizagem supervisionada.

Na aprendizagem supervisionada os dados contêm variáveis de saída (*labels*) que guiam o processo de aprendizagem. Os modelos são portanto desenvolvidos com base no par de variáveis de entrada (*input*) e de saída (*output*), ao contrário do que acontece por exemplo na segmentação, onde a aprendizagem é não-supervisionada, visto que os modelos são desenvolvidos somente com base nas variáveis de input.

No capítulo 2 desta dissertação descreveram-se alguns trabalhos de previsão em ETAR e as respetivas técnicas de DM que são normalmente aplicadas. Como neste trabalho focamo-nos em tarefas de previsão, e dado que algumas das técnicas já mencionadas consistem em algoritmos de classificação, apresenta-se em seguida as técnicas de classificação mais comuns. A apontar que a caracterização destas técnicas baseiam-se nas definições demonstradas por Witten et al. (2011).

- **Árvores de Decisão –** Os algoritmos baseados em árvores seguem a estratégia "*divide-and-conquer*", ou seja, separar as instâncias mediante certas condições para posteriormente as classificar. As sucessivas divisões realizadas, fazem com que seja criada uma estrutura representativa do modelo em árvore, daí o nome de árvores de decisão. Os nós das árvores de decisão comportam testes lógicos de um determinado atributo, realizando normalmente comparações com uma constante. As folhas das árvores, por sua vez, atribuem uma determinada classificação às instâncias que atingem essas folhas. Sendo assim, para cada instância a classificar é percorrido um trajeto descendente na árvore, respeitando as condições lógicas dos nós, até que seja alcançada uma folha da árvore com a respectiva classe a ser atribuída. Para concluir, de referir que estes

algoritmos oferecem uma grande facilidade interpretação, pois a representação do modelo através de uma estrutura em árvore é bastante intuitiva.

- **Regras de Classificação –** As regras de classificação são uma alternativa às árvores de decisão. Na verdade, as pré-condições das regras correspondem aos testes lógicos dos nós das árvores, sendo que as conclusão correspondem às folhas das árvores com as classes. Dá para perceber que facilmente se podem ler regras diretamente das árvores de decisão. Geralmente as pré-condições das regras são conjuntivas, ou seja, são testes lógicos com o operador (E ; $\wedge$). No entanto, em alguns algoritmos de formulação de regras são geradas expressões lógicas gerais, que incluem o operador (OU ; $\vee$).

- **Classificadores baseados em Instâncias –** Uma das formas mais simples de aprendizagem é a memorização literal das instâncias. Assim, no momento de classificação de uma nova instância, esta será comparada com as instâncias de treino em memória e será classificada segundo a instância que mais se assemelha. Esta é uma abordagem "*lazy*" (preguiçosa), pois deixa todo o trabalho para a fase de teste. As comparações executadas baseiam-se em funções de distância, que medem o grau de semelhança das instâncias. Dentro deste tipo de técnicas, os algoritmos mais conhecidos são os vizinhos mais próximos (do Inglês, *nearest neighbors*). Tal como o nome sugere, a classe atribuída à nova instância a classificar é a classe da instância de treino mais próxima. Uma versão mais avançada deste algoritmo classifica as novas instâncias com base nos k vizinhos mais próximos (*K-Nearest-Neighbor – k-NN*), neste caso, é escolhida a classe maioritária ou a média (no caso numérico) dentre as k instâncias mais próximas.

- **Classificadores Bayesianos –** Até ao momento, todos os tipos de classificadores apresentados realizam uma classificação rígida, ou seja, para cada exemplo novo a classificar é atribuída uma classe. No caso dos classificadores bayesianos, para cada exemplo novo são atribuídas as probabilidades de este pertencer às várias classes, sendo naturalmente escolhida a classe com maior probabilidade. As probabilidades são calculadas com base no teorema de Bayes, e uma particularidade destas técnicas é a assunção "ingénua" de independência entre os atributos. Daí o algoritmo ser vulgarmente chamado de "Naive Bayes". Todavia, apesar de ser assumida a independência dos atributos, quando sabemos que na vida real isso não acontece, a verdade é que o Naive Bayes traz bons resultados na prática.

- **Redes Neuronais (*Neural Networks* - NN) –** As NN evoluíram a partir da regra de aprendizagem do *perceptron*. Este algoritmo mais simplista, consiste em realizar várias iterações até que seja encontrado o híper-plano que separa todas as instâncias sem erros, ou seja, que as classifique todas corretamente. Sabendo que na equação do híper-plano são atribuídos pesos aos demais atributos do conjunto de dados, sempre que uma instância é mal classificada o algoritmo altera os pesos, movendo assim o híper-plano de forma a classificar corretamente essas instâncias. Neste caso mais simples, apenas com dados linearmente separáveis é que se alcança a solução ótima, caso contrário o algoritmo não terminará. Pode dizer-se que o híper-plano resultante é chamado de perceptron, na figura 3.4 podemos verificar a representação do perceptron como uma rede de neurónios.

Repare-se que apenas existem duas camadas, uma de input e outra de output, e ainda que para cada atributo (*a)* existe um peso associado (*w)*.

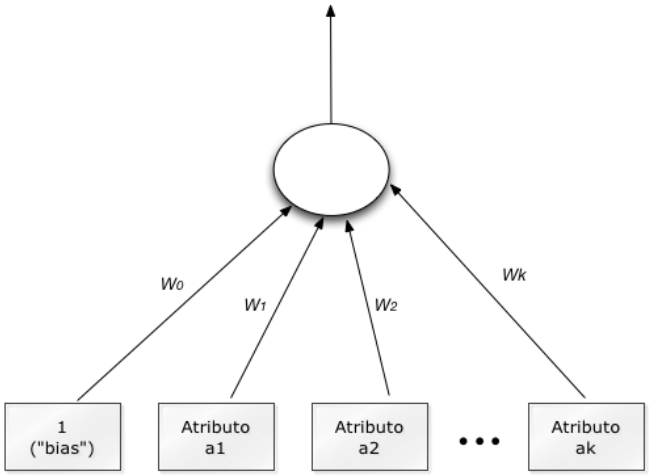


Figura 3.4 - Perceptron representado como uma rede neuronal - adaptado de (Witten et al., 2011)

A palavra neurónios (ou neurónios artificiais) é atribuída a este tipo de técnicas, uma vez que a sua representação é inspirada pela grande entre ligação dos neurónios do cérebro humano. Por exemplo, nós somos capazes de realizar inúmeras tarefas de classificação no dia-a-dia ao identificarmos os objetos que vemos, no entanto um simples neurónio não o faz. Assim, só com um conjunto de neurónios, onde cada um realiza tarefas simples, é que permite que no seu todo sejam realizadas tarefas mais complexas. Isto acontece precisamente nas NN, como vimos, o caso mais simples do perceptron perante dados que não sejam linearmente separáveis não encontra solução. Surge então a necessidade de estender este modelo linear para um patamar não linear, para isso são adicionadas várias camadas de neurónios. O algoritmo *Multilayer Perceptron Neural Networks (MLP-NN*) implementa exatamente essa ideia, para além das camadas de input e output, são adicionados neurónios numa camada denominada oculta (*Hidden Layer*). De referir que esta camada oculta pode mesmo conter várias subcamadas, ao contrário das camadas de input e output que são únicas. Em relação à arquitetura demonstrada na figura 3.4, a grande diferença passa pela adição de uma camada intermédia com os respectivos nós ocultos. Quanto à aprendizagem nas MLP, estas diferem da regra de aprendizagem do perceptron mencionada acima, pois neste caso temos nós ocultos. Nas NN, o algoritmo de aprendizagem normalmente adotado é o *Backpropagation.* De grosso modo, a solução passa por modificar os pesos das ligações que passam pelos nós ocultos, com base na força da contribuição que cada unidade tem para a previsão final. Para atingir esse objetivo é usado o algoritmo matemático de otimização denominado Gradiente Descendente. Adicionalmente, em tarefas de classificação são usadas funções de conversão da soma dos pesos em outputs categóricos (i.e. 0 e 1 na classificação binária),

como é o caso da função sigmoide. Segue-se portanto o processo de otimização da rede, onde é calculado o erro de classificação em cada iteração e são ajustados os pesos até se atingir o erro mínimo. De notar que nas NN, os algoritmos do gradiente descendente apenas encontram os mínimos locais e, no caso da função ter vários mínimos, a solução encontrada poderá não ser a ótima. Esta é uma grande desvantagem das NN, mas não só, ao contrário por exemplo das árvores, o modelo representativo das NN não é nada intuitivo, aliás é até referido como sendo uma "caixa-negra". De referir ainda que existem vários tipos de NN, o exemplo aqui descrito pertence à classe de NN chamadas de *feed-forward,* pois não têm ciclos e dependem apenas dos dados de input. Dentro este tipo de NN existe ainda outra técnica muito popular denominada *Radial Basis Function Network.* Mais haveria a dizer sobre este tipo de técnicas, no entanto este não é o foco desta dissertação. Todavia, esta técnica ficou aqui descrita de forma mais aprofundada que as restantes técnicas de classificação, devido não só ao facto das NN serem várias vezes comparadas com as SVM e vice-versa, mas também ao facto de serem a técnica de eleição em vários trabalhos de previsão nas ETAR. Por causa, naturalmente, dos bons resultados preditivos que se obtém com as NN.

- **Support Vector Machines (SVM) –** As SVM são muito usadas em classificação, aliás, originalmente, este tipo de algoritmos foi desenvolvido para desempenhar tarefas de classificação binária. Resumidamente, as SVM para classificação consistem num problema de maximização da margem do híper-plano que separa as classes, sendo que as instâncias mais próximas da margem são chamadas de vetores de suporte. Como esta dissertação centra-se neste tipo de técnicas, e é até objetivo do projeto explorar as SVM, voltaremos a abordar esta técnica em maior detalhe na secção 3.2.

Todos estes algoritmos, bastante diferentes entre si, com variadas formas de representatividade do conhecimento obtido e também diferentes abordagens de aprendizagem, têm no entanto em comum a habilidade de prever as classe dos exemplos novos a classificar. Existe mais informação acerca destes algoritmos que não foi aqui apresentada. Porém, mais detalhes, bem como referências para trabalhos com discussões mais amplas sobre este tipo de técnicas, podem ser encontrados e.g. no livro de Witten et al. (2011).

Regressão

A regressão consiste na aprendizagem de uma função que mapeia uma instância para um valor real. Como se pode constatar, a regressão distingue-se da classificação no tipo de dados da variável de saída. Em classificação é realizada a previsão de valores nominais, enquanto que na regressão é realizada previsão de valores numéricos. Também nos problemas de regressão é realizada aprendizagem supervisionada, pois o algoritmo "aprende" com base num conjunto de

treino com variáveis de entrada e de saída. Posteriormente, na fase de teste, poderá prever-se então o valor da variável de previsão dos exemplos novos.

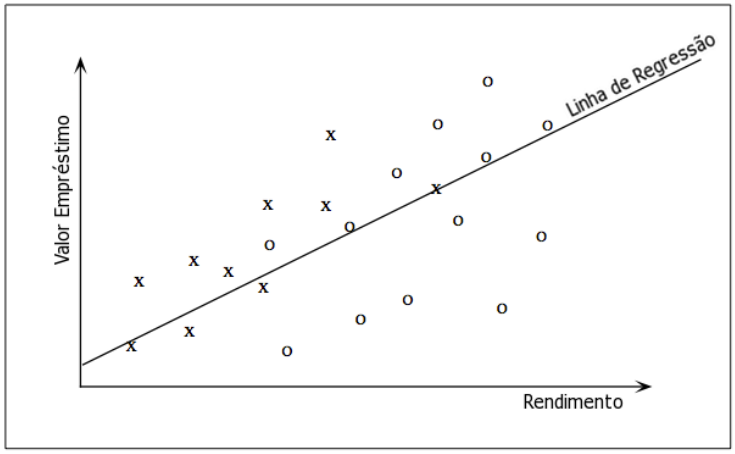


Figura 3.5 - Exemplo de regressão - adaptado de (Fayyad et al., 1996b)

Na figura 3.5 podemos observar um exemplo simples de regressão, no qual a linha de regressão procura prever o atributo "Valor do Empréstimo" com base no atributo "Rendimento" dos clientes do banco. Neste caso, olhando para a distribuição dos exemplos, vê-se claramente que estes estão muito afastados da linha de regressão. Daí que este modelo de regressão linear não preveja bem o valor do empréstimo. Aliás, como temos apenas uma variável de entrada e uma de saída, podemos até constatar que existe uma fraca correlação linear entre estas duas varáveis. De qualquer forma dá para verificar que existe um aumento do atributo "Valor de Empréstimo" quando o atributo "Rendimento" é maior. O que faz sentido na realidade, e mostra que a função de regressão ajusta-se aos dados de treino.

A regressão linear consiste simplesmente numa função que procura estimar o valor do atributo a prever, segundo uma combinação linear de atributos com pesos associados (Witten et al., 2011).

$$y = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k \qquad \textbf{(3.1)}$$

Como podemos observar na equação acima apresentada (y) é o valor a prever, onde (a_1, a_2) é o valor dos atributos do exemplo novo e w corresponde os pesos pré-determinados na fase de treino. Para pré-determinar os pesos é realizado um cálculo a partir dos dados de treino. Normalmente, na regressão linear é minimizada a soma do quadrado dos erros, procurando assim as melhores constantes para os pesos de forma a obter o mínimo erro de previsão.

$$\sum_{i=1}^{n}\left(x^{(i)} - \sum_{j=0}^{k} w_j\, a_j{}^{(i)}\right)^2 \qquad \textbf{(3.2)}$$

Na equação 3.2, é demonstrada a função que mede a soma do quadrado dos erros. Basicamente para *n* instâncias de treino, denotadas com *(i)*, é calculada a diferença entre o valor real da instância (*x*) e o valor de previsão da classe. Como se pode perceber, o cálculo do valor de previsão da instância (*y*), consiste no somatório que foi apresentado na equação 3.1. Desta forma, sabendo o valor do erro de previsão, bastará resolver o problema de minimização por forma a encontrar os melhores coeficientes (*w*). Os métodos de regressão linear parecem assim bastante bons para previsão numérica e também muito simples. No entanto, o seu grande problema é a linearidade. Aqui, quando se está perante um conjunto de dados cujos relacionamentos entre atributos são complexos, os modelos lineares não conseguem "capturar" esses relacionamentos ditos não-lineares.

De forma análoga à mencionada no casos das NN, onde os modelos lineares não conseguem classificar corretamente dados não separáveis, também em problemas de regressão não é possível ajustar a função de previsão corretamente quando se está perante dados complexos. Existem vários algoritmos de regressão não linear, alguns deles surgem com base no chamado "*Kernel Trick*", como é o caso dos algoritmos SVM onde esta técnica teve origem. De grosso modo, é realizado um mapeamento do espaço real para um espaço de maior dimensionalidade, onde a função de decisão linear no espaço multidimensional corresponde a uma função não linear no espaço de input. Voltaremos a este tema em breve, nomeadamente na secção 3.2 onde serão detalhados os algoritmos SVM.

As NN, tal como as SVM, também são algoritmos de modelação não linear. O seu funcionamento é muito parecido com o descrito já para a classificação, sendo apenas alteradas as funções de

ativação como a sigmoide, que servem para transformar o output em probabilidade de classes. De referir ainda que uma rede neuronal sem nós ocultos (perceptron), é na verdade bastante semelhante ao método de regressão linear.

Existem ainda algoritmos baseados em árvores que são muito usados em tarefas de regressão. Esses algoritmos são denominados árvores de regressão e o seu funcionamento é exatamente igual às árvores de decisão na classificação, excepto nas folhas das árvores. Nesta situação, ou as folhas contêm o valor médio da classe das instâncias de treino que atingiram a folha (árvores de regressão), ou contêm modelos lineares que preveem o valor numérico da classe da instância que atinge a folha (árvores de modelos de "regressão"). De notar, que este tipo de algoritmos é intrinsecamente não linear, pois para cada folha da árvore são calculados diferentes modelos de regressão. Mediante o caminho percorrido na árvore, que depende das características da instância a prever, será aplicado o modelo de regressão que melhor se adapta a essas instâncias. Desta forma, percebe-se que este é um bom algoritmo para dados complexos, uma vez que as divisões dos dados por vários modelos lineares permitem capturar os relacionamentos não lineares dos dados e, consequentemente, o comportamento da variável dependente (Atanasova & Kompare, 2002b). Adicionalmente têm a vantagem de ser um modelo representativo bastante intuitivo, à imagem das árvores de decisão. Existem outros algoritmos de regressão não linear. Porém os três algoritmos aqui descritos são os mais utilizados, em particular nos trabalhos de investigação no domínio desta dissertação.

3.1.3 Dificuldades Típicas de Data Mining

Na aplicação de técnicas de DM, existem algumas dificuldades que são típicas do processo de desenvolvimento em questão. Essas dificuldades representam ao mesmo tempo desafios que os analistas de dados procuram superar na elaboração dos seus projetos. Apresentam-se abaixo alguns dos problemas mais comuns em DM segundo Fayyad et al. (1996b).

- **Grandes Bases de Dados** – As BD com centenas de atributos, muitas tabelas, milhões de registos, vários gigabytes/terabytes de tamanho representam um problema, visto que é bastante difícil lidar com toda esta informação. Alguns métodos comuns para contornar

este problema são a amostragem dos dados, processamento paralelo e a utilização de algoritmos mais eficientes em termos de desempenho computacional.

- **Alta Dimensionalidade** – Não existem só grandes bases de dados com um enorme número de registos, existem também BD com um grande número de atributos, o que implica um problema de alta dimensionalidade. A alta dimensionalidade dos conjuntos de dados criam dificuldades na procura de "padrões" pelos algoritmos de DM. Pois, a complexidade é aumentada de uma forma combinatória explosiva. Este problema aumenta também a probabilidade dos algoritmos encontrarem padrões supérfluos. Técnicas de seleção de atributos e um bom conhecimento prévio do problema, por forma a eliminar atributos irrelevantes, são algumas das abordagens adotadas neste tipo problema.

- **Sobre Ajustamento (*Overfitting*)** – Quando um algoritmo procura os melhores parâmetros para um determinado modelo, usando um conjunto de dados limitado, ele pode modelar não só o padrões gerais, mas também o ruído específico do conjunto de dados de treino. Quando isto acontece, os modelos ficam sobre ajustados aos dados de treino e, quando esse modelo é aplicado na previsão de exemplos novos, obtêm-se maus resultados. Métodos de avaliação de modelos como a validação cruzada, técnicas de regularização ou outras estratégias estatísticas podem ajudar a evitar este problema.

- **Valores Nulos e Ruído nos Dados** – Muitos registos presentes nas BD estão repletos de erros, inclusive alguns atributos contêm grandes taxas de erros. Outro problema é o dos valores nulos, visto que muitos formulários são submetidos com campos vazios. Em DM este é um grande problema, pois há algoritmos que não têm capacidade de lidar com valores nulos. As soluções para estes problemas passam por aplicar técnicas comuns à fase de preparação de dados - algumas dessas técnicas serão descritas no capítulo seguinte.

- **Interação do Utilizador e Conhecimento Prévio** – Muitas das técnicas de DM não são muito interativas e nem oferecem a possibilidade de introduzir conhecimento prévio. Este é um problema, uma vez que a aplicação de conhecimento do domínio do problema é muito importante em todas as fases de desenvolvimento.

- **Integração com Outros Sistemas** – Um sistema de descoberta de conhecimento por si só pode não ser muito útil, por isso, a sua integração com outros sistemas de informação é de mais valia. Algumas dificuldades passam pela integração com os SGBD, com ferramentas de visualização, com folhas de cálculo, e também a integração com sensores em tempo real.

Algumas destas dificuldades surgiram, como é natural, no desenvolvimento deste trabalho. A descrição desses problemas, assim como os métodos que foram adotados para os ultrapassar serão descritos ao longo deste documento.

3.2 Support Vector Machines

Olhando para a designação "Support Vector Machines", esta sugere que estamos perante um qualquer tipo de "máquinas". Porém esse é um termo enganador, pois as SVM são algoritmos e não máquinas. Todavia, como esta é uma técnica do campo da aprendizagem máquina (*Machine Learning*), esse termo justifica-se facilmente. A aprendizagem máquina consiste basicamente na capacidade de um computador aprender segundo um dado conjunto de dados. De reparar que este conceito é utilizado pelos algoritmos que foram anteriormente descritos neste capitulo, daí que o DM e a aprendizagem máquina estejam interligados. Na verdade, as técnicas de DM que foram descritas são chamadas técnicas de aprendizagem máquina, enquanto que o DM é o tópico que envolve a aprendizagem na prática. Melhor dizendo, o DM interessa-se em utilizar técnicas para descobrir e descrever padrões dos dados, como uma ferramenta que ajuda a descrever esses dados e a realizar previsões a partir destes (Witten et al., 2011).

As SVM fundamentam-se na teoria de aprendizagem estatística, que é também conhecida por teoria VC (Vapnik e Chervonenkis). Esta começou a ser desenvolvida no fim da década de 1960 até meados dos anos 1990, por Vapnik e Chervonenkis (1974) e Vapnik (1982, 1995), conforme citado em (Smola & Schölkopf, 2004). No início dos anos 90, foi proposto por Vapnik e colaboradores um novo algoritmo de aprendizagem baseado na teoria VC (SVM). Deste modo, as SVM surgiram inicialmente como uma técnica para classificação, tendo sucesso em várias aplicações, particularmente no domínio do reconhecimento de padrões e.g. (Boser et al., 1992; Cortes & Vapnik, 1995). Contudo, pouco tempo depois, as SVM começaram a ser aplicadas em problemas de regressão, e.g. (Drucker et al., 1997).

3.2.1 Teoria de Aprendizagem Estatística

Resumidamente, a teoria de aprendizagem estatística caracteriza as propriedades das máquinas de aprendizagem, que as permite generalizar bem perante novos dados (Smola & Schölkopf, 2004). Antes de avançarmos para uma descrição da teoria VC, deve ser abordado o conceito de generalização. Na aprendizagem, particularmente a supervisionada, o problema consiste em escolher a função que preveja a resposta do supervisor da forma mais acertada possível (Vapnik, 1999). Nas tarefas de aprendizagem supervisionada, os algoritmos têm de ter um bom desempenho de generalização para obter bons resultados. Para se alcançar uma boa capacidade de generalização, tem de haver um equilíbrio entre a precisão obtida ao prever as instâncias de treino de um conjunto de dados específico e da "capacidade" do algoritmo aprender com qualquer outro conjunto de dados que venha a surgir. Um exemplo que retrata bem esta situação é a seguinte. Um algoritmo com muita capacidade é como um botânico com memória fotográfica, que perante uma nova árvore diz que essa não é uma árvore, pois o número de folhas é diferente de todas as árvores que ele já viu. Por outro lado, um algoritmo com pouca capacidade é como o irmão preguiçoso do botânico, que declara que qualquer coisa que seja verde é uma árvore (Burges, 1998). Esta ideia descreve bem o problema que motivou o desenvolvimento das SVM e que aparece em diferentes guias referido como sobre ajustamento (Montgomery & Peck, 1992), capacidade de controle (Guyon et al., 1992) ou ainda como um compromisso entre viés e variância (Geman, Bienenstock & Doursat, 1992), conforme citado em (Burges, 1998). Como já foi referido esta é uma dificuldade típica de DM, tendo a exploração e formalização destes conceitos resultado no ponto forte da teoria de aprendizagem estatística.

Problema de Minimização do Risco

Assumindo que existe uma distribuição de probabilidade desconhecida $P(\mathbf{x}, \mathrm{y})$ de um conjunto de dados de treino, e sendo esses dados independentes e identicamente distribuídos, vamos supor uma máquina de aprendizagem que faz o mapeamento $\boldsymbol{x_i} \mapsto y_i$. Isto considerando $\mathbf{x}$ como o vector de atributos dos dados de entrada e *y* como o valor de previsão da classe. A máquina pode ainda ser definida pelo conjunto de mapeamentos $\boldsymbol{x_i} \mapsto f(\boldsymbol{x}, \alpha)$, em que α corresponde aos

coeficientes da função (i.e. pesos ω na regressão linear). O erro esperado da máquina é portanto definido pela seguinte equação (Burges, 1998):

$$R(\alpha) = \int \frac{1}{2} |y - f(\boldsymbol{x}, \alpha)| \, dP(\mathbf{x}, \mathrm{y}) \quad \textbf{(3.3)}$$

A quantidade $R(\alpha)$, que representa o risco esperado ou risco atual, é o valor de erro que mais interessa para medir se um algoritmo tem uma boa capacidade preditiva. O objetivo passa assim por encontrar a função que minimize o risco atual, contudo não é possível faze-lo diretamente, uma vez que em geral a distribuição de probabilidade $P(\mathbf{x}, \mathrm{y})$ é desconhecida.

O risco empírico (R_{emp}), por sua vez, consiste na medida do erro de aprendizagem que é relativo às instancias de treino:

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^{l} |y_i - f(\boldsymbol{x_i}, \alpha)| \quad \textbf{(3.4)}$$

Repare-se que, neste caso, não está presente $P(\mathbf{x}, \mathrm{y})$, logo o valor de R_{emp} é um número fixo que depende dos valores de α e do conjunto de treino $\{\boldsymbol{x_i}; y_i\}$. A componente $|y_i - f(\boldsymbol{x_i}, \alpha)|$ é a chamada perda e, se repararmos bem, estamos perante uma função de cálculo do erro semelhante à demonstrada na equação 3.2 (soma do quadrado dos erros). Todavia, neste exemplo vamos assumir que ela apenas assume os valores 0 e 1, e escolhemos um valor η que tome valores dentro desse intervalo. Então, para perdas entre esses valores com probabilidade de $1 - \eta$, têm-se o seguinte limite:

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}\right)} \quad \textbf{(3.5)}$$

Sendo que l é o número de exemplos do conjunto de dados, e em que *h* é o inteiro chamado de dimensão Vapnik e Chervonenkis (VC). A dimensão VC trata-se portanto do termo que mede a noção de capacidade descrita acima. Todo o lado direito da inequação é denominado de "limite do risco", já a segunda parcela do lado direito da inequação é conhecida pela "confidência VC". Como já vimos, $R(\alpha)$ não pode ser calculado diretamente mas, sabendo *h*, pode ser calculado o lado direito da inequação (3.5) (Burges, 1998).

A dimensão VC é uma propriedade do conjunto de funções F, no qual o valor *h* corresponde ao número máximo de exemplos de treino que podem ser separados, isto seguindo o exemplo de classificação binária. Para exemplificar este conceito, sabe-se que tendo uma dimensão VC de valor *h*, então pelo menos *h* pontos podem ser separados. No caso de uma função linear percebe-se que uma recta consegue separar 3 pontos, independentemente das $(2^3 = 8)$ diferentes combinações de classes que possam existir. Desta forma, a dimensão VC do conjunto de funções lineares no espaço bidimensional é igual a 3 (Lorena & Carvalho, 2007).

O princípio de minimização do risco empírico é projetado para lidar com conjuntos de dados com muitos exemplos. Quantos mais exemplos de treino existirem menor é o erro de aprendizagem, sendo que no limite ($l \longmapsto \infty$), pode ser provado que os valores de risco empírico convergem para os mesmos valores do risco atual. Olhando para a inequação 3.5, se o valor de (l/h) for grande isso implica que l também é grande, logo a confidência VC será pequena e o risco atual ficaria assim próximo do risco empírico. Contudo, no caso contrário com (l/h) pequeno, mesmo tendo um risco empírico baixo, não implica que o risco atual também seja baixo. Surge assim o princípio de *minimização do risco estrutural*, que procura minimizar o risco empírico e ao mesmo tempo controlar a confidência VC de um conjunto de funções (Vapnik, 1999).

Minimização do Risco Estrutural

Sabendo que a dimensão VC diz respeito à capacidade de um conjunto de funções F, enquanto que o risco empírico mede o erro de uma máquina de aprendizagem f em particular, para minimizar ambas as parcelas é necessário dividir F em subconjuntos de estruturas. Considerando esses subconjuntos, $\mathrm{F}_0 \subset \mathrm{F}_1 \subset \ldots \subset \mathrm{F}_\mathrm{q} \subset \mathrm{F}$, em que com o aumento do índice também aumenta a capacidade das funções $h_0 < h_1 < \ldots < h_q < h$, então, sendo f_k a máquina com menor risco empírico, à medida que k aumenta o risco empírico diminui, uma vez que a complexidade da máquina é maior. Porém, o termo de capacidade h também aumenta com k. Conclui-se então que deve existir um ponto ótimo, no qual se obtém o valor mínimo da soma do risco empírico com a confidência VC, ou seja, o mínimo risco atual (R). A figura 3.6 demonstra, precisamente, este conceito de minimização do risco estrutural.

Apesar do conceito de minimização do risco estrutural ser útil para a compreensão da teoria VC, na verdade existem algumas limitações que fazem deste conceito pouco usado na prática. Por exemplo, calcular a dimensão VC pode ser uma tarefa complexa, ou até impossível, quando se tem uma dimensão VC infinita (Lorena & Carvalho, 2007). Dado que a minimização do risco estrutural consiste numa base teórica, ao contrário e.g. da validação cruzada, é de salientar que na prática o método de validação cruzada é normalmente usado para este efeito, uma vez que a validação cruzada procura o modelo que melhor generalize os dados, ou seja, evita o sobre ajustamento.

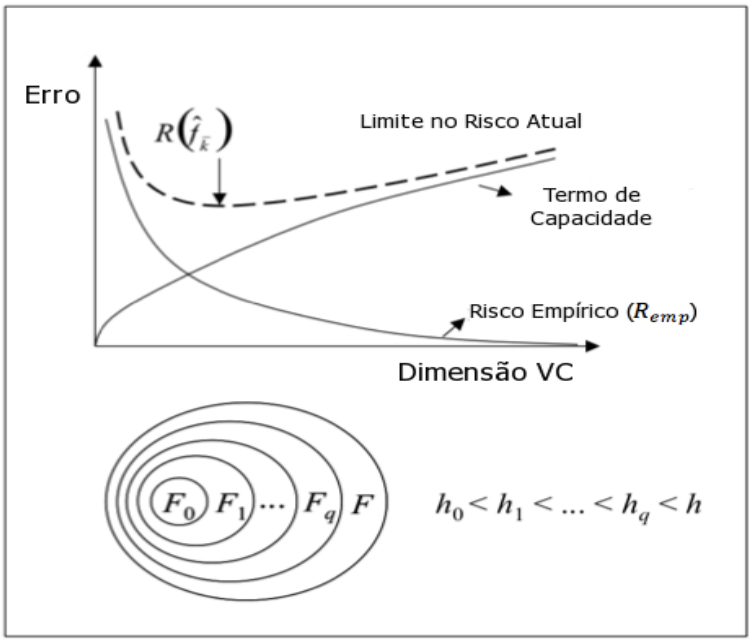


Figura 3.6 - Princípio de minimização do risco estrutural - adaptado de (Smola & Schölkopf, 2002), conforme citado em (Lorena & Carvalho, 2007)

Com a descrição da minimização do risco estrutural, temos a base da teoria de aprendizagem estatística que está na origem das SVM. Sabendo estes conceitos, vamos passar a introduzir o algoritmo SVM para assim o relacionarmos posteriormente com a teoria VC.

3.2.2 Support Vector Machines para Classificação

Para percebermos bem o conceito das SVM é importante começarmos pelo algoritmo que deu origem a esta técnica, ou seja, as SVM para classificação. Em classificação o problema consiste em encontrar o melhor híper-plano que separa as classes. Como se pode ver na figura 3.7, o híper-plano ótimo é a função de decisão que separa as classes por forma a maximizar as margens entre os vetores de suporte, ou seja, as classes sobre as margens.

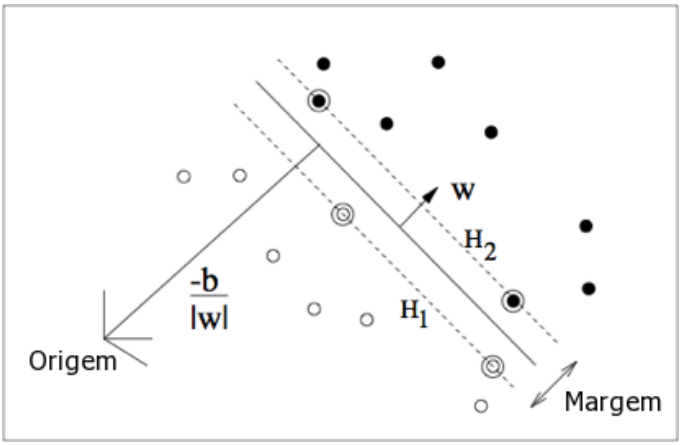


Figura 3.7 - Híper-plano ótimo de separação das duas classes. Círculos indicam os vetores de suporte - adaptado de (Burges, 1998)

Considerando um conjunto de dados como

$$(y_1, x_1) \dots (y_l, x_l) \quad y \in \{-1,1\} \tag{3.6}$$

os dados podem ser separados pelo híper-plano $(w.x + b = 0)$. Seguindo a notação de x como dados de input e de y como as classes a prever, diz-se que os dados são linearmente separáveis se existir um vetor w e uma constante b, tal que as inequações 3.7 sejam satisfeitas para todos os exemplos do conjunto de dados**.**

$$\begin{aligned} w.x_i + b &\geq 1 \quad se \quad y_i = 1 \\ w.x_i + b &\leq -1 \quad se \quad y_i = -1 \end{aligned} \tag{3.7}$$

Estas inequações podem ser escritas sob seguinte forma (Cortes & Vapnik, 1995):

$$y_i(w.x_i + b) \geq 1 \qquad i = 1, \dots, l \tag{3.8}$$

Olhando para a figura 3.7, considere-se os pontos que estão sobre as margens, ou seja, os vetores de suporte que respeitam as igualdades das inequações 3.7. Esses pontos estão sobre os híper-planos das margens $H_1: x_i \,.w + b = 1 \;\; e \;\; H_2: x_i \,.w + b = -1$, ambos com normal w e distância perpendicular à origem de $|1 - b|/\|\boldsymbol{w}\| \;\; e \;\; |-1 - b|/\|\boldsymbol{w}\|$ respetivamente. Como não existem classes entre as margens, as distâncias mínimas entre ambos os híper-planos são $d_+ = d_- = 1/\|\boldsymbol{w}\|$, concluindo assim que a distância da margem é igual a $2/\|\boldsymbol{w}\|$. O objetivo passa então por procurar o híper-plano ótimo que tenha a margem com máxima distância, onde para isso se minimiza $\|\boldsymbol{w}\|^2$ sob as restrições (3.8) (Burges, 1998). Por conseguinte, estamos perante um problema de minimização quadrático, com a seguinte formulação:

$$Minimizar_{w,b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$Segundo\ as\ restrições, \quad y_i(w.x_i + b)\ - 1 \geq 0 \qquad \forall i = 1, \dots, l \tag{3.9}$$

Executando este problema de minimização convexo, encontra-se assim o híper-plano que separa todas as instâncias de treino, isto no caso de estas serem linearmente separáveis. Contudo, como sabemos, na prática os dados são mais complexos e nem sempre são linearmente separáveis, surgindo assim as SVM com margens suaves. Imaginando os casos em que as classes não são linearmente separáveis, o problema de otimização do algoritmo para classes separáveis pode não ter solução. São por isso introduzidas variáveis de folga nas restrições, permitindo que algumas classes permaneçam entre as margens e permitindo também alguns erros de classificação. Considerando as variáveis não negativas ($\xi_i \geq 0, \;\; i = 1, \dots, l$), procura-se agora realizar o mesmo problema de minimização descrito acima, mas procurando também minimizar os erros de classificação. Assim, às restrições apresentadas em 3.8, são adicionadas as seguintes variáveis de folga:

$$y_i(w.x_i + b) \geq 1 - \xi_i \;, \qquad i = 1, \dots, l$$
$$\xi_i \geq 0 \,, \qquad\qquad i = 1, \dots, l \tag{3.10}$$

Ficando assim com o seguinte problema de minimização:

$$Minimizar_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l} \xi_i\right) \qquad \textbf{(3.11)}$$

Isto está sujeito às restrições (3.10) em que C é a constante de regularização. O parâmetro C é escolhido pelo utilizador e, como podemos ver pela equação, para valores maiores de C corresponde uma maior penalização dos erros de classificação.

3.2.3 Support Vector Machines para Regressão

Nos algoritmos SVM para regressão, também conhecidos como *Support Vector Regression* (SVR), a ideia passa por imaginar um tubo à volta da linha da função de aproximação (Figura 3.8, lado esquerdo). O objetivo da ϵ-SVR é assim encontrar a função que tenha no máximo um desvio de ϵ sobre todos os exemplos, procurando-se assim obter um tubo o mais fino possível. Continuando a seguir notação usada na formulação das SVM para classificação, descreve-se a função linear de aproximação (Smola & Schölkopf, 2004).

$$f(x) = \langle w, x\rangle + b, \qquad com\ w \in \mathrm{X}\,, b \in \mathbb{R} \qquad \textbf{(3.12)}$$

De sublinhar que $\langle .,. \rangle$ denota o produto escalar. Aqui, tal como na classificação, o objetivo passa por minimizar $\|\boldsymbol{w}\|^2$ segundo o seguinte problema de otimização:

$$\begin{aligned} &Minimizar_{w,b} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &segundo\ as\ restrições, \quad \begin{cases} y_i - \langle w, x_i\rangle - b \le \epsilon\,, & i = 1, \dots, l \\ \langle w, x_i\rangle + b - y_i \le \epsilon\,, & i = 1, \dots, l \end{cases} \end{aligned} \qquad \textbf{(3.13)}$$

Como se pode observar, relativamente à classificação apenas as restrições são alteradas, particularmente na inclusão de ϵ. A estratégia consiste em encontrar os pesos que aproximem $f$às classes com precisão ϵ e só mediante estas condições é que o problema de otimização tem

solução. De referir, ainda, que o problema de otimização é convexo, o que implica que tem um único mínimo global, não tendo portanto o problema de mínimos locais que ocorre e.g. nas NN.

De forma análoga às *margens suaves* das SVM para classificação, nas SVR são também introduzidas variáveis de folga nas restrições, permitindo assim alguns erros por forma a tornar o problema de otimização possível. Chega-se assim à seguinte formulação do problema de otimização com variáveis de folga:

$$Minimizar_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l} \xi_i + \xi_i^*\right)$$

$$segundo\ as\ restrições, \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i\,, & i = 1, \dots, l \\ \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^*\,, & i = 1, \dots, l \\ \xi_i, \xi_i^* \geq 0\,, & i = 1, \dots, l \end{cases} \tag{3.14}$$

O parâmetro C, factor de penalização dos erros, faz o balanço entre espessura da função e o valor dos desvios para além de ϵ sobre os quais são tolerados erros. Na figura 3.8 pode-se observar que para valores entre $-\epsilon$ e $+\epsilon$ (tubo) não existe penalização, ou seja, apenas os valores fora do tubo são penalizados de forma linear, isto segundo a função de perda *ε-insensitive*:

$$|\xi|_\epsilon := \begin{cases} 0 & ,se \quad |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & ,caso\ contrário \end{cases} \tag{3.15}$$

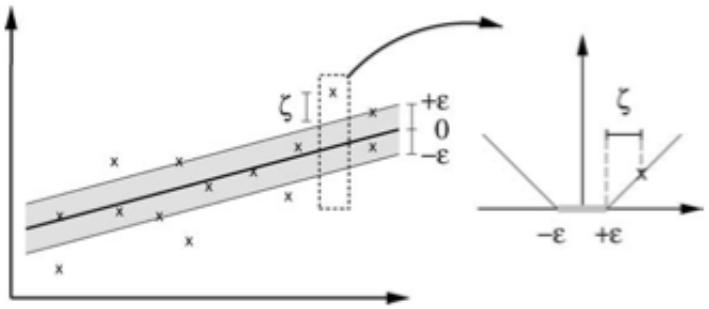


Figura 3.8 - Função de aproximação ϵ-SVR (lado esquerdo), gráfico função de perda ε-insensitive (lado direito) - adaptado de (Smola & Schölkopf, 2004)

O problema de otimização que caracteriza as SVR pode ser apresentado na forma dual através da função de Lagrange, em que certas características das SVM são levantadas. Passando do problema primal (Equação 3.14) para a formulação dual, serão introduzidos alguns multiplicadores de Lagrange não negativos (α_i, α_i^*). De referir, que os detalhes sobre a formulação matemática da transição entre a forma primal e a dual, podem ser encontrados em e.g. (Vapnik, 1995) ou (Smola & Schölkopf, 2004). Descreve-se então o problema, agora de maximização, na sua forma dual:

$$Maximizar \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases}$$
$$segundo\ as\ restrições,\ \ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0\ \ e\quad \alpha_i, \alpha_i^* \in [0, C] \qquad \textbf{(3.16)}$$

Assim, a variável w da forma primal passa a ser definida pela equação

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i$$

chegando à função $$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b\ . \qquad \textbf{(3.17)}$$

A equação 3.17 mostra o conceito da chamada "*support vector expansion*", que demonstra que a complexidade da função de aproximação não depende da dimensionalidade do conjunto de dados de treino, mas sim do número de vetores de suporte. Isto leva-nos a introduzir as condições Karush–Kohen–Tucker (KKT) que, para além de serem úteis na computação do parâmetro b, permitem retirar algumas conclusões interessantes. Resumidamente, estas condições mostram que só para exemplos fora do ϵ-tubo (região não sombreada da figura 3.8, lado esquerdo), é que os multiplicadores de Lagrange (α_i, α_i^*) podem ser diferentes de zero. Isto implica que para exemplos dentro do tubo, (α_i, α_i^*) são iguais a zero e logo desaparecem. Daí advém que os exemplos dentro do tubo não são necessários no cálculo de w, olhando para a equação 3.17, vê-se que $(\alpha_i - \alpha_i^*)$=0 vai anular x_i. Desta forma, explica-se a esparsidade da expansão de w, pois só os exemplos que não são descartados é que constam no problema de otimização e, na verdade, esses são chamados de *vetores de suporte*.

Não Linearidade das Support Vector Regression

As SVM consistem na ideia de mapear os vetores de input x, num espaço de características de maior dimensão Z, isto usando um mapeamento não linear Φ (Figura 3.9). Pela imagem percebe-se claramente que uma linha reta no espaço de características, consegue separar os dados que no espaço real não são linearmente separáveis. É esta propriedade que permite às SVM lidar com dados mais complexos. De acrescentar, que este conceito aplica-se de igual forma tanto para o caso de classificação como para o caso de regressão.

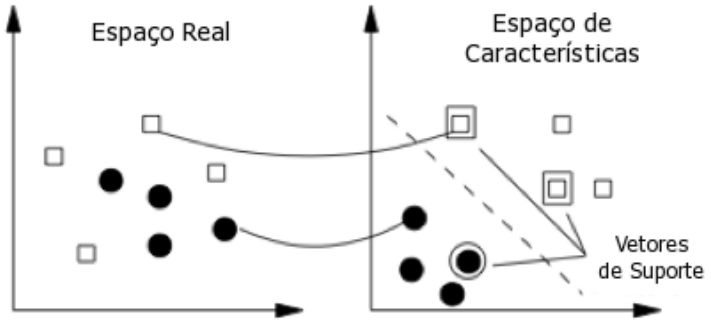


Figura 3.9 - Mapeamento de dados de input (lado esquerdo), para espaço de características (lado direito) - adaptado de (Cortez, 2011)

Tal como foi realçado na equação 3.12, $\langle w, x \rangle$ denota o produto escalar. Como o algoritmo SVM apenas depende do produto escalar, assumimos que $K(x, x') := \langle \Phi(x), \Phi(x') \rangle$, isto é, existe uma função K que recebe dois pontos $(x,\ x')$ e calcula o seu produto escalar num espaço de características. Este processo é conhecido como o "*kernel trick*" , cuja sua introdução no problema de otimização faz com que este passe a ser descrito da seguinte forma:

$$Maximizar \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) \\ -\epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases}$$
$$segundo\ as\ restrições,\ \ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0\ \ e\ \ \ \alpha_i, \alpha_i^* \in [0, C] \qquad \textbf{(3.18)}$$

Sendo o cálculo de w e a função de aproximação das SVR não lineares descritas como :

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\Phi x_i \quad \text{e} \quad f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b \ . \qquad \textbf{(3.19)}$$

Note-se que as diferenças para o caso linear consistem essencialmente na inclusão do *kernel*, que faz com que o problema de otimização passe a ser realizado no espaço de características e não no espaço real.

Com a inclusão do *kernel*, as SVM obtêm a característica descrita por Cortes & Vapnik (1995) de universalidade da máquina. Isto deve-se ao facto de ser possível aplicar várias funções de *kernel* no cálculo do produto interno. Contudo, para assegurar que as funções de *kernel* realizam o produto escalar no espaço de características, estas têm de respeitar as condições de Mercer. As funções de *kernel* mais usadas são a Polinomial, RBF (*Radial-Basis Function*) e Sigmoides. De acrescentar, que inclusão do *kernel* permite a introdução de conhecimento à priori do problema em questão, pois poderão ser desenvolvidas diferentes funções de cálculo do produto escalar. A descrição do caso de regressão das SVM aqui apresentada, seguiu essencialmente o tutorial de (Smola & Schölkopf, 2004). Mais detalhes sobre esta variante das SVM, como notas de implementação do algoritmo ou outras informações, podem ser encontrados nesse guia.

3.2.4 Dimensão VC e Capacidade de Generalização das SVM

Como já referido, a teoria de aprendizagem estatística motivou o desenvolvimento das SVM. Depois da descrição da teoria VC e dos algoritmos SVM, nesta secção vamos agora relacionar estes dois conceitos. A capacidade de generalização de uma máquina de aprendizagem consiste no controle de dois factores, o risco empírico e a confidência VC (Equação 3.5). Onde a confidência VC depende do termo de capacidade, do conjunto de dados de treino e do parâmetro η. Na equação 3.5, é preciso existir um equilíbrio entre os dois termos do lado direito da equação. Quanto menor for dimensão VC menor será a confidência VC, mas neste caso o risco empírico será maior, portanto, o princípio de minimização do risco estrutural, consiste exatamente na procura deste ponto de equilíbrio. Existe ainda um caso particular da minimização do risco estrutural que

segue o princípio "*Occam-Razor*", este consiste de grosso modo na ideia de que os sistemas mais simples são os melhores.

O princípio *Occam-Razor* indica-nos que a equação 3.5 deve manter o primeiro termo da soma a zero ($R_{emp}(\alpha)=0$), minimizando assim o segundo termo *(Confidência VC*). Desta forma, segundo Cortes & Vapnik (1995), a igualdade do primeiro termo deve obedecer às restrições das inequações 3.7, enquanto que o segundo termo deve ser minimizado funcionalmente em (w^2). Deste modo, é possível provar que a maximização da margem do hiperplano implica a minimização do erro estrutural, prevenindo assim problemas de sobre ajustamento. No entanto, mesmo no caso de dados separáveis, poderá ser alcançada uma melhor generalização ao minimizar a confidência VC, mesmo que isso implique alguns erros de classificação nos dados de treino. Como vimos, o ajuste entre penalizar mais os erros é feito segundo o parâmetro C, que faz o compromisso entre a complexidade da função de aprendizagem e os erros de aprendizagem. Com este conceito, justifica-se assim que as SVM conseguem controlar ambos os fatores que determinam a capacidade de generalização de uma máquina de aprendizagem.

3.2.5 Sequential Minimal Optimization (SMO)

A variante de SVM introduzida por Platt (1999) consiste num algoritmo de aprendizagem chamado *Sequential Minimal Optimization* (SMO). Este algoritmo consta na decomposição do problema de minimização quadrático em vários subproblemas de programação quadrática (PQ) de tamanho fixo. O problema de PQ das SVM não pode ser resolvido facilmente pela generalidade das técnicas de PQ. Para se ter uma ideia, para um conjunto de dados com mais de 4000 exemplos, não chegavam 128 Megabytes para alojar a matriz Hessiana do problema.

Na resolução PQ das SVM, Vapnik (1982), conforme citado em (Platt, 1999), introduziu o método na altura conhecido por "*chunking*". O *chunking* consiste simplesmente em remover os exemplos que não têm qualquer influência na solução ($\alpha_i = 0$), ou seja, eliminar todos os exemplos que não sejam vetores de suporte. Com base neste facto, o *chunking* separa o problema PQ em subproblemas PQ menores, com o objetivo de identificar os exemplos que possuem α_i diferente de

zero. A dimensão da matriz do problema de otimização é assim reduzida, passando do número de exemplos de treino para o número de vetores de suporte. No entanto, se forem usadas técnicas usuais de PQ, o *chunking* não consegue lidar com problemas de grande escala, visto que mesmo uma matriz reduzida pode não caber em memória. A técnica de decomposição introduzida por Osuna et al. (1997), conforme citado em (Platt, 1999), é similar ao *chunking*, porém a diferença reside nas matrizes dos subproblemas que têm tamanho fixo, e na inclusão ou remoção de exemplos que violem as condições KKT.

As SMO diferem das técnicas de decomposição anteriores, essencialmente na medida em que não são usados métodos numéricos na resolução do problema PQ. A principal característica do algoritmo SMO, consiste assim na resolução do menor problema de otimização possível em cada passo. Nas SVM simples, o menor problema possível considera apenas dois elementos. Assim, em cada passo, o algoritmo SMO escolhe dois α_i para otimizar, encontrando assim os valores ótimos para essas variáveis e atualizando depois as SVM. A grande vantagem está na resolução analítica do problema, evitando a otimização segundo métodos numéricos de PQ. Existem assim duas componentes principais do algoritmo SMO, o método analítico para calcular os dois alfas e a heurística de escolha dos multiplicadores a otimizar.

A versão de SMO para regressão foi inicialmente apresentada em (Smola & Schölkopf, 1998) e consiste, basicamente, numa extensão da versão inicial para classificação. Contudo, surgiram algumas críticas e melhorias à proposta inicial de Platt e, consequentemente, à extensão SMO para regressão de Smola e Schölkopf, estas são apresentadas no trabalho de Shevade et al. (2000).

3.3 Sumário

Com o constante crescimento do volume de informação no nosso meio, surge a necessidade de desenvolver técnicas que permitam obter conhecimento a partir desses grandes volumes de dados. O DM surge nesse âmbito, podendo ser definido como o processo de extração de informação implícita, previamente desconhecida e potencialmente útil dos dados. Esse processo é também

conhecido por descoberta de conhecimento em BD (KDD). Contudo, existe alguma controvérsia entre o termo DM e KDD. Na verdade, DM é referido por Fayyad et al. (1996a) como uma das fases do processo KDD, no entanto esse mesmo processo é chamado de DM por grande parte da comunidade científica. O processo de desenvolvimento de aplicações de DM adotado neste trabalho foi o CRISP-DM. Esta metodologia, bastante utilizada em projetos de DM, consiste em seis fases de desenvolvimento: análise de negócio; análise de dados; preparação de dados; modelação; avaliação; e implementação. A fase de modelação, que segundo Fayyad et al. (1996a) corresponde à fase de DM do processo KDD, consiste na aplicação de métodos ou algoritmos de DM. Existem portanto dois tipos de modelação, a descritiva e a preditiva. Dentro da modelação descritiva encontramos técnicas como a segmentação e a sumarização, enquanto que no caso da modelação preditiva, as tarefas de DM envolvem essencialmente problemas de classificação e de regressão. Estas duas tarefas distinguem-se no tipo de dados da variável de previsão, para classificação preveem-se dados nominais ou categóricos, enquanto que na regressão preveem-se dados numéricos ou reais. Em cada uma destas tarefas, existem vários algoritmos com diferentes abordagens de aprendizagem e também diferentes formas de representatividade dos modelos desenvolvidos.

Os algoritmos que são alvo de estudo desta dissertação são conhecidos por *Support Vector Machines*. As SVM fundamentam-se na teoria de aprendizagem estatística, também conhecida por teoria VC (Vapnik e Chervonenkis). De grosso modo, a teoria VC caracteriza as propriedades das máquinas de aprendizagem que as permite generalizar bem perante dados novos. É portanto introduzido o conceito de limite de risco e, com base neste limite, o principio de minimização do risco estrutural. Este princípio envolve a procura de um ponto ótimo, encontrando assim um equilíbrio entre o erro de aprendizagem e a capacidade de uma máquina de aprendizagem (e.g. classificador). Desta forma, esta teoria tem como objetivo encontrar a máquina de aprendizagem com maior capacidade de generalização.

Em tarefas de classificação, as SVM procuram encontrar o melhor híper-plano que separe os dados, por forma a maximizar a margem entre o híper-plano e as classes mais próximas deste. Com base nisso, está-se perante um problema de otimização quadrático. No caso de regressão, o processo é análogo. No entanto, em vez do híper-plano de separação temos um tubo que envolve

a função de aproximação. Neste caso, o objectivo passa por encontrar um tubo o mais fino possível, mediante um determinado desvio ϵ. As SVM são ainda conhecidas pelo mapeamento dos dados de *input* para um espaço de características de maior dimensão. Com base neste mapeamento, é introduzido nas SVM o designado "*Kernel Trick"*, que proporciona a capacidade de lidar com dados não lineares. Outra característica principal das SVM é a generalização, relacionando o limite do risco com o algoritmo SVM, consegue-se mostrar a capacidade de generalização destas máquinas de aprendizagem. Adicionalmente, uma extensão do algoritmo SVM que consiste na decomposição do problema de otimização quadrático, foi apresentada por Platt (1999) e é conhecido como algoritmo SMO.

Capítulo 4

Análise e Preparação de Dados

4.1 Análise de Negócio

"Uma correta preparação de dados prepara ambos, os dados e o *miner*. Enquanto que a preparação de dados significa que o modelo é construído corretamente, a preparação do *miner* significa que o modelo correto é construído." (Pyle, 1999, p. 9). Esta citação ajusta-se perfeitamente ao conteúdo deste capítulo. O termo modelação tem sido várias vezes referido ao longo deste documento. Porém, o processo de desenvolvimento de um modelo de previsão deve ser antecedido pela a análise e preparação de dados. Esta fase é crucial principalmente em dois aspetos, por um lado existe um conjunto de técnicas de preparação que têm de ser aplicadas aos dados por forma que o modelo consiga lidar com estes, por outro lado é igualmente importante que o analista de dados ("*miner*") compreenda os dados e o problema com o qual está a lidar.

O domínio do problema das ETAR, conforme foi descrito no capítulo 2, consiste em traços gerais no seguinte: após o afluente bruto dar entrada na ETAR, são executadas várias fases de tratamento sequenciais, por forma a obter-se no final do tratamento um efluente com qualidade e pronto a ser descarregado no meio-ambiente. Como vimos, existem inúmeras variáveis que condicionam este processo, cujo o controle é fulcral para um bom funcionamento da planta de tratamento. Diferentes técnicas de DM podem ser aplicadas nestes casos, sempre com o intuito de melhorar o comportamento das ETAR. Basta para isso lembrar as motivações e técnicas de DM que

podem ser aplicadas nas ETAR, e que foram descritas anteriormente no capítulo 2. Dentro das tarefas descritas, neste caso, o problema envolverá essencialmente tarefas de previsão.

4.1.1 Objetivos de Negócio

Os primeiros passos de um projeto de DM devem passar pela análise, do ponto de vista do negócio, de quais os objectivos que o cliente pretende alcançar. Com base nesta análise, apresentam-se em seguida dois objetivos de negócio que representam também o foco a partir do qual se centrou o desenvolvimento deste trabalho.

- **Modelo de Previsão –** Considera-se o real objectivo deste trabalho, encontrar um modelo de previsão que capture a dinâmica do processo de tratamento da ETAR em estudo. Neste sentido, para o "cliente" é importante obter uma previsão quantitativa e eficaz do desempenho dos tratamentos realizados. Deste modo, com base nos resultados obtidos, pretende-se provar que é possível prever o comportamento da ETAR com um bom nível de eficácia. Isto, por forma a ponderar a implementação de um futuro sistema que dê suporte às tomadas de decisão da ETAR.

- **Informação Útil –** Adicionalmente, a apresentação de informações que possam ser úteis para uma melhor compreensão do processo de tratamento, pode ser uma mais valia para a ETAR. Informações como, por exemplo, apresentar os atributos mais influentes no processo de tratamento.

Convém no entanto referir que na realidade estes objetivos de "negócio" não foram impostos por nenhuma empresa, mas constam sim dos objectivos que foram definidos no início do projeto pelos intervenientes desta dissertação. Contudo, sendo o termo "negócio" vulgarmente usado nestes casos e, em particular, no mundo da engenharia de software, optou-se por esta terminologia.

4.1.2 Objetivos de Data Mining

Os objetivos de negócio mostram uma visão muito geral, e da perspetiva de um hipotético cliente, do que se pretende com este trabalho. Apresenta-se em seguida uma perspectiva mais detalhada dos objetivos deste trabalho, já com termos técnicos do domínio de DM:

- Para se alcançar uma modelação eficaz, é necessário preparar bem o conjunto de dados de trabalho, por forma a facilitar a modelação realizada pelos algoritmos SVM. Pretende-se assim estudar as técnicas de preparação de dados que são normalmente utilizadas, e também recomendadas na literatura de DM.

- As SVM são os algoritmos adotados neste trabalho. Pretende-se explorar estes algoritmos em busca de uma modelação eficaz e ainda comprovar a sua capacidade de generalização, uma vez que este é um ponto forte destas técnicas do ponto de vista teórico.

- Dentro das técnicas de SVM existem algumas variantes como o algoritmo SMO. Assim, pretende-se explorar, e posteriormente comparar, o desempenho de diferentes variantes do algoritmo em estudo.

- A modelação com SVM permite a utilização de diferentes *kernels*. Também neste sentido procura-se analisar qual será a melhor escolha do *kernel* para este problema.

- É objectivo, também, analisar qual o método de avaliação de modelos mais correto. Este é um ponto importante uma vez que, como iremos ver, estamos perante um conjunto de dados relativamente pequeno.

- Na previsão do desempenho do tratamento das ETAR, serão selecionadas como variáveis de previsão os parâmetros que permitem medir a qualidade do efluente final. Assim, é objetivo prever os parâmetros de tratamento do último ponto de amostragem.

- É objetivo utilizar técnicas de seleção de características (atributos), não só para melhorar a eficácia dos modelos de previsão, mas também para analisar quais os parâmetros de tratamento mais importantes. Desta forma, extrai-se assim informação com potencial utilidade para os analistas da ETAR.

Critérios de Sucesso

Sabendo que este projeto é desenvolvido essencialmente com propósitos académicos, os critérios de sucesso passam pela obtenção de uma modelação final que ofereça resultados plausíveis. Ou seja, não existe qualquer tipo de meta ou nível de eficácia de previsão que tenha de ser atingido ou melhorado. Contudo, espera-se retirar conclusões que sejam relevantes, tanto ao nível do funcionamento da ETAR, como dos algoritmos SVM que são também alvo de estudo desta dissertação. Esta passagem pela análise do domínio do problema do ponto de vista de negócio, também implica um planeamento inicial do projeto, contudo este não será aqui apresentado visto considerar-se fora do âmbito deste documento de dissertação.

4.1.3 A Ferramenta de Data Mining Utilizada

A ferramenta de DM utilizada neste estudo foi o *RapidMiner* da empresa *Rapid-I.* Existem várias outras soluções de ferramentas de DM no mercado, no entanto a grande maioria trata-se de software comercial - para este projeto apenas se consideraram como opção ferramentas de DM livres. Dentro deste grupo destacam-se essencialmente três ferramentas: o *RapidMiner*, o *WEKA* e o *R*. Não querendo no âmbito deste trabalho de dissertação entrar em grandes comparações e especificidades, podemos dizer que, de grosso modo, todas elas oferecem um vasto conjunto de algoritmos que permitem realizar variadíssimas tarefas de DM. Porém, neste ponto a ferramenta R é superior, na medida em que contém um vasto repositório de *packages*, oferecendo assim aos utilizadores bastantes mais métodos estatísticos e algoritmos de DM. De acrescentar ainda a vantagem de ser a comunidade do *R* a desenvolver os *packages*, o que possibilita um acesso mais rápido a algoritmos e métodos mais recentes (Cortez, 2010). Contudo a desvantagem do *R* consiste na sua interface via linha de comados, uma vez que requer alguma experiência e familiaridade com a ferramenta para se tirar o máximo proveito desta. A ferramenta do *WEKA*, nomeadamente o *WEKA explorer,* talvez seja das três ferramentas a mais intuitiva. Contudo, esta não oferece tanta liberdade para desenvolver fluxos sequenciais de operadores de modelação, transformação, tratamento de dados, entre outros, como acontece no *RapidMiner.*

Das três ferramentas, o *RapidMiner* é, assim, a que apresenta uma interface mais apelativa. Porém isso não significa que seja a mais simples de usar. Ambos os critérios utilizados são subjetivos. No entanto, no nosso ponto de vista, parece-nos que a nível visual o RapidMiner supera tanto o *R* como o *WEKA*. Porém, ao nível da aprendizagem necessária para usar-se a ferramenta, o RapidMiner fica no meio entre o *R* (mais exigente) e o *WEKA* (mais simples). Estes dois critérios, apesar de importantes, não foram determinantes na escolha do *RapidMiner*. Na verdade, existiram dois fatores que foram decisivos para escolha que fizemos. O primeiro tem a ver com a capacidade do *RapidMiner* em permitir projetar fluxos de processamento com vários operadores. De forma análoga às ferramentas de ETL (*Extract, Transform, Load*), com o *RapidMiner* é possível desenhar uma sequência de operadores de seleção, transformação, modelação, avaliação, entre outros métodos. Esta capacidade permite obter no final um processo completo, que vai desde recolha de dados até à geração de gráficos e avaliação dos modelos. Na figura 4.1 podemos ver um exemplo de um processo de mineração de dados, no qual as "caixas" correspondem aos operadores, e o fluxo do processo segue as ligações da esquerda para a direita. Na verdade, o *RapidMiner* é até referido como uma ferramenta de integração e de ETL, mas também se sabe que esse não é o seu principal objetivo de trabalho (Rapid-I, n.d.). Todavia, esta característica tem vantagens como, por exemplo, a boa percepção do processo que se está a desenvolver e a facilidade de alteração de operadores e respetivos parâmetros. Por outro lado, tem a desvantagem de todas as funcionalidades exigirem um operador, em que muitas delas têm de estar obrigatoriamente conectadas com outros operadores específicos. Isto exige algum conhecimento adicional aos operadores do *RapidMiner*.

A segunda característica é a integração das ferramentas do *R* e do *WEKA* no *RapidMiner*. Este foi o factor chave para a decisão de se optar pelo *RapidMiner* pois, com esta funcionalidade, obtemos o "melhor de dois mundos". Assim, podemos usar os algoritmos do *WEKA* como operadores do *RapidMiner*, bem como incluir um script de *R* no processo de DM que se está a desenvolver. Tendo em conta que este é um trabalho que envolve pesquisa e, como iremos ver, várias experimentações de técnicas de DM diferentes, esta funcionalidade torna-se muito interessante por forma a centralizar tudo numa única ferramenta de trabalho.

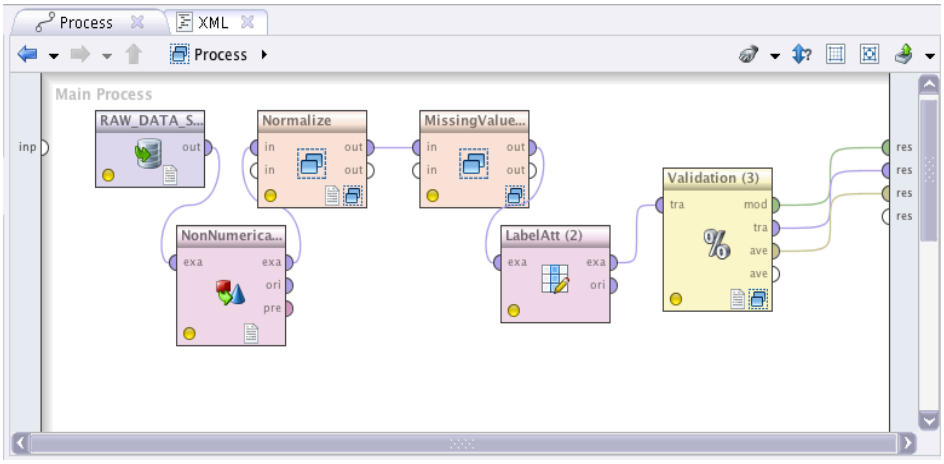


Figura 4.1 - Exemplo de um processo de DM no RapidMiner

Para concluir, na opinião da generalidade da comunidade de DM, segundo o ranking publicado em (KDnuggets, 2011), estas ferramentas livres são das mais utilizadas, juntamente com o *KNIME*. Ainda neste ranking, o *RapidMiner* surge em primeiro lugar como a ferramenta de DM mais usada no ano de 2011, ranking este que também incluí ferramentas comerciais. O *R* segue no segundo lugar com uma adesão muito próxima do *RapidMiner*. Numa sondagem mais recente, para o ano 2012 (KDnuggets, 2012), apesar do *RapidMiner* manter sensivelmente a mesma percentagem de utilizadores, no entanto aumentou a adesão da comunidade de DM à ferramenta *R*. Por conseguinte, o *R* passou a ocupar o primeiro lugar do ranking, passando a ser a ferramenta de DM mais utilizada. Este facto não é surpreendente, uma vez que o *R* oferece um largo repositório de *packages* e, consequentemente, possui uma comunidade muito ativa, o que origina provavelmente esta crescente adesão. De sublinhar ainda a presença do Excel nos lugares cimeiros destas tabelas, esta ferramenta bastante versátil continua a ser muito usada, inclusivamente no universo de DM. Neste trabalho o Excel também foi utilizado, essencialmente nas fase iniciais de desenvolvimento, recolha e análise de dados.

4.2 Análise dos dados

Após a análise do problema do ponto de vista do negócio, analisa-se em seguida o problema na perspectiva dos dados que foram fornecidos. Este processo de análise de dados, pressupõe, assim, a exploração e compreensão do conjunto de dados da ETAR em estudo.

4.2.1 Recolha de dados

Antes de passarmos à caracterização do conjunto de dados de trabalho, convém referir primeiro o processo de recolha de dados executado pela ETAR. Na figura 2.1 foi apresentado o diagrama geral da ETAR em estudo. Nesse diagrama apenas foram indicados os processos de tratamento correspondentes a uma linha de tratamento, para que fosse possível simplificar a descrição da estrutura da própria ETAR. Todavia, a ETAR contem duas linhas de tratamento paralelas, nomeadamente nas fases de tratamento primária e secundária. Na figura 4.2 pode-se observar o diagrama da ETAR em questão, onde estão indicados os tratamentos efectuados, bem como os pontos de amostragem (PA) dos dados, que representam os locais de medição dos diferentes parâmetros de tratamento.

O primeiro ponto de amostragem situa-se no local de entrada do afluente bruto (PA1). Após a entrada das águas residuais na ETAR, são inicialmente removidos os sólidos mais grossos, areias e gorduras. De seguida, o tratamento passa então a ser executado em duas linhas de tratamento. Tanto na primeira como na segunda linha, os tratamentos são processados em três tanques principais: o seletor biológico, o tanque de arejamento (licor misto) e o decantador secundário. No fim de cada um destes tratamentos são realizadas as amostragens de (PA2, PA2'), (PA3, PA3') e (PA5, PA5') respetivamente. São também realizadas ainda as amostragens de recirculação em ambas as linhas (PA4, PA4'). Após o tratamento secundário, as linhas convergem para a última etapa de tratamento, o terciário, sendo depois medido o efluente final já tratado (PA6). Em paralelo a este processo é ainda executado o tratamento da linha sólida, ou seja, o tratamento de lamas (PA7).

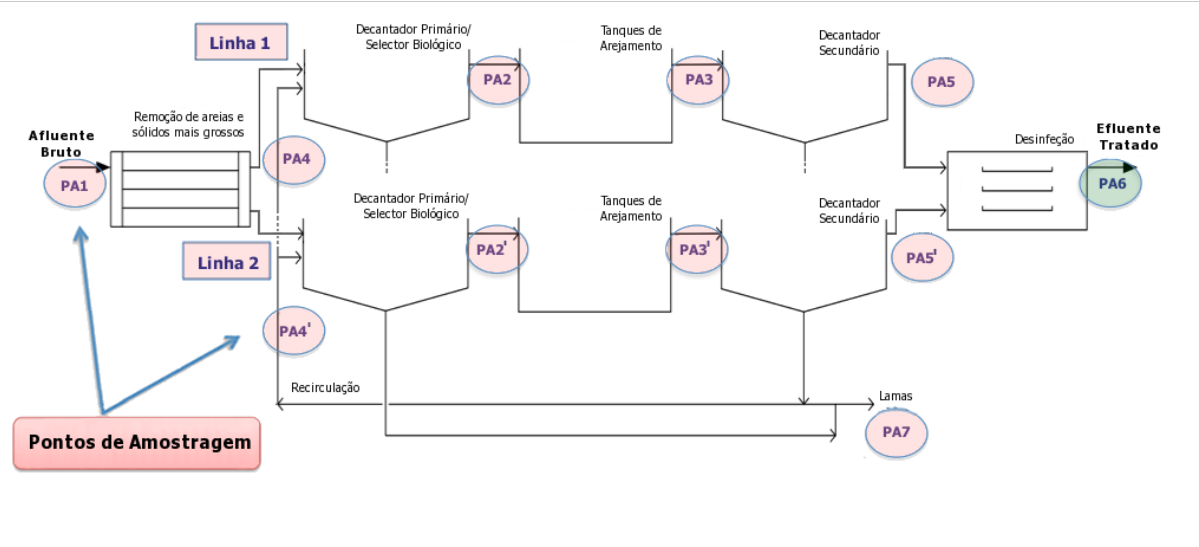


Figura 4.2 - Diagrama da ETAR em estudo e respetivos pontos de amostragem de dados

A ETAR que serviu de caso de estudo neste trabalho está localizada no norte Portugal e serve uma população de cerca de 45 mil habitantes. Os dados recolhidos resultaram das médias diárias dos parâmetros, que foram medidos ao longo dos vários pontos de amostragem. Estes dados foram, portanto, registados numa base diária, implicando que a granularidade do conjunto de dados correspondesse a um dia tratamento. O registo e armazenamento de dados foi realizado com recurso a folhas de cálculo, em que foram compilados os registos mensais do tratamento. Assim, para cada mês existe uma folha de cálculo com os tratamentos respeitantes a esse mesmo mês.

Na fase inicial do projeto, as primeiras tarefas consistiram em organizar toda esta informação por forma a obtermos um conjunto de dados único. Este processo envolveu a transposição dos registos para uma única tabela, centralizando desta forma os dados e obtendo assim a BD (não relacional) da ETAR para este estudo.

4.2.2 Características da Base de Dados

A base da dados resultante do processo de recolha de dados descrito acima, contém um total de 92 exemplos correspondente ao período de um ano de tratamentos. Outra característica da BD é a grande dimensionalidade apresentada, cerca de 122 atributos. Como já referido, a granularidade dos dados corresponde à medida temporal diária, ou seja, uma instância corresponde ao registo de um dia de tratamentos, e os valores de um registo resultam da média das várias medições que são realizadas durante um dia de tratamento, isto para cada parâmetro. Como temos cerca de 122 atributos na BD, estes não serão aqui descritos na sua totalidade, porém será realizada ainda nesta fase uma pré-seleção dos atributos mais relevantes. Deste modo os atributos resultantes e mais relevantes para o problema, apresentar-se-ão mais à frente no conjunto de dados de trabalho inicial.

Atributos e Enriquecimento de Dados

A grande maioria dos atributos da BD são relativos aos parâmetros de tratamento quantitativos e qualitativos apresentados anteriormente na tabela 2.2. Adicionalmente, também constam da BD atributos como a data, a estação do ano e as observações sobre o processo de tratamento. Para além destes atributos, o conjunto de dados foi ainda "enriquecido" com dados externos de cariz meteorológico, como o estado do tempo ("chuva" ou "sol") e a temperatura média. Como se sabe, as condições climatéricas influenciam o processo de tratamento, daí o interesse em adicionar esta informação ao problema. Se por um lado estes dados são um potencial fator de auxílio para uma melhor modelação, por outro lado pode proporcionar a extração de informação útil. Estes dados externos foram retirados da BD meteorológica online em (Wunderground, n.d.), e são respectivos à estação meteorológica automática do Porto – Pedra Rubras (Aeroporto).

Consistência, Integridade e Poluição de Dados

No processo de exploração de dados deve-se investigar a existência de problemas nos dados, pois estes podem afetar bastante o desenvolvimento de um modelo. Em termos de consistência, tendo em conta que o domínio dos dados é predominantemente numérico (valores de concentrações),

encontraram-se alguns valores bem distantes da média dos restantes valores, isto dentro do mesmo atributo. Contudo, a maioria destes casos correspondem a valores fora dos padrões normais (*outliers*), que, na sua maioria, foram devidamente identificados nas observações dos analistas da ETAR como problemas que ocorreram durante o tratamento. Por outro lado, comprova-se que são *outliers* pois a alteração dos padrões dos valores ocorre em mais do que um atributo em simultâneo. Assim, o único problema de inconsistência encontrado, deve-se ao facto dos tratamentos relativos ao Decantador Primário/Seletor Biológico serem alternados. Melhor dizendo, por vezes quando os tratamentos são realizados com recurso ao decantador primário, existem variáveis que têm em média valor x, enquanto que no caso do tratamento ser realizado pelo seletor, na mesma variável têm-se o valor médio de $10x$. Perante este problema foram tomadas algumas posições que serão descritas mais abaixo ainda nesta secção.

Quanto aos problemas de poluição e integridade, estes não foram identificados na BD. Ou seja, para o caso de poluição de dados, não foram encontrados valores que não respeitassem o domínio do atributo, nem valores fora do contexto do atributo que indicassem um erro de input ou de transposição de dados. Também no relacionamento entre atributos não se encontraram situações gritantes que indicassem um erro de integridade. Contudo, deve ser destacado que devido à generalidade dos atributos serem numéricos e à dimensionalidade ser alta, foi difícil analisar com precisão todos os relacionamentos. Ainda assim, no que diz respeito aos atributos nominais, não se encontraram relacionamentos descabidos, como e.g. "mês de tratamento Agosto e estação do ano Inverno".

Valores Nulos

Os valores nulos são uma das dificuldades típicas de DM. Este problema está intrinsecamente ligado à presença de dados esparsos na BD. Como a generalidade dos algoritmos não conseguem lidar com valores nulos, e também porque variáveis bastante esparsas normalmente não trazem informação relevante ao modelo, este é um factor importante de eliminação de atributos. Convém assim definir uma percentagem a partir da qual se eliminam as variáveis, contudo esse valor não pode ser escolhido de forma rígida. Deste modo, é importante ter em conta o "valor" ou relevância

que cada atributo tem no problema e conjugar esse facto com taxa de valores nulos que estes apresentam.

No totalidade das células da BD (11224), 35% delas correspondem a valores nulos, ou seja, em média existem cerca 35% de valores nulos nos atributos da BD. Este facto ajuda a perceber que, caso fosse imposto um limite de aceitação de atributos, com taxas de valores nulos baixas (i.e. 10%), significaria que ficávamos com pouquíssimos atributos para modelar. Por conseguinte, fazendo uma apreciação global dos dados, definiu-se então um limite não rígido de 55%. Deste modo, está-se a contemplar grande parte dos atributos, nomeadamente os parâmetros relativos aos nutrientes, que têm mais de 50% de nulos. Apresenta-se em seguida os atributos com taxas de nulos acima do limite, e ainda os critérios que levaram à sua remoção.

Tratamento de Lamas

Os atributos relativos à fase de tratamento sólida da ETAR foram todos descartados, ou seja, não os integrámos no conjunto de dados, diga-se, inicial. Esta categoria de parâmetros referente ao tratamento das lamas foi descartada devido essencialmente a dois fatores. Primeiro a taxa de nulos destes atributos está, na generalidade, bem acima dos 50%, sucedendo mesmo que alguns deles têm taxas acima de 80%. A segunda razão passa pelo facto deste tratamento ser realizado aparte do tratamento líquido, sendo este último o tratamento a modelar. Mesmo sabendo que poderão existir eventuais relacionamentos entre as variáveis da fase sólida e líquida que se mostrem úteis para a modelação, porém, devido aos valores nulos e à necessidade de reduzir-se a dimensionalidade, optou-se por descartar toda a fase de tratamento de lamas. Adicionalmente, os atributos relativos as fossas sépticas também foram descartados. Neste caso, a taxa de nulos de todos os atributos deste processo de tratamento é superior a 85%.

Microfauna

A BD da ETAR contém ainda registos relativos aos dados qualitativos da microfauna. Estes dados apresentam algumas dificuldades ao ponto de serem usados para modelação. A primeira dificuldade passa pelo tipo de dados que é nominal, em que numa única instância são registados vários de elementos de microfauna. Estes elementos são divididos por dois grupos de organismos,

os protozoários/metazoários e ainda as bactérias filamentosas. Perante estes factos, sabendo que os algoritmos de regressão apenas aceitam valores numéricos, pode-se imaginar que a conversão desta informação para um sistema numérico não seria trivial. Outro problema dos dados relativos à microfauna, tem que ver com a data destes registos, uma vez que em alguns casos estes não correspondem às datas de registo dos dados quantitativos. Também neste caso teriam de ser analisadas algumas estratégias, por forma a introduzir estes registos com datas diferentes no conjunto de dados geral, agregando assim esta informação com os restantes atributos.

Para concluir, a última dificuldade destes dados são os valores nulos. Os dois grupos de microfauna têm taxas de nulos elevadas, onde, no caso mais grave, acima dos 80%. Consequentemente, com base neste conjunto de fatores, tomou-se a decisão de não considerar-se os dados da microfauna para a modelação. Convém no entanto realçar que em estudos de classificação dos estados de funcionamento das ETAR, estes dados seriam de mais valia, assim como as observações dos analistas das ETAR. Mas não só, o trabalho de Belanche et al. (1999b) mostra como se pode usar a microfauna em tarefas de previsão. Todavia, para esse efeito, é necessário ter acesso a um conjunto de informação qualitativa mais ampla.

Eficiência de Remoção

Encontram-se ainda na BD variáveis relativas à eficiência de remoção de impurezas do tratamento, em particular em PA2 e PA6. A eficiência de remoção é calculada com base na diferença entre as concentrações desses pontos (PA2 ou PA6), e do ponto de entrada do afluente bruto (PA1), isto segundo uma percentagem (i.e. $\frac{PA1_{CBO}-PA2_{CBO}}{PA1_{CBO}} * 100$). Em PA2 estas variáveis apresentam uma taxa de valores nulos na ordem dos 70%, enquanto que no final do tratamento (PA6), as taxas de nulos são inferiores a 55%. Contudo, como as variáveis de previsão pertencem a este ponto, não faz sentido prever PA6_CBO e ao mesmo tempo usar PA6_CBO% como input. No caso de dados novos, cujo valor de PA6_CBO é desconhecido, implica que PA6_CBO% também será desconhecido. Note-se que será levantada mais à frente a problemática de seleção de atributos de PA6, pois, sabendo que os valores de CBO demoram 5 dias a ser obtidos, será equacionada a utilização dos outros parâmetros de qualidade de PA6 na sua previsão.

Em todo o caso, nesta fase de análise dos dados relativos à eficiência de remoção, descartaram-se apenas os atributos de eficiência de remoção de PA2, devido à elevada taxa de nulos. Os restantes atributos de PA6, serão posteriormente equacionados na fase de preparação de dados, em específico no tratamento de atributos redundantes e, ainda, na filtragem de atributos para modelação. Convém destacar que como estamos perante uma representação diferente de informação, que consta de outros atributos já presentes na BD (i.e. PA1_CBO e PA6_CBO), estes poderão ser considerados atributos redundantes (Pyle, 1999).

Para concluir a análise de valores nulos e da consequente remoção de atributos da BD, falta descrever o caso de duas instâncias em que se detectou uma evidente falta de valores ao longo dos demais parâmetros de tratamento. Estes registos correspondem aos dias 9 e 11 de Novembro, onde é justificado pelos analistas das ETAR que não foram realizadas amostragens devido à mudança de linhas. Mediante este facto, a presença destes dois registos torna-se irrelevante, sendo portanto descartados. Consequentemente, o conjunto já pequeno de dados (92) é reduzido, passando assim a contar com 90 instâncias.

Outras Características e Problemas dos Dados

Para além das características já levantadas, existem dois pontos que gostaríamos de realçar, visto serem relevantes para a projeção deste problema. Estes problemas, para além de serem perceptíveis na fase de exploração de dados, sendo portanto características intrínsecas da BD, têm que ver também com a definição e análise do problema.

Linhas de Tratamento Paralelas

Tal como foi introduzido na descrição da recolha de dados, a ETAR em estudo contêm duas linhas de tratamento nas fases primária e secundária (PA2, PA3, PA4 e PA5). Em primeiro lugar é de referir que dos vários trabalhos relacionados com previsão em ETAR, nenhum deles menciona esta problemática de forma explícita. No entanto, adianta-se que em dois deles encontraram-se duas abordagens distintas na modelação de linhas de tratamento paralelas.

O tratamento com duas linhas levanta algumas questões, por vezes uma das linhas encontra-se inativa, não sendo assim registados os dados dessa linha, o que implica mais valores nulos. Se por um lado ao considerar-se duas linhas de tratamento tem-se uma maior dimensionalidade e mais valores nulos, por outro lado, ao considerar-se apenas uma linha de tratamento, reduz-se a dimensionalidade e a taxa de valores nulos. Isto fazendo as médias das duas linhas e, no caso de inatividade de uma delas, selecionar somente os dados da linha ativa. Contudo, ao realizar-se a média das duas linhas de tratamento está-se a distorcer os dados, mesmo sabendo que em PA6 o efluente converge das duas linhas de tratamento. Por conseguinte, decidiu-se analisar estas duas abordagens distintas.

Pegando no caso com maior dimensionalidade, considerámos inicialmente a abordagem com duas linhas de tratamento, contabilizando assim todos os parâmetros de ambas as linhas (ETAR_2L). Esta abordagem foi também usada por Dürrenmatt (2011), numa das aplicações investigadas no seu trabalho, na qual são considerados todos os parâmetros relativos aos vários tanques de arejamento, isto numa tarefa de previsão. Será ainda estudada uma outra abordagem deste problema, na qual se considera apenas uma linha de tratamento, cujos valores são as médias dos parâmetros tratados em paralelo (ETAR_1L). Ou seja, por exemplo para L1_PA2_SST e L2_PA2_SST, será realizada a média dos valores destes dois atributos, passando o resultado para o mesmo parâmetro ("PA2_SST") do conjunto de dados ETAR_1L. Os trabalhos relacionados, na sua grande maioria, sugerem apenas uma linha de tratamento. No entanto, não se consegue perceber se omitem o facto da ETAR em questão ter na realidade apenas uma linha de tratamento ou se o problema foi simplificado. Uma exceção é o trabalho de Hamed et al. (2004), no qual é referido que os parâmetros, cujo o tratamento é paralelo, contêm o valor médio das várias linhas de tratamento paralelas. Usando portanto a abordagem ETAR_1L, a explorar neste trabalho.

A análise destas duas abordagens é de grande interesse, uma vez que desta forma podemos compará-las e verificar as diferenças do desempenho de previsão entre as duas abordagens. Por outro lado, permite-nos também analisar se a redução de dimensionalidade proporcionará melhores desempenhos, ou então, se a distorção introduzida pela média afecta ou não as tarefas de modelação.

Decantador Primário / Seletor Biológico

Outro problema que surgiu durante a análise de dados, e que já foi apontado na consistência de dados, envolve o alternar de tratamentos entre o decantador primário e o seletor biológico. A inconsistência dos dados encontrada faz com que certas medidas tenham de ser tomadas por forma a não comprometer a modelação. Analisando o problema, a utilização de dois tratamentos diferentes no mesmo PA é, como se pode perceber, um problema de cariz técnico do tratamento das ETAR. Só um interveniente com grande conhecimento do processos de tratamento das águas residuais, como o analista da ETAR, é que terá a capacidade de o explicar. Contudo, sabe-se que são tratamentos distintos e que para os mesmos atributos são registados diferentes intervalos de valores. Adicionalmente, em alguns atributos só são registados os valores referentes ao tratamento com o decantador, enquanto que outros atributos apenas se referem aos tratamentos com seletor.

Constatou-se que os atributos somente referentes ao decantador, contêm uma taxa superior a 70% de valores nulos. No caso dos atributos somente referentes ao tratamento com seletor, as taxas são bem inferiores, na ordem dos 40%. Assim, conclui-se que o processo de tratamento com recurso ao seletor biológico é realizado com maior frequência. Sendo este problema muito próprio da ETAR em estudo, como se pode calcular não existe informação sobre esta problemática. Foram portanto tomadas algumas decisões por forma atenuar os efeitos deste problema. Em primeiro lugar decidiu-se considerar apenas os dados referentes ao seletor biológico, com base no facto da grande maioria dos dados serem relativos a este tratamento. Isto implica que as variáveis apenas referentes ao tratamento com decantador devem ser descartadas. No caso das variáveis com valores dos dois tipos de tratamento, que neste caso são apenas duas (PA2_SST e PA2_SSV), foram removidos os valores relativos ao decantador, eliminando assim a inconsistência dos dados. Esta foi a alternativa que nos pareceu mais simples e viável, uma vez que não traz complicações para o problema, nomeadamente o aumento de dimensionalidade. Adicionalmente, esta solução procura também minimizar a perda de informação, visto que são descartados os atributos com maior taxa de nulos. Outras estratégias poderiam ser equacionadas e até exploradas mas, estando perante um dos sete PA, poderemos afirmar que em princípio este facto não será determinante no desempenho da modelação. Ao contrário do problema das linhas de tratamento paralelas, que envolve quatro PA, correspondendo à maioria dos atributos usados na modelação.

4.2.3 Conjunto de Dados de Trabalho

Descrição do Conjunto de Dados

Com base na caracterização da BD que foi descrita na secção anterior, como vimos, foram removidas algumas variáveis consideradas menos importantes, que nos conduziram à obtenção do conjunto de dados de trabalho inicial. Considera-se este conjunto de dados inicial, como o conjunto de dados a partir do qual se irá iniciar o processo de tratamento e preparação de dados. Assim, após uma primeira análise à BD, a dimensionalidade foi reduzida na ordem dos 33%, passando de 122 a 82 atributos, enquanto que a percentagem de valores nulos na BD, passou de cerca de 35% para 20%, uma redução na ordem dos 43%. Na tabela 4.1 apresentam-se os atributos que constituem o conjunto de dados do problema, mencionando a sua descrição e também os seus PA.

Tabela 4.1 - Atributos do conjunto de dados inicial do problema

Parâmetros	Descrição	Pontos de Amostragem
Data	Data em que foi realizado o tratamento	Externo
Estação	Estação do Ano	Externo
Temperatura	Temperatura média ambiente, em graus celsius *(Cº)*	Externo
Estado Tempo	Condições climatéricas ou estado do tempo (sol, chuva)	Externo
Q_{en}	Caudal de Entrada - *(m^3/dia)*	PA1
Q_r	Caudal de Recirculação - *(m^3/dia)*	PA4; PA4'
Q_p	Caudal de Purga - *(m^3/dia)*	PA7
pH	Indica a acidez, neutralidade ou alcalinidade de uma solução aquosa.	PA1; PA2; PA2'; PA3; PA3'; PA5; PA5'; PA6
P_{redox}	Potencial de redução (Redox) - *(mV)*	PA1; PA3; PA3'; PA4; PA4'
P_{redox_ZA}	Potencial redox em zona aeróbia - *(mV)*	PA2; PA2'
P_{redox_ZAx}	Potencial redox em zona anóxica - *(mV)*	PA2; PA2'
O_2	Oxigénio Dissolvido - *(mg/l)*	PA3

O_{2_AZ}	Oxigénio Dissolvido em zona aeróbia - *(mg/l)*	PA2; PA2'
O_{2_ZAx}	Oxigénio Dissolvido em zona anóxica - *(mg/l)*	PA2; PA2'
CQO	Carência Química de Oxigénio - *(mg/l)*	PA1; PA5; PA5'; PA6
CBO	Carência Bioquímica de Oxigénio - *(mg/l)*	PA1; PA5; PA5'; PA6
SST	Sólidos Suspensos Totais - *(mg/l)*	Todos PA excepto PA7
SSV	Sólidos Suspensos Voláteis - *(mg/l)*	Todos PA excepto PA7
P_{Total}	Total Fósforo - *(mg/l)*	PA1; PA6
N_{Total}	Total Kjedldhal Nitrogénio - *(mg/l)*	PA1; PA6
$N\text{-}NH_4$	Amónio - *(mg/l)*	PA1; PA6
$N\text{-}NO_{3-}$	Nitrato - *(mg/l)*	PA6
CQO_eff	Eficiência de remoção de CQO - *(%)*	PA6
CBO_eff	Eficiência de remoção de CBO - *(%)*	PA6
SST_eff	Eficiência de remoção de SST - *(%)*	PA6
N_{Total}	Eficiência de remoção de N_{Total} - *(%)*	PA6
V30	Volume do lodo sedimentado após 30 min – *(ml/l)*	PA3
ISS*	Rácio entre SSV/SST - *(%)*	PA4
RAS*	Razão de absorção de sódio - (Q_r/Q_{en})	PA4
IVL*	Índice volumétrico de lodo – *(ml/g)*	PA3;PA3'
SRT*	Idade das lamas, ou tempo de retenção de sólidos – *(dias)*	PA4; PA4'
CBO/CQO*	Biodegradabilidade – *(CBO/CQO)*	PA1
FM*	Taxa de alimentação dos microrganismos	PA3; PA3'

* Variáveis de processo do tratamento, que geralmente correspondem a fórmulas que incluem alguns parâmetros de tratamento. Os PA descritos correspondem aos parâmetros incluídos nas fórmulas, cujo local de medição é o último a ser efectuado (só nesse momento poderá ser calculada a variável do processo).

Como se pode visualizar na tabela 4.1, existem parâmetros que são medidos em vários PA, logo cada PA implica um atributo do conjunto de dados. Assim, torna-se impraticável demonstrar aqui todos os dados estatísticos das variáveis. Porém, encontram-se em anexo duas tabelas com esses dados estatísticos básicos, tanto para o caso ETAR_2L como ETAR_1L (ver Anexo A).

Convém também referir neste ponto, a nomenclatura usada ao longo do documento para referir as variáveis do conjunto de dados. A nomenclatura adotada é, em geral, do tipo *"L_i_PA_j_VAR"* — onde L_i representa a linha de tratamento i (1 ou 2), apenas mencionado nos parâmetros com 2 linhas de tratamento; PA_j é o ponto de amostragem j (1, 2 ... 7); e o termo "VAR" representa o nome da variável ou parâmetro de tratamento. Por exemplo, o parâmetro SST da segunda linha de tratamento recolhido no PA3, é descrito como *"L2_PA3_SST"*.

Relacionamento entre Atributos

O estudo da interdependência dos dados é importante na medida em que proporciona não só uma melhor compreensão dos dados com que se está a lidar, como permite identificar variáveis redundantes. Com recurso à ferramenta *RapidMiner*, gerou-se a matriz de correlação dos atributos do conjunto de dados. Portanto, segundo essa matriz, foi possível identificar os atributos colineares. Apenas vão referir-se os atributos cujos coeficientes de correlação sejam elevados, pois, como se pode imaginar, a matriz gerada (82x82) fornece imensa informação.

Os valores da correlação quadrada R^2, também conhecida como coeficiente de determinação, vão de 0 a 1, onde 0 indica a ausência de corelacionamento linear, enquanto que 1 indica que os atributos são extremamente colineares. O relacionamento mais notório foi o de SST e SSV, estes atributos tiveram coeficientes de correlação entre si superiores a 0.9 em todos os PA, com exceção de PA6, onde o corelacionamento foi de 0.7. Também as variáveis PA2_Redox$_{ZA}$ e PA2_Redox$_{ZAx}$, têm coeficientes de correlação de cerca de 0.9. Todas estas variáveis devem ser portanto analisadas quanto à sua redundância na fase de tratamento de dados.

Foram adicionalmente encontrados relacionamentos interessantes entre variáveis de diferentes linhas de tratamento. Esse é o caso de PA2_Redox$_{ZAx/Za}$ e PA3_Redox que se correlacionam bem entre linhas, ou seja, por exemplo, L1_PA2_Redox$_{ZAx}$ e L2_PA2_Redox$_{ZAx}$ têm 0,7 de correlação. Note-se que estes corelacionamentos poderão ser importantes para o tratamento de nulos, uma vez que uma variável poderá estimar bem os valores nulos de outra variável de outra linha que lhe seja colinear.

Entre os parâmetros de PA6 os relacionamentos não são muito fortes, porém destacam-se os relacionamentos entre CQO e CBO com R^2= 0.4, entre CQO e SST/SSV com R^2= 0.3, e entre CBO e P_{Total} com R^2= 0.3. De acrescentar ainda, que PA6_CQO e PA6_CBO têm uma boa correlação com os mesmos parâmetros da fase anterior (PA5), cerca de 0.3 (L1 e L2), e 0.6 (L1) 0.4 (L2) respectivamente, isto no caso ETAR_2L.

No conjunto de dados ETAR_1L os valores de correlação são idênticos ao já demonstrados, quer para SST/SSV ou PA2_$Redox_{ZAx/Za}$. De acrescentar que as correlações em geral aumentaram ligeiramente, como no caso do relacionamento entre PA6_CBO e PA5_CBO que passou a ter 0.7 de correlação. É interessante ainda sublinhar que, apesar do forte relacionamento de SST/SSV em cada PA, no entanto a correlação entre diferentes PA não é considerável, em especial entre PA5 e PA6 onde é quase nula.

Uma nota final para os atributos relativos à eficácia de remoção em PA6. Como já foi referido, tendo em conta que estes atributos derivam de um cálculo com base em outros dois parâmetros do conjunto de dados, estes podem ser considerados redundantes. Todavia, analisando os valores de correlação desses atributos nota-se que não existe uma correlação muito forte, quer em relação aos respectivos atributos de PA1, quer aos de PA6. Em ambos os casos nenhum dos parâmetros tem valores de R^2 superiores a 0.3. Isto apenas nos indica que o corelacionamento linear é fraco, contudo não nos indica que não existe qualquer tipo de relacionamento. Note-se que variáveis linearmente correlacionadas, são também correlacionadas perante estimadores não lineares, no entanto o contrário não se verifica. Existem aliás algumas dificuldades práticas na descoberta de relacionamentos não lineares (Pyle, 1999).

Histogramas

Os histogramas são gráficos bastante úteis para a análise dos dados, para além de mostrarem a distribuição de uma variável, permitem ainda detectar facilmente a presença de *outliers*. Estes gráficos representam a contagem ou frequência de um atributo num determinado intervalo de valores, sendo que o número de intervalos (*bins*) a considerar são configuráveis. Apresentam-se em seguida alguns histogramas referentes a algumas variáveis de PA6 (CQO, CBO, SST, SSV, N_{Total} e P_{Total}), que foram consideradas importantes para a análise das variáveis de previsão (Figura 4.3).

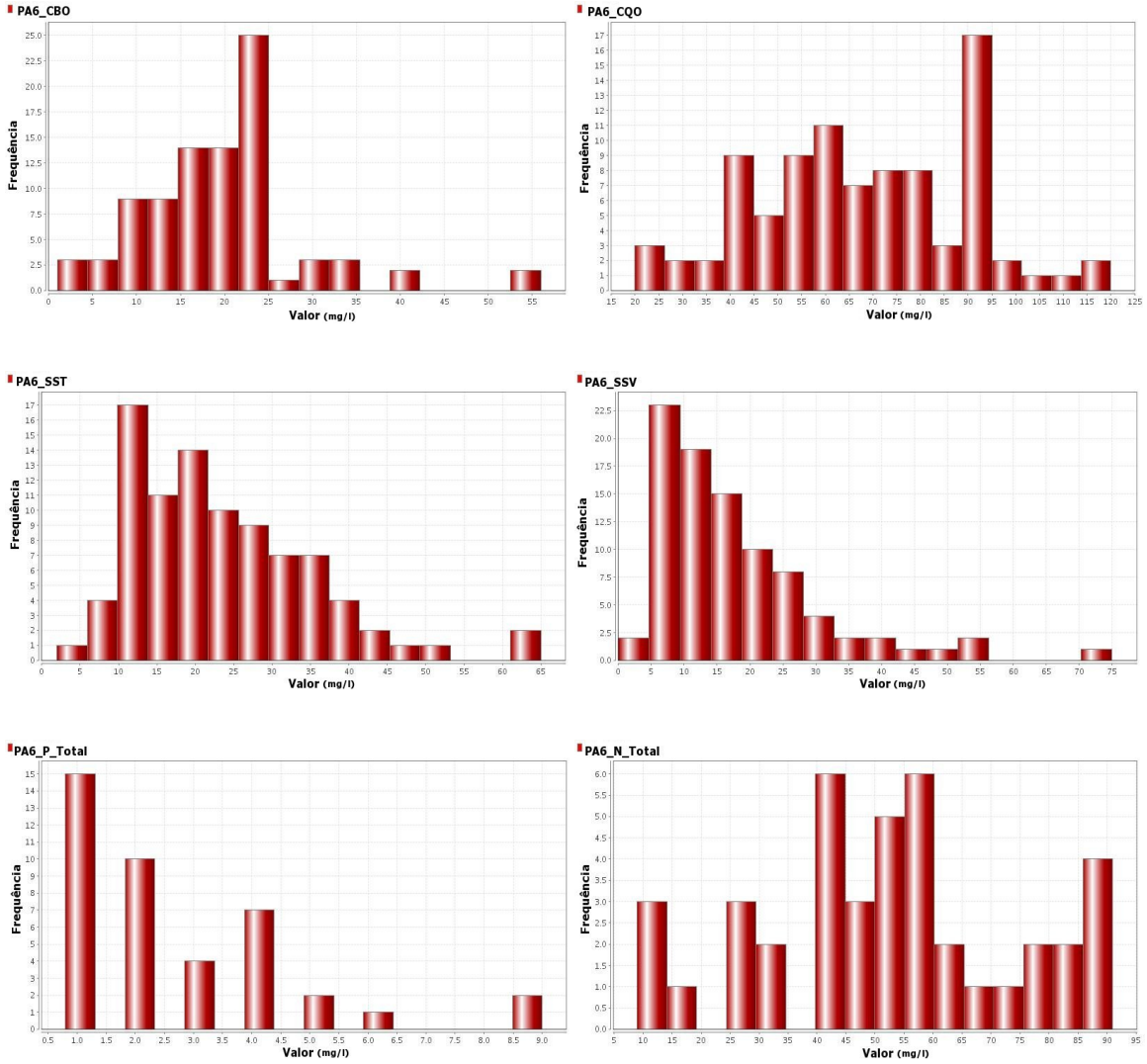


Figura 4.3 - Histogramas das variáveis CBO, CQO, SST, SSV, N_{Total} e P_{Total}, medidas em PA6

De referir que, perante os histogramas, podemos reparar na presença e *outliers* em CBO, SST, SSV e P_{Total}. Curiosamente as datas de tratamento dos dois *outliers* de CBO coincidem com um dos dois valores mais altos de P_{Total}, nessa altura, foi observado pelos técnicos da ETAR a ocorrência de problemas nos tanques de arejamento. Apesar da grande correlação entre SST e SSV, os respetivos *outliers* são referentes a registos de tratamento distintos. Ainda assim, tanto no caso do *outlier* de SSV como no de SST, são relatadas algumas dificuldades de tratamento nessa altura, o

que poderá indicar a presença destes valores anormalmente altos. Porém, só o entendimento de um especialista é que permitiria clarificar estas possibilidades.

Note-se ainda que em CQO os valores não ultrapassam os limites impostos por lei ($125_{mg/l}$), ao contrário de SST e CBO que ultrapassam os respectivos limites ($35_{mg/l}$ e $25_{mg/l}$). As distribuições de SST e SSV são muito idênticas, o que reforça o que já foi dito sobre a redundância destes atributos. Quanto aos nutrientes (N e P), apenas de referir que, olhando para as suas frequências, nota-se claramente que existem menos instâncias em comparação com as outras variáveis, isto deve-se à taxa de valores nulos mais elevada.

Foi realizada uma análise superficial aos histogramas da generalidade das variáveis do conjunto de dados, porém apenas se demostram neste documento estas seis variáveis, uma vez que será abordada em seguida a análise dos atributos de previsão, com base precisamente nestes atributos.

Varáveis de Previsão

É considerado como objetivo deste projeto, prever as variáveis que medem a qualidade do efluente final resultante do tratamento realizado pela ETAR. Isto é, pretende-se prever os parâmetros referentes ao PA6, que é o ponto imediatamente anterior à descarga do efluente para o meio-ambiente. Dentro deste grupo destacamos seis parâmetros CQO, CBO, SST, SSV, N_{Total} e P_{Total}, visto que, com a exceção de SSV, são todos parâmetros regulados por lei. Na verdade, o atributo SSV tem uma grande correlação com SST, daí, como no quadro de lei apenas constam os limites do parâmetro SST, será dada prioridade à previsão de SST sobre SSV.

Olhando agora para os nutrientes, estes são atributos interessantes de serem previstos, até porque medem o comportamento e eficácia do tratamento terciário. Todavia, tanto N_{Total} como P_{Total} contêm uma taxa de valores nulos na ordem dos 50%, o que implica uma redução para metade do conjunto de dados na sua previsão — lembrando que só faz sentido incluir os exemplos que permitam ao modelo "aprender" com a variável de previsão. Portanto, este facto relega os nutrientes para baixo na lista de variáveis com maior interesse de previsão.

Os restantes parâmetros CBO, CQO e SST, no que diz respeito aos nulos não apresentam grandes problemas, apenas CBO contêm dois valores nulos, cerca de 2%. Relativamente ao parâmetro CQO, reparou-se que nenhum dos valores ultrapassa os limites impostos por lei, o que significa que não se conseguirá modelar maus desempenhos da ETAR relativamente a este parâmetro. Ao contrário de CQO, temos que CBO e SST incluem vários registos fora dos limites exigidos. É de realçar ainda que as análises do parâmetro CBO demoram cerca de 5 dias a serem obtidas, o que reforça o interesse de prever este parâmetro. Com base nestes pressupostos, consideraram-se então os atributos CBO e SST como as variáveis de previsão a modelar neste trabalho.

4.3 Preparação de dados

Depois da análise de dados que nos conduziu ao conjunto de dados acima demonstrado, vamos agora descrever as técnicas de preparação de dados efectuadas. Convém referir nesta fase, quais as condições necessárias para que os algoritmos de modelação SVM possam lidar com o conjunto de dados. Em relação ao tipo de dados, o algoritmo requer dados de input numéricos. Logo todos os atributos nominais têm de ser tratados. Outra característica destes algoritmos é que não conseguem lidar com valores nulos, o que implica que deve ser feito um tratamento de nulos antes da modelação. Existem ainda outras técnicas de preparação de dados que, embora não sejam estritamente necessárias, são recomendadas para se obter uma boa modelação. Convém ainda referir que o processo de preparação de dados foi iniciado com o conjunto de dados ETAR_2L. Contudo, pode-se desde já adiantar que para o caso ETAR_1L o processo foi em geral bastante semelhante.

4.3.1 Conversão de Dados Nominais para Numéricos

Os algoritmos SVM a explorar neste trabalho pertencem ao grupo de algoritmos de DM que não suportam dados de tipo não numérico. Por conseguinte, sabendo que as SVM apenas lidam com dados numéricos, é perceptível a necessidade de converter todos os atributos nominais para um

tipo de dados numérico. Deste modo, é importante realçar que deve ser realizado um mapeamento de valores nominais para valores numéricos apropriados, procurando assim não distorcer a informação contida nos valores nominais. Neste caso de estudo apenas constam do conjunto de dados três variáveis nominais: a data, a estação do ano e o estado do tempo.

Data

Esta variável não apresenta qualquer tipo de problema na preparação de dados, pois trata-se na realidade da variável que identifica um registo de tratamento (*id*), uma vez que um registo corresponde a um dia de tratamento. Por conseguinte, esta não interfere na modelação, pois o *id* não é utilizado pelo algoritmo de modelação.

Estação do Ano

Esta variável contém, como é evidente pela sua própria designação, as estações do ano, "Inverno", "Primavera", "Verão" e "Outono", seguindo exatamente esta ordem devido ao conjunto de dados ser anual. Aqui, como estamos perante quatro valores categóricos, poderá ser equacionado um mapeamento de valores nominais para numéricos. Poderia se aplicar aqui um mapeamento *one-of-n,* que implica uma variável binária para cada estação, sendo que uma variável só ficaria igual a verdadeiro (1), no caso de se tratar da respetiva estação do ano. Outra possibilidade seria representar com duas variáveis o efeito cíclico que, como sabemos, ocorre nas estações do ano (Pyle, 1999). Uma solução mais prática, mas também mais ingénua, seria numerar as estações de forma incremental (0, 1, 2 e 3), ficando-se assim apenas com uma variável.

Como apenas temos registos referentes a um ano o efeito cíclico não se chega a verificar, então foram apenas a analisadas as outras duas alternativas. À imagem das decisões que foram tomadas ao longo desta fase de preparação de dados, a regra de ouro para saber qual das alternativas a melhor, é experimentar. Seguindo assim um processo empírico, criaram-se alguns modelos SVM simples e compararam-se os resultados. O mapeamento *one-of-n* não revelou nenhumas melhorias em comparação com o método de numeração incremental. Na verdade os resultados foram inferiores, embora de forma bastante ligeira. Assim, optou-se pela conversão: "Inverno" —> 0;

"Primavera" —> 1; "Verão" —> 2; "Outono" —> 3. É também de acréscimo para esta decisão, o facto de não aumentar-se a cardinalidade dos dados.

Estado do Tempo

A variável "Estado do Tempo" contêm dois valores, "Chuva" e "Sol". Neste caso a conversão é mais intuitiva, pois com uma variável numérica binária (0 e 1), consegue-se caracterizar perfeitamente as duas categorias "Chuva" e "Sol". Esta foi portanto a medida adotada na conversão desta variável nominal para o tipo numérico.

4.3.2 Normalização de Dados

Neste contexto, a normalização de dados significa alterar o intervalo de valores das variáveis do conjunto de dados, para que estas fiquem todas dentro do mesmo limite de valores — o que não tem nada em comum com a normalização das BD. Deste modo, a normalização faz com que todas as dimensões do espaço de dados fiquem dentro do mesmo intervalo de valores. Este formato do espaço de dados permite, assim, que o conjunto de dados fique mais fácil de ser "interpretado" pelos algoritmos de modelação. Deste modo, com uma maior facilidade de aprendizagem, os algoritmos poderão atingir níveis de eficácia superiores. Alguns algoritmos (i.e. NN), requerem mesmo que o conjunto de dados esteja normalizado, porém esse não é o caso das SVM. Não obstante, mesmo nos algoritmos que não requerem a normalização dos dados, estas técnicas poderão trazer benefícios, por vezes de forma significativa (Pyle, 1999).

Assim, realizaram-se alguns testes com modelos SVM simples, a fim de medir os efeitos que a normalização tem no conjunto de dados. Destacaram-se duas técnicas de normalização: $MinMax_{-1:1}$, que normaliza os atributos para intervalo de valores [-1:1]; e $MinMax_{0:1}$, que é igual à anterior, mas para o intervalo [0:1]. Estas duas técnicas apresentam no entanto um problema. Como os valores são normalizados com base nos dados de treino, o que é que acontecerá se na fase de teste os dados tiverem valores fora dos limites dos dados de treino? Esta situação torna-se portanto um problema, principalmente na implementação de um sistema de previsão. No entanto,

existem técnicas de normalização que, mesmo normalizando todos os atributos para o mesmo intervalo de valores, fazem com que estes nunca atinjam os extremos, deixando assim "espaço" para novos valores fora dos limites dos valores dos dados de treino. Este é o caso da técnica *SoftMax Scalling*, apresentada por Pyle (1999). O *RapidMiner* não disponibiliza esta técnica mas, com recurso ao *package DMwR* (Torgo, 2010) e à integração do *R* com o *RapidMiner,* torna-se possível testar esta técnica. Contudo, este software não disponibiliza o método de desnormalizarão, que é muito importante principalmente na fase de modelação para se analisar os resultados.

Com base nos testes realizados verificou-se em primeiro lugar que a normalização trouxe melhorias significativas aos modelos SVM de teste. Também se verificou que todas as técnicas de normalização superaram a não normalização. No caso da técnica *SoftMax Scalling* comparando-a segundo a medida de correlação, que é independente da escala, os resultados são muito próximos de $MinMax_{-1:1}$, chegando a supera-lo ligeiramente em alguns casos (SMO). A técnica $MinMax_{-1:1}$ mostrou-se a mais consistente perante os modelos de testes SVM (SVR e SMO), sendo a melhor no primeiro caso de forma até destacada. Todavia, no segundo caso, teve resultados muito próximos de $MinMax_{0:1}$ e ligeiramente inferiores à normalização *SoftMax.* Com base nestes factos e como este trabalho se trata de um estudo, não sendo por isso muito grave o problema de valores fora dos limites, foi adotada a técnica de normalização $MinMax_{-1:1}$.

Este estudo comprovou ainda que a normalização traz benefícios aos modelos. De salientar que no trabalho de Ali & Smith-Miles (2006) é testada a eficácia da normalização em modelos SVM num total de 112 conjuntos de dados diferentes. Esta investigação demonstra que na grande maioria dos conjuntos de dados testados, a normalização aumentou o desempenho dos modelos, tanto na eficácia de previsão, como no tempo de processamento dos algoritmos. Destaque-se ainda as técnicas *MinMax* que também nesse estudo apresentaram bons resultados.

4.3.3 Tratamento de Valores Nulos

Os valores nulos representam geralmente um grande problema dos conjuntos de dados. Aqui o nosso caso também não é exceção. Como grande parte dos algoritmos, incluindo SVM, não conseguem lidar com valores nulos, estes têm de ser tratados. Duas posições podem ser assumidas perante este problema, remover todos os casos com nulos, ou então substituir os nulos por valores estimados (imputação de nulos). A remoção de exemplos ou atributos com nulos não parece ser a melhor solução, uma vez que o conjunto de dados apenas contêm 16 casos completos, ou seja, 16 instâncias sem nulos. Ficaríamos assim com uma amostra de dados que captura apenas uma pequena parte de problema. Assim, neste caso, os valores nulos devem ser imputados através da estimação do seu valor. Contudo, deve-se ter em atenção os valores de substituição dos nulos, pois estes podem introduzir ruído ou viés no conjunto de dados. A imputação de nulos segundo valores por defeito que não representem minimamente a realidade do atributo, podem distorcer os dados. O objetivo passa então por capturar a informação presente nos dados, tanto das variáveis com nulos como do relacionamento entre variáveis, e então estimar os valores a imputar. Note-se que este procedimento deve ser realizado de modo a introduzir o mínimo ruído possível. Isto, por forma a não distorcer a informação já presente no conjunto de dados (Pyle, 1999).

No entanto, deve-se ter em atenção que nem todos os valores nulos podem ser substituídos. Os valores nulos podem ser divididos em três tipos (Little & Rubin, 1987), conforme citado em (Batista & Monard, 2003):

- **Completamente ao Acaso (*Missing Completely at Random - MCAR*)** – A probabilidade de uma instância ter valores nulos num determinado atributo, não depende dos valores conhecidos nem dos próprios valores nulos. Neste caso qualquer método de imputação de nulos pode ser aplicado.

- **Ao Acaso (*Missing at Random - MAR*)** – A probabilidade de uma instância ter valores nulos num determinado atributo, pode depender de valores conhecidos, mas não dos próprios valores nulos.

- **Não ao Acaso (*Not Missing at Random - NMAR*)** – A probabilidade de uma instância ter valores nulos num determinado atributo pode depender dos próprios valores desse atributo.

Convém apresentar dois exemplos simples para ilustrar estes conceitos. Considere-se que existem valores nulos numa variável "salário". Imaginando que as pessoas mais idosas não gostam de divulgar o seu salário, então, para este caso, os valores nulos de salário dependem da variável idade, estamos perante nulos do tipo MAR. Agora, imaginando um caso em que se indicia que só existem valores nulos em salário quando os salários são baixos, então, como depende do próprio atributo, estamos perante nulos do tipo NMAR. Nos dados da ETAR, como os valores nulos correspondem essencialmente à ausência de processos de tratamento num determinado dia, que é bem diferente dos dois conceitos exemplificados acima, então podemos concluir que os valores nulos são do tipo MCAR. Logo podem ser aplicadas qualquer tipo de técnicas de imputação de nulos.

O problema passa agora por saber qual a técnica de imputação de nulos mais indicada para o nosso problema. No trabalho de Luengo et al. (2011) é realizada uma extensa bateria de testes onde são experimentadas diversas técnicas de imputação de nulos, todas elas aplicadas em 21 conjuntos de dados diferentes e testadas em 23 classificadores distintos. Nesse estudo são tiradas várias conclusões importantes, porém duas delas assumem particular relevância. A primeira diz que não imputar os nulos, ou seja, não considerar os exemplos com nulos, é geralmente superada pelos métodos de imputação de nulos. A segunda diz que não existe um método de imputação que seja melhor em todos os casos. Com isto em mente, foram testados alguns métodos de imputação de nulos no nosso estudo. Embora duas técnicas de imputação de nulos (*Fuzzy K-means Imputation* e *Eventcovering,*) tenham-se mostrado superiores na generalidade dos testes efetuados em (Luengo et al., 2011), o *RapidMiner* não as disponibiliza. Foram então testados alguns métodos de imputação de nulos, com recurso ao *RapidMiner*.

O MMI (*Mean Mode Imputation*), consiste em imputar os nulos com a média dos valores do respectivo atributo. Esta é uma abordagem bastante usada e, embora substitua todos os nulos de um atributo pelo mesmo valor padrão, é referida como uma técnica que traz bons resultados na

prática. Esse facto verifica-se no trabalho de Luengo et al. (2011), no qual em vários testes o MMI revelou ser um dos melhores métodos de imputação. Os outros métodos testados consistem em algoritmos executados juntamente com o operador (*Impute Missing Values*) do *RapidMiner.* Este operador apenas suporta alguns algoritmos de regressão, foram assim testados o k-NN, NN e SMO. Com este operador os algoritmos internos desenvolvem modelos de previsão para cada atributo a imputar, excluindo o atributo de previsão, a partir dos quais são posteriormente previstos os valores nulos com base nos modelos que foram desenvolvidos. Dependendo da capacidade do algoritmo lidar com nulos, estes podem aprender apenas com base nos casos completos ou não. Por exemplo, o k-NNI consegue lidar com valores nulos, logo não é obrigatório que realize a aprendizagem com casos completos, enquanto que SMOI e NNI só conseguem aprender com base nos casos completos. Contudo, pode-se já adiantar que a aprendizagem com base nos casos completos revelou melhores resultados com k-NNI. Adicionalmente, foi testada outra técnica que consiste na exploração do algoritmo SVR para imputação de valores nulos, isto segundo a proposta de Feng et al. (2005). Esta abordagem é semelhante à descrita no operador do *RapidMiner* para imputação de nulos, ou seja, realiza a aprendizagem dos modelos através dos casos completos, alternando as variáveis de input e de output, e por fim são previstos todos os valores nulos. Como o operador do *RapidMiner* para imputar nulos não suporta algoritmos SVR, foi desenvolvido um subprocesso que desempenha exatamente este algoritmo.

Os resultados dos testes realizados revelaram que nem todos os métodos superaram a imputação da média do atributo (MMI). Adicionalmente, é curioso que os resultados das técnicas testadas não são muito diferentes, porém, ainda assim, o k-NNI mostrou-se o mais equilibrado em ambos os modelos de teste (SVR e SMO). Das restantes técnicas o SMOI e SVRI revelaram resultados semelhantes ao K-NNI, enquanto que as redes neuronais (NNI), tiveram resultados ligeiramente superiores nos modelos de teste SMO, mas inferiores com SVR. Note-se ainda que a alteração do número de vizinhos mais próximos a considerar (k) tem influência no desempenho, tal como os parâmetros de configuração dos restantes algoritmos. Sendo o k-NNI bastante simples de configurar (apenas um parâmetro), e como os seus resultados são de uma maneira geral superiores, este foi o método de imputação adoptado para tratamento de nulos. De acrescentar que dependendo da modelação a realizar, neste ponto deve ser ajustado o parâmetro *k* de k-NNI, isto, para cada uma das tarefas de modelação.

Para finalizar esta secção, deixemos uma nota para dois valores nulos correspondentes a dois exemplos do parâmetro de previsão PA6_CBO. Como se trata da variável de previsão, estes dois exemplos devem ser descartados, pois é a partir desta que é realizada a aprendizagem e não faz sentido imputar estes valores porque poderíamos comprometer a modelação. Assim, nas modelações cujo parâmetro de previsão é CBO, fica-se com 88 exemplos.

4.3.4 Remoção de Atributos Redundantes

Antes de se passar à fase de modelação, onde será inclusivamente realizada uma seleção de atributos, decidiu-se ainda na preparação de dados analisar os atributos redundantes que possam afetar a modelação. Os atributos redundantes, quando em demasia, tornam o processamento mais lento, porém este pode ser um mal menor. Estas variáveis redundantes também prejudicam os algoritmos na sua tarefa de modelação, por um lado porque mais dimensionalidade implica maior dificuldade de modelação, por outro lado porque alguns algoritmos, especialmente métodos de regressão, podem mesmo ter problemas em lidar com estes dados (Pyle, 1999).

Na secção anterior foram levantados alguns possíveis atributos redundantes, nomeadamente os pares SST/SSV, PA2_Redox$_{ZA}$/PA2_Redox$_{ZAx.}$ e os atributos de eficiência de remoção, sendo estes últimos atributos derivados. De referir ainda que as variáveis de processo também são atributos derivados, embora as suas formulações sejam em geral bem mais complexas, contemplando por vezes mais de dois atributos. Ao analisar a potencialidade de redundância dos atributos de eficiência de remoção, constatou-se que o modelo de teste SVR piorou, o que leva a concluir que estes atributos não devem ser removidos nesta fase. Quanto aos atributos que se revelaram colineares, a remoção de SSV trouxe melhorias na previsão de CBO em ambos os modelos de teste. No entanto, apesar de os atributos serem colineares entre si, estes podem também ter um bom relacionamento com os atributos de previsão. Essa situação acontece na previsão de SST, na qual a remoção dos atributos redundantes SSV mostrou resultados inferiores. Assim, em função da modelação a ser realizada, os atributos redundantes são removidos ou não. Esta situação aplica-se também aos atributos redundantes PA2_Redox$_{ZA}$/PA2_Redox$_{ZAx}$. Não obstante, já na fase de modelação, uma seleção de atributos mais avançada irá filtrar convenientemente os atributos

resultantes da preparação de dados. Contudo, a sua remoção ainda nesta fase ajuda bastante a aliviar a complexidade, que é principalmente notória ao executar os algoritmos de seleção de atributos.

Em cada tarefa de modelação a realizar, será revista a preparação de dados em função dos objetivos da modelação em estudo. Esta é, até, uma situação prevista no modelo de desenvolvimento CRISP-DM. Assim, em jeito de conclusão, devem ser revistos principalmente os pontos referentes ao tratamento de nulos, no ajuste do parâmetro *k*, e à remoção de atributos redundantes.

4.4 Sumário

Olhando para o problema de tratamento de águas residuais apresentado anteriormente no capítulo 2, considera-se o principal objetivo deste projeto, segundo uma perspetiva de negócio, encontrar um modelo de previsão que capture toda a dinâmica do processo de tratamento. Adicionalmente, qualquer tipo de informação que possa ser útil para a ETAR, deve ser extraída e posteriormente apresentada. Para realizar estes objetivos adotou-se a ferramenta de DM *RapidMiner*, que foi utilizada ao longo do processo de desenvolvimento deste projeto.

No início de desenvolvimento deste trabalho de DM, foram centralizados os dados recolhidos pela ETAR numa única BD. Esses dados correspondem aos registos diários de tratamentos realizados durante o período de um ano, onde em cada registo são medidos cerca de 122 parâmetros correspondentes a sete PA, sendo que quatro deles correspondem a linhas de tratamento paralelas. Após este processo de recolha de dados, foi realizado um processo de análise dos dados. Assim, olhando para a BD foram analisadas as suas características em termos de: atributos, consistência, integridade, poluição dos dados e valores nulos. Algumas destas características, especialmente a de valores nulos, levam com que sejam descartados atributos com taxas de valores nulos muito elevadas.

Levantaram-se ainda outros problemas característicos da BD que são importantes para a definição do problema, logo algumas posições tiveram de ser adotadas por forma a contornar tais problemas. Uma dessas posições consiste no estudo de duas abordagens distintas, uma que considera um conjunto de dados que inclui as duas linhas de tratamento paralelas (ETAR_2L) e outra abordagem que simplifica a dimensionalidade considerando apenas uma linha de tratamento através da média das duas linhas (ETAR_1L). Depois de uma análise inicial, na qual algumas variáveis foram descartadas, chegou-se ao conjunto de dados de trabalho inicial. Esse foi o conjunto de dados a partir do qual realizou-se o processo de preparação e tratamento de dados. Após a análise de dados, tanto a dimensionalidade do conjunto de dados foi reduzida para 82 atributos, como a taxa de valores nulos diminuiu em cerca de 43%. Com base neste conjunto de dados, analisaram-se ainda as potenciais variáveis de previsão do problema. Por conseguinte, concluiu-se que serão previstos neste problema os parâmetros de qualidade CBO e SST.

Antes de se iniciar o processo de modelação, é necessário que os dados estejam preparados para os algoritmos de modelação lidarem com eles. Deste modo, foram aplicadas técnicas de tratamento de dados ao conjunto de dados de trabalho. Como os algoritmos SVM apenas lidam com atributos numéricos, então todos os atributos nominais foram convertidos para valores numéricos. Posteriormente, aplicaram-se técnicas de normalização de dados, por forma aos algoritmos interpretarem mais facilmente o dados. Uma limitação dos algoritmos SVM passa por não conseguirem lidar com valores nulos, torna-se assim necessário tratar os valores nulos do conjunto de dados. Para este efeito foram testadas várias técnicas de imputação de nulos, com o intuito de encontrar a que melhor se ajusta ao nosso problema. O algoritmo k-NN juntamente com o operador de imputação de nulos do *RapidMiner*, revelaram bons resultados, sendo portanto a técnica de imputação de nulos adotada. Para terminar a fase de preparação de dados, foram eliminados os atributos redundantes. Neste ponto, chegou-se à conclusão que a remoção dos atributos redundantes não foi benéfica em todas as tarefas. Então, para diferentes tarefas de modelação (i.e. CBO ou SST), este passo deve ser revisto, por forma a remover somente os atributos redundantes que não prejudiquem os modelos. Adicionalmente, também as restantes tarefas de preparação de dados podem ser revistas, em particular no tratamento de nulos, em que para cada modelo pode-se ajustar o parâmetro *k* do algoritmo de imputação.

Capítulo 5

Modelação da ETAR

O processo de modelação preditiva da ETAR, pressupôs uma estratégia concreta de modelação. Este processo é composto por um conjunto de técnicas que devem ser adotadas com base no problema a modelar e que, no seu todo, produzem modelos de previsão mais eficientes e ajustados à realidade do problema. Após obter-se um conjunto de dados já preparado para os algoritmos de modelação, segue-se a fase de modelação da ETAR. Seguindo a descrição da preparação de dados, apresenta-se neste capítulo o desenvolvimento da modelação preditiva referente ao conjunto de dados ETAR_2L. Quanto à abordagem ETAR_1L, esta será apresentada mais tarde no capítulo seguinte. Apesar disso, refere-se desde já que o processo de desenvolvimento de modelação foi análogo nas duas abordagens.

5.1 Avaliação de Modelos

Iniciamos esta fase de modelação com a análise dos métodos de avaliação mais indicados para o nosso problema. A avaliação ou estimativa de eficácia dos modelos, não só é importante para estimar a futura capacidade de previsão de um modelo, como é também importante para selecionar o melhor modelo de previsão (Kohavi, 1995). Assim, atendendo aos objetivos deste trabalho, é crucial utilizarmos métodos de avaliação que nos permitam obter resultados fiáveis, o que nos reforçará a validade deste trabalho.

No capítulo 3 foi introduzido o princípio de minimização do erro estrutural. Na verdade, este principio é um método de avaliação de máquinas de aprendizagem, das quais derivam modelos de previsão, cujo objetivo é encontrar o modelo com menor erro esperado e, consequentemente, que generalize melhor (não sobre ajustado). Na prática o princípio de minimização do risco estrutural é pouco utilizado, porém o fundamento de minimização do erro esperado persiste nos métodos de avaliação mais utilizados na prática. Como sabemos, é estimado o erro dos modelos por forma a medir o seu desempenho preditivo. O erro resume-se, assim, essencialmente em duas componentes: a viés e a variância. A viés ou distorção, consiste na diferença entre o valor esperado e o valor estimado pelo modelo. A segunda componente indica a variância estimada pelos modelos, ou seja, a variação (distância) da média dos desvios dos valores estimados. Portanto, para medir a eficácia dos modelos de previsão, pretende-se um método de avaliação com viés baixa e variância também baixa.

5.1.1 Métodos de Avaliação de Modelos

Nas aplicações de DM são normalmente usados os seguintes métodos de avaliação: *Holdout, Cross-Validation e Bootstrap.* A utilização destes métodos de avaliação depende no entanto do tamanho do conjunto de dados de trabalho, sendo uns métodos mais indicados para determinados conjuntos de dados. Como já foi observado, o conjunto de dados deste caso de estudo é pequeno. Contém apenas 90 exemplos válidos. Este facto requer métodos de avaliação apropriados ao tamanho do conjunto de dados, por forma a não sobre ajustar demasiado os modelos aos dados de treino e também a capturar o máximo número de exemplos possível na modelação. Perante esta situação, foram procuradas as melhores práticas que são geralmente usadas em problemas com conjuntos de dados pequenos.

Em primeiro lugar, em vários trabalhos de investigação sobre este problema, é referido que o método de avaliação *Holdout* não é o mais indicado para conjuntos de dados pequenos. Este método consiste em dividir os exemplos do conjunto de dados em dois subconjuntos, um para treino e outro para teste, normalmente na proporção de 2/3 e 1/3 respectivamente. Acontece que ao prescindirmos de cerca de 30% dos dados na fase de treino, não se captura toda a informação disponível. Esse facto leva ao desenvolvimento de modelos incompletos, porém quando o

problema tem bastantes exemplos (i.e. >5000) a falta desses valores não se torna tão crítica na modelação. Como se pode imaginar, com cerca de 3300 exemplos captura-se muita mais informação do que com 60 exemplos, daí que este método de avaliação só é aconselhado para conjuntos de dados com alguma dimensão.

A validação cruzada, conhecida do Inglês como k-*fold Cross-Validation*, consiste em distribuir n exemplos por k partições. Sendo que o número de exemplos nas partições é aproximadamente o mesmo, e igual a n/k. Assim, os algoritmos de previsão realizam k vezes o processo de treino e de teste, rodando as partições de forma a que em cada iteração fique uma partição de fora, para teste, e as restantes sejam usadas na aprendizagem do modelo. No final, é estimada a média das k avaliações realizadas nas partições de teste. Este método de validação, ao contrário do *Holdout*, garante que todos os exemplos sejam utilizados na construção do modelo como dados de treino, devido à rotação das partições. Uma variante muito conhecida de validação cruzada é o *leave-one-out* (LOO), que consiste, basicamente, numa validação cruzada, cujo o número de partições é igual ao número de exemplos *(k=n)*. Em cada iteração só é testado um exemplo, sendo os restantes exemplos usados em treino, e as k rotações das partições, neste caso, indicam que todos os exemplos são testados e consequentemente avaliados individualmente.

O método *Bootstrap* consiste em retirar de um conjunto de dados com n exemplos, uma amostra de tamanho também n, mas cujos exemplos da amostra podem ser repetidos (amostragem com repetição). Deste modo, os exemplos que não entrarem na amostragem são usados para teste. Tal como na validação cruzada, no *Bootstrap* este processo é realizado *k* vezes. A taxa de acerto é calculada assim com base na média das avaliações dos *k* conjuntos de treino+teste, em que para cada amostragem o erro é calculado segundo a fórmula característica do *Bootstrap.632, que* contempla as probabilidades das instâncias não entrarem na amostra (0.382) e de serem distintas (0.632) (Kohavi, 1995).

Como o método *Houldout* não é recomendado para conjuntos de dados pequenos, a grande maioria dos trabalhos de investigação sobre os métodos de avaliação, compara essencialmente os métodos baseados em *Cross-Validation* e *Bootstrap.* A avaliação 10-CV (*10-fold Cross-Validation)* é afirmada em vários estudos como sendo um bom método de avaliação de uma forma geral (Kohavi, 1995; Molinaro et al., 2005). Por sua vez o *Bootstrap*, apesar de ter menor variância que o 10-CV, em alguns caso tem mais viés. Por conseguinte, avaliação 10-CV é melhor na

generalidade dos casos, sendo até recomendada para seleção de modelos em (Kohavi, 1995). Adicionalmente, a repetição de 10-CV permite reduzir a variância, superando assim a avaliação 10-CV normal. Com base nisto, segundo Kim (2009), a avaliação de modelos com repetição de 10-CV é recomendada para uso geral, pese embora o *Bootstrap* ser o melhor método de avaliação em alguns casos, porém noutros revela comportamentos divergentes. Por conseguinte, com base nas recomendações dos estudos referidos acima, foi adotado o método de avaliação com 10 repetições de 10-CV (10R-10CV), que perfaz um total de 100 conjuntos diferentes de treino+teste que foram testados. Note-se que as repetições, consistem em diferentes amostragens dos exemplos que são distribuídos pelas 10 partições. Este método de avaliação será portanto adoptado em toda a fase de modelação, procurando assim reduzir o risco de sobre ajustamento que é bastante comum em problemas com poucos exemplos. Convém ainda destacar que será apresentado no final deste documento, nomeadamente nas considerações finais, uma análise comparativa de três métodos de avaliação recomendados em conjunto de dados pequenos (10R-10CV, LOO e Bootstrapping). Esta análise tem o intuito de estudar o possível sobre ajustamento dos dados e, consequente, a possibilidade dos resultados das avaliações estarem sobre estimados ou até subestimados.

5.1.2 Métricas de Desempenho

Ainda na avaliação dos modelos, é importante referir as métricas de desempenho, ou de avaliação, utilizadas para medir a capacidade preditiva dos modelos. Neste estudo foram analisadas essencialmente três métricas, a raiz do erro quadrático médio, a raiz do erro quadrático médio relativo e o coeficiente de correlação.

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

$$RRSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \cdots + (a_n - \overline{a})^2}}$$

$$R = \frac{\sum_i (p_i - \overline{p})\,(a_i - \overline{a})/n-1}{\sqrt{(\sum_i (p_i - \overline{p})^2/n-1) * (\sum_i (a_i - \overline{a})^2/n-1)}} \quad \textbf{(5.1)}$$

Estas três medidas são apresentadas no livro de Witten et al. (2011), e serão descritas neste documento segundo as siglas da terminologia apresentada nesse trabalho, que são também mais familiares no universo de DM.

A raiz do erro quadrático médio (do Inglês, *Root Mean Squared Error* - RMSE), é uma medida de desempenho bastante usada, que consiste, como o próprio nome indica, no cálculo da raiz quadrada aplicada ao erro médio quadrado, que foi introduzido anteriormente na equação 3.2. Neste caso, sabendo que p corresponde aos valores previstos, a indica os valores reais e n o número de instâncias, o erro é calculado com base na média das diferenças, entre os valores previstos e esperados, elevada ao quadrado. Ao aplicar a raiz quadrada aos erros médios quadrados, estes ficam na mesma escala que os valores do parâmetro de previsão. Daí, esta ser uma métrica bastante usada, pois mostra o valor do erro à escala real do atributo de previsão. Convém ainda referir, que a métrica RMSE foi a métrica utilizada como critério principal na seleção de atributos e otimização de parâmetros.

A raiz do erro quadrático médio relativo (do Inglês, *Root Relative-Squared Error* - RRSE), é uma métrica interessante, uma vez que mostra o valor do erro relativamente ao erro cometido por um preditor simples, cujas previsões são sempre a média do atributo de previsão. Isto permite-nos comparar a performance dos modelos em estudo, com um modelo que pode ser considerado como ingénuo. Note-se ainda, que valores de 100% de erro relativo indicam um desempenho igual ao modelo simples, logo, quanto mais baixos são os valores de RRSE, melhor é o desempenho do modelo em estudo.

O coeficiente de correlação (do Inglês, *Correlation Coefficient*), usualmente apresentado como *R*, é uma métrica interessante, pois distingue-se das restantes pela independência de escala, o que foi importante para analisar os modelos de teste na fase de preparação de dados. Contrariamente às métricas anteriores, cujos valores mais baixos indicam erros menores e logo melhor desempenho, na métrica R os valores mais altos e mais próximos de 1 significam melhores resultados. Os valores desta métrica, consistem assim no intervalo entre 0 e 1, tal como foi apresentado na análise de relacionamentos entre variáveis.

5.2 Seleção de Atributos

Desde o momento em que se apresentaram os dados da ETAR, tem sido referido por diversas vezes a questão relacionada com a sua alta dimensionalidade. A grande dimensionalidade de um conjunto de dados traz várias dificuldades à modelação, algumas já referidas anteriormente no capítulo 3. Assim, para atenuar os efeitos da alta dimensionalidade, em particular, da complexidade de processamento e da dificuldade de aprendizagem dos algoritmos, deve-se proceder à remoção de variáveis. Este processo de redução de dimensionalidade iniciou-se na análise de dados, removendo as variáveis menos úteis para o problema e, em seguida, na eliminação de variáveis redundantes. Contudo, ainda assim temos muitas variáveis para o reduzido número de exemplos, portanto torna-se importante analisar técnicas de seleção de atributos.

As técnicas de seleção de atributos têm como objetivo geral melhorar o algoritmo de aprendizagem, quer em termos de velocidade, de capacidade generalização ou ainda de simplicidade de representação. É assim possível obter melhores resultados preditivos através da diminuição do volume de armazenamento, da redução do ruído gerado por atributos irrelevantes e redundantes, e da eliminação de conhecimento irrelevante (Molina et al., 2002). Por conseguinte, aplicaram-se técnicas de seleção de atributos com dois objetivos distintos. O primeiro é reduzir a dimensionalidade por forma a aumentar a capacidade preditiva dos modelos, enquanto que o segundo passa por encontrar os atributos mais importante na previsão de CBO e SST, adquirindo desta forma algum conhecimento sobre os dados e, consequentemente, dos seus processos de tratamento.

5.2.1 Técnicas de Seleção de Atributos

Como sabemos, existem várias técnicas de seleção de atributos, que podem ser divididas em métodos *Wrappers, Filters,* e *Embedded.* Os métodos *Wrappers* utilizam os algoritmos de modelação como uma caixa-negra, avaliando assim a importância ou utilidade relativa de um subconjunto de atributos com base no poder preditivo do algoritmo de modelação. Para este efeito são testados vários subconjuntos de atributos, em que consoante o desempenho preditivo dos

modelos melhorem ou não, os atributos são portanto selecionados como úteis ou então são descartados. Nos métodos *Wrappers*, torna-se importante definir três processos: qual o método de procura de atributos a utilizar; qual o método de avaliação do desempenho preditivo dos modelos a adotar; e qual o algoritmo de previsão a usar na modelação. Os dois últimos processos são escolhidos em função do problema de DM a modelar, uma vez que os métodos *Wrappers* têm capacidade de lidar com qualquer tipo de algoritmos preditivos e que podem também ser aplicados diferentes métodos de avaliação de modelos. Quanto ao processo de procura dos métodos *Wrappers*, este pode ser crescente (*Forward*), no qual se começa com um conjunto de atributos vazio e vai-se adicionando apenas os atributos que melhorem o desempenho do modelo. Uma outra abordagem passa pela procura decrescente (*Backward*), que é exatamente o oposto do *Forward*, em que se inicia com o conjunto de dados completo e vai-se removendo os atributos que prejudiquem o modelo. Existem ainda variantes de força bruta, que testam todas as combinações de atributos possíveis e, ainda, variantes baseadas em algoritmos genéticos.

Por sua vez, os métodos *Filters* selecionam um subconjunto de atributos como se fosse um pré-processamento independente dos algoritmos de modelação. Na generalidade destes métodos, são atribuídos pesos que indicam a importância dos atributos, e a seleção de atributos é realizada com base na variável de previsão. Quanto aos métodos *Embedded*, estes realizam a seleção de atributos dentro de um algoritmo específico (i.e. seleção de atributos interno das árvores de decisão). Adicionalmente, utilizam-se também métodos de redução de dimensionalidade (i.e. *Principal Component Analysis* - PCA). De grosso modo estes métodos envolvem uma transformação dos dados, sendo ainda caracterizados pela independência do algoritmo de previsão e, ao contrário dos métodos já mencionados, são não supervisionados (Guyon & Elisseeff, 2003).

5.2.2 Escolha do Método de Seleção de Atributos

Alguns dos métodos já apresentados foram testados com recurso à ferramenta *RapidMiner*. Começando pelos métodos de redução de dimensionalidade, nesta linha, foram testados alguns métodos PCA. Dos métodos *Filters*, foram testadas as pesagens de atributos (*weight by* - w) por correlação, *Relief* e SVM. Destas técnicas, testadas com diferentes capturas de variância ou limites de aceitação (i.e. top 20%), apresentam-se na tabela 5.1 as melhores configurações alcançadas

para cada modelo (SVR e SMO). Por fim, foram testados ainda os métodos *Wrappers*, nomeadamente a variante *Forward Selection* e o método *Wrapper* baseado em algoritmos genéticos (*Genetic Evolutionary*). A tabela 5.1 apresenta os resultados obtidos segundo as métricas de desempenho RMSE e R. Note-se que os valores dos erros são normalizados e lembra-se que estes testes correspondem à modelação de CBO na abordagem ETAR_2L.

Tabela 5.1 - Comparação dos métodos de seleção de atributos testados (valores de RMSE normalizados)

Método de Seleção de Atributos	**SVR**		**SMO**	
	RMSE	*R*	*RMSE*	*R*
Sem seleção de atributos	0.239	0.71	0.235	0.73
PCA	0.237	0.72	0.211	0.73
w-Correlação	0.225	0.76	0.186	0.81
w-Relief	0.225	0.77	0.189	0.80
w-SVM	0.220	0.79	0.168	0.84
Forward Selection	0.213	0.81	0.153	0.86
Genetic Evolutionary	**0.206**	**0.84**	**0.131**	**0.88**

Com base nos resultados apresentados na tabela 5.1, podemos verificar que o método PCA não revelou grandes melhorias em relação à não seleção de atributos. Contudo, a generalidade dos restantes métodos melhoraram o desempenho dos modelos. Os métodos *Filters* (*w-SVM, w-Relief e w-Correlação*) demonstraram resultados semelhantes entre si, com uma ligeira vantagem do método *w-SVM*. Quanto aos métodos de seleção de atributos *Wrappers*, estes foram, de forma evidente, superiores aos restantes. De acrescentar que no trabalho de Kohavi & John (1997), vários métodos de seleção de atributos são comparados com a abordagem *Wrapper*. Neste estudo, é demonstrada a superioridade destes métodos, nomeadamente em comparação com métodos *Filters e Embedded.* Outra nota sobre este estudo são as desvantagens destes métodos em termos de esforço computacional, porém como temos um conjunto de dados pequenos esse problema não é crucial. Quanto à questão do sobre ajustamento de dados, é referido que esse é um problema de menor importância nestes métodos. No entanto, ele ocorre especialmente em conjuntos de dados

pequenos, mas esse é um problema inerente aos conjuntos de dados pequenos de uma forma geral.

Sabendo que este é o caso do nosso conjunto de dados, cujos problemas de sobre ajustamento foram levantados na secção anterior, a solução passou por aplicar o método de avaliação exaustivo 10R-10CV. Embora seja referido que se deve retirar uma percentagem dos dados, ficando com um conjunto de dados independente do processo de seleção de atributos, que inclui a avaliação de modelos (treino+teste), no nosso caso isso pode implicar que variáveis importantes sejam descartadas. Principalmente, devido a exemplos com determinados tipos de tratamento ficarem de fora da modelação e, nesse caso, também fora da seleção de atributos. Só com um conjunto de dados maior se evitaria este problema, porém assume-se que uma avaliação de modelos exaustiva, cujo o resultado da média das 100 avaliações é que permitirá ao método *Wrapper* selecionar os atributos mais importantes, deverá atenuar bastante o efeito do sobre ajustamento. De sublinhar, contudo, que esta avaliação exaustiva só é exequível devido ao baixo número de exemplos, pois, como se pode imaginar, o elevado número de avaliações aliada à complexidade do próprio método *Wrapper* implica um processo de execução computacionalmente pesado.

Dentre os métodos *Wrappers* testados, a seleção de atributos *Optimize Selection (Genetic Evolutionary)* do *RapidMiner* teve resultados superiores à variante *Forward Selection* (Tabela 5.1). Deste modo, foi adoptado o método *Wrapper Genetic Evolutionary*, cujo método de procura dos atributos mais importantes é baseado em algoritmos genéticos (Rapid-i, 2010). Também no trabalho de Huang & Wang (2006), pode ser encontrado um estudo sobre a aplicação de técnicas de seleção de atributos baseada em algoritmos genéticos, isto em tarefas de modelação com algoritmos SVM.

5.3 Algoritmos de Previsão

Como já foi referido ao logo deste documento serão utilizadas técnicas de modelação preditivas, em particular algoritmos baseados em SVM. Por conseguinte, investigaram-se neste estudo duas implementações distintas de SVM. A primeira delas consistiu na implementação baseada na biblioteca LibSVM (Chang & Lin, 2011), mais concretamente do algoritmo SVM para regressão

(SVR) que é disponibilizado no *RapidMiner.* Adicionalmente, foi explorado o algoritmo SMOReg, doravante referido como SMO, com recurso à extensão do WEKA integrada no *RapidMiner*, cuja implementação é baseada nos trabalhos de Shevade et al. (2000) e Smola & Schölkopf (1998), conforme citado na documentação em (WEKA, n.d.).

O algoritmo SVR (LibSVM) inclui ainda duas classes de algoritmos: o ε-SVR e o ν-SVR. Enquanto que o algoritmo ε-SVR foi apresentado no capítulo 3 e consiste na versão clássica das SVM para regressão, o ν-SVR é uma nova classe de SVR apresentada por Schölkopf et al. (2000). Na verdade, esta nova classe é similar ao ε-SVR, porém é introduzido um novo parâmetro ν que substitui ε. É referido pelos autores que a extensão ν-SVR do algoritmo clássico, em particular o novo parâmetro ν, permite controlar o número de vetores de suporte e os erros. Assim, esta extensão traz algumas vantagens do ponto de vista prático e teórico. Após experimentar-se estas duas classes de algoritmos, chegou-se à conlcusão que os resultados não divergem muito entre si. Porém, como ν-SVR apresentou resultados ligeiramente superiores, especialmente nos modelos com a configuração de *kernel* adotada (RBF), este foi portanto o escolhido e utilizado nas várias tarefas de modelação.

5.3.1 Escolha do Kernel

As SVM possuem uma caracteristica referida por Vapnik como a "universalidade da máquina", devido ao facto de diferentes funções de *kernel* poderem ser utilizadas na modelação com SVM. Das duas variantes em estudo (SVR e SMO), ambas permitem a utilização de vários *kernels*, como por exemplo o RBF, o polinomial, o sigmoid e o pré-computado. Dentro deste grupo destacaram-se os *kernels* RBF e polinomial, pois demonstraram resultados acima dos restantes. Comparando ambos os *kernels*, podemos dizer que não existiram grandes diferenças que permitam indicar que um *kernel* é superior ao outro. Por exemplo, em SMO o *kernel* polinomial foi ligeiramente superior, enquanto que em SVR foi o *kernel* RBF que teve melhores resultados. Curiosamente, notou-se que o tempo de processamento no caso de SMO diminui com a aplicação do *kernel* RBF.

Vários autores na literatura sobre as SVM, sugerem o RBF como uma boa primeira escolha de *kernel*. Algumas das razões para esta escolha são, em primeiro lugar, o facto de tratar-se de um

kernel que permite mapeamentos não lineares, o que é interessante para capturar a complexidade dos dados da ETAR. Outra razão é a maior facilidade de otimização de hiperparâmetros, pois o *kernel* RBF tem menos hiperparâmetros de configuração do que e.g. o *kernel* polinomial (Hsu et al., 2010). Com base nestas recomendações e nos resultados empíricos obtidos, optou-se pela utilização do *kernel* RBF em ambos os algoritmos de modelação.

5.3.2 Otimização de Parâmetros

Os algoritmos SVM em estudo contêm vários parâmetros de configuração, é importante calibrar esses parâmetros por forma a retirar todo o potencial da capacidade preditiva destes algoritmos. Entenda-se por capacidade preditiva não como a eficácia previsão das instâncias de treino, mas sim de previsão das futuras instâncias desconhecidas a prever. Esta questão também pressupõe evitar o problema de sobre ajustamento dos dados de treino. Em Flexer (1996) é recomendado como uma boa prática retirar 20% das instâncias para um conjunto de dados independente do processo de modelação, que ocorre durante a otimização dos parâmetros. Também neste caso persiste o mesmo dilema referido na seleção de atributos, devido ao facto do conjunto de dados do problema ser pequeno. Assim, mantem-se a posição de colmatar este problema através da avaliação dos modelos segundo o método 10R-10CV.

Relativamente à abordagem ou estratégia de otimização dos parâmetros, utlizou-se o método de procura em grelha (*Grid Search*). Esta técnica consiste em testar a performance do modelos perante todas as combinações de parâmetros, cujos intervalos de valores e números de parâmetros a explorar é previamente definido pelo utilizador. Este processo pode facilmente tornar-se num fardo computacional. Por exemplo, procurar 10 valores diferentes de 3 parâmetros, implica executar o processo de modelação num total de, $10 \times 10 \times 10 = 1000$ vezes. Deste modo, foi executado um processo de procura em grelha fasedo, em que em primeiro lugar são procurados os melhores valores para os parâmetros segundo intervalos mais esparsos, refinando-se depois a procura em intervalos mais curtos. Este processo de otimização de paramêtros em grelha é recomendo por Hsu et al. (2010), que refere ainda que se deve avaliar os modelos através de métodos de validação cruzada. Convém ainda acrescentar que os valores de alguns parâmetros podem ser selecionados de uma forma prática. Algumas dessas heurísticas ou métodos analítos,

nomedamente para os parâmetros C e ϵ, são apresentados em (Cherkassky & Ma, 2004), onde estes são calculados com base no conjunto de dados e na estimação de ruído desses dados.

5.4 Processo de Desenvolvimento

A sequência de execução dos processos de modelação seguiu em cada tarefa a seguinte ordem: filtraram-se os atributos mais interessantes ao problema de modelação (i.e. atributos de PA6 podem ser removidos); depois executou-se a seleção de atributos através do método *Wrapper Genetic Evolutionary*; de seguida, ao conjunto de dados resultantes foi aplicada a otimização de parâmetros, através do método de procura em grelha; e, finalmente, foram calibrados os modelos com os valores ótimos dos respectivos parâmetros, voltando-se a modelar e a avaliar os modelos segundo o método 10R-10CV. Ainda uma pequena nota para as fases de seleção de atributos e de otimização de parâmetros, que levantam a questão de qual deve ser executado primeiro. Se por um lado sem o conjunto de atributos mais importantes não se estão a encontrar os melhores parâmetros, que permitam ao modelo generalizar melhor. Por outro lado, no caso contrário, não se está a selecionar os atributos mais importantes. Assim, este processo de seleção de atributos conjugado com a otimização de parâmetros poderá ser reiterado e, em função dos resultados obtidos, este poderá ser considerado ou não. Desde já, pode-se adiantar que este processo não revelou melhorias em todas as tarefas de modelação. Inclusive, algumas delas, apenas com uma passagem atingiram o ponto ótimo, e, também, em nenhuma das tarefas foi produtivo executar mais de 2 iterações.

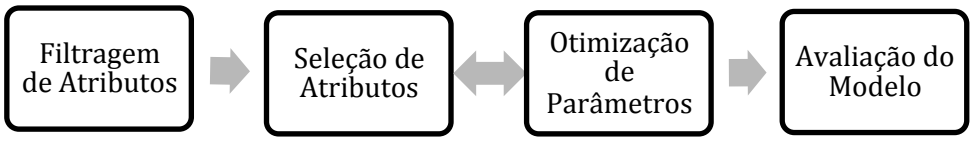


Figura 5.1 - Processo de desenvolvimento da fase de modelação

5.5 Sumário

A fase modelação envolve um conjunto de técnicas que devem ser adotadas com base no problema a modelar, portanto com o propósito de obter-se assim modelos eficientes e ajustados à realidade do problema. Existem diferentes métodos de avaliação de modelos, porém, em função do conjunto de dados do problema, é importante selecionar o método mais adequado. Após um estudo sobre qual o método de avaliação mais adequado para o nosso problema, selecionou-se um método de validação cruzada, mais especificamente 10 repetições de 10 validações cruzadas (10R-10CV), que é um método apropriado para conjuntos de dados pequenos e que procura reduzir o sobre ajustamento dos dados.

O conjunto de dados de trabalho apresenta um problema de alta dimensionalidade, surge portanto a necessidade de explorar técnicas de seleção dos atributos mais importantes. A aplicação de técnicas de seleção de atributos, não só permite obter melhores resultados de modelação, como possibilita também extrair informação acerca dos parâmetros mais relevantes ao problema. Por conseguinte, foram testadas algumas técnicas de seleção de atributos. Dos métodos analisados, as técnicas do tipo *Wrapper* mostraram-se superiores, daí a escolha do método *Wrapper Genetic Evolutionary*.

Os algoritmos de modelação utilizados neste problema consistem em implementações SVM para regressão, nomeadamente o LibSVM (SVR) e o SMOReg do WEKA (SMO). A calibração destes algoritmos envolveu assim: a seleção do tipo de *kernel* a utilizar, na qual se optou pelo *kernel* RBF; e ainda a otimização dos híper-parâmetros dos algoritmos SVM, em que foi adoptada a técnica de procura em grelha. O processo de modelação utilizado neste problema segue assim o seguinte fluxo: filtragem de atributos; seleção de atributos mais importantes; otimização de parâmetros; e, por fim, a avaliação do modelo resultante.

Capítulo 6

Experiências e Resultados

Mediante o processo de desenvolvimento de modelação descrito anteriormente no capítulo 5, apresentam-se neste capítulo as experiências ou tarefas de DM investigadas. As várias tarefas de DM exploradas, estão organizadas segundo as abordagens ETAR_2L e ETAR_1L. Em cada um das abordagens experimentaram-se duas tarefas de previsão distintas, a primeira na previsão do parâmetro de qualidade CBO e a segunda na previsão de SST. Note-se ainda que na previsão de cada parâmetro, são analisados os desempenhos de duas variantes de algoritmos SVM, o SVR e o SMO.

6.1 Previsão da ETAR Duas Linhas de Tratamento (ETAR_2L)

Continuando a descrever o processo de desenvolvimento da abordagem ETAR_2L, a partir da qual se iniciou este estudo, apresentam-se assim os resultados das tarefas de previsão dos parâmetros CBO e SST para os dois algoritmos SVM e SMO, respetivamente. Todavia, antes de apresentarmos estas tarefas, será realizada uma análise à potencial utilização dos parâmetros de PA6, em particular na previsão de CBO.

6.1.1 Análise de Previsão com Atributos de PA6

Como já foi referido neste documento, as medições do parâmetro PA6_CBO, por vezes referido como CBO_5, demoram pelo menos 5 dias a ser obtidas. Perante este facto, e assumindo que as medições dos restantes parâmetros de PA6 são obtidas previamente, analisou-se a contribuição dos parâmetros de PA6 para esta tarefa de previsão. As diferenças de performance dos modelos SVR e SMO na previsão de CBO são apresentadas no gráfico da figura 6.1, no qual é analisada a inclusão ou não inclusão dos atributos de PA6 nos dados de input.

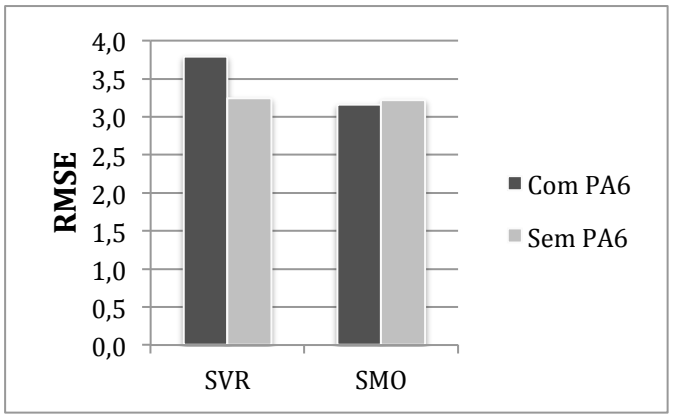


Figura 6.1 - Comparação dos modelos de previsão de CBO com/sem atributos de PA6

Olhando para o gráfico apresentado na figura 6.1, podemos reparar que os valor dos erros (RMSE) para o modelo SVR baixam quando são retirados os atributos de PA6. Este facto é curioso pois existem alguns atributos de PA6 (i.e. CQO) que se correlacionam com CBO, porém, esses coeficientes de correlação não são muito fortes (ver secção 4.2.3). No caso de SMO, os erros subiram quando se retiraram os atributos de PA6, porém, essa quebra de desempenho é muito ligeira, sendo até insignificante do ponto de vista estatístico.

Tabela 6.1 - Desempenho dos modelos de previsão para CBO com/sem atributos de PA6

Tarefa	SVR	SMO
com PA6	3.795 ± 0.59*	$\mathbf{3.161 \pm 0.47}$
sem PA6	$\mathbf{3.241 \pm 0.42}^{\blacklozenge}$	3.221 ± 0.42

* Diferença estatisticamente significativa sobre comparação com o par SMO na modelação com atributos PA6.

◆ Diferença estatisticamente significativa sobre comparação com o mesmo algoritmo na modelação com atributos PA6.

Na tabela 6.1 pode-se visualizar os valores dos erros. Note-se que é realizada a comparação de pares segundo o teste estatístico *Student's T-test*, com intervalo de confiança de 95% das medidas de desempenho RMSE. Neste documento, ao longo da apresentação dos resultados são realizados testes estatísticos para comparação de modelos, visto que esta é uma boa prática recomendada em (Flexer, 1996).

Na tarefa de modelação com atributos de PA6, os modelos SVR e SMO tiveram resultados diferentes (estatisticamente significativos), com um desempenho superior por parte de SMO. De notar que no método SVR, os resultados da tarefa de modelação sem atributos de PA6 melhoraram de forma significativa, isto relativamente ao mesmo algoritmo na tarefa de previsão com atributos de PA6. Em suma, como o algoritmo SMO manteve aproximadamente a mesma performance preditiva nas duas tarefas, e como SVR progrediu com a não inclusão dos atributos de PA6, conclui-se que os atributos de PA6 podem ser filtrados na modelação. Deste modo, como estes atributos não trazem melhorias significativas de performance à modelação de CBO, tomou-se a decisão de não os considerar nas tarefas de modelação em estudo. Para reforçar esta posição, de realçar que desta forma serão efectuadas análises mais equitativas entre as tarefas de previsão de CBO e SST. Uma vez que da parte de SST não faria sentido utilizar atributos de PA6, pois, perante dados novos, não se podia incluir nos atributos de *input* os parâmetros cujas recolhas são efectuadas no mesmo momento que o atributo de *output*. No entanto, podemos acrescentar que devido à forte correlação apresentada entre PA6_SST e PA6_SSV, conjugada ainda com correlações médias entre outros atributos (i.e. PA6_CQO), é calculável que os atributos de PA6 contribuam imenso para a previsão de SST. Aliás, nos modelos de teste, ainda na fase de preparação de dados, notou-se uma capacidade preditiva de SST altamente eficaz, que quebrou naturalmente na fase de modelação devido à ausência dos parâmetros de PA6. Voltaremos a estas

questões sobre escolha de parâmetros de PA6 para *input* da modelação na parte final desta dissertação, em particular nas conclusões do estudo realizado.

Para concluir, retêm-se essencialmente desta análise que os atributos de PA6 não se mostraram interessantes na previsão de CBO – este seria o parâmetro de qualidade com alguma potencialidade de ser previsto segundo estes atributos. Deste modo, na filtragem de atributos (Figura 6.1), são removidos todos os atributos de PA6, com exceção como óbvio dos próprios atributos de previsão. A filtragem de atributos deve ser portanto configurada em função de cada tarefa de modelação a desempenhar.

6.1.2 Previsão de CBO

Após a fase de modelação descrita na capítulo anterior, foi realizada uma avaliação sobre o desempenho dos modelos desenvolvidos. Começando pelos modelos relativos à previsão de CBO, segundo a abordagem ETAR_2L, na tabela 6.2 mostram-se os resultados obtidos. A eficácia preditiva dos modelos foi avaliada através das três métricas de desempenho adotadas neste estudo (RMSE, RRSE e R). Na tabela 6.2, podemos reparar que os resultados dos dois algoritmos de modelação, SVR e SMO, são bastante idênticos, não havendo portanto diferenças significativas. No entanto, destaca-se também os bons resultados de previsão obtidos, com a correlação muito próxima do máximo e do erro relativo que é cerca de 60% inferior ao erro cometido ao prever-se sempre a média do atributo PA6_CBO. Olhando para o RMSE, verifica-se que o erro médio cometido por ambos os modelos de previsão é de cerca de 3.2*(mg/l)*. Sabendo que os valores médios da variável PA6_CBO são de 19.682 +/- 9.369 *(mg/l)*, pode-se perceber a boa eficácia destes modelos.

Tabela 6.2 - Resultados dos modelos SVR e SMO, na previsão de CBO segundo a abordagem ETAR_2L

Tarefa	SVR			SMO		
	RMSE	*RRSE*	*R*	*RMSE*	*RRSE*	*R*
CBO	3.241±0.42	39.5% ±9.3%	0.94±0.03	**3.221±0.42**	**39.2%±9.1%**	**0.93±0.04**

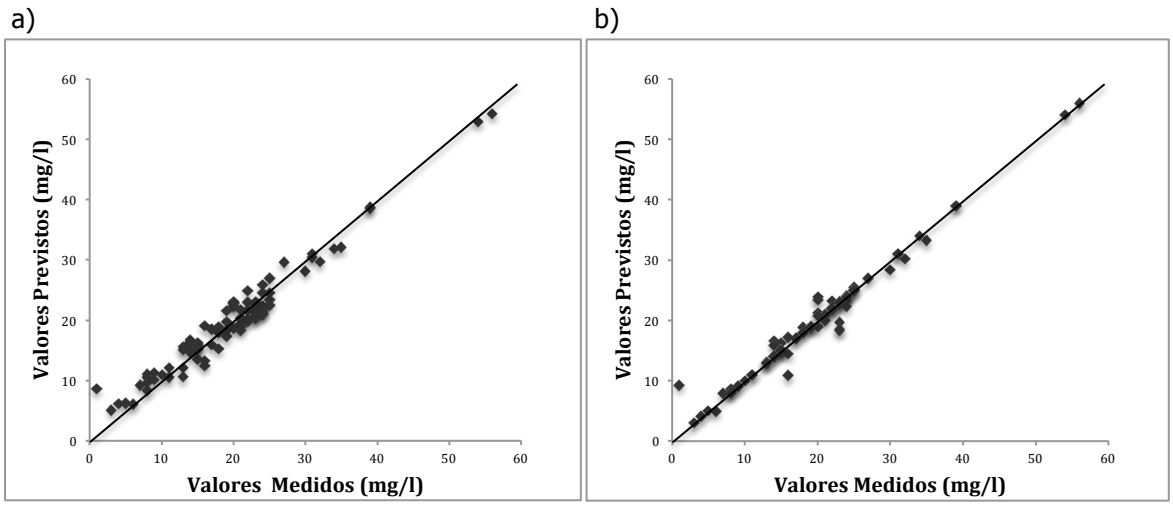


Figura 6.2 – Gráficos de dispersão dos valores de CBO medidos (eixo-x) e os valores previstos (eixo-y). a) modelação SVR-CBO; b) modelação SMO-CBO (ETAR_2L)

Pegando nos modelos desenvolvidos procedeu-se ao teste da sua capacidade preditiva no conjunto de dados de trabalho, uma que vez não tínhamos dados de teste. Na figura 6.2 são apresentados dois gráficos de dispersão, nos quais os pontos marcam a relação entre as concentrações medidas (valores reais) e os valores previstos pelos modelos SVR e SMO. A linha diagonal representa uma previsão perfeita, ou seja, valores previstos iguais aos valores reais. Nota-se aqui que a generalidade das instâncias aproximam-se bastante da linha em ambos os modelos. Porém, em SMO verifica-se uma distribuição dos dados mais fina em redor da linha diagonal. Mesmo sabendo que os coeficientes de correlação na avaliação dos modelos são similares, percebe-se com esta menor espessura que existem previsões mais próximas do valor ideal em SMO. Note-se ainda que o exemplo medido com o valor mais baixo de CBO, é previsto bem acima do valor real - curiosamente este registo corresponde à entrada na ETAR de um afluente bruto anómalo. Relativamente aos dois *outliers* cujos valores são bastante mais altos que as restantes instâncias, os modelos conseguem prevê-los eficazmente, mas isso pode significar um sobre ajustamento.

Contudo, ambos os modelos preveem bastante bem o conjunto de dados. Na figura 6.3 têm-se essa perspectiva, em que se vê que ambos os modelos seguem a linha correspondente aos valores reais. Convêm no entanto sublinhar que estas previsões bastante precisas devem-se, obviamente, ao facto dos modelos estarem a prever os dados segundo os quais realizaram a aprendizagem.

Perante novos dados seguramente que as aproximações não seriam tão precisas. Aliás, a performance obtida na avaliação 10R-10CV indica precisamente isso. Os valores dos erros e da correlação da tabela 6.2 são bem mais pessimistas, mas também mais realistas face à previsão de novos exemplos desconhecidos. Daí que esta análise teria mais valor se aplicada a um conjunto de dados de teste. Todavia será interessante comparar-se a aprendizagem dos modelos nas diferentes tarefas.

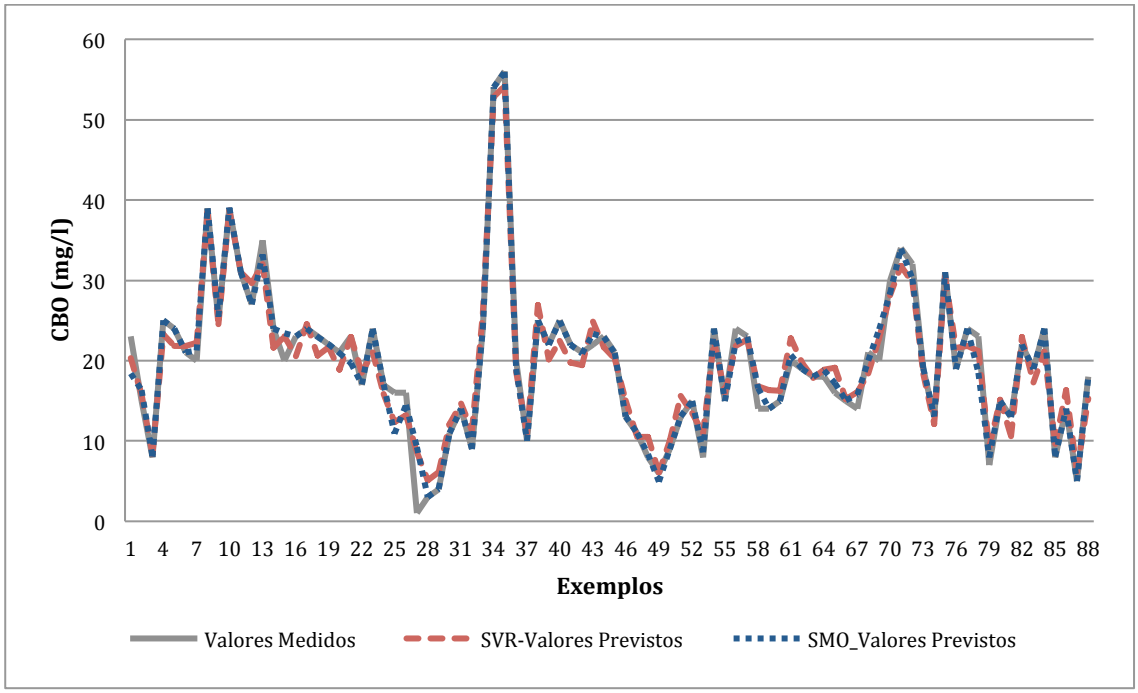


Figura 6.3 - Gráfico comparativo dos valores de CBO medidos, com os valores previstos por SVR e SMO (ETAR_2L)

6.1.3 Previsão de SST

A tarefa de previsão de SST desenrolou-se de modo análogo à tarefa de CBO. Na fase de preparação de dados foi ajustado o parâmetro *k* do algoritmo de imputação de nulos, e não se removeram os pares redundantes SSV, visto que estes relacionam-se com a variável de previsão. Adicionalmente, como os dois exemplos com nulos em PA6_CBO não são descartados nas tarefas

de previsão de SST, fica-se com 90 instâncias. Quanto ao processo de modelação, foram removidos os atributos de PA6 e na seleção de atributos utilizaram-se os mesmos métodos utilizados na previsão de CBO. Visto que a seleção de atributos através de métodos *Wrappers* também se mostrou mais eficaz na tarefa de previsão de SST, daí a sua escolha.

Na tabela 6.3 demonstram-se os resultados obtidos na modelação preditiva de SST. Olhando para os resultados, sobressai logo o facto da capacidade preditiva dos modelos para o parâmetro SST ser inferior à obtida com CBO. Isto verifica-se comparando os valores de correlação com a tabela 6.2, pois a correlação é independente da escala. O erro relativo piorou, sendo ainda assim melhor que um modelo preditivo simples em cerca de 15% e 25% de redução dos erros para SVR e SMO, respetivamente. Quanto ao RMSE estes valores são mais altos que em CBO, porém, o parâmetro PA6_SST tem valores médios na ordem de 23.178 +/- 12.182 *(mg/l)*, ou seja, valor médio superior a CBO e também uma maior variabilidade dos dados. Comparando agora os dois algoritmos de modelação, nota-se claramente um melhor desempenho de SMO relativamente a SVR, todavia esta diferença não é significativa segundo teste estatísticos com intervalo de confiança de 95%.

Tabela 6.3 - Resultados dos modelos SVR e SMO, na previsão de SST segundo a abordagem ETAR_2L

Tarefa	**SVR**			**SMO**		
	RMSE	*RRSE*	*R*	*RMSE*	*RRSE*	*R*
SST	8.639±1.70	84.2% ±21.1%	0.66±0.16	**7.791±1.71**	**75.2%±16.2%**	**0.75±0.07**

Nos gráficos de dispersão da figura 6.4, o modelo SMO demonstra uma melhor aproximação à linha diagonal que SVR. Em SVR é curioso que as instâncias mantêm uma aproximação uniforme ao longo da linha, tal como acontece em CBO (figura 6.2a), só que neste caso essa disposição dos dados é mais larga. Em SMO os dados são mais dispersos, dando a entender que existem duas posições, ou eles são previstos idealmente ou então têm um erro considerável, mas a realidade é que SMO prevê eficazmente muitas mais instâncias que aquelas que falha. Quanto aos *outliers* de SST, comparativamente com CBO, não são previstos de forma tão eficaz pelos dois modelos, o que poderá ser até um sinal positivo. O valor mínimo de SST, à imagem do que aconteceu com CBO, também é mal previsto, curiosamente corresponde ao mesmo mês de tratamento embora em dias

distintos. A acrescentar que durante o mês em questão, foi observado pelos analistas alguns problemas técnicos na ETAR, tal como o seletor e medição de caudal desativados. Isto poderá justificar as dificuldades dos modelos preverem estes comportamentos anómalos da ETAR.

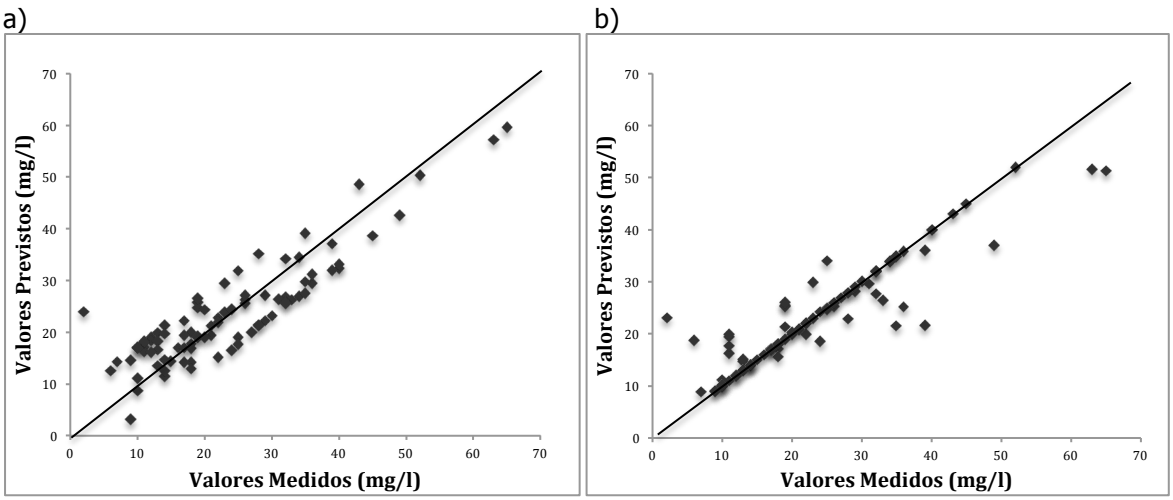


Figura 6.4 - Gráficos de dispersão dos valores de SST medidos (eixo-x) e os valores previstos (eixo-y). a) modelação SVR-SST; b) modelação SMO-SST (ETAR_2L)

Na figura 6.5 pode-se observar que os modelos não acompanham a linha dos valores reais de forma tão ajustada como em CBO (Figura 6.3), principalmente quando esta apresenta valores extremos. Visualiza-se ainda neste gráfico a melhor proximidade dos valores previstos por SMO aos valores reais, tendo assim um melhor ajuste que o modelo SVR. Vendo que a previsão dos dados de aprendizagem é pior que no caso de CBO, é natural que as avaliações do desempenho dos modelos na previsão de SST também sejam piores. Em suma, retêm-se da comparação destes resultados com os resultados dos modelos de previsão de CBO que a tarefa de previsão de SST é mais complexa e difícil, tendo portanto pior desempenho. Sabendo que SST revelou fortes relacionamentos com outros atributos, em particular do mesmo PA, os relacionamentos entre PA distintos mostraram-se assim mais fracos ou de maior complexidade.

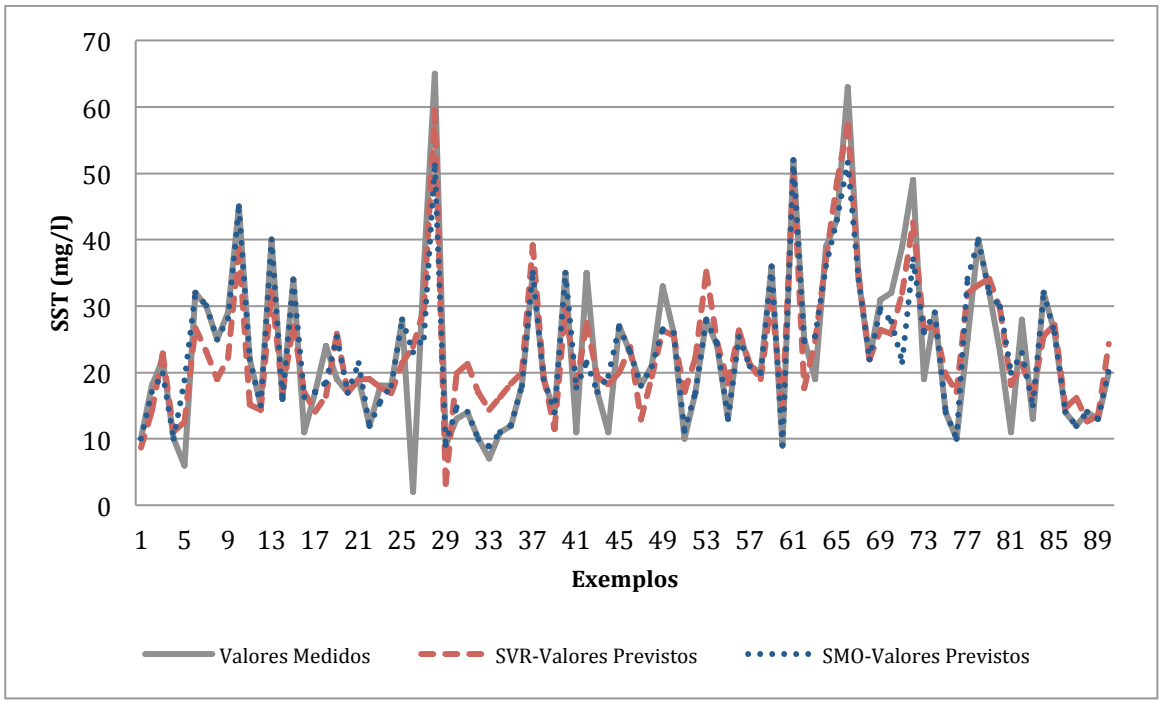


Figura 6.5 - Gráfico comparativo dos valores de SST medidos, com os valores previstos por SVR e SMO (ETAR_2L)

6.2 Previsão da ETAR Uma Linha de Tratamento (ETAR_1L)

Após a descrição de todo o processo que engloba a previsão dos parâmetros de qualidade segundo a abordagem que considera duas linhas de tratamento, será agora analisada a abordagem simplificada que contempla somente uma linha de tratamento. Pretende-se portanto analisar a redução de dimensionalidade e simplificação do problema da abordagem ETAR_1L, com a abordagem de maior dimensionalidade e mais complexa de ETAR_2L. Relembra-se que esta análise comparativa das duas abordagens é também objetivo deste trabalho.

6.2.1 Processo de Desenvolvimento

Antes de passarmos à apresentação das experiências e resultados obtidos nesta abordagem, convém referir primeiro o processo de desenvolvimento, nomeadamente, das fases preparação de dados e de modelação.

Na preparação do conjunto de dados ETAR_1L, foi realizada em primeiro lugar uma transformação do conjunto de dados de trabalho inicial. Esta transformação consistiu em calcular a média de todos os parâmetros relativos aos PA com duas linhas de tratamento, passando o valor resultante para um "novo" parâmetro. Pegando no exemplo de um parâmetro (i.e. PA2_SST), é realizada a média de L1_PA2_SST e L2_PA2_SST, onde valor resultante é atribuído ao parâmetro PA2_SST do conjunto de dados ETAR_1L. Realizando este processo para todos os parâmetros dos PA 2, 3, 4 e 5, chegou-se ao conjunto de dados ETAR_1L, cujas estatísticas podem ser visualizadas no anexo A.2. O importante factor a reter desta transformação e simplificação do problema, é a redução da dimensionalidade, passando de 82 para 56 atributos. A preparação dos dados relativos à abordagem ETAR_1L, foi em todo semelhante à adoptada para a abordagem ETAR_2L. Na conversão de atributos nominais para numéricos o processo é o mesmo, visto que este atributos são externos aos PA paralelos da ETAR. Na normalização de dados, também o método $MinMax_{-1:1}$ revelou superioridade em ambos os modelos de teste da fase de preparação de dados. A imputação de dados foi executada através do método k-NNI, embora as diferenças entre os métodos de imputação não tenham sido muito acentuadas, em particular comparando com MMI. Quanto ao parâmetro *k,* este foi calibrado em função dos melhores resultados obtidos pelos modelos de teste. Manteve-se ainda a posição de remover os atributos redundantes PA2_SSV e $PA2\_Redox_{ZAx}$ na previsão de CBO, e de os incluir no conjunto de input para a modelação de SST.

O processo de desenvolvimento de modelação executado nesta abordagem, consiste no apresentado no capítulo 5 (figura 5.1). Recapitulando este processo: foram filtrados os atributos de PA6, com exceção dos atributos de previsão; selecionados os atributos mais relevantes para os modelos, também na abordagem ETAR_1L os métodos *Wrappers* foram superiores aos restantes; otimizados os parâmetros através de método de procura em grelha; e, por fim, os modelos foram avaliados. O método de avaliação adoptado foi o 10R-10CV, devido às mesmas razões apontadas no capítulo anterior.

6.2.2 Previsão de CBO

A previsão do parâmetro de qualidade CBO segundo o conjunto de dados ETAR_1L, mostrou-se eficaz nos dois modelos SVR e SMO. Olhando para a tabela 6.4, verifica-se que os dois modelos tiveram performances muito similares, onde os valores de correlação ficaram muito próximos de 0.9, os erros relativos superam em 50% os erros de um preditor simples e o RMSE apresenta valores de perto de $4_{(mg/l)}$. Comparando com os resultados das mesmas tarefas de modelação na abordagem ETAR_2L (Tabela 6.2), podemos constatar que os resultados pioraram de forma geral. Estes são em inclusive significativamente inferiores, por exemplo, os erros (*RMSE*) aumentaram passando de cerca de $3.2_{(mg/l)}$ para $4_{(mg/l)}$, mas também o erro relativo aumentou cerca de 10% e o coeficiente de correlação piorou para 0.88. Por conseguinte, verifica-se que a capacidade preditiva dos modelos ETAR_1L é inferior aos da abordagem ETAR_2L, porém destaca-se que o desempenho destes modelos na previsão de CBO é ainda assim bastante positivo.

Tabela 6.4 - Resultados dos modelos SVR e SMO, na previsão de CBO segundo a abordagem ETAR_1L

Tarefa	SVR			SMO		
	RMSE	*RRSE*	*R*	*RMSE*	*RRSE*	*R*
CBO	4.167 ± 1.08◆	$49.5\% \pm 14.5\%$	0.88 ± 0.09	**4.020 ± 0.93**◆	**$48.6\% \pm 15.1\%$**	**0.88 ± 0.09**

◆ Diferença estatisticamente significativa sobre comparação com a mesma tarefa de modelação na abordagem ETAR_2L.

Analisando agora os gráficos de dispersão da figura 6.6, pode-se reparar que os modelos SVR e SMO cometem mais erros na previsão das concentrações com valores mais baixos e mais altos. Por conseguinte, observa-se que os valores intermédios são em geral previstos eficazmente. Assim, não se notam diferenças assinaláveis entre os dois modelos. Contudo, quando comparados com os gráficos das mesmas tarefas para ETAR_2L (Figura 6.2), nota-se que existem bastantes exemplos mais distantes da linha diagonal, o que significa uma pior aprendizagem dos dados de treino. Olhando para os *outliers,* vemos que neste caso os modelos não os preveem eficazmente, o que pode indicar uma boa generalização. No entanto, existem alguns valores intermédios com erros de previsão consideráveis.

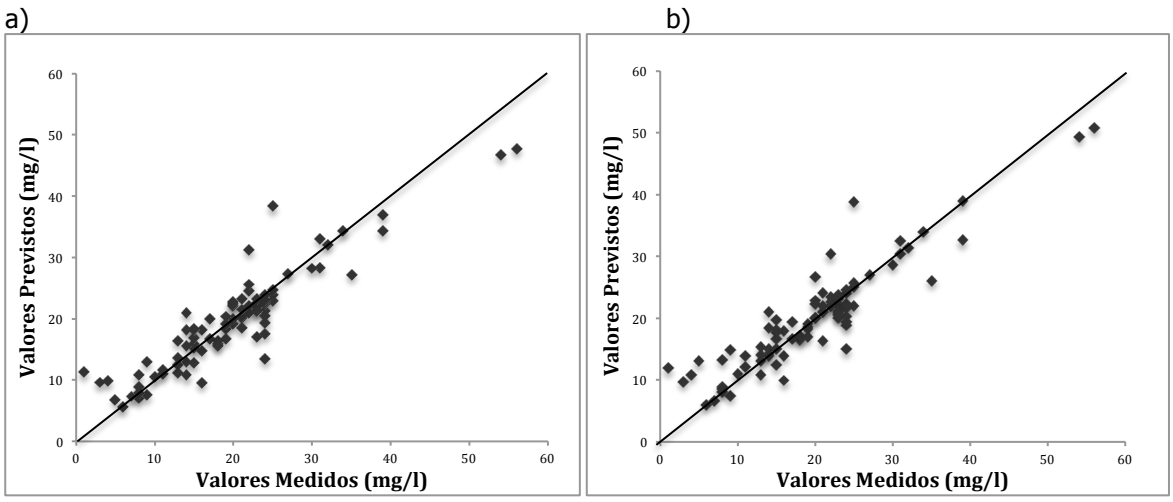


Figura 6.6 - Gráficos de dispersão dos valores de CBO medidos (eixo-x) e os valores previstos (eixo-y). a) modelação SVR-CBO; b) modelação SMO-CBO (ETAR_1L)

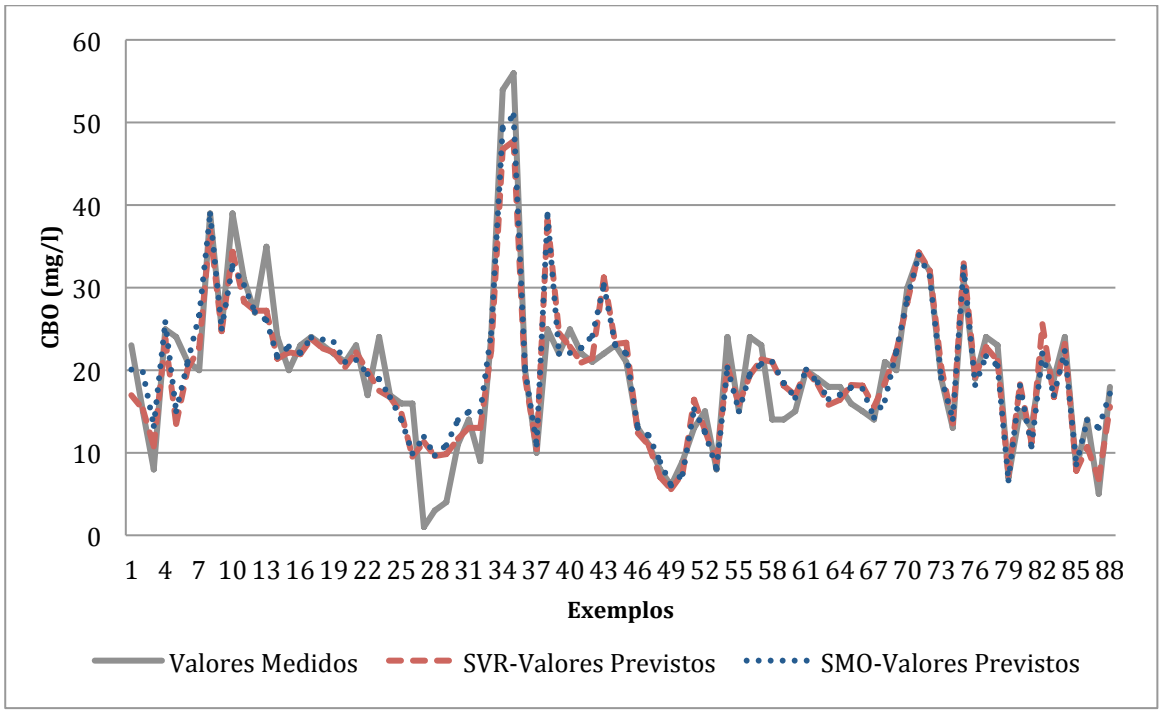


Figura 6.7 - Gráfico comparativo dos valores de CBO medidos, com os valores previstos por SVR e SMO (ETAR_1L)

Na figura 6.7 pode-se visualizar estes factos, pois os valores previstos pelos modelos não acompanham os extremos dos valores medidos, e ainda, nos exemplos 38-43, observam-se previsões bem acima das concentrações reais. Nos restantes exemplos, pode-se observar que ambos os modelos se ajustam relativamente bem à linha dos valores reais. Embora, quando comparado com a figura 6.3 da abordagem ETAR_2L, é perceptível que as linhas de SVR e SMO não se aproximam de forma tão precisa à linha dos valores medidos.

6.2.3 Previsão de SST

A tarefa de previsão de SST na abordagem ETAR_1L revelou o pior desempenho preditivo de todas as tarefas aqui analisadas. Na tabela 6.5 são apresentados os resultados desta tarefa de previsão. A métrica de desempenho RMSE tem valores na ordem dos $9.6_{(mg/l)}$ e $9.4_{(mg/l)}$, os erros relativos apenas superam um preditor simples em cerca de 10% e coeficiente de correlação baixou relativamente aos outros modelos para cerca de 0.56. Nesta modelação os resultados alcançados por SVR e SMO também são idênticos, tal como aconteceu na tarefa de modelação anterior. Comparando a previsão de SST com a previsão de CBO, verifica-se que os coeficientes de correlação em SST são bem menores. O que indica uma maior dificuldade de previsão de SST, tal como verificou-se na abordagem ETAR_2L.

Tabela 6.5 – Resultados dos modelos SVR e SMO, na previsão de SST segundo a abordagem ETAR_1L

Tarefa	SVR			SMO		
	RMSE	*RRSE*	*R*	*RMSE*	*RRSE*	*R*
SST	9.683±2.83	90.7% ±12.0%	0.56±0.15	**9.410±2.42**◆	**89.1%±13.3%**	**0.57±0.09**

◆ Diferença estatisticamente significativa sobre comparação com a mesma tarefa de modelação na abordagem ETAR_2L.

Para a mesma tarefa de previsão na abordagem ETAR_2L (Tabela 6.3), nota-se que os resultados de ETAR_1L são inferiores, porém, só no caso de SMO é que as diferenças são estatisticamente significativas. Em SVR o RMSE piorou passando de cerca de $8.6_{(mg/l)}$ para $9.6_{(mg/l)}$, o erro relativo passou de cerca 85% para 90% e a correlação desceu de 0.66 para 0.56. No caso de SMO essas

diferenças são ainda mais acentuadas, o RMSE aumentou de cerca $7.8_{(mg/l)}$ para $9.4_{(mg/l)}$, o RRSE aumentou cerca de 15% e a correlação desceu de 0.75 para 0.57.

Analisando os gráficos da figura 6.8, vê-se claramente que existem muitos exemplos com erros consideráveis de previsão. É curioso que o modelo SMO prevê alguns exemplos de forma perfeita (sobre a linha), enquanto que em SVR essas situações são mais escassas. Comparando estes gráficos de dispersão, com os das mesmas tarefas na abordagem ETAR_2L (Figura 6.4), é notória a diferença do número de exemplos distantes da linha diagonal e ainda da enorme distância de alguns exemplos à linha, o que indica erros de previsão muito elevados.

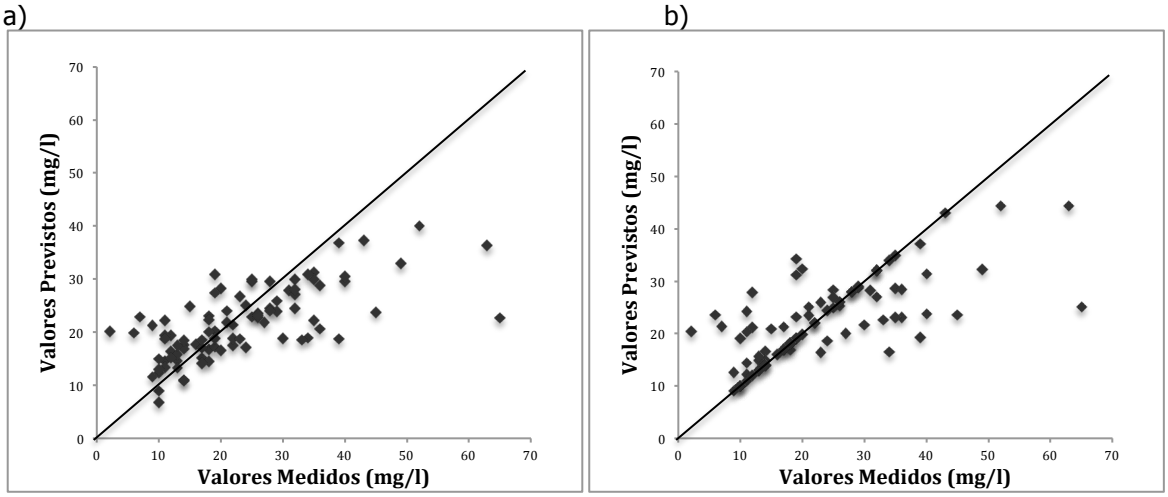


Figura 6.8 - Gráficos de dispersão dos valores de SST medidos (eixo-x) e os valores previstos (eixo-y). a) modelação SVR-SST; b) modelação SMO-SST (ETAR_1L)

Visualizando a figura 6.9, comprova-se que nestas tarefas de modelação acentuaram-se as diferenças entre os valores previstos e os valores medidos, principalmente nos exemplos cujos valores são mais distantes do valor médio. Comparando com a figura 6.5 da abordagem ETAR_2L, é bem perceptível a discrepância entre as linhas relativas às previsões dos dois modelos e a linha dos valores medidos. Assim, nesta abordagem ETAR_1L, verifica-se uma grande dificuldade dos modelos ajustarem-se à linha dos valores reais, o que mostra a dificuldade dos modelos derivados em prever os próprios dados de aprendizagem. Isto pode indicar que a previsão de dados

desconhecidos seja ainda menos precisa, como é de resto estimado pela avaliação dos modelos e cujos resultados foram descritos na tabela 6.2.

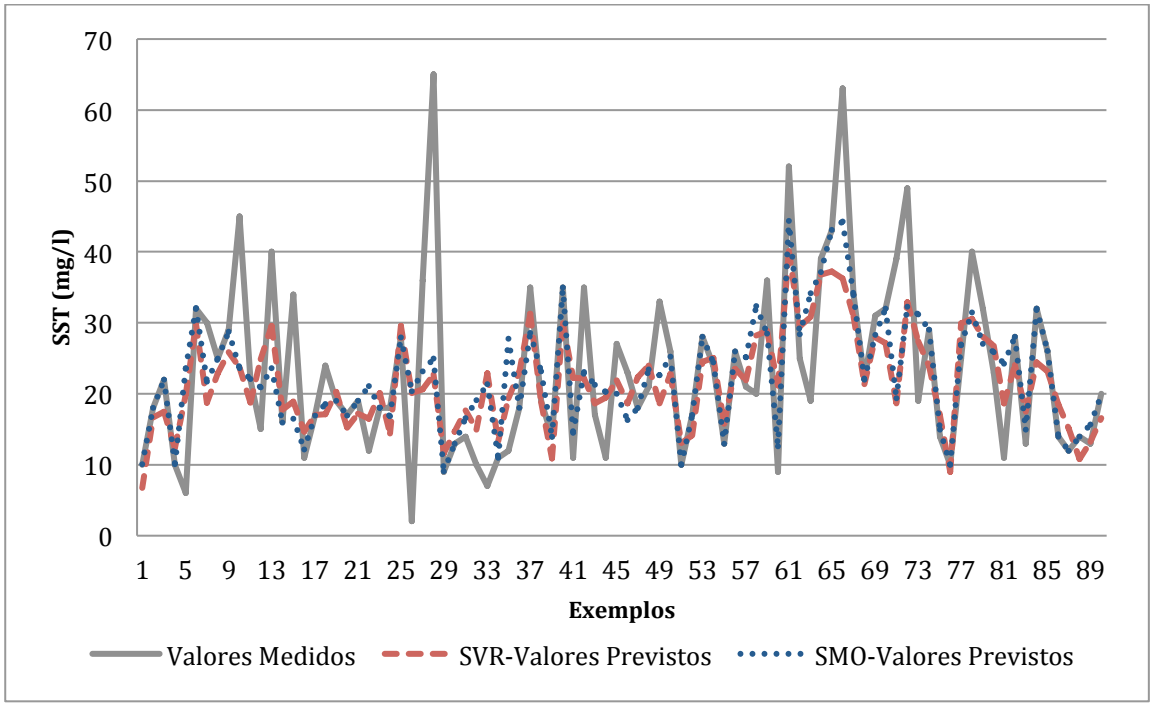


Figura 6.9 - Gráfico comparativo dos valores de SST medidos, com os valores previstos por SVR e SMO (ETAR_1L)

6.3 Atributos Mais Importantes

A seleção de atributos realizada na fase da modelação, para além de proporcionar melhores desempenhos aos modelos preditivos, permite ainda retirar informação potencialmente útil sobre os atributos mais importantes à modelação. Deste modo, serão analisados os atributos que foram selecionados através do método *Wrapper* em ambos os algoritmos de modelação, isto, para cada uma das tarefas de previsão de CBO e SST. Como neste ponto já foram apresentados todos os modelos preditivos, tanto para a abordagem ETAR_1L como para a abordagem ETAR_2L, a análise de atributos será realizada com bases nos modelos gerados para ambas as abordagens. Deste

modo, descrevem-se assim os atributos com maior frequência de seleção nas várias tarefas executadas. Ainda de referir que não sendo este o objetivo principal do projeto, será apenas realizada esta análise dos parâmetros mais influentes na previsão de CBO e de SST. Outras técnicas poderiam ser exploradas neste âmbito, como por exemplo analisar o peso relativo à importância dos atributos, mediante as técnicas de seleção de atributos como as *Filters* (w-Correlação, w-Relief e w-SVM). Não obstante, a definição de importância de um atributo para o tratamento, requer um conhecimento aprofundado dos processo da ETAR que só os seus analistas possuem. Por conseguinte, a informação será extraída com o intuito de ser posteriormente analisada por um analista da ETAR, por forma a que este comprove a sua relevância.

Da seleção de atributos resultante dos métodos *Wrappers,* foram incluídos na tabela 6.6 aqueles que foram selecionados em pelo menos três dos quatro modelos desenvolvidos para as tarefas de previsão de CBO e SST, isto incluindo as duas abordagens exploradas. Restringe-se assim todos os atributos resultantes dos métodos de seleção, a um grupo que compreende apenas os parâmetros de tratamento mais proeminentes.

Tabela 6.6 – Principais atributos selecionados nas tarefas de modelação de CBO e SST, em ambas as abordagens ETAR_2L e ETAR_1L

Tarefa	Atributos
CBO	PA1_CQO; PA1_CBO; PA1_SST; PA2_O_{2_ZAx}; PA3_V30; PA3_O_2; PA4_P_{redox}; PA5_CBO; Caudal_Purga; IVL; SRT.
SST	Estação; PA1_N_{Total}; PA1_N-NH_4; PA2_$P_{redox_ZA/ZAx}$; PA2_SST; PA3_P_{redox}; PA5_pH; PA5_SST; PA5_CQO; PA5_CBO; Biodegradabilidade.

Algumas observações interessantes podem ser retiradas olhando para a tabela 6.6. Começando com os atributos selecionados pelos modelos na previsão de CBO, é perceptível que três atributos são relativos às medições realizadas pré-tratamento (PA1). Note-se ainda a presença de PA5_CBO em todos os modelos de previsão de CBO, na verdade este parâmetro demonstrou um forte corelacionamento com PA6_CBO, sendo portanto natural esta seleção. Ainda em CBO constam dois

atributos referentes a variáveis de processo (IVL e SRT), o que pode indicar que este tipo de variáveis contêm informação importante para os modelos. Nota ainda para o relacionamento entre os parâmetros PA3_V30 e IVL, ambos referentes a medições do volume do lodo.

Passando agora para os principais atributos na previsão de SST, saliente-se a presença do parâmetro referente à estação do ano, que pode indicar que mediante as estações do ano, diferentes as águas residuais dão entrada e/ou diferentes eficácias de tratamento conseguem ser alcançadas para o parâmetro SST. É interessante ainda a presença de atributos relativos aos nutrientes (PA1_N_{Total}, PA1_N-NH_4) que, apesar de terem muitos valores nulos, mostraram-se ainda assim relevantes nestes modelos. Na previsão de SST, destaca-se ainda a seleção de quatro parâmetros de PA5, que é o ponto anterior ao final do tratamento. Quanto aos parâmetros SST e SSV, embora apenas SST tenha sido selecionado por pelo menos 3 modelos, particularmente em PA2 e PA5, porém, SSV também foi selecionado pelos modelos em vários PA, alguns desses casos em simultâneo com SST. Outro factor interessante é a seleção do parâmetro Biodegradabilidade (PA1_CBO/PA1_CQO), que nos modelos de SST mostrou-se mais relevante que os respectivos parâmetros PA1_CBO e PA1_CQO – lembrando que estes foram selecionados frequentemente na previsão de CBO. Isto poderá indicar que nos modelos de SST, a informação derivada da biodegradabilidade é preferível aos parâmetros originais que a derivam.

Para finalizar esta análise, de referir que os parâmetros relativos ao oxigénio dissolvido (O_2) e potencial de redução (P_{Redox}) surgiram em vários modelos, tanto na previsão de CBO como SST. No entanto, destacam-se os parâmetros PA2_O_{2_ZAx}, que surgiram nos quatro modelos de previsão de CBO e, ainda na tarefa CBO, de salientar a presença do parâmetro PA4_P_{Redox}. Na previsão de SST realçam-se os parâmetros PA2_$P_{redox_ZA/ZAx}$, que apresentaram uma grande correlação entre si, e foram selecionados em todos os modelos de SST — em dois deles constam até os dois parâmetros em simultâneo. Concluindo, convém realçar que este tipo de informação deve ser portanto analisada por um especialista da ETAR, cujo conhecimento da plataforma e dos processos de tratamento permitirá determinar a utilidade e importância desta informação. Em todo o caso, ilustra-se aqui a capacidade de extração de informação mediante os modelos preditivos e respectiva seleção de atributos.

6.4 Sumário

Mediante o processo de modelação descrito no capítulo 5, foram realizadas várias tarefas de previsão. Das experiências realizadas e respetivos resultados, começamos por descrever a abordagem ETAR_2L. Inicialmente foi realizada uma análise à utilidade dos parâmetros de PA6 na modelação de CBO, os resultados demonstraram que estes parâmetros não são uma mais valia para os modelos de CBO. Por conseguinte, optou-se por descartar os parâmetros de PA6 de todas as tarefas de modelação, que inclui as duas abordagens em estudo. Ainda na abordagem ETAR_2L, apresentaram-se os resultados da modelação de CBO segundo os algoritmos SVR e SMO, onde os resultados mostram uma boa eficácia de previsão de ambos os modelos. Quanto à previsão de SST, nota-se que existe uma quebra da capacidade preditiva em relação CBO, e ainda que o algoritmo SMO superou o algoritmo SVR nesta tarefa de previsão em particular. Passando à abordagem ETAR_1L, cujo processo de preparação de dados e modelação foi em geral semelhante ao da ETAR_2L, também apresentaram-se os resultados da previsão com CBO e SST. Tanto na previsão de CBO como de SST, os dois modelos SVR e SMO revelaram um desempenho preditivo inferior aos mesmos modelos da abordagem ETAR_2L. De realçar, pelo lado negativo, os fracos resultados apresentados por ambos os modelos de SST, uma vez que ficaram bem abaixo das restantes tarefas de previsão.

Com base na seleção de atributos realizada na fase de modelação, foram analisados os atributos que se mostraram mais importantes na previsão de SST e CBO. Esta análise possibilitou identificar alguns parâmetros, cuja presença na maioria dos modelos de previsão fornece-nos informação de potencial interesse. Contudo, o relevo deste tipo de informações para o problema deverá ser comprovado por um especialista da ETAR, devido ao seu conhecimento profundo da planta e dos respetivos processos de tratamento.

Capítulo 7

Conclusões e Trabalho Futuro

7.1 Considerações Finais

Definiram-se inicialmente como meio de orientação deste projeto alguns objetivos a alcançar. Esses objetivos foram introduzidos no início desta dissertação e posteriormente detalhados no capítulo 4. Com isto, quer sobre uma perspetiva de negócio mais abstrata, quer sobre uma perspectiva mais orientada aos objetivos da aplicação de técnicas de DM, surgiram mais alguns objetivos durante a fase de análise de dados (i.e. análise das abordagens ETAR_2L e ETAR_1L). Por conseguinte, achamos que ao concluir-se este projeto tais objetivos devem ser confrontados com os resultados obtidos.

Desempenho dos Modelos de Previsão

Considerando que o principal objetivo deste projeto foi a procura de um modelo de previsão que capturasse a dinâmica de tratamento de uma ETAR, e que, consequentemente, tivesse uma boa capacidade preditiva dos parâmetros de qualidade, este estudo demonstrou que é possível aplicar modelos de previsão com sucesso numa ETAR. Em especial na previsão do parâmetro de qualidade CBO, no qual o desempenho preditivo foi elevado nas duas abordagens. No melhor modelo de previsão de CBO, obteve-se um erro relativo de 40%, que significa um desempenho que supera um preditor simples em 60%. Obteve-se, ainda, um coeficiente de correlação de 0.94, ou seja, muito próximo do valor máximo de 1. Apesar dos modelos de SST demonstrarem resultados mais baixos,

em particular na abordagem ETAR_1L, porém o melhor modelo preditivo de SST demonstrou resultados aceitáveis, com uma redução de 25% do erro relativo a um preditor simples e uma correlação de 0.75, isto na abordagem de modelação ETAR_2L. Pode-se ainda acrescentar, que esta diferença de desempenho preditivo de CBO em relação a SST, verifica-se também no trabalho apresentado em (Hamed et al., 2004). Similarmente, no trabalho de Atanasova & Kompare (2002b), SST revela uma grande dificuldade em ser previsto quando comparado com os restantes modelos. Estes factos reforçam a confiança nos resultados obtidos, que demonstram a dificuldade de prever o parâmetro de qualidade SST.

Comparação dos Algoritmos de Modelação (SVR e SMO)

O estudo dos algoritmos SVM, nomeadamente em tarefas de regressão, foi também alvo de estudo deste trabalho. Entre os dois algoritmos utilizados (SVR e SMO) os resultados aproximam-se bastante, isto, apesar dos modelos selecionarem atributos de input diferentes através do método *Wrapper*. Assim, os desempenhos preditivos dos algoritmos foram semelhantes em todas tarefas executadas, à exceção da previsão de SST na abordagem ETAR_2L. Nesta situação o algoritmo SMO demonstrou uma maior capacidade de previsão, que é visível pelos resultados da avaliação 10R-10CV e também pelos gráficos de dispersão e de linhas. De uma maneira geral, não se distingue com clareza qual dos algoritmos SVR ou SMO é o melhor. Todavia, atribuímos uma ligeira vantagem a SMO, essencialmente por duas razões: a primeira, pelos melhores resultados obtidos na tarefa de previsão de SST (ETAR_2L); e a segunda, pela maior precisão demonstrada nos gráficos de dispersão e de linhas, visto que os valores previstos por SMO mostraram uma maior frequência de valores muito próximos dos valores reais.

Generalização das SVM

De uma maneira geral, analisando a performance dos algoritmos SVM no problema, podemos dizer que as suas vantagens teóricas, em particular a generalização e o mapeamento não linear, comprovam-se com base nos resultados obtidos. O elevado desempenho demonstrado, principalmente na previsão de CBO, através de um processo de avaliação que consistiu em 100 conjuntos de dados de treino+teste distintos, permitiu demonstrar uma boa capacidade de generalização das SVM. Contudo, pode-se levantar mais uma vez a questão da generalização

perante dados desconhecidos e a consequente limitação de não se utilizar dados de teste devido ao pequeno conjunto de dados do problema. Para reforçar a credibilidade dos resultados, vamos considerar os resultados da métrica RMSE obtidos nas diferentes tarefas de previsão, segundo a validação 10R-10CV. Esses resultados serão comparados com outros dois métodos de avaliação apropriados para conjuntos de dados pequenos: o Bootstrap e o LOO. Assim, pegando nos modelos já com os respectivos atributos mais importantes e com os parâmetros optimizados segundo a avaliação 10R-10CV, foi estimado o desempenho desses mesmos modelos perante dois métodos de avaliação distintos. Na tabela 7.1 é resumida essa informação que nos permite analisar se os modelos estão sobre ajustados aos dados de treino, o que implicará resultados sobre estimados, ou então se o método de avaliação 10R-10CV está a subestimar os erros.

Tabela 7.1 – Comparação de métodos de avaliação, valores de RMSE.

Tarefa		**SVR**			**SMO**		
		10R-10CV	*LOO*	*Bootstrap*	*10R-10CV*	*LOO*	*Bootstrap*
ETAR_2L	CBO	3.241 ± 0.42	2.665 ± 1.71	3.763 ± 0.57	3.221 ± 0.42	2.546 ± 1.72	3.830 ± 0.46
	SST	8.639 ± 1.70	6.647 ± 5.25	10.759 ± 1.30	7.791 ± 1.71	5.854 ± 5.39	9.654 ± 1.44
ETAR_1L	CBO	4.167 ± 1.08	3.012 ± 3.11	4.856 ± 0.48	4.020 ± 0.93	2.974 ± 2.84	4.528 ± 0.74
	SST	9.683 ± 2.83	7.431 ± 7.105	10.757 ± 2.09	9.410 ± 2.42	6.459 ± 7.08	11.963 ± 2.12

Olhando para a tabela 7.1, podemos constatar alguns factos relatados na literatura, como o facto do Bootstraping ter menor variância que os métodos de validação cruzada, e ainda o facto de LOO apresentar uma grande variabilidade - neste caso representada pelo desvio padrão da média da precisão obtida. Contudo, para a análise da capacidade de generalização dos algoritmos SVM utilizados, é interessante realçar quatro pontos:

- Todas as avaliações confirmam o desempenho dos modelos, ou seja, nos modelos com desempenho preditivo elevado o erro estimado é baixo em todas as avaliações, e quando o desempenho é mau, o erro estimado é alto em todas as avaliações. Isto significa que a tendência dos resultados serem melhores ou piores mantem-se em todas as avaliações, comprovando a comparação realizada neste estudo entre os vários desempenhos obtidos.

- O método LOO apresenta uma enorme variabilidade, porém este método normalmente tem boas performances de estimação do RMSE, incluindo uma viés baixa (Molinaro et al., 2005). Isto pode-nos indicar que as outras avaliações possam estar subestimadas. Lembrando que em LOO são usadas n-1 instâncias para treino, o que implica que este método de avaliação é o que captura mais informação na construção dos modelos, este facto pode ajudar a explicar os erros mais baixos de LOO.

- O Bootstrapping foi o método mais pessimista dos três, que poderá revelar um sobre-ajustamento dos modelos aos dados de treino. Porém, no caso de CBO estas diferenças não parecem ser expressivas, ao passo que em SST, aí sim, existem diferenças mais consideráveis. Lembrando que este método realiza a avaliação com uma probabilidade de em cerca de 38.2% das instâncias não serem utilizadas em treino, é natural que esta falta de captura de informação vá baixar o desempenho preditivo. Assim, tendo em conta que essa percentagem de instâncias poderão ser avaliadas em teste sem serem usadas na modelação, e, olhando para os resultados de 10R-10CV, não nos parece que estes estejam a ser sobre estimados, principalmente no caso de CBO.

- Uma última nota para a escolha de 10R-10CV. Ao se observar os resultados da tabela repara-se que 10R-10CV fica no meio, ou seja, mostra-se o mais equilibrado quer em termos de variância quer em precisão. Adicionalmente, sabendo que a rotação da validação cruzada permite tirar o máximo partido dos dados de treino, que a variância de 10R-10CV é baixa, e como a tendência de resultados foi coerente entre os três modelos, o método 10R-10CV provou ser uma boa escolha de avaliação de modelos neste problema.

Com base nesta análise podemos reforçar a ideia que os algoritmos SVM demonstraram uma boa capacidade de generalização. Adianta-se ainda que todo o processo de preparação de dados e de modelação adotado contribuiu para este feito. Por um lado, na preparação de dados, foi evidente a melhoria obtida em cada passo executado. Por outro lado, na modelação, essa melhoria também se verificou na seleção de atributos e na otimização de parâmetros.

Abordagens ETAR_2L e ETAR_1L

Nas duas abordagens do problema que foram investigadas, de forma geral, verificou-se que a modelação com duas linhas de tratamento teve resultados superiores, o que pode indicar que a modelação ETAR_1L distorce os dados ao realizar a média dos parâmetros. No entanto, a abordagem ETAR_1L traz uma maior facilidade de interpretação de resultados, devido à menor dimensionalidade e consequente simplificação do problema. Nesta abordagem, o conjunto de atributos resultantes do método de seleção de atributos foi menor, em especial no caso de CBO com cerca de 12 atributos, o que permite identificar melhor os parâmetro que influenciam o tratamento. Durante o processo de desenvolvimento foi notória a redução dos tempos de processamento na abordagem ETAR_1L, em especial na seleção de atributos. Este facto indica que em futuros problemas com maior volume e/ou dimensionalidade de dados, a abordagem ETAR_1L deverá equacionada, até porque no caso de CBO os resultados podem ser considerados bons. Contudo, realça-se que neste problema, apesar da maior dimensionalidade ao considerar-se duas linhas, estamos a modelar a ETAR de forma mais clara, não introduzindo ruído, e, por conseguinte, os modelos revelaram uma melhor aprendizagem e um consequente desempenho preditivo elevado — mesmo perante um maior número de valores nulos presentes em ETAR_2L que tiveram de ser tratados.

Informação Útil

Com a análise realizada aos atributos mais relevantes, identificou-se um grupo restrito de atributos que se mostraram, assim, mais influentes na previsão de CBO e SST. Não obstante, também na fase de análise de dados e preparação de dados, observaram-se várias características dos parâmetros que podem ser de potencial interesse, e até alguns problemas da ETAR. Este foi o caso da análise de relacionamentos de atributos, que revelou algumas dependências fortes entre atributos, mas também a análise de valores nulos, inconsistências dos dados, e histogramas, que demonstraram a existência de várias dificuldades e problemas nos tratamentos da ETAR, em particular, sobre um ponto de vista da análise da aptidão dos dados a serem modelados. Convém no entanto relembrar que a confirmação e definição da utilidade desta informação, deverá ser atribuída por especialistas do domínio das ETAR, em particular, com um grande conhecimento da ETAR em estudo.

Ferramenta de DM Utilizada (RapidMiner)

Nesta altura, pensa-se ser adequado rematar este estudo, apontando algumas considerações finais sobre a ferramenta de DM utilizada em todo este processo de desenvolvimento e aplicação de técnicas de DM numa ETAR. É de referir que a versatilidade do *RapidMiner*, devido às extensões do R e do WEKA revelaram-se bastante úteis. Especialmente na utilização de algumas técnicas de DM originárias destas duas ferramentas, como o SMOReg, o *SoftMax Scalling,* entre outras, que foram experimentadas tanto na fase de preparação de dados como na modelação, possibilitando assim um estudo mais abrangente. Outro factor positivo foi a completude do *RapidMiner*, uma vez que proporcionou realizar todas as diferentes tarefas apresentadas neste estudo, desde a apresentação das estatísticas base do conjunto de dados, até à avaliação dos modelos com três métodos distintos. Por último, de referir que os operadores ETL disponibilizados mostraram-se úteis, particularmente em ligeiros ajustes ao conjunto dados, como a remoção de alguns atributos ou exemplos, sem que se tivesse de recorrer ao processo manual através de folhas de cálculo. Assim, conclui-se que a ferramenta *RapidMiner* mostrou-se adequada à realização deste projeto de DM.

7.2 Trabalho Futuro

Várias dificuldades e limitações características do problema em estudo, foram apontadas ao longo deste documento. É por isso relevante equacionar essas limitações e indicar possíveis caminhos a seguir por forma a contorná-las. Isso implica de facto, que deve ser realizada investigação adicional sobre este problema. Em primeiro lugar, deve-se referir aquela que pode ser considerada a principal limitação deste estudo, ou seja, o facto de estarmos perante um problema cujo conjunto de dados é pequeno. Apesar de se usarem técnicas conhecidas pela boa generalização dos dados, é inevitável ter-se que assumir que devido ao reduzido número de exemplos não se consegue obter uma captura de informação completa do problema real. Note-se, também, que os dados do problema correspondem ao período de apenas um ano de tratamento e com uma frequência de registos baixa durante esse período. Tal facto mostra que não se está a capturar todo o problema que envolve o tratamento de águas residuais da ETAR. Portanto, corre-se o risco de não se estar a produzir modelos válidos para o problema que se projetou (Pyle, 1999). Por conseguinte, é de suma

importância que em trabalhos futuros, sobretudo antes de um processo de implementação, se investigue este problema perante um cenário com um maior volume de dados, quer em termos do número de registos, quer na duração dos períodos temporais referentes aos tratamentos realizados. Outra limitação deste conjunto de dados é a sua granularidade, o facto dos registos conterem a média dos valores diários pode afectar a eficácia dos modelos, pois não capturam de forma efetiva a dinâmica do processo de tratamento. Um conjunto de dados com um grão mais fino de medições (e.g. intervalos de 2 horas), provavelmente traria uma melhor eficácia na previsão, tal como é referido no trabalho relacionado de Atanasova & Kompare (2002b). Deste modo, como trabalho futuro deverá ser investigado este problema com base em registos mais frequentes e sem valores médios das várias medições, deste facto resulta também um maior volume de informação.

Como trabalho futuro, poderá ser ainda explorada a inclusão de parâmetros na modelação do sistema de tratamento que não foram considerados neste trabalho, essencialmente devido à grande taxa de valores nulos apresentada, como aconteceu no caso do tratamento da fase sólida e ainda dos parâmetros relativos à microfauna. Deste modo, perante mais dados, seria interessante analisar o contributo que estes parâmetros poderão trazer à modelação do comportamento da ETAR. Adicionalmente, a previsão de outros parâmetros de qualidade, para além de SST e CBO, trariam com certeza mais informação interessante sobre o processo de tratamento da ETAR. Mas não só a previsão de outros parâmetros de qualidade poderá ser de mais valia, o estudo adicional da previsão de parâmetros relativos a outros PA poderiam ser de grande utilidade para a ETAR. Contudo, as decisões sobre quais as tarefas de modelação de maior importância para a ETAR deverão ser especificadas pelos agentes de decisão das ETAR.

O mesmo acontece nas decisões sobre quais os atributos de *input* a considerar e quais os respectivos momentos de recolha de dados. Isto, por forma a idealizar um modelo de previsão que seja exequível em tempo útil, ou seja, que preveja com alguma antecedência o parâmetro de previsão. Estas questões devem ser levantadas especialmente em processos de implementação de aplicações de DM para produção, que requerem um conhecimento e levantamento de requisitos mais detalhados sobre todo processo de recolha e previsão de dados. Não obstante, nesta dissertação elaborou-se um estudo que se baseia essencialmente na capacidade de se prever eficazmente, ou não, a qualidade dos tratamentos realizados na ETAR. Para este efeito, procurou-se sempre adoptar as decisões que nos pareceram mais ajustadas à realidade da ETAR em questão, e

a uma futura implementação das tarefas de modelação estudadas. Contudo, sabe-se que é necessário conhecimento prático e profundo da ETAR a modelar, por forma a encontrar as melhores decisões.

Com base neste estudo de um caso real, poderá ser assim equacionada futuramente a implementação de um sistema de suporte à decisão que possibilite, por exemplo, aos analistas da ETAR obterem as previsões da qualidade final do tratamento no momento que introduzem os dados de input das fases de tratamento anteriores. Porém, será necessária uma investigação adicional de aplicações reais de DM e a da sua integração na ETAR, bem como toda uma estratégia de modelação adaptada à ETAR em questão — um exemplo de um sistema deste género é apresentado em (Poch et al., 2000). Conclui-se assim, que é evidente que mais investigação sobre esta ETAR, inclusive sobre as limitações mencionadas anteriormente, deverá ser realizada por forma a desenvolver um sistema útil e credível. Em todo o caso, o estudo aqui relatado deve ser considerado como um ponto de partida, sobre o qual se poderá avançar para um processo de implementação prática. Por conseguinte, o resultado final dessa implementação será com certeza uma mais valia para o tratamento da ETAR, possibilitando o aumento da qualidade das águas tratadas que são descarregadas no nosso meio-ambiente dia após dia.

Referências

Ali, S. & Smith-Miles, K.A., 2006. Improved Support Vector Machine Generalization Using Normalized Input Space. *AI 2006 Advances in Artificial Intelligence*, 4304(2), pp.362-371.

Atanasova, N. & Kompare, B., 2002a. Modelling of Wastewater Treatment Plant with Decision and Regression Trees. Em: *3rd Workshop on Binding Environmental Sciences and Artificial Intelligence*, 2002, pp.52-60.

Atanasova, N. & Kompare, B., 2002b. Modelling of Wastewater Treatment Plant with Regression Trees. Em: *Proc. of the Third International Conference on Data Mining*, pp.867-879. Bologna, Italy, 2002. WIT Press.

Azevedo, R.T.d., n.d. *Tecnologias de Tratamento de Águas Residuais Urbanas*. [Online] Naturlink & PTC. Disponível em: http://naturlink.sapo.pt/Natureza-e-Ambiente/Gestao-Ambiental/content/Tecnologias-de-Tratamento-de-aguas-Residuais-Urbanas/section/2?bl=1&viewall=true#Go_2 [Acedido a 20 de Maio 2012].

Batista, G.E.A.P.A. & Monard, M.C., 2003. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, pp.519-533.

Belanche, L.A. et al., 1999a. Towards a Model of Input-Output Behaviour of Wastewater Treatment Plants using Soft Computing Techniques. *Environmental Modeling and Software*, 14(5), pp.409-419.

Belanche, L.A. et al., 1999b. A Study of Qualitative and Missing Information in Wastewater Treatment Plants. Em: *EDSSAI'99: Environmental Decision Support Systems and Artificial Intelligence*, pp.30-38. Florida, 1999. Papers from the AAAI'99 Workshop.

Boser, B.E., Guyon, I.M. & Vapnik, V.N., 1992. A Training Algorithm for Optimal Margin Classifiers. Em: D Haussler, ed. *Proceedings of the fifth annual workshop on Computational learning theory COLT 92*, 1992, 6(8), pp.144-152. ACM Press.

Burges, C.J.C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), pp.121-167.

Cărbureanu, M., 2010. Pollution Level Analysis of a Wastewater Treatment Plant Emissary using Data Mining. *Petroleum-Gas University of Ploieşti Bulletin Mathematics-Informatics-Physics Series*, LXII(1), pp.69-78.

Chang, C.-C. & Lin, C.-J., 2011. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp.1-27.

Cherkassky, V. & Ma, Y., 2004. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks*, 17(1), pp.113-126.

Comas, J. et al., 2001. Knowledge Discovery by Means of Inductive Methods in Wastewater Treatment Plant Data. *AI Commun.*, 14(1), pp.45-62.

Connoly, T. & Berg, C., 2010. *Database Systems: An Pratical Aprouch to Design, Implementation and Management*. 5th ed. Addison-Wesley.

Cortes, C. & Vapnik, V., 1995. Support-Vector Networks. *Machine Learning*, 20(3), pp.273-297.

Cortez, P., 2010. Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. Em: Perner, P., ed. *Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects (ICDM'10)*, pp.572-583. Berlin, Heidelberg, 2010. Springer-Verlag.

Cortez, P., 2011. Data Mining with Multilayer Perceptrons and Support Vector Machines. *Data Mining Foundations and Intelligent Paradigms*, p.9–25.

CRISP-DM, 2000. *Cross Industry Standard Process for Data Mining, Process Model*. [Online] Disponível em: http://www.crisp-dm.org/download.htm [Acedido a 28 de Junho 2012].

CRISP-DM, 2006. *Cross Industry Standard Process for Data Mining, Process Model*. [Online] Disponível em: http://crispdm.wordpress.com/process-model/ [Acedido a 28 de Junho 2012].

Dürrenmatt, D.J., 2011. *Data Mining and Data-Driven Modelling Approaches to Support Wastewater Treatment Plant Operation*. PhD Thesis. Zurique: ETH.

Decreto de Lei n.º 152/97 de 19 de Junho, 1997. *DIÁRIO DA REPÚBLICA — I SÉRIE-A — Nº139*. [Online] Disponível em: http://dre.pt/pdf1s/1997/06/139A00/29592967.pdf [Acedido a 20 de Maio 2012].

Dixon, M. et al., 2007. Data Mining to Support Anaerobic WWTP Monitoring. *Control Engineering Practice*, 15(8), pp.987-999.

Drucker, H. et al., 1997. Support Vector Regression Machines. *Electronic Engineering*, 1, pp.155-161.

Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P., 1996a. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11), pp.27-39.

Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P., 1996b. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), pp.37-54.

Feng, H., Chen, G., Yin, C. & Yang, B.C.Y., 2005. A SVM Regression Based Approach to Filling in Missing Values. Em: R.K.e. al., ed. *Knowledge-Based Intelligent Information and Engineering Systems*. Springer Berlin / Heidelberg. pp.581–587.

Flexer, A., 1996. Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice. Em: *3th European Meeting on Cybernetics and Systems Research,* 1996, 2, pp.1005-1008.

Gallop, J.R. et al., 2004. The Use of Data Mining for the Monitoring and Control of Anaerobic Wastewater Plants. Em: *4th International Workshop on Environmental Applications of Machine Learning (EAML)*, 2004.

Golfareli, M. & Rizzi, S., 2009. *Data Warehouse Design: Modern Principies and Methodologies.* McGraw-Hill.

Guyon, I. & Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(7-8), pp.1157-1182.

Hamed, M.M., Khalafallah, M.G. & Hassanien, E.A., 2004. Prediction of Wastewater Treatment Plant Performance Using Artificial Neural Networks. *Environmental Modelling Software*, 19(10), pp.919-28.

Hong, Y., Fei, L., Yuge, X. & Jin, L., 2008. GA Based LS-SVM Classifier for Waste Water Treatment Process. Em: *27th Chinese Control Conference,* 2008, pp.436-439.

Hsu, C.-w., Chang, C.-c. & Lin, C.-j., 2010. A Practical Guide to Support Vector Classification. *Bioinformatics*, 1(1), pp.1-16.

Huang, Z., Luo, J., Li, X. & Zhou, Y., 2009. Prediction of Effluent Parameters of Wastewater Treatment Plant Based on Improved Least Square Support Vector Machine with PSO. Em: *1st International Conference on Information Science and Engineering (ICISE)*, 2009, pp.4058-4061.

Huang, C.-L. & Wang, C.-J., 2006. A GA-based Feature Selection and Parameters Optimization for Support Vector Machines. *Expert Systems with Applications*, 31, pp.231–240.

KDnuggets, 2007. *KDnuggets : Polls : Data Mining Methodology*. [Online] Disponível em: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm [Acedido a 28 de Junho 2012].

KDnuggets, 2011. *KDnuggets Home; Polls; Data Mining/Analytic Tools Used*. [Online] Disponível em: http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html [Acedido a 26 de Maio 2012].

KDnuggets, 2012. *KDnuggets Home; Polls; Analytics, Data mining, Big Data software used*. [Online] Disponível em: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html [Acedido a 16 de Agosto 2012].

Kim, J.-H., 2009. Estimating Classification Error Rate: Repeated Cross-Validation, Repeated Hold-out and Bootstrap. *Computational Statistics & Data Analysis,* 53(11), pp.3735-3745.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Em: *International Joint Conference on Artificial Intelligence,* 14(2), pp.1137-1143. Montreal, Canada, 1995.

Kohavi, R. & John, G.H., 1997. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2), pp.273-324.

Lardon, L. et al., 2002. Specifications of Modular Internet-Based Remote Supervision Systems for Wastewater Treatment Plants. Em: *3rd Workshop on Binding Environmental Sciences and Artificial Intelligence*, 2002, pp.45-50.

Levy, J.d.Q., 2000. *Aspectos Determinantes na Concepção de Estação de Tratamento de Águas Residuais*. [Online] Disponível em: http://www.ecoservicos.pt/index_htm_files/Aspectos_determ_concep_ETAR.pdf [Acedido a 7 de Maio 2012].

Lorena, A.C. & Carvalho, A.C.P.L.F., 2007. Uma Introdução às Support Vector Machines. *RITA*, XIV(2), pp.43-67.

Luengo, J., García, S. & Herrera, F., 2011. On the Choice of the Best Imputation Methods for Missing Values Considering Three Groups of Classification Methods. *Knowledge and Information Systems* , pp.1-32.

Luo, F., Yu, R.-h., Xu, Y.-g. & Li, Y., 2009. Effluent Quality Prediction of Wastewater Treatment Plant Based on Fuzzy-Rough Sets and Artificial Neural Networks. Em: *Sixth International Conference on Fuzzy Systems and Knowledge Discovery - FSKD'09*, 2009, 5, pp.47-51.

Metcafl & Eddy, 2003. *Wastewater Engineering: Treatment and Reuse*. 4th ed. McGraw-Hill.

Molina, L.C., Belanche, L. & Nebot, À., 2002. Feature Selection Algorithms: A Survey and Experimental Evaluation. Em: *2002 IEEE International Conference on Data Mining (ICDM '02).* Washington DC, USA, 2002. IEEE Computer Society.

Molinaro, A.M., Simon, R. & Pfeiffer, R.M., 2005. Prediction Error Estimation: A Comparison of Resampling Methods. *Bioinformatics*, 21(15), p.3301–3307.

Platt, J.C., 1999. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. *Optimization*, 11, pp.1-8.

Poch, M. et al., 2000. Wastewater Treatment Improvement Through an Intelligent Integrated Supervisory System. *Contributions to Science*, 1(4), pp.451-462.

Pyle, D., 1999. *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Rapid-i, 2010. *Optimize Selection (Evolutionary)*. [Online] Rapid-i. Disponível em: http://rapid-i.com/wiki/index.php?title=Optimize_Selection_(Evolutionary) [Acedido a 24 de Junho 2012].

Rapid-I, n.d. *RapidMiner*. [Online] Disponível em: http://rapid-i.com/content/view/181/196/ [Acedido a 4 Julho 2012].

Schölkopf, B., Smola, A., Williamson, R. & Bartlett, P.L., 2000. New Support Vector Algorithms. *Neural Computation*, 12(5), pp.1207-1245.

Shevade, S.K., Keerthi, S.S., Bhattacharyya, C. & Murthy, 2000. Improvements to the SMO Algorithm for SVM Regression. *IEEE Transactions on Neural Networks*, 11(5), pp.1188-1193.

Smola, A.J. & Schölkopf, B., 1998. *A Tutorial on Support Vector Regression*. NeuroColt2 Technical Report Series.

Smola, A.J. & Schölkopf, B., 2004. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3), pp.199-222.

Sosik, S.J., n.d. *SCADA Systems in Wastewater Treatment*. [Online] Disponível em: http://www.process-logic.com/content/images/SCADA.pdf [Acedido a 14 de Maio 2012].

Spellman, F.R., 2003. *Handbook of Water & Wastewater Treatment Plant Operations*. LEWIS Publishers.

Sperling, M.v., 2007. *Wastewater characteristics, treatment and disposal*. London: IWA Publishing.

Torgo, L., 2010. *DMwR*. [Online] Disponível em: http://cran.r-project.org/web/packages/DMwR/DMwR.pdf [Acedido a 12 de Maio 2012].

UN-Water, n.d. *UN-Water Statistics*. [Online] Disponível em: http://www.unwater.org/statistics.html [Acedido a 8 de Julho 2012].

Vanrolleghem, P.A. & Lee, D.S., 2003. On-line Monitoring Equipment for Wastewater Treatment Processes: State of the Art. *Water Science and Technology*, 47(2), pp.1-34.

Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. 2nd ed. New York: Springer-Verlag.

Vapnik, V.N., 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), pp.988-999.

Wang, L.-j. & Chen, C.-b., 2008. Support Vector Machine Applying in the Prediction of Effluent Quality of Sewage Treatment Plant with Cyclic Activated Sludge System Process. Em: *IEEE International Symposium on Knowledge Acquisition and Modeling Workshop. KAM Workshop* 2008, pp.647-650.

WEKA, n.d. *Class SMOReg*. [Online] Disponível em: http://weka.sourceforge.net/doc/weka/classifiers/functions/SMOreg.html [Acedido a 2 de Julho 2012].

WHO & UNICEF, 2012. *Progress on Drinking Water and Sanitation: 2012 Update*. [Online] Disponível em: http://www.wssinfo.org/fileadmin/user_upload/resources/JMP-report-2012-en.pdf [Acedido a 20 de Julho 2012].

Witten, I.H., Frank, E. & Hall, M.A., 2011. *Data Mining: Pratical Machine Learnign Tools and Tecniques*. 3rd ed. Morgan Kaufmann.

Wunderground, n.d. *Histórico para Porto, Portugal*. [Online] Disponível em: http://portuguese.wunderground.com/history/airport/LPPR/2010/1/4/MonthlyHistory.html?req_city=NA&req_state=NA&req_statename=NA [Acedido a 10 de Abril 2012].

Yang, Y., 2005. *Support Vector Machines for Environmental Informatics : Application to Biological Nitrogen Removal in Wastewater Treatment Plants*. M.Sc. Dissertation. Ryerson University.

Yang, B.-l., Zhao, D.-a. & Zhang, J., 2011. Prediction System of Sewage Outflow COD Based on LS-SVM. Em: *2nd International Conference on Intelligent Control and Information Processing (ICICIP)*, 2011, 1, pp.399-402.

Lista de Siglas e Acrónimos

- *BD* - Base de Dados
- *BI* - *Business Intelligence*
- *CBO* - Carência Bioquímica de Oxigénio
- *CQO* - Carência Química de Oxigénio
- *CRISP-DM* - *Cross Industry Standard Process for Data Mining*
- *DM* - Prospecção ou Mineração de Dados (do Inglês - *Data Mining*)
- *DW* - *Data Warehouse*
- *ETAR* - Estações de Tratamento de Águas Residuais
- *ETL* - *Extract, Transform, Load*
- *JMP* - *Joint Monitoring Programme*
- *KDD* - *Knowledge Discovery in Databases*
- *KKT* - *Karush–Kuhn–Tucker*
- *k-NN* - *K-Nearest-Neighbor*
- *MAR* - *Missing at Random*
- *MCAR* - *Missing Completely at Random*
- *MDG* - *Millennium Development Goals*
- *MLP-NN* - *Multilayer Perceptron - Neural Networks*
- *NMAR* - *Not Missing at Random*
- *NN* - Redes Neuronais (do Inglês – *Neural Networks*)
- *PA* - Pontos de Amostragem
- *PQ* - Programação Quadrática
- *RBF* - *Radial-Basis Function*
- *RMSE* - Raiz do Erro Quadrático Médio (do Inglês - *Root Mean Squared Error)*

- *RRSE* - Raiz do Erro Quadrático Médio Relativo (do Inglês - *Root Relative-Squared Error)*
- *SCADA* - *Supervisory Control and Data Acquisition*
- *SEMMA* - *Sample, Explore, Modify, Model, Assess*
- *SGBD* - Sistemas de Gestão de Base de Dados
- *SMO* - *Sequential Minimal Optimization*
- *SS* - Sólidos Suspensos
- *SSD* - Sistemas de Suporte à Decisão
- *SST* - Sólidos Suspensos Totais
- *SSV* - Sólidos Suspensos Voláteis
- *SVM* - Máquinas de Vetores de Suporte (do Inglês – *Support Vector Machines*)
- *SVR* - *Support Vector Regression*
- *VC* - *Vapnik e Chervonenkis*
- *WWW* - *World Wide Web*

ANEXO A

Estatísticas do Conjunto de Dados

Neste anexo apresentam-se duas tabelas com os dados estatísticos do conjunto de dados de trabalho. Tal como foi referido no capítulo 4, foram analisadas duas abordagens diferentes do problema, uma em que se considerou os dados relativos às duas linhas de tratamento na integra (ETAR_2L), e uma segunda em que foi feita a média dos valores da duas linhas de tratamento (ETAR_1L).

Foram ainda usadas duas abreviaturas na tabela: "min", que corresponde ao valor mínimo e "med" que corresponde ao valor médio. De acrescentar que as tabelas foram geradas com recurso à ferramenta *RapidMiner*, cujos cálculos estatísticos derivam, portanto, da funcionalidade oferecida por esta ferramenta.

A.1 ETAR_2L

Papel	Parâmetro	Tipo	Estatísticas	Intervalo de Valores	Nulos
id	Data	nominal	moda = 2010/01/04 (1), min = 2010/01/04 (1)	[04/01/2010 : 30/12/2010]	0.0
regular	Estação	nominal	moda = Winter (24), min = Autumn (20)	Winter (24), Spring (23), Summer (23), Autumn (20)	0.0
regular	Temperatura	real	med = 15.478 +/- 5.008	[6.000 ; 30.000]	0.0
regular	Estado_Tempo	Bi-nominal	moda = Sun (51), min = Rain (39)	Rain (39), Sun (51)	0.0
regular	PA1_Redox	real	med = -145.844 +/- 122.097	[-315.000 ; 245.000]	0.0
regular	PA1_pH	real	med = 7.469 +/- 0.324	[6.690 ; 8.520]	0.0
regular	PA1_CQO	real	med = 756.678 +/- 336.063	[170.000 ; 2280.000]	0.0
regular	PA1_CBO	real	med = 343.633 +/- 188.091	[25.000 ; 950.000]	0.0
regular	PA1_SST	real	med = 282.467 +/- 248.029	[48.000 ; 2084.000]	0.0
regular	PA1_SSV	real	med = 235.011 +/- 222.333	[35.000 ; 1884.000]	0.0
regular	PA1_P_Total	real	med = 8.595 +/- 3.976	[2.000 ; 23.000]	48.0
regular	PA1_N_Total	real	med = 98 +/- 41.139	[12.000 ; 210.000]	48.0
regular	PA1_N_NH4	real	med = 63.220 +/- 37.168	[11.000 ; 164.000]	49.0
regular	L1_PA2_pH	real	med = 7.193 +/- 0.151	[6.820 ; 7.840]	19.0
regular	L1_PA2_$O2_{ZA}$	real	med = 0.505 +/- 0.238	[0.140 ; 1.420]	37.0
regular	L1_PA2_$O2_{ZAx}$	real	med = 0.137 +/- 0.130	[0.030 ; 0.910]	37.0
regular	L1_PA2_$Redox_{ZA}$	real	med = -177.585 +/- 61.107	[-298.000 ; -64.000]	37.0
regular	L1_PA2_$Redox_{ZAx}$	real	med = -203.442 +/- 66.630	[-347.000 ; -81.000]	38.0
regular	L1_PA2_SST	real	med = 2571.830 +/- 821.975	[810.000 ; 4730.000]	37.0
regular	L1_PA2_SSV	real	med = 2079.151 +/- 714.333	[390.000 ; 4050.000]	37.0
regular	L1_PA3_pH	real	med = 7.455 +/- 0.278	[6.730 ; 7.930]	13.0
regular	L1_PA3_O2	real	med = 2.549 +/- 1.639	[0.490 ; 8.120]	13.0
regular	L1_PA3_Redox	real	med = -17.623 +/- 76.782	[-183.000 ; 234.000]	13.0

regular	L1_PA3_V30	real	med = 434.740 +/- 286.252	[60.000 ; 1000.000]	13.0
regular	L1_PA3_SST	real	med = 2728.092 +/- 935.805	[930.000 ; 5070.000]	14.0
regular	L1_PA3_SSV	real	med = 2188.592 +/- 811.661	[740.000 ; 4170.000]	14.0
regular	L1_PA4_Redox	real	med = -72.481 +/- 102.672	[-260.000 ; 141.000]	13.0
regular	L1_PA4_SST	real	med = 6831.382 +/- 2893.280	[1780.000 ; 12700.000]	14.0
regular	L1_PA4_SSV	real	med = 5437.763 +/- 2393.482	[1330.000 ; 10135.000]	14.0
regular	L1_PA4_SSV_SST	real	med = 79.014 +/- 3.950	[69.665 ; 89.026]	14.0
regular	L1_PA5_pH	real	med = 7.325 +/- 0.277	[6.590 ; 7.920]	13.0
regular	L1_PA5_CQO	real	med = 71.766 +/- 25.698	[20.000 ; 161.000]	13.0
regular	L1_PA5_CBO	real	med = 21.571 +/- 11.482	[7.000 ; 64.000]	13.0
regular	L1_PA5_SST	real	med = 25.104 +/- 12.681	[9.000 ; 65.000]	13.0
regular	L1_PA5_SSV	real	med = 19.377 +/- 11.579	[6.000 ; 58.000]	13.0
regular	L2_PA2_pH	real	med = 7.208 +/- 0.114	[7.020 ; 7.520]	22.0
regular	L2_PA2_$O2_{ZA}$	real	med = 0.563 +/- 0.239	[0.100 ; 1.270]	46.0
regular	L2_PA2_$O2_{ZAx}$	real	med = 0.147 +/- 0.120	[0.020 ; 0.790]	46.0
regular	L2_PA2_$Redox_{ZA}$	real	med = -174.773 +/- 76.820	[-307.000 ; -19.000]	46.0
regular	L2_PA2_$Redox_{ZAx}$	real	med = -199.068 +/- 75.684	[-331.000 ; -68.000]	46.0
regular	L2_PA2_SST	real	med = 2792.545 +/- 707.811	[1300.000 ; 4320.000]	46.0
regular	L2_PA2_SSV	real	med = 2314.773 +/- 630.277	[1070.000 ; 4010.000]	46.0
regular	L2_PA3_pH	real	med = 7.441 +/- 0.300	[6.290 ; 7.890]	14.0
regular	L2_PA3_O2	real	med = 2.254 +/- 1.901	[0.270 ; 9.180]	14.0
regular	L2_PA3_Redox	real	med = -16.961 +/- 88.987	[-168.000 ; 254.000]	14.0
regular	L2_PA3_V30	real	med = 525.133 +/- 280.798	[120.000 ; 1000.000]	15.0
regular	L2_PA3_SST	real	med = 2692.200 +/- 1002.992	[525.000 ; 5085.000]	15.0
regular	L2_PA3_SSV	real	med = 2072.067 +/- 803.608	[320.000 ; 4050.000]	15.0
regular	L2_PA4_Redox	real	med = -87.197 +/- 109.688	[-282.000 ; 335.000]	14.0

regular	L2_PA4_SST	real	med = 7008 +/- 3444.623	[1130.000 ; 19035.000]	15.0
regular	L2_PA4_SSV	real	med = 5499.840 +/- 2826.769	[960.000 ; 15740.000]	15.0
regular	L2_PA4_SSV_SST	real	med = 78.492 +/- 7.231	[40.494 ; 95.184]	15.0
regular	L2_PA5_pH	real	med = 7.337 +/- 0.278	[6.360 ; 7.890]	14.0
regular	L2_PA5_CQO	real	med = 75.579 +/- 22.584	[22.000 ; 158.000]	14.0
regular	L2_PA5_CBO	real	med = 23.158 +/- 12.023	[3.000 ; 77.000]	14.0
regular	L2_PA5_SST	real	med = 51.882 +/- 149.368	[8.000 ; 1239.000]	14.0
regular	L2_PA5_SSV	real	med = 41.553 +/- 122.880	[4.000 ; 1004.000]	14.0
regular	PA6_pH	real	med = 7.475 +/- 0.337	[6.590 ; 8.050]	0.0
regular	PA6_CQO	real	med = 67.733 +/- 22.151	[20.000 ; 120.000]	0.0
label	PA6_CBO	real	med = 19.682 +/- 9.369	[1.000 ; 56.000]	2.0
label	PA6_SST	real	med = 23.178 +/- 12.182	[2.000 ; 65.000]	0.0
regular	PA6_SSV	real	med = 17.667 +/- 12.666	[0.000 ; 75.000]	0.0
regular	PA6_P_Total	real	med = 2.644 +/- 2.021	[0.800 ; 9.000]	49.0
regular	PA6_N_Total	real	med = 52.098 +/- 22.414	[9.000 ; 91.000]	49.0
regular	PA6_N_NH4	real	med = 43.122 +/- 23.727	[5.000 ; 100.000]	49.0
regular	PA6_N_NO3	real	med = 8.951 +/- 9.450	[2.000 ; 45.000]	49.0
regular	PA6_CQO_eff	real	med = 89.257 +/- 6.700	[60.656 ; 97.500]	0.0
regular	PA6_CBO_eff	real	med = 90.860 +/- 11.873	[27.586 ; 99.630]	2.0
regular	PA6_SST_eff	real	med = 87.355 +/- 11.831	[32.292 ; 98.802]	0.0
regular	PA6_N-Total_eff	real	med = 42.443 +/- 24.023	[-10.526 ; 84.615]	49.0
regular	Caudal_Entrada	real	med = 5204.250 +/- 1208.705	[2193.000 ; 9629.000]	2.0
regular	Caudal_Recirculacao	real	med = 7205.556 +/- 2604.981	[4000.000 ; 14000.000]	0.0
regular	RAS	real	med = 1.422 +/- 0.466	[0.555 ; 2.900]	2.0
regular	L1_Caudal_purga	real	med = 130.862 +/- 106.801	[0.000 ; 480.000]	10.0
regular	L2_Caudal_purga	real	med = 114.988 +/- 88.449	[0.000 ; 340.000]	10.0

regular	L1_IVL	real	med = 178.655 +/- 147.749	[33.679 ; 714.286]	14.0
regular	L2_IVL	real	med = 265.063 +/- 315.873	[0.000 ; 1809.524]	15.0
regular	L1_SRT	real	med = 8.854 +/- 17.580	[0.553 ; 85.885]	22.0
regular	L2_SRT	real	med = 2.565 +/- 2.809	[0.000 ; 11.926]	26.0
regular	Biodegradabilidade	real	med = 0.443 +/- 0.155	[0.093 ; 0.906]	0.0
regular	L1_FM	real	med = 0.390 +/- 0.350	[0.044 ; 2.515]	15.0
regular	L2_FM	real	med = 0.395 +/- 0.324	[0.074 ; 2.089]	15.0

A.2 ETAR_1L

Papel	Parâmetro	Tipo	Estatísticas	Intervalo de Valores	Nulos
id	Data	nominal	moda = 04/01/10 (1), min = 04/01/10 (1)	[04/01/2010;30/12/2010]	0.0
regular	Estação	nominal	moda = Winter (24), min = Autumn (20)	Winter (24), Spring (23), Summer (23), Autumn (20)	0.0
regular	Temperatura	real	med = 15.478 +/- 5.008	[6.000 ; 30.000]	0.0
regular	Estado_Tempo	Bi-nominal	moda = Sun (51), min = Rain (39)	Rain (39), Sun (51)	0.0
regular	PA1_Redox	real	med = -145.844 +/- 122.097	[-315.000 ; 245.000]	0.0
regular	PA1_pH	real	med = 7.469 +/- 0.324	[6.690 ; 8.520]	0.0
regular	PA1_CQO	real	med = 756.678 +/- 336.063	[170.000 ; 2280.000]	0.0
regular	PA1_CBO	real	med = 343.633 +/- 188.091	[25.000 ; 950.000]	0.0
regular	PA1_SST	real	med = 282.467 +/- 248.029	[48.000 ; 2084.000]	0.0
regular	PA1_SSV	real	med = 235.011 +/- 222.333	[35.000 ; 1884.000]	0.0
regular	PA1_P_Total	real	med = 8.595 +/- 3.976	[2.000 ; 23.000]	48.0
regular	PA1_N_Total	real	med = 98 +/- 41.139	[12.000 ; 210.000]	48.0
regular	PA1_N_NH4	real	med = 63.220 +/- 37.168	[11.000 ; 164.000]	49.0
regular	PA2_pH	real	med = 7.188 +/- 0.118	[6.820 ; 7.520]	6.0
regular	PA2_O2_ZA	real	med = 0.541 +/- 0.239	[0.120 ; 1.420]	34.0
regular	PA2_O2_ZAx	real	med = 0.140 +/- 0.114	[0.030 ; 0.850]	34.0
regular	PA2_Redox_ZA	real	med = -172.536 +/- 64.080	[-302.500 ; -71.000]	34.0
regular	PA2_Redox_ZAx	real	med = -197.164 +/- 67.275	[-339.000 ; -80.000]	35.0
regular	PA2_SST	real	med = 2674.634 +/- 734.603	[1105.000 ; 4730.000]	34.0
regular	PA2_SSV	real	med = 2189.821 +/- 643.661	[745.000 ; 4050.000]	34.0
regular	PA3_pH	real	med = 7.417 +/- 0.280	[6.525 ; 7.890]	0.0

regular	PA3_O2	real	med = 2.512 +/- 1.737	[0.270 ; 8.120]	0.0
regular	PA3_Redox	real	med = -19.206 +/- 76.044	[-169.000 ; 221.000]	0.0
regular	PA3_V30	real	med = 518.667 +/- 267.202	[112.500 ; 1000.000]	0.0
regular	PA3_SST	real	med = 2625.694 +/- 864.604	[930.000 ; 5077.500]	0.0
regular	PA3_SSV	real	med = 2054.239 +/- 723.249	[740.000 ; 4110.000]	0.0
regular	PA4_Redox	real	med = -79.772 +/- 87.593	[-250.500 ; 141.500]	0.0
regular	PA4_SST	real	med = 6589.917 +/- 2811.518	[1780.000 ; 14735.000]	0.0
regular	PA4_SSV	real	med = 5185.906 +/- 2323.475	[1330.000 ; 12020.000]	0.0
regular	PA4_SSV_SST	real	med = 78.303 +/- 5.032	[53.110 ; 86.965]	0.0
regular	PA5_pH	real	med = 7.313 +/- 0.276	[6.590 ; 7.890]	0.0
regular	PA5_CQO	real	med = 72.978 +/- 18.622	[21.000 ; 108.000]	0.0
regular	PA5_CBO	real	med = 22.056 +/- 9.457	[7.000 ; 56.500]	0.0
regular	PA5_SST	real	med = 36.483 +/- 68.807	[10.000 ; 627.500]	0.0
regular	PA5_SSV	real	med = 28.583 +/- 56.836	[4.000 ; 508.500]	0.0
regular	PA6_pH	real	med = 7.475 +/- 0.337	[6.590 ; 8.050]	0.0
regular	PA6_CQO	real	med = 67.733 +/- 22.151	[20.000 ; 120.000]	0.0
label	PA6_CBO	real	med = 19.682 +/- 9.369	[1.000 ; 56.000]	2.0
label	PA6_SST	real	med = 23.178 +/- 12.182	[2.000 ; 65.000]	0.0
regular	PA6_SSV	real	med = 17.667 +/- 12.666	[0.000 ; 75.000]	0.0
regular	PA6_P_Total	real	med = 2.644 +/- 2.021	[0.800 ; 9.000]	49.0
regular	PA6_N_Total	real	med = 52.098 +/- 22.414	[9.000 ; 91.000]	49.0
regular	PA6_N_NH4	real	med = 43.122 +/- 23.727	[5.000 ; 100.000]	49.0
regular	PA6_N_NO3	real	med = 8.951 +/- 9.450	[2.000 ; 45.000]	49.0
regular	PA6_CQO_eff	real	med = 89.257 +/- 6.700	[60.656 ; 97.500]	0.0
regular	PA6_CBO_eff	real	med = 90.860 +/- 11.873	[27.586 ; 99.630]	2.0
regular	PA6_SST_eff	real	med = 87.355 +/- 11.831	[32.292 ; 98.802]	0.0
regular	PA6_N-Total_eff	real	med = 42.443 +/- 24.023	[-10.526 ; 84.615]	49.0

regular	Caudal_Entrada	real	med = 5204.250 +/- 1208.705	[2193.000 ; 9629.000]	2.0
regular	Caudal Recirculacao	real	med = 7205.556 +/- 2604.981	[4000.000 ; 14000.000]	0.0
regular	RAS	real	med = 1.422 +/- 0.466	[0.555 ; 2.900]	2.0
regular	Caudal_Purga	real	med = 114.739 +/- 77.964	[0.000 ; 300.000]	0.0
regular	IVL	real	med = 242.918 +/- 192.788	[45.083 ; 962.505]	0.0
regular	SRT	real	med = 7.414 +/- 16.202	[0.000 ; 85.885]	7.0
regular	Biodegradabilidade	real	med = 0.443 +/- 0.155	[0.093 ; 0.906]	0.0
regular	FM	real	med = 0.403 +/- 0.328	[0.044 ; 2.515]	2.0