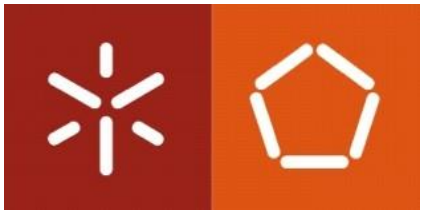
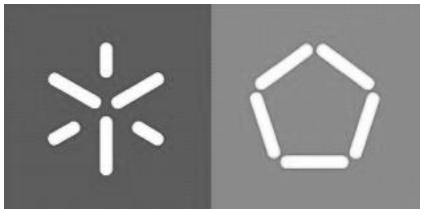


Universidade do Minho
Escola de Engenharia

Tiago Mendes Ferreira

Descoberta de Padrões de Consumo de Energia Elétrica

Abril de 2013



Universidade do Minho
Escola de Engenharia

Tiago Mendes Ferreira

Descoberta de Padrões de Consumo de Energia Elétrica

Dissertação de Mestrado em Engenharia Informática

Trabalho efetuado sob a orientação do
Professor Doutor Orlando Manuel Oliveira Belo

Abril de 2013

Universidade do Minho, 15/04/2013

Assinatura: 

Universidade do Minho

Escola de Engenharia

Departamento de Informática

Descoberta de Padrões de Consumo de Energia Elétrica

Tiago Mendes Ferreira

Dissertação de Mestrado

2013

Descoberta de Padrões de Consumo de Energia Elétrica

Tiago Mendes Ferreira

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre em Engenharia Informática elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2013

Aos meus pais, por tudo.

Agradecimentos

Ao Departamento de Informática da Universidade do Minho e a todo o seu corpo docente por todo o conhecimento de excelência que me transmitiram ao longo da minha Licenciatura e Mestrado em Engenharia Informática.

Ao professor Orlando Belo, por me ter orientado ao longo deste projeto e por me ter transmitido durante todo o Mestrado em Engenharia Informática os conhecimentos necessários, tanto para a elaboração desta dissertação como para o meu percurso profissional.

Aos meus pais e irmão, por todo o apoio e dedicação que demonstraram ao longo de toda a minha vida académica. A vossa missão, que agora termina, foi bem sucedida.

A todos os meus amigos, pelas palavras de incentivo, apoio e confiança quando nem tudo corria pelo melhor.

Ao Paulo Festa, em especial, pelos sábios e ponderados conselhos que sempre me foi dando ao longo desta jornada. A tua presença nos momentos fulcrais foi essencial para levar este último projeto académico a bom porto.

À Ana, por ter estado sempre presente e disponível em todas as fases boas e menos boas desta etapa da minha vida académica. Sem o teu apoio teria sido tudo muito mais difícil.

Resumo

Descoberta de Padrões de Consumo de Energia Elétrica

Desde a década de 60, o consumo de energia elétrica em Portugal tem vindo a aumentar de uma forma consideravelmente acentuada e constante, tendo como consequência o aumento da despesa, principalmente ao nível empresarial. Este mesmo aumento do consumo de energia elétrica também se verifica a nível mundial, tendo como principais responsáveis o aumento da população e a evolução tecnológica. Estas duas condicionantes ajudam desde já a compreender que o consumo de energia elétrica mundial tenha praticamente triplicado nas últimas quatro décadas.

É possível analisar o consumo de energia elétrica recorrendo a técnicas de mineração de dados, que irão ajudar a encontrar padrões e anomalias numa quantidade substancial de informação, o que, por sua vez, se pode revelar bastante útil para as pessoas que enfrentam este tipo de problemas energéticos.

Nesta dissertação, pretende-se recorrer a técnicas de mineração de dados como as *support vector machines* e as redes neuronais (MLP), de forma a construir modelos capazes de prever o consumo de energia elétrica em habitações domésticas.

Palavras-chave: Mineração de dados, Sistemas de Energia Elétrica, *Support Vector Machines*, Redes Neuronais, Previsão de Consumo Doméstico de Energia Elétrica.

Abstract

Discovering Electrical Energy Consumption Patterns

Since the 60's decade, the electric power consumption in Portugal has been rising considerably in a constant manner, resulting in the significant increase of costs over the years, mainly at business level. This increase of electric power consumption is also verified worldwide level and the main causes for this phenomenon are the increasing population and the technological evolution. These two factors help to explain why the electric power consumption in the whole world has almost tripled in the last four decades.

It is possible to analyse the electric power consumption by using data mining techniques, which will help finding patterns and anomalies in a substantial amount of information which can turn out to be very useful to people who face this kind of energetic issues.

In this dissertation, it is intended to use data mining techniques such as support vector machines and artificial neural networks (MLP) in order to build models capable of predict the electric power consumption in domestic residences.

Keywords: Data Mining, Power Systems, Support Vector Machines, Neural Networks, Prediction of Domestic Electric Energy Consumption.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 O Consumo de Energia Elétrica

Nos últimos 50 anos, o consumo de energia elétrica em Portugal aumentou de uma forma muito consistente (Figura 1). Esse mesmo consumo é relevante se tivermos em conta que desde os anos 60 até 2009, o consumo de eletricidade quase quintuplicou, crescendo sempre de uma forma praticamente linear (Google Public Data Explorer, 2013).

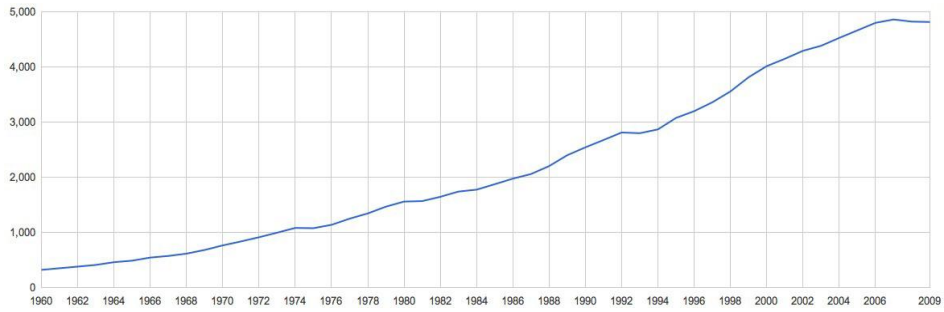


Figura 1 - Consumo de energia elétrica (kWh) per capita em Portugal (Google Public Data Explorer, 2013)

A problemática do aumento do consumo de energia elétrica em Portugal deve-se a um extenso número de variáveis, como o aumento praticamente constante da população desde as primeiras

estatísticas registadas. Esta tendência é apenas quebrada com um ligeiro decréscimo nos anos anteriores à revolução de Abril de 1974 (Pordata, 2013), sobretudo devido às medíocres condições de vida que se verificavam naqueles tempos.

Um outro fator contribuinte para o aumento do consumo deste bem indispensável é igualmente a evolução tecnológica conjugada com o aumento das condições financeiras, pois verificou-se um acréscimo significativo do número de equipamentos domésticos por agregado familiar (Pordata, 2013). A título de exemplo, entre 1995 e 2005, a percentagem de agregados familiares que possuíam computador passou de cerca de 10% para praticamente 44%. Se analisarmos exaustivamente os dados, poderemos concluir que esta situação se reflete também na maioria dos eletrodomésticos, televisões e outros tipos de bens adquiridos. Mais, numa indústria que cada vez menos recorre a mão-de-obra humana em detrimento das máquinas, é relativamente fácil perceber os problemas que daí poderão advir no que diz respeito ao consumo de energia elétrica.

Ainda outra variável que deverá alterar os índices de consumo de energia elétrica é o clima. Em Portugal, o clima está cada vez mais adverso. Os dias apresentam-se mais quentes no verão, o que torna praticamente indispensável o recurso ao ar condicionado. Já os invernos mostram-se substancialmente mais frios, sendo que o uso de aparelhos térmicos se revela uma maior necessidade. Todas estas situações vão, portanto, ao encontro de um aumento exacerbado do consumo de eletricidade, tendo como consequência direta a despesa monetária associada e os problemas ambientais. Contudo, ao longo dos anos tem-se verificado um esforço significativo para a redução da utilização de energia elétrica, bem como tornar o seu consumo mais barato e eficiente. Veja-se o exemplo das empresas de distribuição de eletricidade - no caso de Portugal, a EDP (EDP, 2013) – que criaram novas tarifas de venda de eletricidade para fomentar o consumo em certas horas do dia. Essas tarifas, bi-horária e tri-horária, têm como objetivo incentivar o consumo de eletricidade nas horas de vazio (horas noturnas e fim-de-semana), proporcionando uma redução na fatura de eletricidade de 45% no preço base do kWh (EDP, 2013). Esta é uma forma de ajudar famílias e empresas a gerirem melhor o seu orçamento. Todavia, existem muitas outras formas para reduzir o consumo de energia elétrica que serão elencadas mais à frente.

No que concerne ao consumo de energia elétrica a nível mundial, é possível verificar, por observação da Figura 2, que desde que existem estatísticas sobre o tema, o consumo tem vindo sempre a aumentar, embora numa velocidade mais gradual comparativamente com os valores

registados em Portugal (Figura 1).

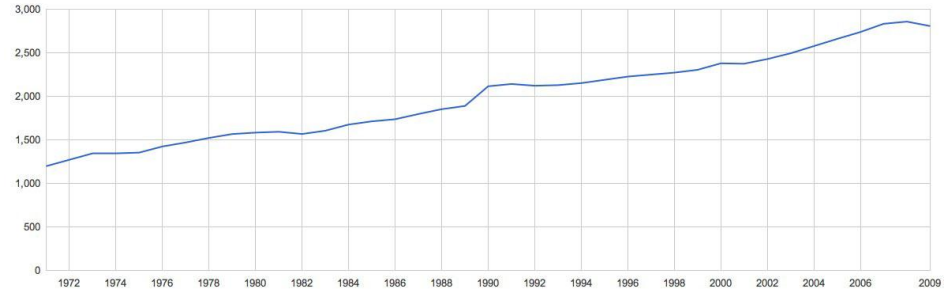


Figura 2 - Consumo de energia elétrica (kWh) per capita no Mundo (Google Public Data Explorer, 2013)

O aumento em questão justifica-se aqui facilmente, na medida em que a população mundial nos últimos 50 anos aumentou de ≅3 biliões de pessoas para ≅7 biliões (Figura 3). Este aspecto conjugado com a evolução tecnológica já anteriormente mencionada, explica o porquê de o consumo de energia elétrica mundial nos últimos 40 anos ter praticamente triplicado.

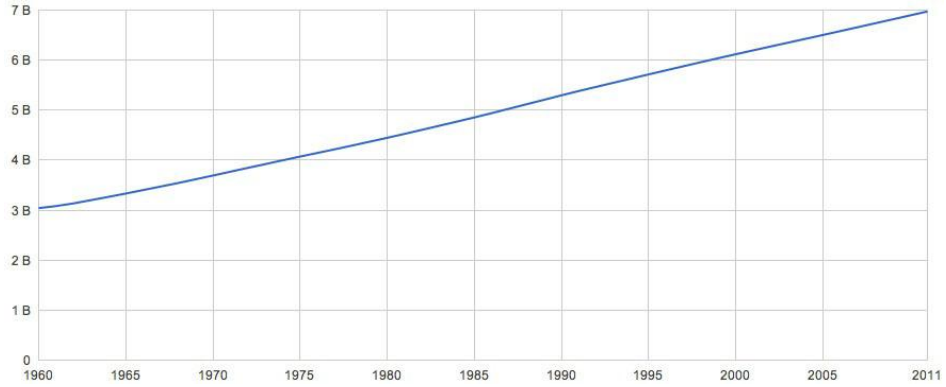


Figura 3 - População mundial (Google Public Data Explorer, 2013)

Vários esforços estão a ser levados a cabo para ajudar a travar o consumo abrupto de energia elétrica em todo o mundo. Veja-se o caso dos computadores: o seu consumo de energia é facilmente reduzido recorrendo a diversos métodos. Desligar o monitor ou reduzir a sua luminosidade quando o mesmo se encontra sem qualquer atividade durante um determinado intervalo temporal é um método possível, por exemplo. Aquando de períodos de grande inatividade

dos sistemas também é possível reduzir o consumo de energia recorrendo à redução da velocidade do relógio do processador ou mesmo desligando-o. Estas e outras formas de redução de energia em sistemas informáticos podem ser encontradas em Fung (1998).

Já no que à iluminação diz respeito, não há dúvidas que esta é uma área cujo desenvolvimento se tem evidenciado positivamente. Desde lâmpadas economizadoras de halogéneo (cuja poupança anunciada se situa em 30%) passando por outras, também de halogéneo (cuja poupança já ascendia aos 50%), tornam-se aqui evidentes algumas das opções que o futuro poderá oferecer. Nos últimos anos, as lâmpadas mais conhecidas são as lâmpadas economizadoras fluorescentes compactas, cuja poupança anunciada é de 80%. Entretanto, e já mais recentemente, chegaram as lâmpadas LED cujo aquecimento é praticamente nulo e cuja poupança é superior a 80%, tendo apenas como desvantagem a destacar o seu elevado preço para quem as pretende adquirir (Lourenço, 2009).

Ainda outro esforço levado a cabo ao longo das últimas décadas corresponde à diminuição do consumo de energia elétrica por parte dos dispositivos que precisam de estar em modo de espera (*Standby mode*) para utilizarem todas as suas funcionalidades, como por exemplo, receber um sinal de um telecomando ou mesmo para alimentar a luz LED que indica o estado de funcionamento do dispositivo (Standby Power, 2013). Felizmente, o consumo de energia para alimentar estas funcionalidades desceu abruptamente nos últimos anos. A título de exemplo, uma televisão poderia gastar cerca de 10W em modo de espera (Standby Power, 2013). No entanto, hoje em dia uma televisão já poderá ter um consumo em modo de espera inferior a 0,1W (Samsung, 2013).

Uma outra área que tem igualmente contribuído, ainda que não de uma igual forma devido aos seus elevados custos de aquisição, instalação e manutenção, é a domótica. A domótica é uma tecnologia que permite a gestão dos recursos numa habitação ou empresa. No que toca à energia elétrica, esta tecnologia permite monitorizar os consumos energéticos; optimizar o desperdício, como desligar a iluminação ou as televisões quando uma divisão se encontra vazia ou substituir o acender de uma lâmpada por uma simples abertura de estore; e até controlar problemas como incêndios, alertando a corporação de bombeiros registada no sistema (Soucek, Russ G., & Tamarit, 2000). O isolamento de um edifício também é importante para a redução do consumo de eletricidade, pois se o mesmo estiver mal isolado, os equipamentos térmicos irão funcionar mais do

que o suposto, elevando assim a fatura de eletricidade. No que diz respeito a este fator, existem já empresas especializadas em verificação de qualidade de isolamento de um edifício para que os seus proprietários possam ter uma casa eficiente no que ao consumo de equipamentos diz respeito.

1.2 Mineração de Dados e Consumo de Energia Elétrica

Uma área do saber que tem estudado os problemas do aumento do consumo de energia elétrica é a mineração de dados (*Data Mining*), que é definida como um processo de descoberta de padrões e de extração de conhecimento numa quantidade substancial de dados (Han & Kamber, 2006). A descoberta destes padrões pode trazer inúmeras vantagens do ponto de vista económico para as organizações (Witten, Frank, & Hall, 2011). A mineração de dados tem contribuído ao longo dos tempos com algumas soluções interessantes, mais ou menos viáveis, para o aumento da eficiência energética nos mais diversos tipos de instalações, recorrendo à análise dos dados de consumo dos edifícios para tentar encontrar padrões e anomalias. A exploração desses dados de consumo pode ser feita com recurso à mineração de dados através de modelos de classificação ou de regressão. Os primeiros consistem, tal como nome indica, em classificar um determinado exemplo dentro de uma determinada categoria (classe) enquanto que os segundos assentam num processo de previsão de uma certa quantidade numérica (Witten *et al.*, 2011). Podem ser ainda utilizadas técnicas como o *clustering* de dados, também conhecido como Segmentação, que é definido como um processo de agrupamento de objetos semelhantes (Han & Kamber, 2006).

O tema da mineração de dados aplicado ao domínio do consumo de energia elétrica é atualmente um dos assuntos abordados pela comunidade científica dentro desta área de estudo. Por exemplo, Tso e Yau (2007) utilizaram três modelos para a previsão do consumo de energia elétrica: análise de regressão, árvores de decisão e rede neuronais. Esta última técnica foi igualmente utilizada por Kalogirou e Bojic (2000), enquanto Chen, Das e Cook (2010) incluíram também as *Naïve Bayes*, *Bayes Net* e *Support Vector Machines* (SVM) no seu estudo.

1.3 Motivação e Objetivos da Dissertação

Como é já sabido, a tecnologia tem vindo a desenvolver-se a um ritmo absolutamente alucinante. Hoje em dia já é possível armazenar uma enorme quantidade de dados em formato digital sem que para isso seja necessário um grande esforço financeiro. É fácil encontrar uma empresa que possua um *data warehouse* (DW) atualmente, que mais não é do que um grande repositório de dados históricos cujo objetivo é suportar os processos de tomada de decisão de uma empresa. É pois aqui que entra a mineração de dados - uma atividade que pode ser definida, segundo Witten *et al.* (2011), como um processo de extração de informação implícita e potencialmente útil, anteriormente desconhecida.

No caso concreto do consumo de energia elétrica, que representa um elevado custo monetário e ambiental, a mineração de dados pode ser uma ferramenta importante para ajudar tantos fornecedores como consumidores de energia. Relativamente aos fornecedores é possível prever o consumo de energia elétrica de uma dada cidade ou região, com a finalidade de se precaverem para o futuro, como por exemplo ao nível das infraestruturas, evitando assim construções e aquisições de equipamentos feitas à pressa que tantas vezes conduzem a situações de grandes prejuízos. Relativamente aos consumidores, é possível, por exemplo, prever qual o seu consumo de energia elétrica anual, de acordo com o conjunto de equipamentos que possui. Desta forma, o consumidor consegue avaliar os seus custos energéticos previamente, o que o poderá ajudar a tomar decisões tão simples como mudar as lâmpadas por outras mais económicas ou simplesmente saber que não deve exceder-se na utilização do ar condicionado, uma vez que este é usualmente um equipamento que consome bastante energia elétrica e, como tal, muito dispendioso.

Neste trabalho pretende-se estudar e avaliar as *support vector machines* (SVM), no sentido de determinar se estas são uma técnica de mineração de dados mais adequada para a previsão do consumo de energia elétrica em habitações domésticas. Para esse fim, delinearam-se os seguintes objetivos:

- Estudar as *support vector machines* e desenvolver modelos capazes de fornecer a informação pretendida;

- Analisar a qualidade da informação fornecida pelos modelos de *support vector machines* construídos, verificando assim se os mesmos modelos se adequam aos objetivos pretendidos;

- Comparar os modelos de *support vector machines* desenvolvidos com outras técnicas de mineração de dados geralmente utilizadas neste tipo de problemas.

1.4 Organização da Dissertação

Para além do presente capítulo, este documento está organizado em mais 5 capítulos. Ao longo de cada um deles irão ser abordados e discutidos os principais tópicos que este trabalho de dissertação envolveu. No Capítulo 2, contextualiza-se o sector energético e descrevem-se os vários estudos e técnicas de mineração de dados existentes na literatura para o cálculo do consumo de energia elétrica. Em particular, no Capítulo 3 apresentam-se as técnicas de mineração de dados que serão efetivamente utilizadas neste trabalho. No Capítulo 4, é descrito o pré-processamento dos dados, que inclui tarefas de compreensão, preparação e consolidação dos dados. Adicionalmente, é feita a análise dos dados disponíveis para os processos de mineração levados a cabo, a avaliação da sua qualidade e a construção do conjunto de dados final. Seguidamente, no Capítulo 5, são apresentados os testes efetuados na fase de modelação e os resultados daí resultantes para cada uma das técnicas de mineração de dados testadas. O documento encerra, no Capítulo 6, com a apresentação das conclusões do trabalho, fazendo-se a avaliação dos resultados obtidos, bem como a análise comparativa dos modelos testados, a avaliação do trabalho efetuado e ainda a apresentação de algumas linhas de orientação para trabalho futuro.

Capítulo 2

O Consumo de Energia Elétrica

2.1 Energia Elétrica e Consumo

O setor da energia é um dos sectores de atividade cuja comunidade científica ainda hoje se debruça seriamente. Entender a progressão ou o modo como este setor é sustentado é compreender as diferentes necessidades da população, das empresas e as próprias questões ambientais.

A energia elétrica pode ser gerada com recurso a diferentes tecnologias, embora as mais comuns sejam a termoelétrica com $\cong 67\%$ de quota de mercado, as energias renováveis como a hidroelétrica e a eólica com $\cong 16\%$ e, por fim, a termonuclear com $\cong 13\%$ (Figura 4). Todas elas objetivam a movimentação das turbinas de um gerador, alterando apenas o método a que recorrem para atingir esse fim. Assim, a termoelétrica utiliza diferentes fontes de calor (e.g. carvão, gás ou madeira) para aquecer caldeiras de água que, por sua vez, libertam vapores que movem as turbinas de um gerador. As restantes utilizam, respetivamente, o potencial hidráulico existente nos rios, o vento e a reação nuclear para movimentar essas mesmas turbinas.

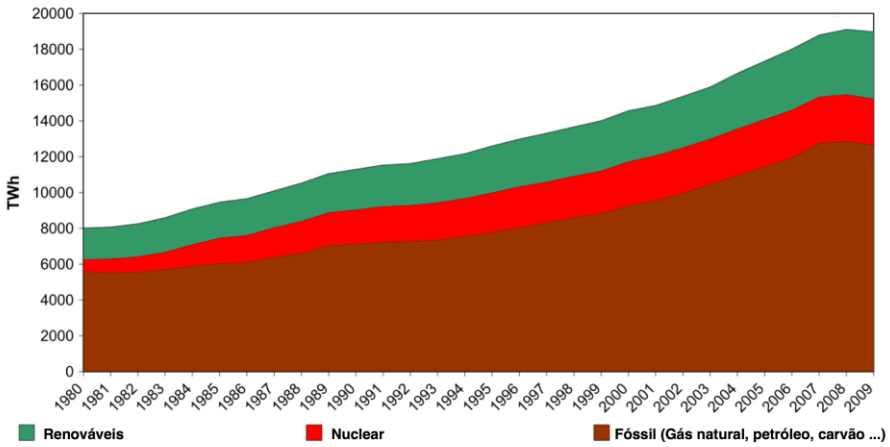


Figura 4 - Fontes de geração de energia (OECD, 2011)

Toda a energia produzida é depois colocada na rede de transporte de modo a conduzi-la até aos centros de consumo, tarefa em Portugal assegurada pela REN (www.ren.pt), sendo sempre sujeita a perdas consideradas insignificantes (Figura 5). O consumo de energia elétrica junto da população é posteriormente medido em kWh (*kilowatt-hour*), embora vejamos com regularidade referências a GWh (*gigawatt-hour*) ou TWh (*terawatt-hour*) associado à produção.

Tal como já mencionado, o crescimento no consumo energético tem sido praticamente constante ao longo das últimas quatro décadas. Atualmente são grandes potências mundiais como Estados Unidos, China, Japão ou Rússia as responsáveis pelos maiores consumos elétricos (Figura 6). Apesar de não ocupar os lugares cimeiros, Portugal consome cerca de 5000 kWh *per capita* ao ano demonstrando que o consumo elétrico é um ponto chave no orçamento da famílias ou empresas (Trading Economics, 2010).

Figura 5 - Perdas registadas em Portugal entre 1997 e 2009 (ERSE, 2009)

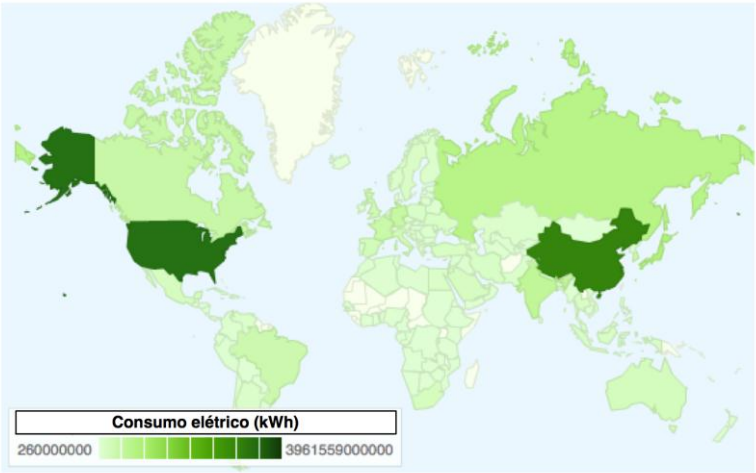


Figura 6 – Consumo de energia elétrica mundial (Lebanese Economy Forum, 2009)

2.2 Avaliação de Consumo de Energia Elétrica

Em Dezembro de 2002 foi aprovada uma diretiva europeia sobre o consumo energético dos edifícios que exige que os países membros desenvolvam metodologias capazes de:

- calcular o consumo energético dos edifícios, dos sistemas de AVAC (HVAC na terminologia anglo-saxónica) – aquecimento, ventilação e ar condicionado – e dos sistemas de iluminação;
- estabelecer requisitos mínimos de eficiência energética para os edifícios que venham a ser construídos no futuro e aplicar esses requisitos nos edifícios existentes;
- desenvolver um sistema de certificação energética para os edifícios;
- inspecionar regularmente os sistemas de aquecimento e ar condicionado.

O consumo energético dos edifícios depende significativamente de determinados critérios estabelecidos para os seus sistemas, critérios esses que também afectam a saúde, produtividade e o conforto dos seus ocupantes. No caso dos sistemas AVAC, esses critérios (i.e. temperatura mínima no inverno e máxima no verão) serão usados para calcular a carga de aquecimento e arrefecimento e assim ajudar a garantir que dentro de determinado departamento de um edifício estes limites de temperatura sejam cumpridos. Para cada tipo de edifício existem diferentes parametrizações destes limites (Tabela 1). Comparando, por exemplo, um escritório e uma loja

comercial, facilmente se consegue compreender que este último, no inverno, não carece de uma temperatura tão elevada uma vez que os seus ocupantes deslocam-se a pé de um lado para o outro, enquanto que no verão precisa de ver a sua temperatura mais reduzida, também pelas mesmas razões. Já nos edifícios que não possuem sistema AVAC, os mesmos critérios deverão ser cumpridos. Contudo, para os controlar será preciso recorrer a outro tipo de estratégias como, por exemplo, criar sombras para que a luz solar não sobreaqueça o edifício ou abrir janelas.

Tabela 1 - Valores recomendados para a temperatura no interior dos edifícios – adaptado de Olesen, Seppanen e Boerstra (2006)

Tipo de Edifício	**Categoria**	**Aquecimento ºC (Inverno)**	**Arrefecimento ºC (Verão)**
Edifícios Residenciais: quarto, sala, cozinha – Sedentário	A	21,0	25,5
	B	20,0	26,0
	C	18,0	27,0
Edifícios Residenciais: hall, sótão, dispensa – Não sedentário	A	18,0	-
	B	16,0	-
	C	14,0	-
Escritório – Sedentário	A	21,0	25,5
	B	20,0	26,0
	C	19,0	27,0
Cafetaria ou Restaurante – Sedentário	A	21,0	25,5
	B	20,0	26,0
	C	19,0	27,0
Grandes superfícies comerciais – Não sedentário	A	17,5	24,0
	B	16,0	25,0
	C	15,0	26,0

As categorias referenciadas na Tabela 1 correspondem ao nível de exigência requerido por determinado edifício. Relativamente ao controlo de iluminação e controlo de humidade existem igualmente critérios a ser cumpridos dependendo do tipo de edifício. No que concerne aos

sistemas de humidificação e de desumidificação do ar, tipicamente apenas são necessários em edifícios de cariz especial como hospitais ou museus. Todos estes critérios deverão ser definidos aquando do processo de desenho dos edifícios. Contudo, nos que já se encontram construídos também poderá existir esta parametrização.

Existem diversos métodos para o cálculo do consumo de energia elétrica. Este cálculo pode ser realizado numa base sazonal, mensal ou até de hora a hora. Olesen *et al.* (2006) explicam que para um cálculo sazonal e mensal são usados os valores parametrizados inicialmente para a temperatura de aquecimento e arrefecimento. Para um cálculo hora à hora deverá ser usado o ponto médio da temperatura durante a hora avaliada.

Para ser feita uma avaliação do consumo de energia elétrica é necessário executar uma medição num determinado período temporal para se poder concluir se há realmente poupança de energia relativamente ao previsto. Birt e Newsham (2009) analisaram vários estudos feitos por variados autores em países distintos e concluíram que as variações que existem relativamente ao consumo de energia elétrica previsto e o que efetivamente foi consumido advém de fatores como:

- o número total de horas de ocupação de determinado edifício pode diferir do que foi inicialmente concebido;
- a construção final do edifício pode diferir significativamente relativamente ao projeto inicial e as suas previsões de consumo;
- as tecnologias propostas para poupança energética podem não ser tão eficazes como inicialmente fora previsto;
- as ligações à corrente são geralmente muito diferentes do assumido;
- a existência de uma lacuna na transferência do conhecimento por parte das equipas de projeto e os utilizadores finais.

Enquanto os efeitos da temperatura no conforto estão perfeitamente estudados, os efeitos da temperatura na performance têm sido desprezados ao longo do tempo. Uma temperatura do ar inadequada pode influenciar a produtividade de forma indireta já que pode causar sintomas como irritação dos olhos, nariz e garganta, irritação da pele e até más sensações relacionadas com o odor ou sabor (Godish, 2001). Este tipo de condições térmicas é bastante complicada de controlar devido a diversos fatores como estruturas de refrigeração insuficientes, áreas demasiados grandes ou a concepção inadequada da infraestrutura. Adicionalmente, estas condições térmicas podem

variar consideravelmente ao longo do tempo, uma vez que as condições exteriores podem mudar (e.g. o espaçamento entre os edifícios após uma reestruturação urbana).

Nesse sentido, Seppänen e Fisk (2005) realizaram um estudo onde demonstraram que a performance decaía ≅2% a cada aumento de 1ºC para um intervalo de temperatura entre os 25ºC e os 32ºC, sendo que a temperatura ideal encontrada se situa entre os 21ºC e os 23ºC (Figura 7).

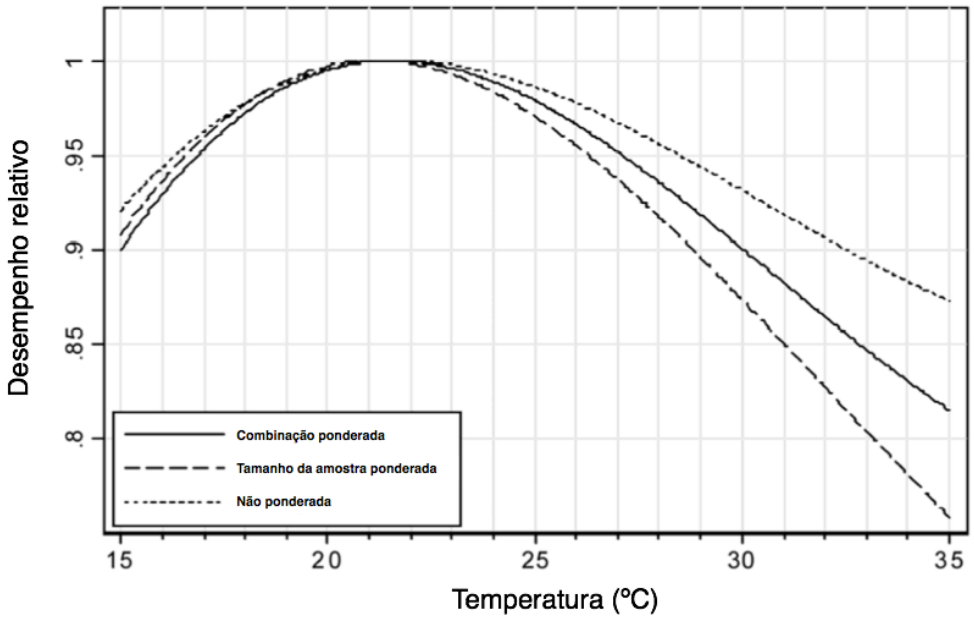


Figura 7 - Efeito da temperatura no desempenho humano – adaptado de Seppänen e Fisk (2005)

Para alcançar estes resultados sobre a relação entre temperatura e produtividade os autores utilizaram dados de sete locais de trabalho, dois referentes a ambientes laboratoriais e os restantes em *call centers*. O desempenho laboral dos primeiros foi medido com base na rapidez e eficácia das ações realizadas, enquanto que nos *call centers* se utilizou unicamente o tempo por chamada necessário pelo operador. A recolha de dados foi realizada em incrementos de $10\ L/s - pessoa$, i.e. a razão entre a taxa de ventilação tipicamente expressa em m^3/s e o número de pessoas no espaço. O desempenho foi calculado com base na variação percentual do mesmo, i.e. $\frac{D_{i-1}-D_i}{D_{i-1}}$ onde D_i representa o desempenho na *i*-nésima observação. Foram também aplicados três tipos de fatores de ponderação para enquadrar os resultados com o contexto real (e.g. uma tarefa pode ser mais relevante que outra), devidamente detalhados em Seppänen, Fisk e Lei (2005).

2.3 Técnicas para a Previsão do Consumo de Energia

Uma vez que o consumo de energia elétrica tem um impacto considerável em termos sociais e ambientais, o interesse da comunidade científica em temáticas da indústria energética tem-se intensificado nas últimas duas décadas. Por outro lado, convém referir que a análise eficiente dos dados armazenados e relacionados com esta temática apenas é possível com recurso a computação, nomeadamente através de técnicas de mineração de dados.

Os principais estudos existentes debruçam-se - para além da previsão de energia elétrica - na classificação e caracterização de perfis de consumo (Ramos & Vale, 2008), na qualidade do sistema de distribuição elétrica (Pang & Ding, 2008) (Lin & Wang, 2006) (Bhende, Mishras, & Panigrahi, 2008) e na classificação de falhas na rede (Silva, Souza, & Brito, 2006) (Costa *et al.*, 2006) (Dola & Chowdhury, 2005).

As técnicas mais utilizadas na literatura para lidar com estas problemáticas são as árvores de regressão (Loh, 2011), a regressão linear múltipla (Mendenhall & Sincich, 1996), as redes neuronais (ANN) (Zhang G. , 2000) e as *support vector machines* (SVM) (Cortez & Vapnik, 1995). Uma vez que as últimas duas técnicas têm um papel importante neste estudo, reserva-se para o Capitulo 3 uma explicação mais detalhada dos mesmos.

A título de exemplo, Figueiredo *et al.* (2005) apresentou um modelo capaz de obter um conjunto de classes que representam diversos perfis de consumo energético, recorrendo a um algoritmo de segmentação denominado de *K-Means* (Berkhin, 2002) e que originou 9 *clusters*. Posteriormente, foi construído um modelo de classificação, representando um conjunto de regras que permitiram classificar eficientemente cada consumidor no *cluster* anteriormente encontrado. Por sua vez, Zhang *et al.* (2009) utilizaram modelos de segmentação para analisarem as perturbações nos sistemas elétricos. Com o objetivo de garantir o sucesso das operações inicialmente planeadas para um determinado sistema, os autores reconheceram anomalias no comportamento dos sistemas, especialmente no que diz respeito à sua proteção e performance. Para tal, classificaram o tipo de perturbação bem como o local onde a mesma ocorreu, de modo a obter rápidas respostas.

Na literatura encontram-se também trabalhos de relevo que utilizam as técnicas de mineração de

dados na tentativa de modelar o consumo de energia elétrica de forma eficiente. Recorrendo a regressões lineares múltiplas, Ranjan e Jain (1999) analisaram dados relativos a cidade de Delhi nas quatro estações do ano, nomeadamente quantidade de população (ψ), temperatura ambiente (α), humidade relativa (β), precipitação (γ) e insolação (δ). Para evidenciar quais destes atributos tem maior influência no consumo de energia elétrica ao longo das diferentes estações os autores recorreram ao teste estatístico *t-test*, formulando como hipótese nula a ausência de dependência entre consumo e determinado atributo (Tabela 2).

Tabela 2 - Equações obtidas Ranjan e Jain (1999) para as diferentes estações

Estação do ano	Equação Formulada	R^2
Primavera	$-0{,}048 * \delta - 0{,}318 * \beta + 43{,}612$	0,3868
Verão	$6{,}992 * \psi + 0{,}244 * \alpha - 44{,}219$	0,9925
Outono	$6{,}069 * \psi + 0{,}349 * \alpha + 0.005 * \gamma - 40{,}842$	0,9854
Inverno	$5{,}964 * \psi - 0{,}338 * \alpha - 25{,}468$	0,9725

Estes resultados foram validados usando o coeficiente de determinação R^2. O R^2 é um valor que varia entre 0 e 1 e quanto maior for o seu valor, maior será a precisão da previsão. Por exemplo, um R^2 igual a 1 significa que o modelo se ajusta na perfeição ao conjunto de teste utilizado (Steel & Torrie, 1960). Verifica-se exceção para os resultados obtidos para a primavera, cujos autores justificam com as constantes quebras de energia nos meses em questão, onde foram obtidos excelentes indicadores de que os atributos considerados podem ser essenciais para a problemática em questão.

Já Kalogirou e Bojic (2000) adoptaram o uso de redes neuronais para estimar o consumo de energia elétrica num único edifício, uma vez que é uma técnica amplamente aceite e que oferece soluções para lidar com o problema de mapeamentos complexos. Os dados utilizados no seu estudo foram gerados tendo como atributos a estação do ano (verão e inverno), o tipo de isolamento (caracterizando a utilização ou ausência de isolamento térmico em todas as paredes), a espessura das paredes, um atributo responsável por definir se o coeficiente de transferência de energia para o cálculo do consumo de energia elétrica é constante ou variável e, também, a hora do dia.

Após vários testes para encontrar qual a melhor arquitetura para a rede neuronal final, chegou-se a uma arquitetura do tipo recorrente, composta por 4 camadas (Figura 8). Esta é caraterizada

pelas duas camadas obrigatórias (i.e. entrada e saída), por uma *hidden layer* e por uma camada denominada de *long term memory*, responsável por reter informações relevantes para o suporte da rede.

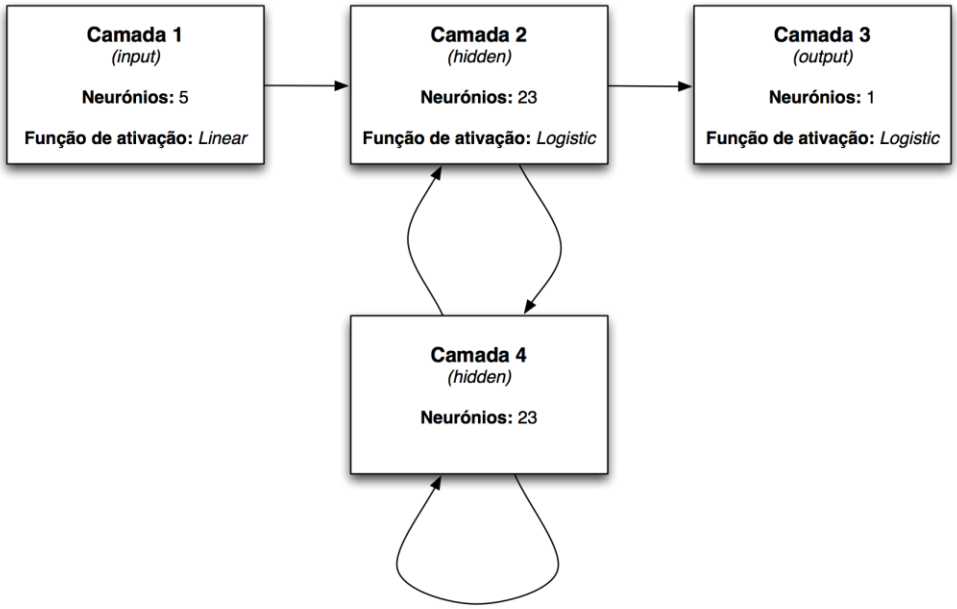


Figura 8 - Arquitetura de redes neuronais usada em Kalogirou e Bojic (2000)

Após se atingir um patamar satisfatório, o processo de treino foi interrompido e um novo conjunto de teste completamente desconhecido foi utilizado para validar a precisão do modelo sendo obtido um R^2 de 0,9991. Se se atentar aos gráficos da Figura 9 é possível perceber que para o caso do inverno a correspondência entre os valores reais e os valores previstos é quase perfeita. Aliás, como se pode verificar, as duas linhas são praticamente indistinguíveis. Já no verão, verifica-se uma ligeira variação.

Em suma, os autores, com este estudo, concluíram que as redes neuronais apresentadas foram capazes de prever o consumo de energia elétrica com uma precisão bastante aceitável, como se pode verificar pelo valor de R^2 encontrado.

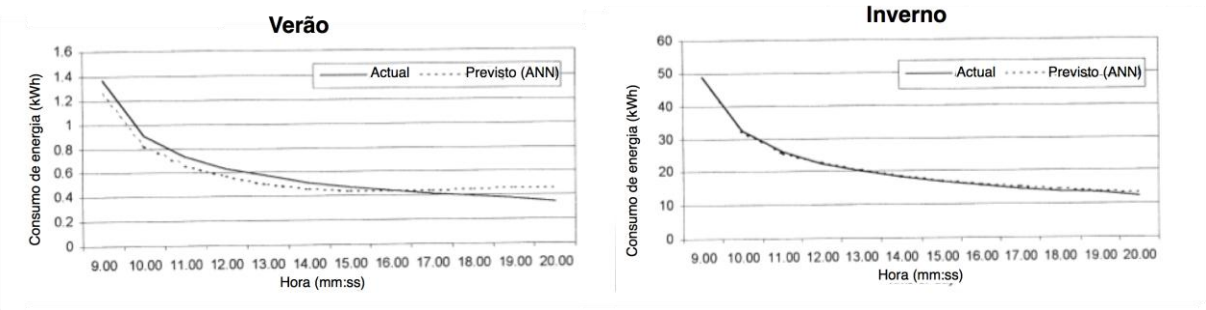


Figura 9 - Comparação entre a previsão e os dados reais - adaptado de Kalogirou e Bojic (2000)

Por sua vez, Tso e Yau (2007) realizaram um comparativo entre diferentes técnicas de previsão de consumo de energia elétrica com base em dados relativos à cidade de Hong Kong. Para tal, foi recolhida informação de cerca de duas mil habitações (e.g. composição do agregado familiar, consumos elétricos semanais de equipamentos e dados gerais do edifício) em dois períodos distintos: verão e inverno. Os métodos selecionados neste estudo foram as regressões lineares múltiplas (Mendenhall & Sincich, 1996), as redes neuronais (Zhang G. , 2000) e árvores de decisão (Breslow & Aha, 1996) que foram implementados através do *SAS Enterprise Miner* (SAS Institute, 2012).

Com as regressões lineares, uma das mais conhecidas técnicas para previsão devido à sua simplicidade e interpretabilidade, obteve-se a estimativa para o valor esperado do consumo elétrico semanal. A equação 1 representa formalmente a regressão linear múltipla, onde γ_i é a variável dependente (i.e. o valor objetivo) e ω_{ij} são as variáveis independentes da *i*-ésima observação, β_j os parâmetros de regressão e ε o erro aleatório segundo uma distribuição normal de média igual a zero e variância constante.

$$\gamma_i = \beta_0 + \beta_1 * \omega_{i1} + \beta_2 * \omega_{i2} + \cdots + \beta_j * \omega_{ij} + \varepsilon \quad \text{(Eq.1)}$$

$$Q = \sum_{i=1}^{n} [\gamma_i - \gamma_{real(i)}]^2 \quad \text{(Eq.2)}$$

Para a estimativa de γ_i foi utilizado o método dos mínimos quadrados que considera os desvio de γ_i em relação ao valor esperado $\gamma_{real(i)}$ (equação 2), sendo o resultado avaliado em função dos *p-values* fornecidos pelas regressões.

Por outro lado, as redes neuronais, introduzidas em detalhe no próximo capítulo, apresentam-se como soluções mais válidas para situações onde não existe linearidade nem grande conhecimento entre a relação das diferentes variáveis (Curram & Mingers, 1994). As principais desvantagens desta técnica estão relacionadas com ausência de *p-values* que nos permitam avaliar os resultados e a necessidade de uma pré-seleção de dados de modo a eliminar informação redundante ou irrelevante. Para esta implementação, os autores utilizaram redes neuronais MLP (*MultiLayer Perceptron*) definidas com base em funções de ativação sigmoidais e de apenas uma *hidden layer*.

O último modelo implementado utiliza as árvores de decisão como algoritmo de classificação, com o *F-test* como critério de partição e um nível de significância de 0.2 para evitar o crescimento desmedido do modelo. Graficamente, as árvores de decisão são representadas por nós que contêm testes unitários a um determinado atributo e ramos descendentes correspondentes a possíveis valores desse mesmo atributo (Figura 10). No contexto deste estudo, as folhas representam um nível de consumo energético semanal e o percurso até alcançar as mesmas representam as regras que definem o modelo de classificação.

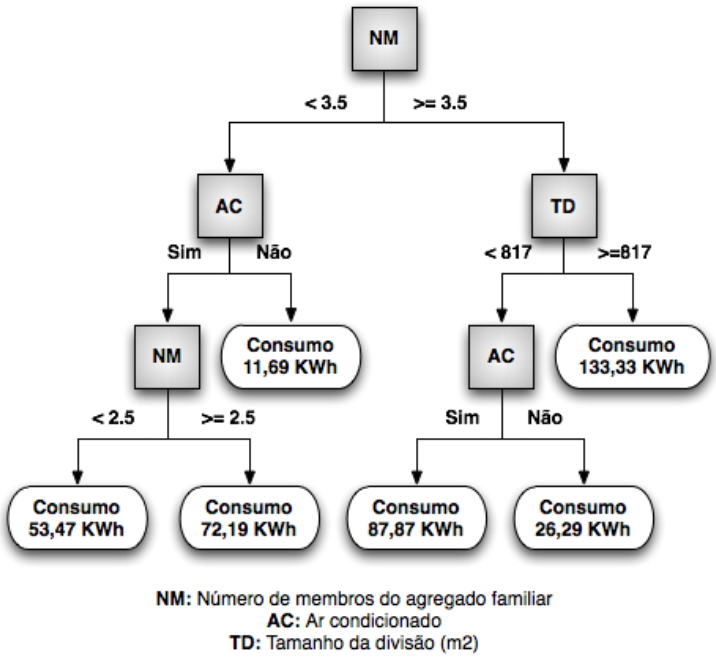


Figura 10 - Modelo de classificação gerado por árvores de decisão para a previsão de consumo de energia elétrica (Tso & Yau, 2007)

As árvores de decisão possuem uma fácil e intuitiva interpretação de resultados, bem como um baixo consumo de recursos quando comparado com as restantes técnicas. No entanto, não se revelam tão eficientes na presença de ruído nos dados ou em atributos numéricos sem padrões suficientemente evidentes (e.g. séries temporais).

Para este estudo os autores concluíram que todas as abordagens tem resultados relevantes, satisfatórios (e.g. *p-values* das regressões com ordens de grandeza de $\cong 10^{-4}$) e muito semelhantes entre si. Nos três métodos utilizados verificou-se que os consumos de energia elétrica variam efetivamente entre estações, sendo os ar condicionados e os equipamentos de aquecimento os principais responsáveis por grande parte do consumo no verão e no inverno, respetivamente. As redes neuronais evidenciaram ainda a tipologia da habitação como uma variável de análise preponderante na previsão de consumos no inverno. É ainda mencionada, embora não comprovada, a melhoria dos modelos com a incorporação de variáveis como a temperatura ou a velocidade do vento.

Com o mesmo objetivo em mente Chen *el al.* (2010) desenvolveram um estudo para previsão do consumo de energia elétrica em casas inteligentes. O conceito de casas inteligentes está intrinsecamente ligado à evolução da domótica, tecnologia esta que permite a gestão automatizada de um edifício com o objetivo de facilitar ações e rentabilizar recursos, como anteriormente referido. Tomando por base a monitorização e reconhecimento de atividades através de sensores, esta tecnologia pode ser aplicada, por exemplo, na identificação da localização dos habitantes dentro de uma habitação (Orr & Abowd, 2000), na automatização de processos com base na atividade humana dos habitantes (Mozer, 1998) ou para gerar alertas de consumo de energia anormais (BeAware, 2009). Este estudo decorreu num edifício preparado para o efeito e com dois habitantes, onde todos os espaços estavam equipados com sensores que permitiam a medição de energia consumida, temperaturas e o estado dos equipamentos (e.g. luzes ligadas, eletrodomésticos em uso). Apenas algumas atividades foram alvo de previsão (Tabela 3) apesar de todas possuírem uma duração temporal considerável e picos de consumo evidentes (Figura 11).

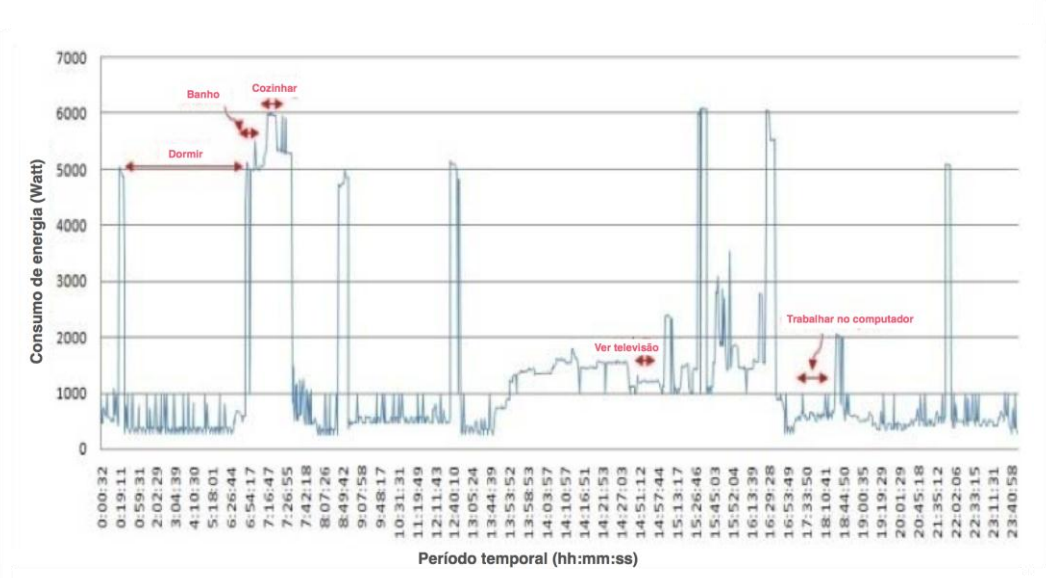


Figura 11 - Consumo de energia elétrica ao longo de um dia - adaptada de (Chen *et al.*, 2010)

Tabela 3 - Atividades consideradas e equipamentos envolvidos (Chen *et al.*, 2010)

Atividade	**Equipamentos diretamente envolvidos**
Trabalhar no computador	Computador e impressora
Dormir	-
Cozinhar	Micro-ondas, fogão e forno
Ver televisão	Televisão ou leitor de DVD
Banho	Esquentador
Aparência	Secador de cabelo

As técnicas utilizadas para previsão foram *Naïve Bayes* (Rish, 2001), redes bayesianas (Pearl, 1988), redes neuronais e SVMs. O algoritmo *Naïve Bayes* assume a independência entre atributos e utiliza o teorema bayesiano para encontrar o valor de classe com maior probabilidade de acerto dado uma instância y. Por seu lado, as redes bayesianas assumem a dependência condicional que normalmente é representada graficamente por um grafo acíclico e orientado. Os detalhes relativos às restantes técnicas serão abordados no decorrer do próximo capítulo.

Para a aprendizagem dos modelos foram considerados dados como as atividades anteriores e seguintes, a duração de cada atividade, o número de sensores ativados e o estado dos equipamentos (i.e. ligado ou desligado). Estes dados foram recolhidos durante os períodos de verão e inverno. O atributo classe, i.e. a quantia de energia necessária para um dia, foi discretizado em n intervalos iguais para todo o $n \in [2..6]$ e validados com *3-fold cross validation*.

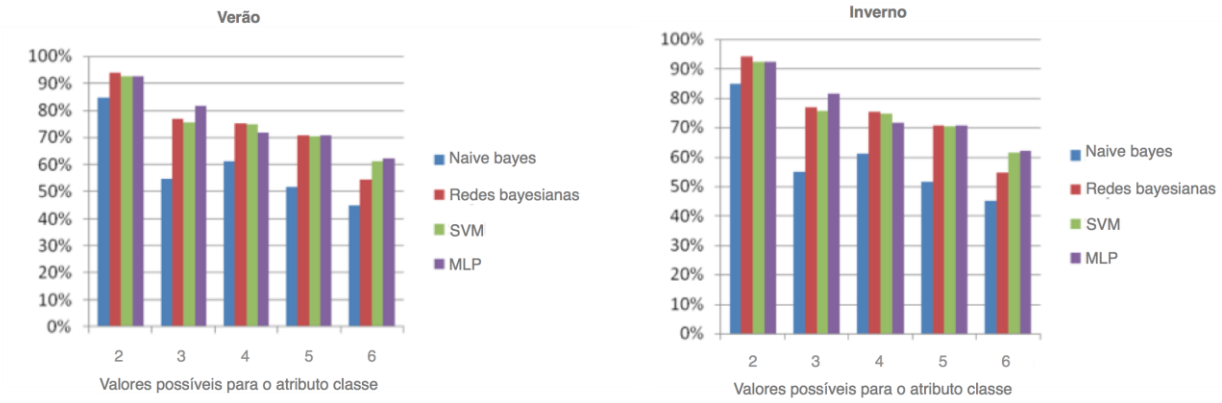


Figura 12 - Comparação da precisão obtida pelos diferentes algoritmos de previsão no estudo de Chen *et al.* (2010)

Os resultados obtidos (Figura 12) demonstram que a precisão dos algoritmos estão enquadrados entre os 60% e os 90%, sendo que quanto maior for o leque de valores possíveis para o atributo classe, menor a precisão dos algoritmos. Já os piores resultados são obtidos pelo algoritmo *Naïve Bayes*, pelo facto de assumirem a independência condicional entre atributos. Os restantes têm resultados semelhantes, apesar dos autores assumirem que são insuficientes para os considerar eficazes perante o contexto em causa.

A principal razão apontada pelos mesmos é a dificuldade em monitorizar e prever consumos de equipamentos como aquecedores, uma vez que depende das temperaturas interiores e exteriores. Por outro lado, a própria atividade humana no interior de uma habitação é um fator imprevisível por si só. A Tabela 4 sintetiza as técnicas usadas pelos diversos autores nos seus trabalhos.

Tabela 4 - Técnicas de Mineração de Dados usadas em Processos de Previsão de Consumo de Energia Elétrica.

Técnicas Usadas	**Trabalhos**
Regressão Linear Múltipla	(Ranjan & Jain, 1999)
Redes Neuronais	(Chen *et al.*, 2010) (Tso & Yau, 2007) (Kalogirou & Bojic, 2000)
Support Vector Machines	(Chen *et al.*, 2010) (Tso & Yau, 2007)
Árvores de Decisão	(Tso & Yau, 2007)
Bayes Net	(Chen *et al.*, 2010)
Naïve Bayes	(Chen *et al.*, 2010)

Capítulo 3

As Técnicas Adotadas

As *support vector machines* (SVM) e as *multilayer perceptron* (MLP) são técnicas de mineração de dados completamente distintas e que apareceram em tempos distintos. No entanto, ambas as técnicas são bastante poderosas e reconhecidas no mundo da mineração de dados, tendo as mesmas provas dadas ao longo da já longa história da mineração de dados. Nas secções seguintes abordar-se-á, com algum pormenor, cada uma das técnicas.

3.1 *Support Vector Machines*

As *support vector machines* (Cortez & Vapnik, 1995) são uma ferramenta de previsão extremamente eficiente para descoberta de conhecimento correspondendo a uma nova abordagem para resolução de problemas de classificação, de regressão, de ranking, etc (Li). As *support vector machines* podem ser definidos como sistemas que usam o espaço de hipóteses de uma função linear no espaço hiperdimensional (Jakkula). Esta técnica de mineração de dados tornou-se famosa quando, usando mapas de pixéis como conjunto de entrada, foi possível obter níveis de precisão comparáveis aos dos sofisticados modelos desenvolvidos sobre redes neuronais (Moore, 2013).

É possível identificar dois tipos distintos de modelos de *support vector machines*: os modelos construídos sobre conjuntos de dados linearmente separáveis e os modelos construídos sobre conjuntos de dados não linearmente separáveis. Estes dois tipos de modelos serão devidamente explorados nas próximas subsecções.

3.1.1 *Support Vector Machines* Lineares

Em relação aos modelos construídos sobre conjuntos de dados linearmente separáveis é possível explicá-los de uma forma bastante simplista. O objetivo da construção do modelo de SVM passa por encontrar um hiperplano que separe os dados de treino de forma a ser possível classificar corretamente todas as novas instâncias. Contudo, existe um número infinito de hiperplanos possíveis que classificam corretamente todos os dados. A figura 13 demonstra, num problema de classificação binária, vários hiperplanos que separam os dados mas apenas um poderá ser o escolhido. Assim sendo, terá de se encontrar aquele que alcance a separação máxima entre os dados, isto é, o plano que consiga uma melhor generalização para classificar os dados futuros. Para encontrar esse hiperplano será necessário seguir um determinada estratégia.

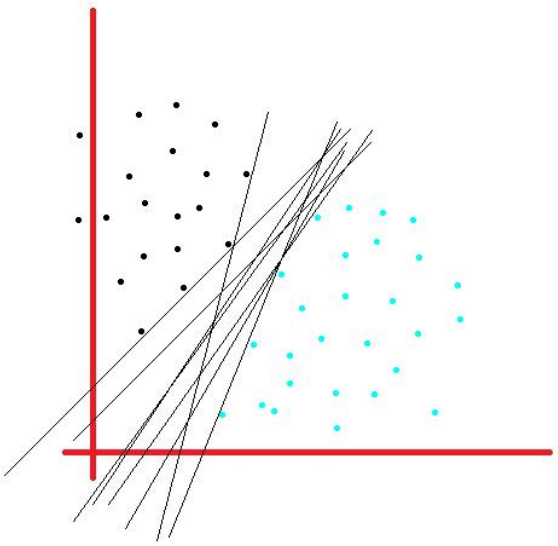


Figura 13 – Hiperplanos de separação – adaptado de Moore (2013)

A figura 14 ilustra a estratégia usada pelas *support vector machines*. Os *support vectors* são os exemplos (instâncias de dados) mais próximos do hiperplano, sendo que o objetivo do algoritmo de *support vector machines* é orientar este hiperplano de maneira a que o mesmo esteja o mais distante possível dos *support vectors* de ambas as classes (Fletcher, 2008). A esta distância é dada a denominação de margem - daí que também seja dada a designação de maximização da margem

ao objetivo pretendido pelo algoritmo de SVM. Mais, mesmo que os dados estejam distribuídos de uma forma aproximadamente linear, isto é, se por exemplo, num dos lados hiperplano existirem casos pontuais de dados que não os da classe predominante, o algoritmo de SVM irá "suavizar" as margens, permitindo assim a presença destes ruídos e *outliers*. Aliás, sendo que é muito pouco provável que em contexto real se trabalhe com conjuntos de dados linearmente separáveis, esta técnica de "suavização" das margens acaba por ser uma mais valia em relação a outras técnicas, como as que irão ser abordadas na secção seguinte.

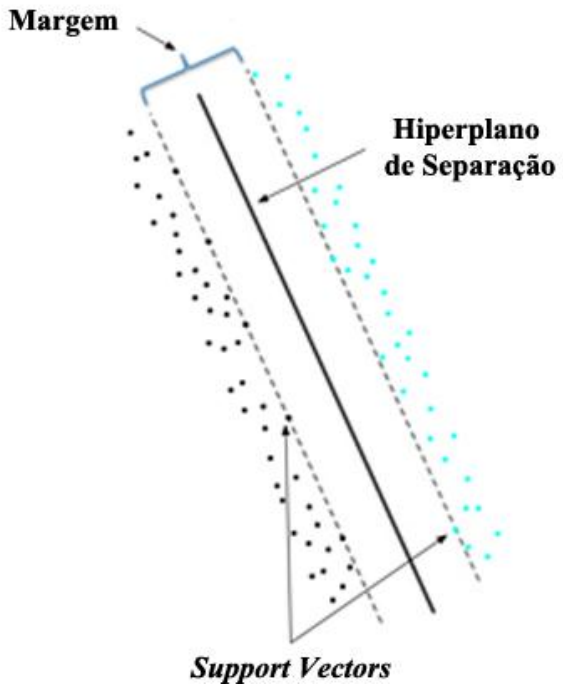


Figura 14 – Funcionamento das *Support Vector Machines* lineares

3.1.2 *Support Vector Machines* Não Lineares

Quando os dados não são linearmente separáveis, as *support vector machines* lidam com este problema, mapeando o conjunto de treino do seu espaço original para um novo espaço de maior dimensão, denominado de espaço de características (*feature space*) (Lorena & Carvalho, 2007). Desta forma, é necessário encontrar uma função F que seja capaz de mapear o conjunto de treino inicial num outro que possa ser representado num espaço de dimensão superior, e a esta função dá-se o nome de função *kernel* (*kernel function*).

Para que se consiga compreender melhor este conceito, considere-se a figura 15a. Dado que o conjunto de dados é não linear será então necessário a aplicação das anteriormente referidas funções de *kernel*. Desta forma, os dados não lineares da figura 15a representados em R^2 (espaço original), após a aplicação da função, passarão a estar representados em R^3 (espaço de características), onde será desta forma possível resolver o problema inicial como se de um normal problema linear se tratasse, pois já será possível encontrar um hiperplano que separe os dados (figura 15b).

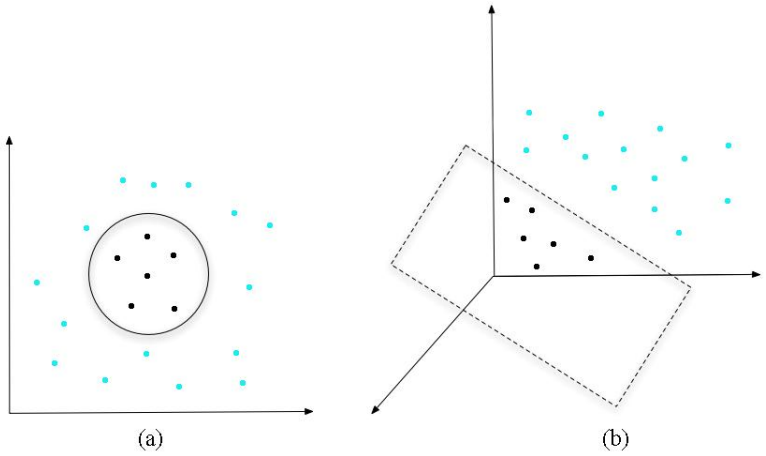


Figura 15 – Separação de dados não lineares no espaço original (a); Separação de dados não lineares no espaço de características (b) – adaptado de Müller et al. (2001)

Entenda-se que a utilização das funções *kernel* tem como principal vantagem a de transformar o problema num simples problema linear, pois os dados conseguirão ser divididos por um hiperplano após a utilização das ditas funções. Este fenómeno é comummente denominado por truque de *kernel* (*kernel trick*).

Existem bastantes funções de *kernel* embora, na prática, as mais utilizadas sejam as funções de *kernel* Polinomiais, as Gaussianas ou RBF (*Radial-Basis Function*) e as Sigmoidais (Lorena & Carvalho, 2007). Observando a tabela 5, é possível visionar quais as funções matemáticas para cada tipo de *kernel*, bem como quais os parâmetros que terão de ser customizados pelo utilizador da função.

Tabela 5 - Funções de *kernel* mais utilizadas

Tipo de Kernel	Função $K(\mathbf{x}_i, \mathbf{x}_j)$	Parâmetros
Polinomial	$(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)^d$	δ, κ e d
Gaussiano	$\exp(-\sigma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$	σ
Sigmoidal	$\tanh(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)$	δ e κ

3.1.3 *Support Vector Regression*

Entretanto, as *support vector machines* são uma técnica de mineração de dados bastante útil para problemas de classificação, cujo objetivo se prende com a previsão do valor de classe para todas as instâncias de um determinado conjunto de teste. Contudo, as *support vector machines* podem também ser aplicados a problemas de regressão, sendo-lhes dada a designação de *support vector regression* (SVR).

Support vector regression é um método usado para encontrar um função que consiga mapear o objeto de entrada num número real baseado no conjunto de treino. Tal como as SVM usados em problemas de classificação, o método SVR têm o mesmo objetivo da maximização da margem e usam, também, o truque de *kernel* para os casos de dados não linearmente separáveis (Yu & Kim, 2010). Qualquer problema de regressão tem uma função de perda (*loss function*) cujo intuito é o de estimar qual o desvio da função encontrada em relação à função real (Farag & Mohamed, 2004), uma vez que a função perfeita pode nunca ser encontrada.

Existem muitas funções de perda – como é o caso da função de perda linear, a quadrática, a exponencial, a ε-insensitive, etc. A título de exemplo, a função de perda ε-insensitive presente na figura 16, que é geralmente utilizada, tem como objetivo encontrar a função que tenha um desvio máximo de ε em relação à diferença entre o valor original (t) e o valor previsto após a aplicação da função g (Farag & Mohamed, 2004).

$$\mathcal{L}(t, g(\mathbf{y})) = \begin{cases} 0 & \text{if } |t - g(\mathbf{y})| \leq \varepsilon \\ |t - g(\mathbf{y})| - \varepsilon & \text{otherwise} \end{cases}$$

Figura 16 - Função de perda (*loss function)* ε-insensitive

Tabela 6 – Aplicações de Support Vector Machines

Aplicações	Trabalhos
Identificação de partículas	(Barabino, Pallavicini, Petrolini, Pontil, & Verri, 1999)
Categorização de texto	(Drucker, Wu, & Vapnik, 1999), (Dumais, Plaat, Sahami, & Heckerman, 1998), (Joachims, 1998)
Deteção facial	(Osuna, Freund, & Girosi, 1997)
Bioinformática	(Brown *et al.*, 2000), (Furey *et al.*, 2000), (Jaakkola, Diekhans, & Haussler, 1999), (Zien et al., 2000), (Mukherjee, et al., 1998)
Database marketing	(Bennet & Bredensteiner, 2000)
Reconhecimento de voz	(Li)
Reconhecimento de palavras	(Rahim, Viard-gaudin, Khalid, & Poisson)
Consumo de energia elétrica	(Chen *et al.*, 2010), (Tso & Yau, 2007)
Solubilidade de fármacos	(Louis, Agrawal, & Khadikar, 2010)

Por outras palavras o algoritmo de regressão não contabiliza os erros desde que não sejam maiores que ε (Farag & Mohamed, 2004).

3.1.4 Aplicações de SVM

Quando falamos de *support vector machines* é importante realçar que, ao longo dos tempos, as mesmas têm vindo a ser aplicadas em inúmeras áreas de conhecimento. Alguns exemplos onde as SVM foram aplicadas com bastante sucesso podem ser vistos na tabela 6.

3.2 Redes Neuronais – *Multilayer Perceptron* (MLP)

Quando mencionámos a redes neuronais artificiais (ANN), referimo-nos a uma técnica de computação não algorítmica caraterizada por sistemas que, de alguma maneira, se assemelham à estrutura do raciocínio de um cérebro humano (de Leon F. de Carvalho, de Pádua Braga, & Ludermir, 1998). As redes neuronais podem ser vistas como grafos orientados e pesados em que os nodos são os neurónios artificiais e os arcos pesados são as ligações entre as saídas de uns neurónios e as entradas dos próximos neurónios. Os pesos das ligações são responsáveis pelo armazenamento do conhecimento armazenado pelas ANN. A solução para um problema de ANN passa inicialmente por uma fase de treino onde os pesos das ligações vão sendo ajustados até perfazerem uma rede neuronal com a capacidade de representar o problema. Após essa fase de treino, os pesos passam a ser fixos e a ANN pode ser utilizada como modelo, permitindo assim estimar os valores de saída para um determinado conjunto de dados de entrada. Posto isto, tendo em conta os padrões das ligações é possível agrupar as redes neuronais em duas categorias (Figura 17):

- redes *feed-forward* (figura 17a), que são caraterizadas por não possuírem *loops* ao longo do grafo.
- redes *feedback* (figura 17b), nas quais os *loops* fazem parte integrante do grafo.

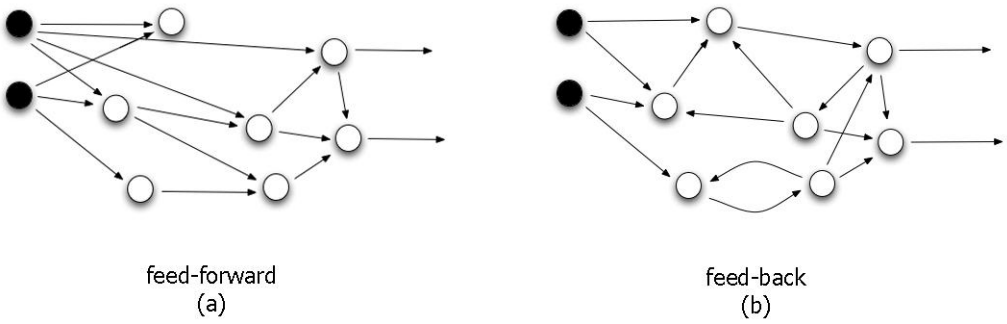


Figura 17 – Tipos de redes neuronais artificiais (ANN)

3.2.1 Redes *Feed-forward*

As redes *feed-forward* são o tipo de rede neuronal que será abordado mais à frente nesta Dissertação. Este tipo de ANNs estão organizadas em camadas (*layers*), podendo ser constituídas por uma ou mais camadas. As redes multicamada são caracterizadas por possuírem um conjunto de dados de entrada, uma camada de saída e uma ou mais camadas intermédias, designadas também por *hidden layers*. O conjunto de entrada não é considerado uma camada da rede pois apenas recebe e passa os dados à camada seguinte, escusando-se a fazer quaisquer cálculos (Wasserman, 1993). As redes neuronais com mais que uma camada são denominadas de *multilayer perceptron* (MLP). A figura 18 mostra uma rede MLP com três entradas, uma *hidden layer* com quatro nodos (preenchidos a cinza) - *hidden nodes* - e uma camada de saída com apenas um nodo, produzindo uma única informação de saída.

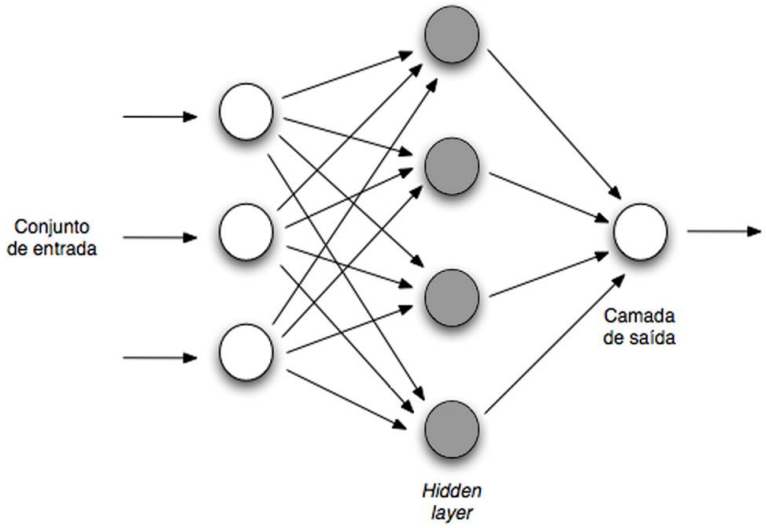


Figura 18 – Exemplo de uma rede MLP com uma *hidden layer*

De entre todas as topologias de ANNs, o modelo em multicamada (MLP) é uma das arquiteturas mais utilizadas nos dias de hoje em diferentes tipos de problemas, incluindo reconhecimento de padrões e interpolação (Noriega, 2005). Segundo Cyberko (1989), uma MLP com apenas uma *hidden layer* pode implementar qualquer função contínua. A utilização de um número suficiente de *hidden layers* e de instâncias de treino consegue aproximar praticamente qualquer função (Han & Kamber, 2006). O número de *hidden layers* e de *hidden nodes* não são determinados à partida pois dependem de problema para problema. Tipicamente, para a maioria das tarefas apenas é

utilizada uma *hidden layer*, sendo que se poderá utilizar um maior número de *hidden layers* para lidar com tarefas mais complexas (Cortez P. , 2012) (Han & Kamber, 2006). Quanto ao número de *hidden nodes* não existe necessariamente um número ótimo, sendo o método de descoberta um assumido processo de tentativa e erro (Witten *et al.*, 2011).

3.2.2 Funcionamento de um nodo (neurónio)

O primeiro neurónio artificial foi definido por McCullogh e Pitts (1988) e caracterizava-se por ser bastante simples, uma vez que à saída apenas era gerado um sinal binário, derivado do somatório das entradas normalizadas que, quando comparado com um determinado limite, produzia o valor zero ou o valor um. A partir desse neurónio foram produzidos outros modelos capazes de produzir um qualquer valor de saída, não necessariamente o valor zero ou o valor um, e com diferentes operações aplicadas ao somatório das entradas normalizadas. Desta forma, cada nodo cinza da rede MLP representada na figura 18 pode ser visto como um neurónio genérico *j* do tipo da figura 19.

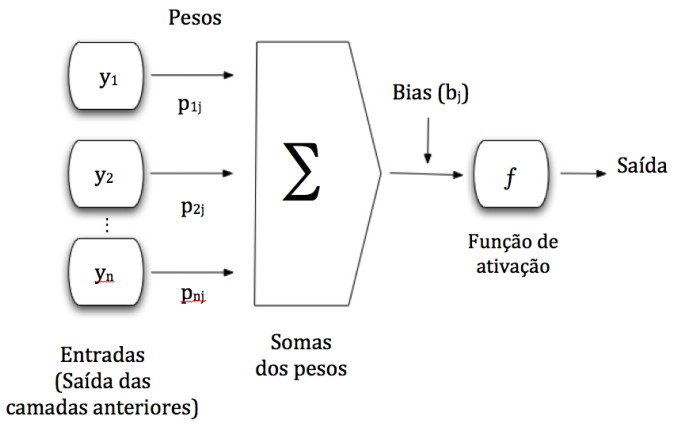


Figura 19 – Modelo de um neurónio de um MLP – baseado em (Han & Kamber, 2006)

Assim, o valor gerado à saída do neurónio corresponde à aplicação de uma função de ativação sobre o somatório de todos os pesos do conjunto de entrada combinado com o *bias* (b_j). O *bias* tem o efeito de variar a atividade do neurónio, funcionando como um limite tal como no neurónio de McCullogh e Pitts (Haykin, 1999).

A função de ativação f define a saída do neurónio de acordo com a atividade produzida pelas suas entradas. Existem várias funções de ativação para além da proposta por McCulloch e Pitts (1988), sendo que as mais conhecidas são a função linear, a sigmoidal ou a gaussiana. No entanto, a função sigmoidal é a função mais utilizada em redes neuronais (Jain, Mao, & Mohiuddin, 1996), devido ao facto de conseguir mapear um vasto domínio de valores de entrada num pequeno intervalo de valores entre zero e um.

3.2.3 Algoritmo *Backpropagation*

Percebido que está o funcionamento de uma rede neuronal artificial (ANN) *feed-forward* é necessário torná-la útil num contexto de mineração de dados - isto é, é preciso aplicar-lhe um algoritmo de aprendizagem para que a ANN seja capaz de resolver problemas de classificação ou de regressão.

Existem vários algoritmos capazes de treinar um rede *multilayer perceptron* (MLP), no entanto, o algoritmo de treino de redes MLP mais difundido e utilizado é o algoritmo *backpropagation* (Rumelhart, Hinton, & Williams, 1986). O *backpropagation* consegue treinar uma rede MLP em duas fases: numa primeira fase ocorre a propagação das entradas ao longo da rede MLP (*feed-forward*) e numa segunda fase sucede a retropropagação do erro (*backpropagation*). Na primeira fase, a partir das entradas, o sinal propaga-se ao longo da rede mantendo fixos os pesos das ligações, aplicando-se todas as operações ao nível do neurónio (ver secção 3.3.2). Na segunda fase, a saída é comparada com o valor original, gerando um sinal de erro. Este sinal propaga-se no sentido contrário (da saída para a entrada da rede) ajustando os pesos de todas as ligações ao longo da rede de forma a minimizar o erro. Embora não seja garantido, os pesos das ligações eventualmente hão-de convergir e o algoritmo termina (Han & Kamber, 2006).

O algoritmo *backpropagation* pode ser visto mais ao pormenor em Han e Kamber (2006).

3.2.4 Aplicações

Com o decorrer dos anos, as redes MLP têm vindo a ser aplicadas em inúmeras áreas de conhecimento. Alguns exemplos que sustentam esta afirmação podem ser vistos na tabela 7:

Tabela 7 – Aplicações de *Multilayer Perceptron*

Aplicações	**Trabalhos**
Reconhecimento de voz	(Li)
Reconhecimento de imagem	(LeCun, et al., 1989)
Solubilidade de fármacos	(Louis *et al.*, 2010)
Consumo de energia elétrica	(Tso & Yau, 2007), (Kalogirou & Bojic, 2000), (Chen *et al.*, 2010)
Bioinformática	(Cortez, Rocha, & Neves, 2002)

Capítulo 4

O Processo de Previsão

4.1 Metodologia de Trabalho

A concepção, planeamento e desenvolvimento dos trabalhos relacionados com este projeto serão feitos de acordo com a metodologia CRISP (Cross Industry Standard Process for Data Mining (Chapman *et al.*, 1999). A aplicação desta metodologia tem como principal objetivo gerir de forma metódica projetos de mineração de dados com alguma dimensão, permitindo reduzir os custos associados com o projeto, com maior grau de confiança e de usabilidade e com maior rapidez (Chapman *et al.*, 1999). A metodologia CRISP proporciona uma vista sobre o ciclo de vida de um projeto de mineração de dados que consiste basicamente em 6 fases (Chapman *et al.*, 1999):

1) compreensão dos objectivos e dos requisitos do negócio em causa;
2) coleção e compreensão dos dados;
3) transformação e limpeza dos dados;
4) seleção e aplicação das técnicas de mineração de dados;
5) confirmação de que todos os objectivos de negócio foram tido em conta com a devida importância;
6) apresentação do resultado final do modelo.

Após a realização do projeto será realizada uma análise detalhada sobre os resultados obtidos e tiradas as devidas conclusões.

Na metodologia CRISP-DM, o processo de KDD (*Knowledge Discovery in Databases*) é dividido em 6 etapas principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e desenvolvimento.

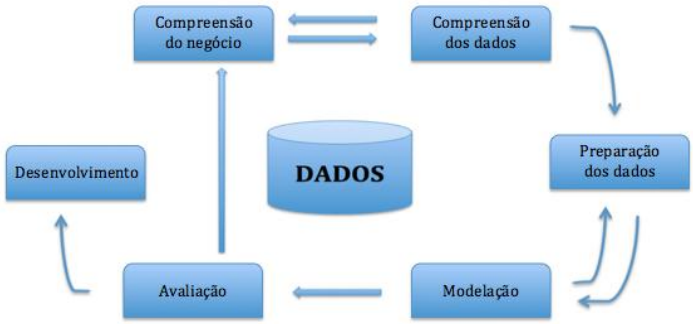


Figura 20 – Metodologia CRISP-DM - baseado em (Chapman *et al.*, 1999)

A figura 20 apresenta o processo de KDD segundo a metodologia CRISP-DM, através da representação das diferentes etapas do processo e das conexões entre elas:

- Compreensão do negócio (*Business Understanding*) – que se foca na compreensão dos objetivos do projeto numa perspetiva de negócio, transformando posteriormente esse conhecimento num problema de mineração de dados e estabelecendo as linhas gerais para a execução do mesmo;
- Compreensão dos dados (*Data Understanding*) – consiste na recolha e exploração dos dados com o intuito de identificar as suas principais características, bem como os seus problemas ao nível da qualidade;
- Preparação dos dados (*Data Preparation*) – nesta fase incluem-se atividades como integração dos dados provenientes de diferentes tabelas, derivação de novos atributos e limpeza dos dados, constituindo assim o conjunto de dados final;
- Modelação (*Modelling*) – consiste na seleção e aplicação de várias técnicas de mineração de dados, tentando parametrizá-las da melhor maneira possível de acordo com o problema e dados em causa;

- Avaliação (*Evaluation*) – trata-se da avaliação dos modelos desenvolvidos, tentando discernir se estes conseguem alcançar todos os objetivos de negócio;
- Desenvolvimento (*Deployment*) – após o desenvolvimento dos modelos de mineração de dados é necessário colocar toda a informação que daí resulta num contexto em que o cliente a possa utilizar.

A execução de cada uma destas etapas é descrita ao longo deste documento. A fase da compreensão do negócio está descrita no capítulo 2, onde é possível compreender os objetivos do negócio e as metodologias que irão ser utilizadas. A compreensão e preparação dos dados é descrita ao longo deste mesmo capítulo. A fase da modelação é explorada ao longo do capítulo 5 bem como a fase de avaliação. A fase de desenvolvimento será a única que não será incluída nesta Dissertação, remetendo-a para possíveis novas análises e trabalhos futuros.

4.2 Os Dados e a sua Preparação

A compreensão e preparação de dados tendo em vista a sua mineração requere dois diferentes tipos de atividades, sendo que primeiro é necessário encontrar e reunir os dados e, de seguida, estes terão de ser manipulados para que se alcance a sua máxima utilidade no que à sua mineração diz respeito (Pyle, 1999). Esta preparação de dados assenta em três passos (Pyle, 1999):

- descoberta dos dados em que se pretende identificar as fontes de dados e os possíveis problemas que possam afetar o seu uso;
- caracterização dos dados que consiste na caraterização dos conjuntos de dados tendo em conta um vasto conjunto de vertentes;
- consolidação do conjunto de dados que tem como objetivo principal encontrar o conjunto de dados final a ser utilizado posteriormente.

4.3 A Descoberta dos Dados

A descoberta dos dados consiste em identificar a(s) fonte(s) dos dados, bem como todos os problemas que possam dificultar o seu uso. Os dados de consumo de energia que servirão posteriormente de suporte a modelos de Data Mining foram encontrados do sítio na internet *US*

Energy Information Administration (EIA, U.S. Energy Information Administration (EIA), 2013), que contém informação de consumo de 1979 até 2009 (inclusive) sendo que os conjuntos de dados só são disponibilizados de 4 em 4 anos. Contudo, apenas será possível aproveitar como fontes de dados para o processo de preparação de dados as de 2005 e de 2001, pois os restantes conjuntos de dados apresentavam problemas - ou lhe faltavam elementos importantes ainda não disponibilizados, como é o caso do consumo de energia elétrica (kWh) no conjunto de 2009 ou estavam já muitos desfasados da realidade, pois de 2001 em diante os equipamentos sofreram uma grande evolução tecnológica e, com isso, o consumo de energia dos mesmos era bastante significativo - algo que não se verificava nos conjuntos de dados anteriores a esse mesmo ano, onde se verificava uma diminuição de equipamentos consumidores de energia elétrica, tornando-se assim *datasets* inúteis, no que aos modelos de Data Mining diz respeito.

Estando desde já identificadas as fontes de dados a utilizar, é necessário identificar e, se possível, resolver todos os problemas que possam dificultar o seu uso. Relativamente a estes problemas que poderão dificultar o uso dos dados, eis alguns que serão levados em linha de conta para análise: legalidade do acesso aos dados, o formato dos dados, a conectividade e as razões de arquitetura (Pyle, 1999). No que concerne à legalidade do acesso aos dados, o sitio na Internet que disponibiliza estes dados é bem claro na sua política de acesso e de uso: "as publicações estão ao serviço do público e não estão sujeitas a qualquer política de proteção de direitos de autor. (…) os dados podem ser usados ou distribuídos através de qualquer suporte existente no sítio (ficheiros, bases de dados, gráficos, etc). Contudo, o seu uso ou distribuição devem ser acompanhados de um reconhecimento da fonte bem como a data de publicação" (EIA, 2013).

Relativamente ao formato dos dados - cuja preocupação se centra na compatibilidade dos formatos dos dados das variadas fontes que serão posteriormente consolidadas (Pyle, 1999) – afortunadamente, as mesmas não constituíram um problema, na medida em que mesmo as fontes estando alojadas em tipos diferentes de ficheiros, o formato dos dados manteve-se intacto. Uma das razões prováveis para este acontecimento será o facto de os dados terem a mesma origem. No que diz respeito à conectividade, esta tem que ver com a disponibilidade futura dos dados, isto é, ao facto de a mesma estar online ou não. Felizmente, isto também não foi um problema pois os dados encontram-se em ficheiros que podem ser descarregados para a máquina que irá desenvolver os modelos de mineração. Poderia, no entanto, representar um problema se os dados se encontrassem em bases de dados externas e daí pudesse advir alguma falta de conectividade.

Por fim, o possível problema de arquitectura - ou seja, e em formato de exemplo, a possibilidade dos dados estarem alojados em diferentes arquiteturas de motores de bases de dados que poderia revelar-se um problema extremamente complicado de resolver devido à dificuldade de traduzir certos formatos de dados (Pyle, 1999). Contudo, como já foi referido anteriormente, as fontes de dados não estão alojadas em bases de dados mas sim em ficheiros, ficheiros esses que, também como já foi reiterado, não possuem diferentes formatos de dados.

Em suma, as fontes de dados utilizados podem ser caracterizadas como externas, uma vez que são dados que não pertencem à organização que pretende fazer mineração de dados. Isto pode muitas vezes representar um problema para as organizações devido ao custo que poderá estar envolvido na aquisição destes dados, tanto a nível monetário como temporal. Antes de se passar à fase seguinte (à caracterização dos dados) foi levada a cabo uma redução do número de atributos nos conjuntos de dados identificados, na medida em que, uma grande parte nada influenciaria no consumo de energia elétrica. Com isto, chegou-se a um número bem mais atrativo de praticamente 160 atributos que contrasta com os mais de 1000 iniciais, sendo que destes 160 serão posteriormente selecionados aqueles que possuírem uma maior significância. Por fim, foi efetuado o carregamento dos dados para uma única tabela, a chamada tabela de factos, concluindo assim a última fase do processo (Figura 21). Todo o processo foi levado a cabo com ajuda da ferramenta open-source Kettle (Pentaho, 2013).

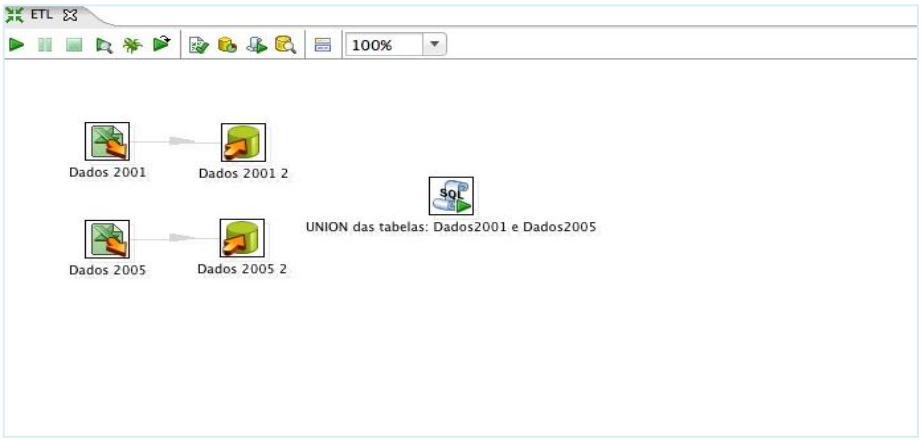


Figura 21 – Fluxo de trabalho do processo de carregamento de dados.

4.4 A Caracterização dos Dados

Nesta fase em que já foram identificadas as fontes de dados a serem usadas, será agora necessário caracterizá-las relativamente a um conjunto de vertentes bastante importantes:

- Nível de detalhe / Nível de agregação – Nas fontes de dados usadas, o nível de detalhe encontrado está ao nível do consumo de energia elétrica por habitação e por ano, pelo que não será possível baixar o nível de detalhe da previsão, pois o nível de detalhe ou de granularidade presente nas fontes de dados irá determinar o nível de detalhe mínimo possível para o output;
- Consistência – Neste ponto, as fontes de dados, tinham algumas inconsistências, pois mesmo que as fontes tenham tido a mesma origem, alguns valores que representavam o mesmo eram diferentes, pelo que os dados tiveram de sofrer uma verificação e consequente alteração para que os conjuntos de dados fossem consistentes aquando da sua junção;
- Poluição – A poluição dos dados tem, normalmente, origem em aplicações que não evoluíram com o percurso natural de determinada atividade. Assim, por vezes, são avistados dados que nada têm a ver com determinado campo, apenas porque não existe outro campo para colocar determinado valor. No caso concreto das fontes de dados que serão utilizadas posteriormente, esse problema simplesmente não se coloca, uma vez que os dados provêm de um estudo e não de uma aplicação real;
- Objetos – A caracterização deste ponto tem que ver com o facto de por vezes existirem nomes parecidos para determinado atributo que podem ser confundidos como sendo o mesmo. No caso concreto das fontes de dados usadas, este é uma questão que também não se coloca, pois as variáveis têm os mesmos nomes;
- Relacionamentos – Este ponto é importante quando consideramos que é necessário fazer junções entre os dados através de determinada chave, tendo todos esses relacionamentos de serem caracterizados. Contudo, esta questão também não se coloca neste caso concreto;
- Domínio – Este ponto é igualmente importante para se verificar se o domínio de valores de todos os atributos está correto. No caso das fontes de dados usadas para esta dissertação, todos os domínios foram verificados e nenhum problema foi encontrado. Mais, as ferramentas de mineração dão uma grande ajuda nestas tarefas, pois apresentam

ferramentas que nos permitem acesso ao domínio de valores de cada um dos atributos, permitindo verificar de uma forma rápida qualquer erro mais grosseiro;

- Valores por omissão – Em caso de falta de valores num conjunto de dados, certos programas de captura podem inserir, nestes locais, valores por omissão. Estes valores podem-se revelar bastantes nocivos para os modelos de mineração. Contudo, nos dados a serem utilizados não se registaram valores em falta, logo não houve necessidade de introduzir qualquer valor por omissão;
- Integridade – A integridade apenas faz sentido caracterizar quando existem relacionamentos entre atributos, o que no caso concreto e como já foi citado em cima, não se verifica;
- Concorrência – A caracterização deste ponto também não será necessário, uma vez que as fontes de dados a serem usadas são estáticas. Assim, não haverá qualquer problema de concorrência, na medida em que qualquer problema de desfasamento temporal na junção dos dados não se coloca;
- Variáveis duplicadas ou redundantes – Pode acontecer existirem variáveis duplicadas ou redundantes nas fontes de dados, logo é necessário estar alerta para esta situação. Um dos problemas será o aumento do tempo de treino, já que a maioria dos modelos são influenciados neste sentido pelo número de variáveis em vez de pelo número de instâncias (Pyle, 1999). No caso das fontes de dados que serão usadas, também não se verificou este problema.

4.5 Consolidação do Conjunto de Dados

Nesta fase de consolidação do conjunto de dados será levado a cabo um processo de seleção de atributos. A seleção de atributos é um processo em que é escolhido um subconjunto do conjunto original de atributos segundo um determinado critério, sendo esta uma importante e usada técnica para reduzir as dimensões de um conjunto de dados (Borges & Nievola, 2005) (Liu, Motoda, Setiono, & Zhao, 2010). Os principais objetivos deste procedimento de redução do número de atributos são a remoção de atributos irrelevantes, redundantes e que introduzem ruído no conjuntos de dados (Liu *et al.*, 2010). Desta forma, garantimos que os dados, quando chegam à fase da sua mineração, são de boa qualidade (Borges & Nievola, 2005) trazendo depois vantagens como, por exemplo, o aumento da velocidade de treino dos variados modelos, o aumento da precisão dos referidos modelos, bem como a uma sua melhor compreensão (Liu *et al.*, 2010).

Os algoritmos usados para a seleção de um subconjunto de atributos podem ser divididos em duas atividades principais: (1) procurar um subconjunto de atributos e (2) avaliar esse subconjunto (Borges & Nievola, 2005). É possível observar o processo na figura 22.

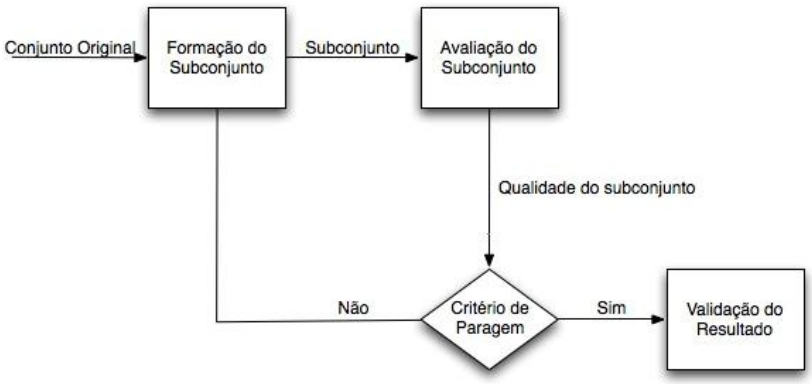


Figura 22 – O processo de seleção de atributos - adaptado de (Liu, Motoda, & Yu, 2003)

Estes algoritmos de procura estão divididos em três grupos principais: exponenciais, aleatórios e sequenciais (Borges & Nievola, 2005). Os algoritmos exponenciais, como é o caso do algoritmo de procura exaustiva, tentam encontrar todas as combinações de atributos possíveis para serem avaliadas posteriormente. Este será, por sua vez, o algoritmo ideal para encontrarmos o melhor subconjunto de atributos. Contudo, podemos considerar que estes algoritmos não simpatizam com a máquina, pois computacionalmente são extremamente pesados, uma vez que o tempo que demoram a correr cresce exponencialmente no número de atributos inicial (Borges & Nievola, 2005). Já os algoritmos de procura genética, um exemplo de algoritmo aleatório, têm como principal vantagem o facto de serem capazes de lidar com o problema da interação entre atributos (Freitas, 2001). Por fim, surgem os algoritmos sequenciais que são relativamente eficientes em problemas em que o número de atributos é elevado. Contudo, têm como desvantagem o facto de não tomarem em linha de consideração a interação entre os mesmos (Borges & Nievola, 2005). Os algoritmos de procura sequenciais mais conhecidos são: *forward selection* e *backward selection*.

Relativamente à avaliação dos subconjuntos, podem ser categorizados em duas abordagens: *filter approach* e *wrapper approach* (Borges & Nievola, 2005). Os primeiros usam as características gerais dos dados para avaliar os atributos, dispensando o uso de qualquer algoritmo de aprendizagem (Borges & Nievola, 2005) (Hall & Holmes, 2003) (Liu *et al.*, 2010). Já os segundos avaliam os atributos usando o valor de precisão obtido pelo uso de um algoritmo de aprendizagem,

previamente escolhido (Borges & Nievola, 2005) (Hall & Holmes, 2003). Ambas têm vantagens e desvantagens, sendo que a abordagem *wrapper* é considerada a melhor, pois tem a tendência para classificar corretamente um maior número de instâncias do que a abordagem *filter*. Contudo, em aplicação no mundo real, a abordagem mais usada é a segunda (Liu *et al.*, 2010), pois a abordagem *wrapper* tem um tempo de processamento do seu algoritmo muitas vezes incomportável (Borges & Nievola, 2005), sendo que num conjunto de dados razoável, em termos de número de atributos e de instâncias, inviabiliza o uso deste método.

No caso concreto do conjunto de dados que será utilizado para construir o modelo de mineração, foram testados vários métodos de procura e de avaliação, de forma a tentar escolher o melhor, tanto em termos de precisão do modelo de mineração como em termos temporais. A escolha dos métodos foi até bastante simples, uma vez que relativamente às abordagens de avaliação, a *wrapper approach* foi completamente incomportável em termos temporais, independentemente do algoritmo de aprendizagem escolhido. Esta abordagem poderá ser óptima quando se está na presença de conjuntos de dados relativamente curtos, em termos de grau e cardinalidade pois, de outra forma, serão necessárias máquinas com elevado poder de processamento. Assim sendo, a abordagem de avaliação escolhida foi a *filter*. Em relação aos algoritmos de procura, foi escolhido um algoritmo sequencial, pois tanto os algoritmos exponenciais como os algoritmos genéticos foram descartados. Os exponenciais devido ao seu tempo de processamento ser intolerável, e os genéticos por retornarem ainda um elevado número de atributos - que não era o objetivo em causa. Desta forma, as hipóteses de algoritmos de procura a considerar eram o *forward selection* e o *backward selection*. Existem opções bastante diversificadas na literatura, pelo que foi necessário testar ambos e desta forma tirar conclusões. O resultado em causa foi que se acabou, mais uma vez, por simplificar o processo de escolha, uma vez que o *backward selection* retornou um conjunto de atributos colossal, pelo que a escolha do *forward selection* se tornou óbvia.

Escolhidas as técnicas a serem usadas neste processo de seleção de atributos, apenas foi necessário decidir mais um pormenor. Aquando da execução do processo, as variáveis escolhidas pelos algoritmos citados anteriormente eram acompanhadas de uma percentagem, que mais não era que uma espécie de valor de confiança. Inicialmente, foram escolhidas todas as variáveis que tinham alguma confiança, ou seja, todas as variáveis cuja confiança fosse superior a 0%, perfazendo um conjunto de 36 atributos. De seguida, foram escolhidos apenas os atributos com uma confiança igual a 100%, reduzindo o conjunto de atributos para 24. Com estes dois conjuntos de atributos foi possível executar alguns algoritmos de aprendizagem e comparar as suas precisões

para decidir qual o conjunto de atributos a seguir para o processo de mineração. A escolha tornou-se evidente uma vez que as precisões eram essencialmente idênticas, pelo que computacionalmente falando era bastante vantajoso seguir com um conjunto de atributos menor. Os atributos escolhidos e as suas respectivas descrições podem ser consultados na Tabela 8.

Tabela 8 - Atributos e suas descrições

NCOMBATH	Número de Casas de Banho Completas
BEDROOMS	Número de Quartos
UGASHERE	Gás Natural Subterrâneo Disponível?
NUMFRIG	Quantos frigoríficos são usados?
ICE	Frigorifico tem congelador interior?
SEPFREEZ	Existe um congelador separado?
NUMFREEZ	Quantos congeladores separados existem?
DRYER	Existe máquina de secar roupa?
WATERBED	Existem aquecedores de colchões de água?
NOWTBDHT	Quantos aquecedores de colchões de água existem?
WTBEDUSE	Quantos aquecedores de colchões de água são usados durantes o ano?
NUMCFAN	Quantas ventoinhas são usadas?
TVCOLOR	Quantas TV são usadas?
WELLPUMP	Motor elétrico para bombear água potável?
AQUARIUM	Aquários aquecidos com mais de 75L?
CELLPHON	Telefone móvel?
COMPCTST	Sistema stereo compacto?
FAX	Fax?
FUELHEAT	Combustível principal usado para aquecer
ACROOMS	Quartos arrefecidos por AC no verão
LGT12	Número de luzes ligadas mais de 12h por dia
ELWARM	Eletricidade é usada para aquecimento?
ELFOOD	Eletricidade é usada para cozinhar?
ELWATER	Eletricidade é usada para a água?
KWH **(classe)**	Eletricidade usada em KWH

Capítulo 5

O Modelo de Mineração de Dados

5.1 Modelação com *Support Vector Machines* (SVM)

5.1.1 Desenho dos Modelos

Tendo desde logo como base os atributos selecionados no capítulo anterior, foram desenvolvidos vários modelos de *Support Vector Machines* de modo a testar algumas combinações no que aos parâmetros de entrada diz respeito, sendo que este tipo de técnica de mineração de dados é, tendencialmente, bastante influenciada pelos mesmos. Antes de ser construído qualquer modelo foi necessário idealizar o conjunto de teste uma vez que modelos poderosos, como é o caso das SVMs, têm propensão para incorrer em *overfitting* (Cortez P. , 2012). O problema de *overfitting* surge quando um modelo memoriza todos os dados do conjunto de treino, incluindo o potencial ruído que possa existir, ao invés de fazer uma generalização através da tendência existente nos dados. A possibilidade de um modelo poder incorrer em *overfitting* surge devido aos diferentes critérios disponíveis para o treino do modelo e para o teste do mesmo. Mais concretamente, os modelos são tipicamente construídos para maximizar a sua performance num determinado conjunto de treino. No entanto, a sua eficácia é determinada não pela sua performance mas pelo melhor ou pior comportamento em contacto com conjuntos de dados desconhecidos. Neste contexto, com um conjunto de treino reduzido não será possível construir um modelo preditivo eficiente, uma vez que não é sustentado nos padrões existentes nos dados. No caso concreto do consumo de energia elétrica, uma previsão feita por um modelo deste género pode ter consequências pesadas em termos económicos e ambientais. Desta forma, a potência adquirida pelos distribuidores de energia elétrica no mercado diário (i.e. no dia anterior à distribuição de energia pelos consumidores finais) será inferior ao necessário, implicando o

recurso a energias não renováveis (e.g. combustíveis fósseis, carvão ou gás natural) no dia de distribuição, para colmatar esse diferencial energético.

Assim, para definir o melhor conjunto de teste, podemos adoptar os seguintes métodos para avaliação dos modelos desenvolvidos:

1) *Holdout*, que consiste em dividir o conjunto de dados inicial de forma aleatória em dois conjuntos independentes: o conjunto de treino e o conjunto de teste. Tipicamente, $2/3$ do conjunto de dados inicial são alocados ao conjunto de treino, ficando o $1/3$ remanescente alocado ao conjunto de teste (Kohavi, 1995) (Han & Kamber, 2006).
2) *k-fold cross-validation,* que divide aleatoriamente o conjunto de dados em *k* partições, $D_1,D_2,\ldots,D_k$, mutualmente exclusivas e de aproximadamente igual tamanho. O modelo é treinado e testado *k* vezes. Em cada iteração i, a partição D_i é alocada para o conjunto de teste sendo que as restantes partições serão usadas, em conjunto, para testar o modelo. Isto é, na primeira iteração as partições $D_2,\ldots,D_k$ são usadas coletivamente como conjunto de treino para obtenção do primeiro modelo, sendo este testado pelos dados existentes na partição D_1. Na interação seguinte o modelo será treinado pelas partições $D_1,D_3,\ldots,D_k$ e testado pela partição D_2 e assim sucessivamente (Kohavi, 1995) (Han & Kamber, 2006). Por fim, a média das estimativas de erro observadas em cada uma das *k* iterações originará a estimativa do erro final (Witten *et al.*, 2011).
3) *Bootstrap* que, dado o conjunto de dados, constrói o conjunto de treino através de amostragens com reposição. Existem várias variantes deste método sendo que a variante mais conhecida é a *0.632 bootstrap*. Este método, para um conjunto de dados com *n* instâncias faz *n* amostragens (com reposição) originando um novo conjunto de dados também com *n* instâncias - o conjunto de treino. Como alguns elementos deste conjunto de treino serão (quase de certeza) repetidos, existirão tuplos do conjunto de dados inicial que não foram escolhidos. O conjunto desses tuplos será usado como conjunto de teste. Este procedimento é repetido várias vezes, sendo que é calculada a média dos resultados obtidos (Efron & Tibshirani , 1993) (Han & Kamber, 2006) (Witten *et al.*, 2011).

Qualquer um dos três métodos enumerados apresentam características distintas, permitindo-lhes obter diferentes resultados para diferentes casos de estudo (Tabela 9). O método *holdout* tem como principal vantagem o facto de ser computacionalmente mais leve em relação aos métodos *k-fold cross-validation* e *bootstrap*. Contudo, tem desvantagens que o tornam igualmente bastante limitado. Tal como foi referido anteriormente, este método divide aleatoriamente o conjunto de dados inicial em dois conjuntos: o conjunto de treino e o

conjunto de teste. Ora para um conjunto de dados relativamente pequeno, retirar-lhe uma parte (tipicamente $1/3$) para testar o modelo é algo desaconselhado pois para um conjunto de dados de pequenas dimensões isso seria retirar, provavelmente, a sua representatividade, dando origem a um modelo pouco credível. Mais, mesmo com um conjunto de dados de grandes dimensões, a divisão desse mesmo conjunto pode não ser a mais indicada, isto é, podemos ter a infelicidade desta divisão originar um conjunto de treino pouco representativo. Por exemplo, num problema de classificação, a divisão do conjunto de dados não consegue garantir que o conjunto de treino originado possua todas as classes com uma proporção aceitável. O método *bootstrap* funciona melhor com conjuntos de dados com uma cardinalidade relativamente reduzida (Han & Kamber, 2006). Uma vez que o caso de estudo abordado neste documento utiliza dados de consumo de energia elétrica relativos a um período de 2 anos (i.e. dados suficientemente representativos), o método *bootstrap* foi então preterido para o nosso estudo. Adicionalmente, este método tem a tendência de ser exageradamente optimista (Han & Kamber, 2006) sendo que no caso concreto da previsão do consumo de energia elétrica é preferível um modelo menos optimista, sendo aconselhável precaução para o pior caso do que o contrário.

O *k-fold cross-validation* tem a vantagem de utilizar o conjunto de dados simultaneamente para treinar e testar o modelo (Cortez P. , 2012). Contudo, este método sofre também de uma desvantagem evidente: o tempo para construir o modelo é bastante maior comparativamente com o *Holdout*, tornando o processo mais moroso. Ponderando todos os fatores acima enumerados, optou-se pelo método *k-fold cross-validation*, uma vez que é o que apresenta a melhor relação entre o tempo, os recursos necessários para execução e a qualidade de previsão dos dados. O número de *folds* (*k*) utilizado será de 10. Justifica-se este número nos numerosos testes que foram efectuados ao longo do tempo, com diferentes tipos de modelos que mostraram que 10 seria o número ideal de *folds* para se obter a melhor estimativa do erro (Witten *et al.*, 2011).

Tabela 9 – Métodos para avaliação dos modelos de Mineração de Dados

Método	Prós	Contras
Holdout	Computacionalmente leve	Separação do conjunto de dados
k-fold Cross-validation	Precisão na estimativa do desempenho	Computacionalmente pesado
Bootstrap	Representa, de certa forma, a vida real	Funciona melhor com conjuntos de dados reduzidos

Tal como indicado no início desta secção, é necessário identificar as combinações de parâmetros das *Support Vector Machines* que melhores resultados garantem na previsão de energia elétrica. O primeiro destes parâmetros, denominado vulgarmente de parâmetro de

complexidade e representado por C, controla o *trade-off* entre a complexidade do modelo e a frequência do erro (Cortez & Vapnik, 1995). O segundo parâmetro, representado por ε e ilustrado na figura 23, está diretamente relacionado com o ruído do conjunto de treino e visa determinar o nível de acerto da função de previsão alcançada pelo modelo (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002).

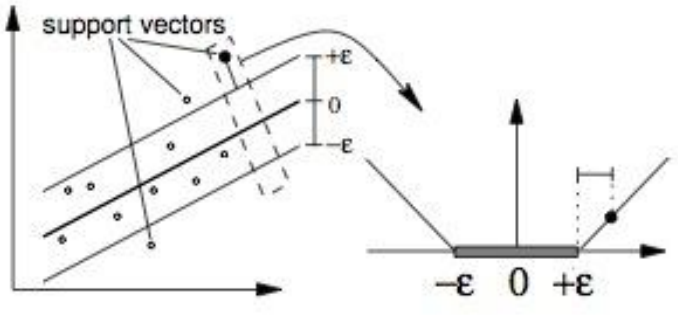


Figura 23 - Representação gráfico da aplicação do parâmetro ε no contexto das SVM

De forma a alcançar os objetivos delineados nesta dissertação, os modelos desenvolvidos recorrem à função de *kernel* polinomial com diferentes combinações dos dois parâmetros acima explanados. Nesse sentido, seguindo as indicações de Hsu, Chang e Lin (2003), foram considerados os valores: $\{0{,}007;\ 0{,}003;\ 0{,}001\}$ e $\{1;\ 0{,}125;\ 0{,}03125\}$ para os parâmetros ε e C, respetivamente.

Os modelos gerados serão objeto de avaliação na secção seguinte.

5.1.2 Avaliação dos Testes

Construídos os modelos de mineração abordados na secção anterior, há que escolher aquele que melhor se adequa ao problema. Para isso, serão analisados os vários critérios acerca dos modelos como o coeficiente de correlação, o erro médio absoluto e o tempo de aprendizagem do modelo. A primeira métrica, o coeficiente de correlação, mede a correlação estatística entre o valor previsto e o valor real sendo que o seu valor varia de 1, onde a correlação é perfeita, passando por 0, onde não existe qualquer correlação até -1, onde a correlação é perfeita negativamente. Obviamente, que para métodos de previsão considerados razoáveis, este valor não deverá ser negativo (Witten *et al.*, 2011). No que tem que ver com o erro médio absoluto, o mesmo é caraterizado por fazer uma média entre todos os erros individuais absolutos sem ter em conta o seu sinal, isto é, a dimensão de cada erro é tratada de igual forma, independentemente da sua magnitude (Witten *et al.*, 2011).

Existem, entre outras, métricas de avaliação de modelos como o erro quadrático médio e a raiz do erro quadrático médio. No entanto, as mesmas não foram consideradas para a análise dos modelos em causa pois conclui-se que estavam altamente correlacionadas com as anteriores ou, alternativamente, não acrescentavam qualquer valor a nossa análise tendo em conta os objetivos traçados inicialmente.

Os resultados dos testes dos vários modelos que consideram estes critérios estão representados nas figuras 24, 25 e 26.

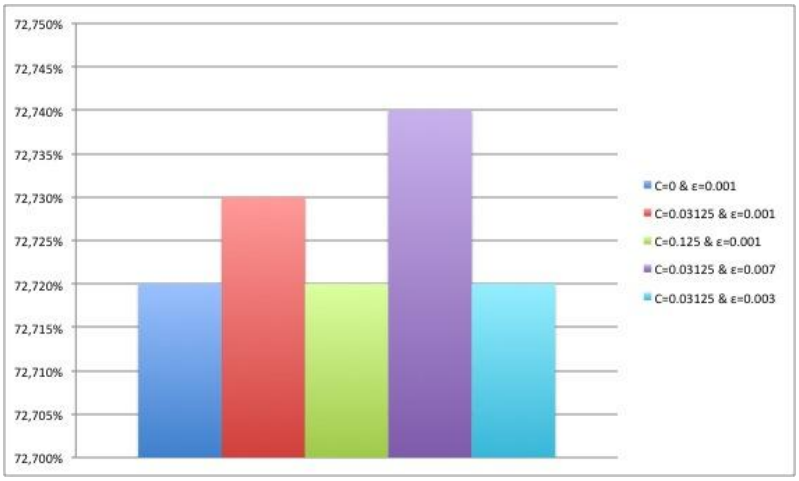


Figura 24 - Coeficiente de correlação dos modelos SVM

Relativamente ao coeficiente de correlação dos modelos, a figura 24 mostra que este valor é idêntico em todos eles. As variações do coeficiente de correlação entre os vários modelos são bastante subtis, na medida em que as diferenças andam na ordem das centésimas. Desta forma, o mais lógico seria escolher o modelo mais preciso, ou seja, o quarto modelo da figura. Todavia, os restantes critérios podem alterar esta escolha pois se o modelo mais preciso tiver, por exemplo, um tempo de treino dez vezes superior, talvez não seja a melhor escolha já que as diferenças entre os coeficientes de correlação são extremamente baixas e pode não compensar o esforço computacional.

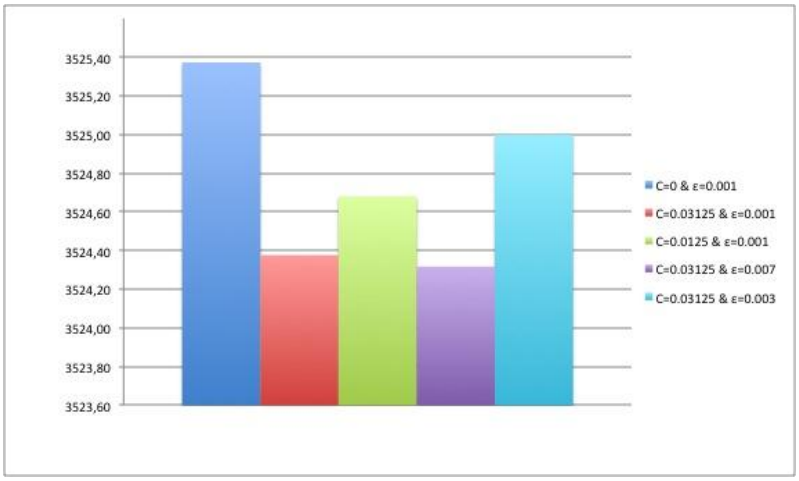


Figure 25 - Erro médio absoluto dos modelos SVM

A análise do erro médio absoluto corrobora a análise feita à figura 24. Como é visível nas figuras e como seria lógico, o erro médio absoluto de um modelo é tanto maior quanto menor for o coeficiente de correlação desse mesmo modelo. Mais, é possível também verificar que a variação do erro médio absoluto é bastante sublime entre todos os modelos, sendo que a variação máxima verificada rondará 1 kWh. Em suma, nesta fase da análise, o quarto modelo representado nas figuras é o que se comporta melhor.

Por fim, é necessário avaliar o tempo de treino dos modelos, tendo em conta que os modelos de SVM são conhecidos por serem robustos mas muito pesados computacionalmente. No caso concreto dos modelos da figura 26, o modelo com o tempo de treino mais baixo é também o modelo mais preciso e com um erro médio absoluto menor. Nesse caso, as dúvidas que poderiam existir inicialmente, no que toca ao modelo a escolher, foram assim extintas.

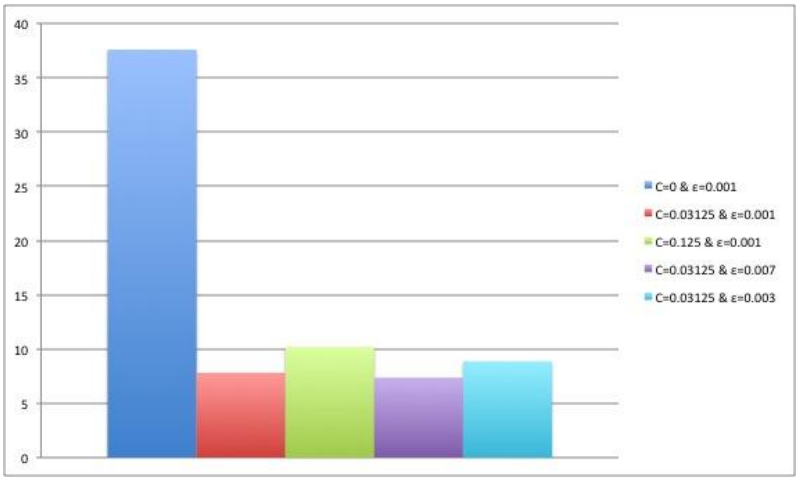


Figura 26 - Tempo de treino dos modelos SVM

5.2 Modelação com Redes Neuronais – *Multilayer Perceptron* (MLP)

5.2.1 Desenho dos Modelos

Antes de se construírem os modelos de redes neuronais para o nosso caso de estudo, foi necessário decidir qual o método a seguir relativamente ao conjunto de teste usado para testar o modelo desejado. Optámos pelo *k-fold cross validation* com 10 partições, tal como na modelação de SVMs, de modo a que a comparação entre as várias técnicas de mineração de dados utilizadas fosse o mais adequada possível. A técnica MLP tem um conjunto de hiperparâmetros que têm que ser definidos – e.g. o número de *hidden layers*, o número de *hidden nodes* ou o *weight decay* –, para que se possa encontrar o melhor modelo possível, tal como referido no Capítulo 3.

Desta forma, relativamente ao número de *hidden layers*, todos os testes realizados recorrerão a uma única camada, pois esta é a abordagem mais utilizada, sendo que apenas para tarefas mais complexas poder-se-á recorrer à utilização de mais camadas (Cortez P. , 2012) (Han & Kamber, 2006). No que concerne o número de *hidden nodes*, não existe um número óptimo, na medida em que é um processo de tentativa e erro (Witten *et al.*, 2011). Assim sendo, foram testados vários modelos com um número de *hidden nodes* diferente com o intuito de verificar aquele que terá melhor comportamento. Por fim, no que diz respeito ao hiperparâmetro *weight decay*, este quando utilizado previne que os pesos dos arcos da rede neuronal cresçam demasiado, sem que tal seja necessário, podendo assim aumentar a generalização da rede (Krogh & Hertz, 1995). Tendo isto em conta, foram também construídos modelos, equivalentes aos anteriores mas utilizando estes hiperparâmetros. Desta forma, será possível avaliar o impacto efetivo desde hiperparâmetro nos modelos construídos. Todos os modelos construídos serão alvo de uma avaliação na secção seguinte.

5.2.2 Avaliação dos Testes

Construídos os modelos de mineração abordados na secção anterior, há que escolher agora aquele que melhor se adequa ao problema. Para isso, à semelhança do que foi realizado com as SVMs, serão analisados os vários critérios acerca dos modelos como o coeficiente de correlação, o erro médio absoluto e o tempo de aprendizagem do modelo. Os resultados dos testes dos vários modelos, tendo em conta estes critérios, estão representados nas próximas figuras.

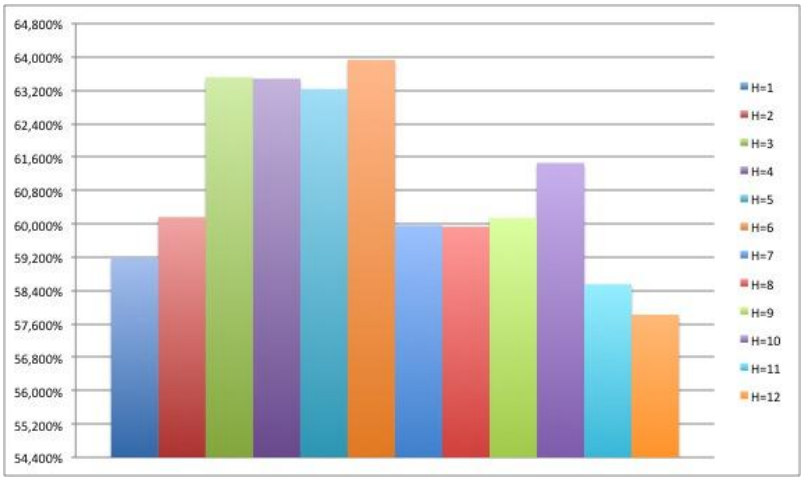


Figura 27 – Coeficiente de correlação dos modelos MLP

Relativamente ao tempo de treino dos modelos, a média andará pelo 70 segundos, pelo que nem sequer será necessário incluir este critério na análise dos modelos construídos não representando um problema.

Em relação ao coeficiente de correlação dos vários modelos, a figura 27 mostra que existem 4 modelos com um coeficiente mais elevado, com ligeira vantagem para o modelo com 6 nodos na *hidden layer* da rede neuronal. Assim sendo, estes modelos que se destacam serão os mais que prováveis para serem escolhidos. Contudo, só depois de uma análise mais profunda com recurso a outros critérios se poderá tomar uma decisão final.

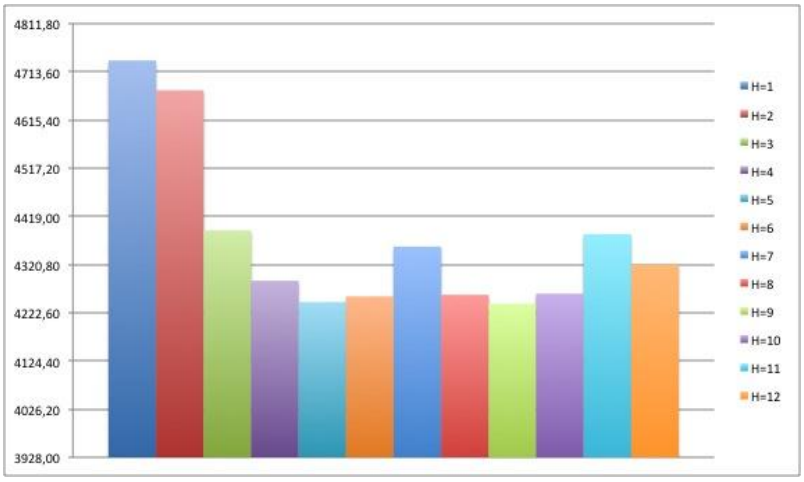


Figura 28 - Erro médio absoluto dos modelos MLP

Analisando agora o erro médio absoluto dos vários modelos, chega-se à conclusão que, para além dos modelos inicialmente selecionados como os prováveis a utilizar, existem outros que surgem como modelos que, mesmo que a sua taxa de acerto seja inferior, quando erram, o

erro registado é menor. Particularizando, o modelo com 9 nodos tem um erro médio absoluto menor que todos os outros. No entanto, a diferença existente entre o modelo com 6 nodos e o de 9 no que ao erro médio absoluto diz respeito é de cerca de 15 kWh, o que será manifestamente insuficiente para justificar ser o modelo escolhido se se tiver em conta que a diferença de precisão entre os dois modelos é ainda bastante significativa.

Quando adicionado o hiperparâmetro *weight decay* (WD) aos modelos, uma considerável melhoria foi verificada na generalidade dos mesmos. Analisando a figura 29, é possível perceber que o coeficiente de correlação de alguns modelos aumentou mais de 10%, o que corresponde a uma melhoria bastante significativa quando comparando com os modelos construídos sem o recurso a este hiperparâmetro. Desta forma, foi possível selecionar outros modelos de redes neuronais, como o caso do modelo que possui 8 nodos na sua *hidden layer* cuja precisão é de 74%, o que corresponde a um brutal aumento de precisão. No entanto, este pode não ser necessariamente o melhor modelo, na medida em que será preciso analisar outros critérios para uma decisão final mais precisa.

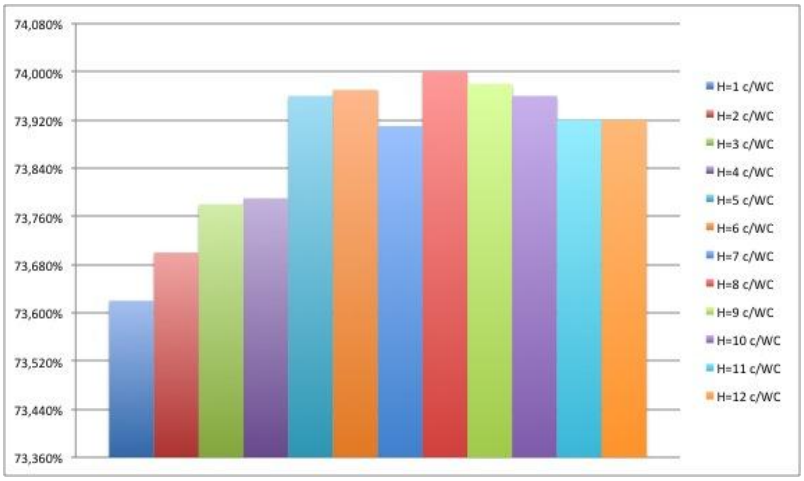


Figura 29 - Coeficiente de correlação dos modelos MLP com WC

Outro critério de análise será o erro médio absoluto dos modelos. Atentando na figura 30, é possível perceber que, ao contrário do que se tinha verificado com os modelos sem *weight decay*, o modelo com maior acerto é também o que possui uma margem de erro mais baixa. Assim sendo, o melhor modelo de redes neuronais para previsão de consumo de energia elétrica encontrado foi o que possui 8 *hidden nodes*.

Ainda relativamente ao número de *hidden layers*, foram construídos alguns modelos com recurso a mais uma camada, apenas para verificar a possibilidade de encontrar um modelo mais capaz. Contudo, nenhuma melhoria significativa foi verificada, fazendo com que a velha máxima de que a simplicidade é o melhor caminho se confirmasse.

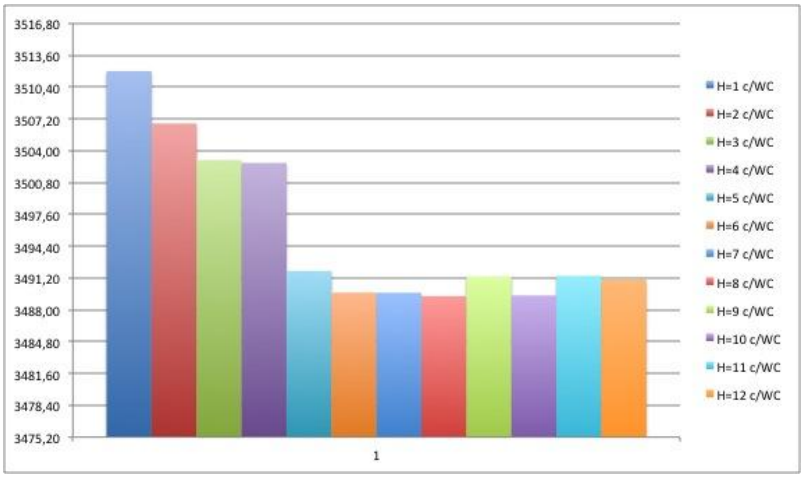


Figura 30 - Erro médio absoluto dos modelos MLP com WC

5.3 Aplicação dos modelos ao caso de estudo

A análise da qualidade dos modelos anteriormente apresentados foi igualmente testada com base na previsão de energia elétrica para o caso de estudo apresentado neste documento. Uma vez que os valores reais de consumo de energia elétrica representam as instâncias de treino dos nossos modelos, foi necessário discretizá-los para uma melhor interpretação dos resultados obtidos (Figura 32). Considerando intervalos de 1000 kWh, é possível verificar que a generalidade das instâncias apresenta consumos entre os 2000 kWh e os 15000 kWh. Assim, assumindo que os dados utilizados representam uma amostragem equilibrada de uma população, podemos apontar este intervalo como sendo o mais frequente e consequentemente o mais crítico.

Os resultados obtidos pelos dois modelos são apresentados na figura 33, onde se analisa o desvio médio (em percentagem) em relação ao consumo real. Apesar de os modelos utilizados apresentarem resultados semelhantes, as SVMs realizaram melhores previsões para baixos-médios consumos, enquanto as ANNs foram superiores na previsão de altos consumos de energia elétrica. Tal como seria expectável, os melhores resultados foram alcançados para os intervalos de consumo onde havia mais instâncias de treino.

Tabela 10 – Desvio médio, em percentagem, para os intervalos de consumo [0,1000[e [1000,2000[

Intervalo	ANN	SVM
[0,1000[	1629,833761	1566,449439
[1000,2000[	182,8049668	137,8111153

No entanto, para facilitar a observação da figura em questão, não foram representados dois intervalos onde ocorreram previsões inesperadas: [0,1000[e [1000,2000[(Tabela 10).

Entretanto, para as 221 instâncias presentes nestes dois intervalos, ambos os modelos tiveram dificuldades para definir padrões, algo que não se sucedeu para outras situações onde, inclusive, havia menos instâncias de teste. Por esta razão, é previsível que grande parte dessas instâncias se tratem de *outliers*. Pela análise das melhores e piores previsões que resultaram da execução das ANNs (Tabela 11) podemos igualmente concluir que, aparentemente, os consumos inferiores são os mais difíceis de prever enquanto os intermédios-altos obtém melhores desvios médios para o consumo real.

Tabela 11 – As 10 melhores e piores previsões resultantes das ANNs

NCOMBATH	BEDROOMS	UGASHERE	NUMFRIG	ICE	SEPFREEZ	NUMFREEZ	DRYER	WATERBED	NOWTBDHT	WTBEDUSE	NUMCFAN	TVCOLOR	WELLPUMP	AQUARIUM	CELLPHON	COMPCTST	FAX	FUELHEAT	ACROOMS	LGT12	ELWARM	ELFOOD	ELWATER	KWH Previsto	KWH	Diferencial	Diferencial (%)
2	4	1	1	0	0	0	1	0	0	0	3	4	0	0	1	0	0	5	7	0	1	1	1	21149,58479	21146	3,584789	0,016952563
1	1	0	1	0	1	1	1	0	0	0	1	3	0	0	1	0	0	5	0	0	1	1	1	12584,19011	12581	3,190109	0,025356561
1	1	1	1	0	0	0	0	0	0	0	0	2	0	0	1	1	1	5	0	0	1	1	1	7023,243289	7028	4,756711	0,067682285
2	3	0	2	1	1	1	1	0	0	0	4	6	1	0	1	1	1	5	6	0	1	1	1	26935,68484	26955	19,315159	0,071657054
1	4	0	1	1	1	2	1	0	0	0	1	2	0	0	0	0	0	5	6	0	1	1	1	20809,57677	20828	18,423226	0,088454129
2	3	0	1	1	1	2	1	0	0	0	2	3	1	0	1	1	1	5	7	2	1	1	1	28615,48335	28588	27,483349	0,096135963
1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	5	3	0	1	1	1	8390,021103	8380	10,021103	0,119583568
1	2	0	1	0	1	1	1	0	0	0	2	3	0	0	0	0	0	7	0	0	1	1	1	14173,58071	14154	19,580714	0,138340497
3	3	1	1	1	1	1	1	0	0	0	5	4	0	0	1	0	0	5	5	0	1	1	1	21214,88844	21248	33,111559	0,155833768
												(...)												(...)	(...)	(...)	(...)
1	2	1	1	0	1	1	0	0	0	0	1	1	1	0	0	0	0	7	0	0	1	1	1	10207,66591	2058	8149,66591	395,9993154
1	3	1	1	0	0	0	1	0	0	0	0	4	0	0	1	0	0	5	0	2	1	1	1	14559,32992	2847	11712,32992	411,3919888
1	1	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	3	0	0	1	1	1	7254,192847	1417	5837,192847	411,940215
1	3	1	1	0	0	0	1	0	0	0	0	2	1	0	1	0	0	5	0	0	1	1	1	13362,09279	2603	10759,09279	413,3343368
2	3	0	2	1	0	0	1	0	0	0	1	2	0	0	1	0	0	5	6	0	1	1	1	18920,02979	3681	15239,02979	413,9915725
1	2	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	5	4	1	1	1	1	12348,67192	1890	10458,67192	553,3688849
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	0	0	1	1	1	6882,300911	981	5901,300911	601,5597259
1	1	1	1	0	0	0	1	0	0	0	1	1	0	0	1	1	0	5	3	0	1	1	1	10950,62642	1447	9503,626418	656,7813696
1	2	1	1	0	0	0	0	0	0	0	1	3	0	0	1	1	0	5	2	1	1	1	1	11484,85015	1112	10372,85015	932,8102652
2	3	0	1	1	1	1	1	0	0	0	2	4	1	0	1	1	1	5	6	0	1	1	1	23608,94528	1341	22267,94528	1660,547746

Numa visão mais geral, por observação da figura 31, constata-se ainda que, em ambos os modelos, 25% das melhores previsões para este caso de estudo obtiveram um desvio de $\cong 13\%$. Se aumentarmos essa referência para 50% ou 75%, os desvios são de $\cong 30\%$ e $\cong 50\%$. Mais uma vez, os valores máximos, já introduzidos na tabela anterior, foram propositadamente excluídos desta figura para melhor leitura dos resultados.

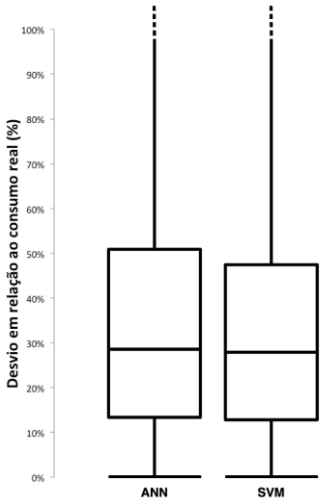


Figura 31 - Quartis para o desvio médio (em percentagem) do consumo previsto

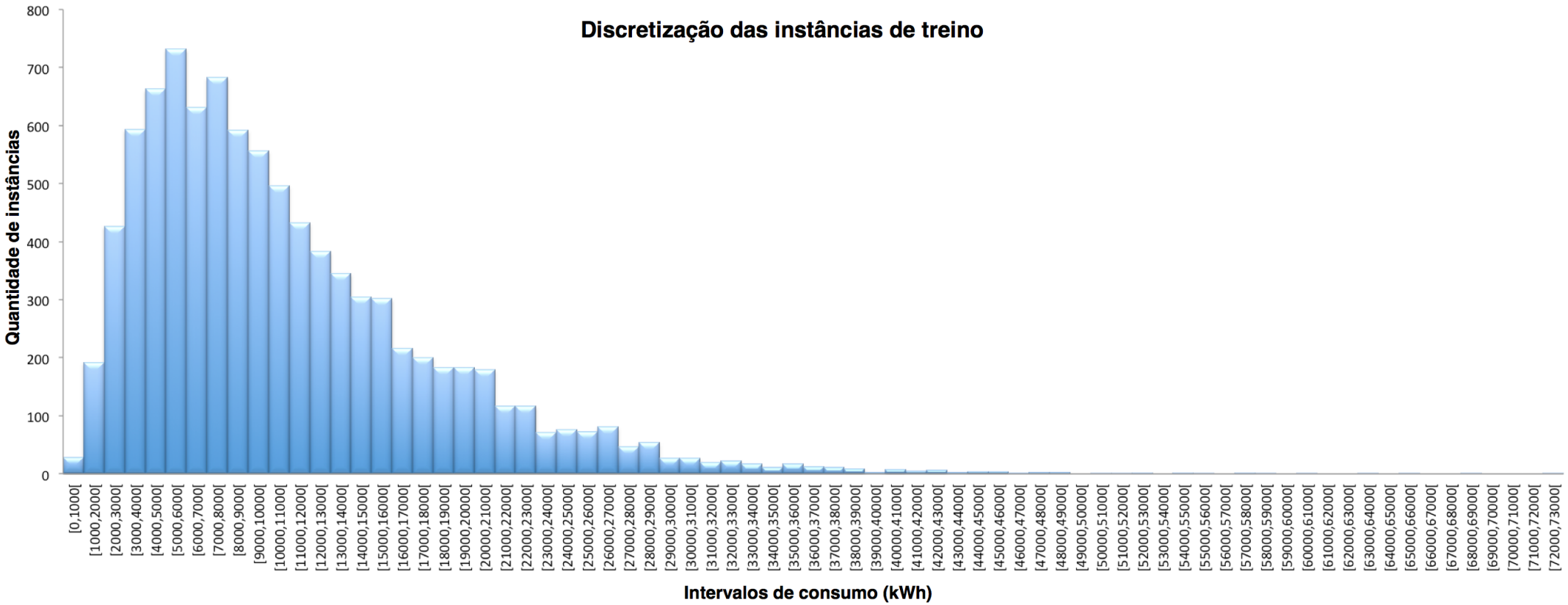


Figura 32 - Discretização das instâncias de treino

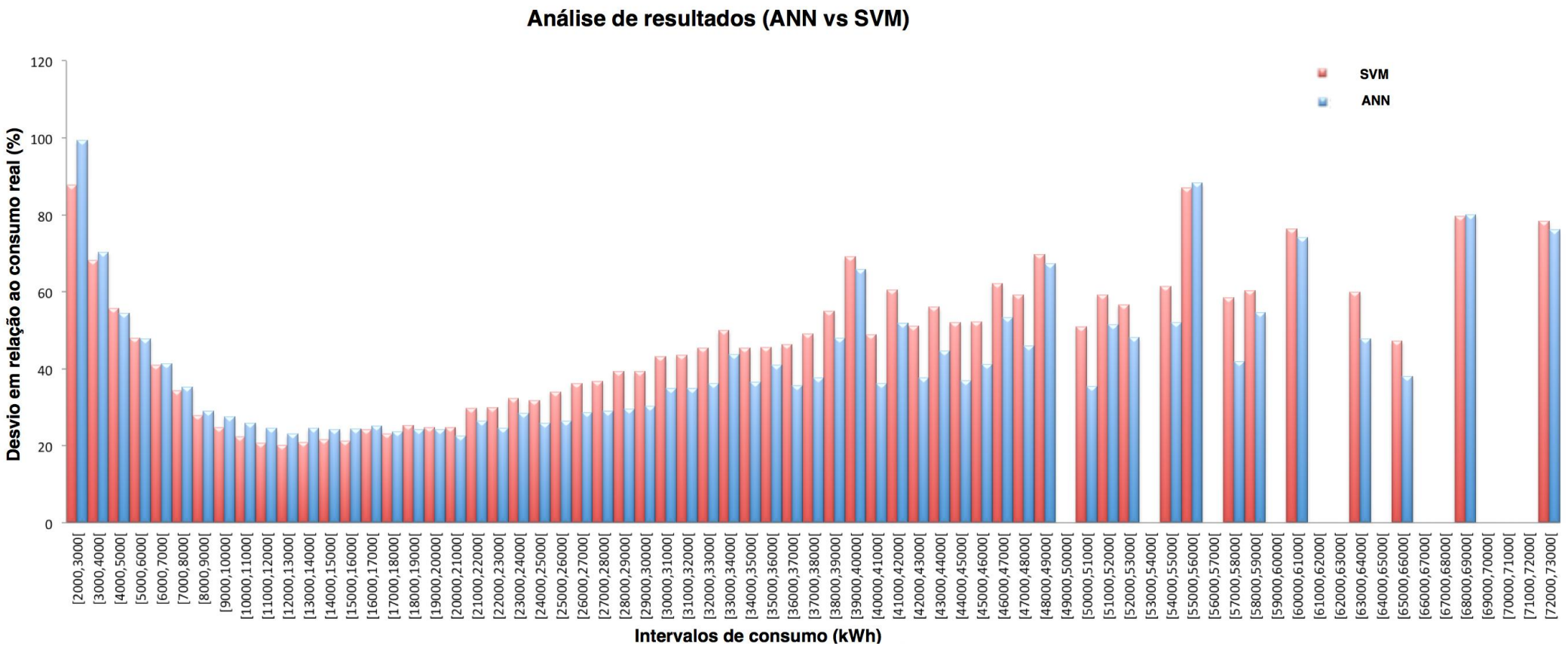


Figura 33 - Análise dos resultados obtidos para os modelos estudados

Capítulo 6

Conclusões e Trabalho Futuro

6.1 Análise crítica dos resultados

Chegada a altura de tecer algumas conclusões, constatamos que, através dos resultados obtidos nos testes efetuados (capítulo 5) as técnicas de mineração de dados utilizadas são capazes de prever, com uma precisão aceitável, o consumo de energia elétrica de uma habitação doméstica. Contudo, apesar dos resultados serem satisfatórios no que diz respeito à precisão do modelo, é ainda possível verificar que mesmo que os modelos preditivos consigam acertar em cerca de 75% dos casos, esses mesmos modelos quando erram, o seu erro médio absoluto é ainda bastante alto, tendo em consideração o tipo de problema que se está a tratar. É necessário ter aqui em conta que estes valores são verificados num contexto de previsão de energia elétrica em habitações particulares.

Há, contudo, algumas explicações que podem ajudar a justificar este mesmo problema. Uma das justificações mais prováveis para este acontecimento é o facto de haver uma insuficiência no que concerne aos dados, tanto em número de atributos pois na realidade o conjunto de dados utilizado não é muito rico nesta matéria, como também em termos de cardinalidade, uma vez que, à partida, se os dados forem de boa qualidade, quanto maior o conjunto, melhor serão os resultados. Posto isto, com um conjunto de dados mais rico nos termos atrás apresentados, seria provável um aumento da generalização dos modelos e, por consequência, um aumento da capacidade preditiva dos mesmos. Contudo, o problema de performance dos modelos pode ter outra explicação. Se o conjunto de dados

utilizado contiver um número substancial de registos considerados pouco ou nada informativos (registos esses que são também apelidados de *outliers*), os modelos podem ficar comprometidos a montante no que à sua performance diz respeito. Pois é necessário compreender que, mesmo que as técnicas de mineração de dados utilizadas nesta dissertação sejam técnicas poderosas, bastante avançadas e com tendência para lidar relativamente bem com o conhecido problema da presença de *outliers* nos conjuntos de dados de treino, não existem técnicas 100% eficazes. Posto isto, é perfeitamente viável que os problemas de precisão e de erro médio dos modelos desenvolvidos tenham proveniência neste possíveis dados enganadores.

6.2 Comparação entre os métodos

Uma comparação entre as diferentes técnicas de mineração de dados usadas ao longo desta dissertação, salientando os seus pontos fortes e fracos, é fundamental para se decidir qual a técnica mais adequada para resolver o problema da previsão de energia elétrica. Para que se consiga comparar estas técnicas, é necessário definir um conjunto de critérios a partir dos quais seja possível tirar uma conclusão fundamentada. Desta forma, os critérios usados para comparar as duas técnicas usadas neste trabalho foram: a precisão do modelo, o erro médio absoluto do modelo, o tempo de treino e a interpretabilidade do modelo.

Em relação à precisão dos modelos construídos neste trabalho, uma comparação e análise do critério de precisão são bastante importantes, na medida em que fornecem uma ideia geral da qualidade dos modelos. Já nos modelos construídos com recurso às *Support Vector Machines* e aos *Multi-layer Perceptron* foi possível observar precisões bastantes aceitáveis, que variam entre os 72,72% e 72,74% nas SVM e entre os 73,60% e 74,00% nas MLP. Como é evidente, em todos os casos foi possível observar uma ligeira supremacia dos modelos construídos sobre as redes neuronais, o que contraria os resultados obtidos por Kotsiantis, S.B. (2007) que constata no seu estudo que a precisão é geralmente superior em modelos construídos com recurso às SVM.

O erro médio absoluto dos modelos construídos é um aspecto crucial de análise, pois não sendo possível na maioria das vezes evitar o erro, é necessário determinar se os modelos, quando erram, se esse mesmo erro é sustentável de se suportar ou não. Nos modelos construídos o erro médio absoluto varia entre 3524,3 kWh e 3525,4 kWh para as *Support Vector Machines* e entre 3490,0 kWh e 3512,5 kWh para as *Multi-layer Perceptron*. Como é possível observar pelos resultados, ambos os métodos

apresentam diferenças pouco significativas, tendo em conta o tipo de problema que estamos a tratar. No entanto, as *Multi-layer Perceptron* conseguem, novamente, alcançar melhores resultados.

Relativamente ao tempo de treino, este é um aspecto fundamental para o sucesso do modelo. Se o tempo de treino for curto, torna-se um aspeto insignificante para a construção, tanto do primeiro modelo como das futuras atualizações do mesmo, isto é, quando existirem mais dados que possam ser adicionados ao modelo para tentar tornar o modelo mais eficaz na sua tarefa. Contudo, quando o tempo de treino do modelo é elevado, este deixa de ser tornar um aspeto insignificante para se tornar num problema a ter em consideração. No caso dos modelos construídos os tempos variam entre os 7 e os 37 minutos no caso das *Support Vector Machines*, já no caso das redes neuronais a média entre todos os modelos construídos rondará os 70 segundos.

A interpretabilidade dos modelos também é um aspeto bastante importante a considerar aqui, pois é esta que permite ter a noção de onde poderão estar os problemas. Por outras palavras e dando um caso concreto, se uma previsão do consumo de energia elétrica for considerada estranha em determinado contexto, é de esperar que num modelo com uma boa interpretabilidade se consiga extrair informação relevante para tentar justificar tal facto. Por exemplo, se estivéssemos na presença de modelos de árvores de decisão, facilmente conseguimos perceber que estas são bastante fáceis de interpretar, pois basta fazer o *backtracking* da árvore para tentar encontrar a justificação para aquele mesmo resultado. Já no caso concreto das *Support Vector Machines* e das *Multi-layer Perceptron* ambas deixam um pouco a desejar neste campo pois, apesar de serem técnicas bastante poderosas, a sua interpretabilidade é bastante fraca para o comum utilizador, sendo necessária alguma perícia e conhecimento para conseguir justificar certos e determinados resultados.

Em suma, de todos os aspetos apresentados, é possível perceber qual a técnica de mineração de dados mais adequada para a previsão de energia elétrica. Com uma precisão de 74% que, embora não possa ser considerada ideal, é bastante aceitável, com um erro médio absoluto de 3490 kWh, erro este que, apesar de ainda bastante elevado, foi o menor valor encontrado em todos os testes e com um tempo de treino insignificante para o contexto que se está a tratar. Isto, atente-se, em contraste com os valores bastante elevados das *Support Vector Machines*, a técnica *Multi-layer Perceptron* revelou-se a mais adequada para satisfazer os objetivos previamente definidos.

6.3 Avaliação e Trabalho futuro

Durante a realização desta dissertação surgiram alguns contratempos que dificultaram a sua realização, de acordo com os trâmites considerados normais neste tipo de trabalhos. O facto dos dados de consumo de energia elétrica, que correspondem à essência do âmbito desta dissertação, se terem revelado difíceis de recolher, foi um problema encarado com alguma preocupação inicial. Com isto, foi-nos possível perceber a dificuldade em arranjar dados em bruto para a realização deste tipo de trabalhos, sendo, contrariamente, fácil de perceber que nestes casos, antes de se decidir avançar em definitivo para o estudo em si, deve-se previamente assegurar que os dados existem e são acedíveis. Caso contrário, reveses e contrariedades poderão acontecer.

Existem ainda variadas situações em que se poderia ter optado por seguir outras direções, que pudessem levar a uma melhoria significativa dos resultados alcançados nesta dissertação. A utilização de modelos de segmentação poderia ser uma mais-valia, pois iria permitir uma parametrização individual para cada um dos modelos dos vários *clusters*, o que poderia resultar numa melhoria significativa de performance para cada um desses modelos. Aliás, seria até possível aplicar diferentes técnicas de mineração de dados a cada *cluster*, o que poderia também significar uma possível melhoria dos resultados. A utilização de outras técnicas de mineração de dados, para além das que foram usadas nesta dissertação, seria uma outra forma de testar se melhores resultados poderiam ser encontrados. Frisa-se aqui que, apesar de terem sido usadas as técnicas teoricamente mais poderosas neste tipo de problemas, não significa que técnicas mais simples não consigam alcançar resultados tão bons ou melhores que os encontrados.

Por fim, um enriquecimento manual do conjunto de dados usado poderia igualmente ser levado a cabo com a finalidade de tentar aumentar a performance dos modelos. Para aumentar o detalhe do referido conjunto de dados poderiam ser acrescentados atributos como a temperatura interior ou exterior de cada edifício ou até a introdução das potências dos aparelhos elétricos utilizados. Desta forma, seria possível tentar aumentar a generalização dos modelos, tendo como consequência imediata o aumento da precisão desses mesmos modelos bem como a diminuição do erro médio absoluto, que representa um *handicap* nos modelos existentes, como foi possível verificar na secção deste mesmo capítulo.

Bibliografia

Barabino, N., Pallavicini, M., Petrolini, A., Pontil, M., & Verri, A. (1999). Support vector machines vs multi-layer perceptrons in particle identification. Proceedings of the European Symposium on Artificial Neural Networks, (pp. 257-262).

BeAware. (2009). BeAware . Obtido em 01 de Janeiro de 2013, de BeAware: http://www.energyawareness.eu/beaware/

Bennet, K., & Bredensteiner, E. (2000). Dyality and Geometru in SVMs. pp. 65-72.

Berkhin, P. (2002). *Survey of clustering data mining techniques.*

Birt, B., & Newsham, G. (2009). Post-occupancy evalutaion of energy and indoor environment quality in green buildings: a revie. *National Research Council Canada* , 1-7.

Bhende, C., Mishras, S., & Panigrahi, B. (2008). Detection and classification of power quality disturbances using S-transform and modular neural network. *Electric Power Systems Research*, *78*, pp. 122-128.

Borges, H. B., & Nievola, J. C. (2005). Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma. *Proceedings of the Fourth International Conference on Machine Learning and Applications* (pp. 201-206). Washington: IEEE Computer Society.

Breslow, L. A., & Aha, D. W. (1996). *Simplifying Decision Trees: A Survey.*

Brown, M., Grundy, W., Lin, N., Cristianini, C., Sugnet, T., & Furey, M. (2000). Knowlede-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences*, *97*, pp. 262-267.

Curram, S., & Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis: an empirical comparison. *Journal of the Operational Research Society , 45* (4), 440-450.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems , 2* (4), 303-314.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearar, C., et al. (1999). *Crisp-dm 1.0 - step-by-step data mining guide.* Obtido em Janeiro de 2013, de http://www.crisp-dm. org/CRISPWP-0800.pdf

Chen, C., Das, B., & J.Cook, D. (2010). Energy Prediction Based on Resident's Activity. *SensorKDD'10.* Washington: ACM.

Costa, F. B., Silva, K. M., Souza, B. A., Dantas, K. M., & Brito, N. S. (2006). A method for fault classification in transmission lines bases on ANN and wavelet coefficients energy. *International Joint Conference Neural Networks*, (pp. 3700-3705).

Cortez, C., & Vapnik, V. (1995). Support-Vector Networks. In *Machine Learning* (pp. 273-297).

Cortez, P. (2012). Data Mining with Multilayer Perceptrons and Support Vector Machines. In *Data Mining: Foundations and Intelligent Paradigms* (Vol. 24, pp. 9-25). Springer Berlin Heidelberg.

Cortez, P., Rocha, M., & Neves, J. (2002). A lamarckian approach for neural network training. *Neural Processing Letters , 15* (2), 105-116.

EDP. (Janeiro de 2013). Obtido em Janeiro de 2013, de http://www.edp.pt/pt/Pages/homepage.aspx

EDP. (Janeiro de 2013). *Tarifas de Baixa Tensão Normal até 20,7 kVA*. Obtido em Janeiro de 2013, de http://www.edpsu.pt/pt/particulares/tarifasehorarios/BTN/Pages/TarifasBTNate20.7kVA.aspx

Efron, B., & Tibshirani , R. (1993). An Introduction to the Bootstrap. In *Monographs on Statistics and Applied Probability* (6th ed., Vol. 57, p. 436). Chapman & Hall.

EIA. (Janeiro de 2013). Obtido em Janeiro de 2013, de U.S. Energy Information Administration (EIA): http://www.eia.gov/

EIA. (Janeiro de 2013). *About EIA – Policies - U.S*. Obtido em Janeiro de 2013, de Energy Information Administration (EIA): http://www.eia.gov/

ERSE. (Janeiro de 2009). *Perdas na rede de transporte* . Obtido em Janeiro de 2013, de ERSE - Entidade Reguladora dos Serviços Energéticos: http://www.erse.pt/pt/electricidade/actividadesdosector/transporte/Paginas/RNT-Perdas.aspx

Dumais, S., Plaat, J., Sahami, M., & Heckerman, D. (1998). Inductive Learning Algorithms and Representations for Text Categorization. pp. 148-155.

de Leon F. de Carvalho, A. C., de Pádua Braga, A., & Ludermir, T. B. (1998). *Fundamentos de redes neurais artificiais.* DCC/IM, COPPE/Sistemas, NCE-UFRJ.

Dola, H., & Chowdhury, B. (2005). Data mining for distribution system fault classification . *Power Symposium, 2005. Proceedings of the 37th Annual North American*, (pp. 457-462).

Drucker, H., Wu, D., & Vapnik, V. (1999). Support vector machines for SPAM categorization. *IEEE Trans. on Neural Networks , 10*, pp. 1048-1054.

Fung, H. (1998). System and method of computer operating mode control for power consumption reduction.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics , 16*, 906-914.

Farag, A., & Mohamed, R. M. (2004). *Regression Using Support Vector Machines: Basic Foundations.*

Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005). An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques . *IEEE Transactions on Power Systems*, *20*, pp. 596-602.

Fletcher, T. (2008). Support Vector Machines Explained.

Freitas, A. A. (2001). Understanding the Crucial Role of Attribute Interaction in Data Mining. *Artificial Intelligence Review , 16*, 177-199.

Godish, T. (2001). Indoor Environmental Quality. *CRC Press* , 196-197.

Google Public Data Explorer. (Janeiro de 2013). *Electricity consumption per capita*. Obtido em Janeiro de 2013, de World Development Indicators and Global Development Finance: http://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&strail=false&bcs=d&nselm=h&met_y=eg_use_elec_kh_pc&scale_y=lin&ind_y=false&rdim=region&idim=country:PRT&ifdim=region&tstart=-297997200000&tend=1311375600000&hl=en&dl=en&ind=false&icfg&iconSize=0

Google Public Data Explorer. (Janeiro de 2013). *Electricity consumption per capita*. Obtido em Janeiro de 2013, de World Developmet Indicators and Global Development Finance: http://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&strail=false&bcs=d&nselm=h&met_y=eg_use_elec_kh_pc&scale_y=lin&ind_y=false&rdim=region&ifdim=region&tdim=true&tstart=298083600000&tend=1311289200000&hl=en&dl=en&ind=false&q=electricity+consumption+in+world

Google Public Data Explorer. (Janeiro de 2013). *Population*. Obtido em Janeiro de 2013, de World Develpment Indicators and Global Development Finance: http://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&strail=false&bcs=d&nselm=h&met_y=sp_pop_totl&scale_y=lin&ind_y=false&rdim=region&ifdim=region&tdim=true&tstart=297997200000&tend=1311375600000&hl=en&dl=en&ind=false

Haykin, S. S. (1999). *Neural networks: a comprehensive foundation* (2nd ed.). Prentice Hall.

Hall, M. A., & Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions On Knowledge And Data Engineering , 15*, 1437-1447.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques* (2 ed.). San Francisco: Morgan Kaufmann.

Hsu, C., Chang, C., & Lin, C. (2003). *A practical guide to support vector classification.* National Taiwan University, Department of Computer Science and Information Engineering, Taiwan.

Jaakkola, T., Diekhans, M., & Haussler, D. (1999). A Discriminative Framework for Detecting Remote Protein Homologies.

Jain, A., Mao, J., & Mohiuddin, K. (1996). Artificial neural networks: a tutorial. *Computer , 29* (3), 31-44.

Jakkula, V. *Tutorial on Support Vector Machine (SVM).* Washington State University, School of EECS.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features.

Kalogirou, S., & Bojic, M. (2000). Artificial neural networks for the prediction of the energy consumption of passive solar building. *Energy , 25*, 479-491.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. (pp. 1137-1143). Morgan Kaufmann.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica , 31* , 249-268.

Krogh, A., & Hertz, J. A. (1995). A Simple Weight Decay Can Improve Generalization. *Advances in Neural Information Processing Systems 4* (pp. 950-957). Morgan Kaufmann.

Lebanese Economy Forum. (Janeiro de 2009). *Electric power consumption (kWh)*. Obtido em Janeiro de 2013, de Lenabese: http://lebanese-economy-forum.com/wdi-gdf-advanced-data-display/?curve=EG-USE-ELEC-KH

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation , 1* (4).

Liu, H., Motoda, H., & Yu, L. (2003). *The Handbook of Data Mining.* Mahwah: Lawrence Erlbaum Associates.

Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining. *ournal of Machine Learning Research - Proceedings Track*, *10*, pp. 4-13.

Li, J. (s.d.). An Empirical Comparison between SVMs e ANNs for Speech Recognition.

Lin, C., & Wang, C. (2006). Adaptive wavelet networks for power-quality detection and discrimination in a power system. *IEEE Transactions on Power Delivery,*, *21*, pp. 1106-1113.

Louis, B., Agrawal, V., & Khadikar, P. (2010). Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. *European Journal of Medicinal Chemistry , 45*, 4018-4025.

Lourenço, J. (2009). *Eficiência Energética na Iluminação - Sector Residencial.* Philips Lightning.

Loh, W. (2011). Classification and regression trees. *WIREs Data Mining Knowl Discov*, (pp. 14-23).

Lorena, A., & Carvalho, A. (2007). Uma introdução às Support Vector Machines. *RITA , 14* (2), pp. 43-67.

Noriega, L. (2005). *Multilayer Perceptron Tutorial.* Staffordshire University, School of Computing.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J. P., et al. (1998). *Support Vector Machine Classification of Microarray Data.* AI Memo 1677, Massachusetts Institute of Technology.

Müller, K., Mika, S., Rätsch, G., Tsuda, K., & Shölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks , 12* (2), pp. 181-201.

McCulloch, W. S., & Pitts, W. (1988). A logical calculus of the ideas immanent in nervous activity. In *Neurocomputing: foundations of research* (pp. 15-27). Cambridge, MA, USA: MIT Press.

Mendenhall, W., & Sincich, T. (1996). *A Second Course in Statistics: Regression Analysis* (Vol. 5). New Jersey: Prentice Hall.

Mozer, M. C. (1998). The Neural Network House: An Environment hat Adapts to its Inhabitants . *AAAI Spring Symp. Intelligent Environments .*

Moore, A. (Janeiro de 2013). *Statistical Data Mining Tutorials*. Obtido em Janeiro de 2013, de http://www.cs.cmu.edu/~awm

OECD. (2011). *OECD Factbook 2011-2012 - Economic, Environmental and Social Statistics.* Organisation for Economic Co-operation and Development.

Olesen, B., Seppanen, O., & Boerstra, A. (2006). Criteria for the indoor environment for energy performance of buildinggs - a new european standard. *Facilities , 24* (11/12), 445-457.

Osuna, E., Freund, R., & Girosi, F. (1997). *Training Support Vector Machines: an Application to Face Detection.*

Orr, R. J., & Abowd, G. D. (2000). The smart floor: A mechanism for natural user identification and tracking . *Conference on Human Factors in Computing Systems*, (pp. 75-76).

Pyle, D. (1999). *Data Preparation for Data Mining* (1st ed.). Morgan Kaufmann.

Pang, P., & Ding, G. (2008). Power quality detection and discrimination in distributed power system based on wavelet transform . *27th Chinese Control Conference*, (pp. 635-638). China.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference . Morgan Kaufmann.

Pentaho. (Janeiro de 2013). *Pentaho Keetle Project*. Obtido em Janeiro de 2013, de http://kettle.pentaho.com/

Pordata. (Janeiro de 2013). *Agregados privados com os principais equipamentos domésticos*. Obtido em Janeiro de 2013, de Pordata - Base de dados de Portugal Contemporâneo: http://www.pordata.pt/Portugal/Agregados+privados+com+os+principais+equipamentos+domesticos+(percentagem)-191

Pordata. (Janeiro de 2013). *População Residente: Total e por sexo*. Obtido em Janeiro de 2012, de Pordata - Base de dados de Portugal Comtemporâneo: http://www.pordata.pt/Portugal/Populacao+residente+total+e+por+sexo-6

Suykens, J., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least Squares Support Vector Machines.* Singapore: World Scientific.

Samsung. (Janeiro de 2013). *Samsung LED 8000 - UE40D8000 - LED TV*. Obtido em Janeiro de 2013, de Samsung: http://www.samsung.com/pt/consumer/tv-audio-video/television/led-tv/UE40D8000YSXXC/index.idx?pagetype=prd_detail&tab=specification

SAS Institute. (2012). *SAS Enterprise Miner*. Obtido em Janeiro de 2013, de SAS Enterprise Miner: http://www.sas.com/technologies/analytics/datamining/miner/

Seppänen, O., & Fisk, W. (2005). Some quantitive relations between indoor environment quality and work performance or health. *Proceedings of 9th International conference on Indoor Air Quality and Climate.* Beijing.

Seppänen, O., Fisk, W. J., & Lei, Q. H. (2005). Ventilation and performance in office work. *International Journal of Indoor Air Quality and Climate*, *16*, pp. 28-36.

Silva, K. M., Souza, B. A., & Brito, N. S. (2006). Fault detection and classification in transmission lines based wavelet transform and ANN . *IEEE Transaction on Power Delivery* , *21*, pp. 2058-2063.

Soucek, S., Russ G., & Tamarit, C. (2000). The smart kitchen project - an application of fieldbus techonology to domotics.

Standby Power. (Janeiro de 2013). *Data*. Obtido em Janeiro de 2013, de Data: http://standby.lbl.gov/summary-table.html

Standby Power. (Janeiro de 2013). *Define & Measure*. Obtido em Janeiro de 2013, de Define & Measure: http://standby.lbl.gov/measure.html

Steel, R. G., & Torrie, J. H. (1960). *Principles and Procedures of Statistics.* New York: McGraw-Hill.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature , 323*, 533-536.

Rahim, A., Viard-gaudin, A. C., Khalid, M., & Poisson, E. (s.d.). Comparison os Support Vector Machine and Neural Network in Character Level Discriminant Training for Online Word Recognition.

Ranjan, M., & Jain, V. (1999). Modelling of electrical energy consumption in Delhi. *Energy. 24*, pp. 351-361. Exergy, An International Journal.

Ramos, S., & Vale, Z. (2008). Data mining techniques application in power distribution utilities. *Transmission and Distribution Conference and Exposition*, (pp. 1-8). Chicago.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI-01 workshop on "Empirical Methods in AI" .*

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy , 32* (9), 1761-1768.

Trading Economics. (Janeiro de 2010). *Electric Power Consumption (kWh per Capita) in Portugal*. Obtido em Janeiro de 2013, de Electric Power Consumption (kWh per Capita) in Portugal: http://www.tradingeconomics.com/portugal/electric-power-consumption-kwh-per-capita-wb-data.html

Wasserman, P. (1993). *Advanced Methods in Neural Computing* (1 ed.). New York, NY, USA: John Wiley & Sons, Inc.

Witten, I., Frank, E., & Hall, M. (2011). *Data mining: practical machine learning tools and techniques* (3 ed.). Burlington: Morgan Kaufmann.

Yu, H., & Kim, S. (2010). SVM Tutorial - Classification, regression and ranking. In *Handbook of Natural Computing.*

Zhang, G. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* , 451-462.

Zhang, Y., Ma, J., Zhang, J., & Wang, Z. (2009). Applications of Data Mining Theory in Electrical Engineering. *Engineering , 1*, 211-215.

Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., & Müller, K. R. (2000). Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites.