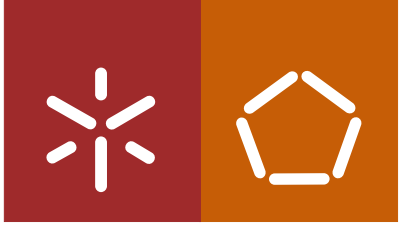
Universidade do Minho
Escola de Engenharia

Hélder Daniel Borges

Exploração de Séries Temporais em Processos de Previsão de Vendas

outubro de 2015

Exploração de Séries Temporais em Processos de Previsão de Vendas
Hélder Daniel Borges
UMinho | 2015



Universidade do Minho
Escola de Engenharia

Hélder Daniel Borges

Exploração de Séries Temporais em Processos de Previsão de Vendas

Dissertação de Mestrado
Mestrado em Engenharia de Sistemas

Trabalho realizado sob orientação do
Professor Orlando Manuel de Oliveira Belo

outubro de 2015

Anexo 3

DECLARAÇÃO

Nome Helder Daniel Borges

Endereço electrónico: helderdanielborges@gmail.com Telefone: 917733769

Número do Bilhete de Identidade:13563929

Título dissertação

Exploração de Series Temporais em Processos de Previsão de Vendas

Orientador: Professor Orlando Manuel de Oliveira Belo Ano de conclusão: 2015

Designação de Mestrado

Mestrado em Engenharia de Sistemas

Universidade do Minho, ___/___/_____

Assinatura: _____

Dedico esta dissertação a todas as pessoas que me acompanharam ao longo deste percurso, principalmente aos meus pais, à minha namorada Patrícia Alves, e aos meus amigos mais próximos que sempre me acompanharam ao longo deste percurso e que nunca me deixaram “cai”

Com o apoio deles aprendi que a razão de cairmos é apenas para aprendermos a levantar-nos. Afinal, a realização de uma dissertação é um caminho longo que nos coloca várias vezes à prova e com a qual aprendemos a superar os mais diversos desafios para ser possível chegar à fase final.

AGRADECIMENTOS

Nenhum grande projeto é realizado apenas por uma pessoa, não sendo este projeto uma exceção.

Assim, gostaria em primeiro lugar de agradecer ao meu orientador Professor Orlando Belo por me ter aceite como orientando e por ter sido um orientador na verdadeira aceção da palavra. Não um professor que impõem filosofias, mas sim um professor que indica um caminho a seguir perante propostas obtidas a partir de investigação. Além de orientador demonstrou sempre boa vontade e boa disposição que sem dúvida me ajudou a superar diversos obstáculos, culminando com o facto de ser um professor de excelência sempre disposto a partilhar conhecimento.

Não sendo este projeto meramente académica, não posso deixar de agradecer à PrimaveraBSS por me ter permitido utilizar parte de trabalho prático lá realizado como base prática para este projeto. Os agradecimentos são em especial para o Lourenço Antunes que, enquanto diretor da Primavera Bss sempre me apoiou e me concedeu a primeira oportunidade de ter uma experiência profissional, e para o António Mendonça, que como primeiro chefe de equipa me ajudou a integrar no ambiente profissional. A ambos um profundo obrigado por me terem ensinado o que são bons exemplos de lideres no mundo empresarial.

Por último mas em nada menos importante quero agradecer à Patrícia Alves por ter sido uma companheira de aventura em que mesmo não sendo da área, sempre se disponibilizou para ler e reler esta dissertação no qual permitiu ter uma opinião de alguém que não da área. A ti, um muito obrigado por estares a meu lado em mais uma etapa da minha vida.

EXPLORAÇÃO DE SÉRIES TEMPORAIS EM PROCESSOS DE PREVISÃO DE VENDAS

RESUMO

Na maior parte das empresas do mundo o seu objetivo final é gerar lucro. Existindo diversas formas de atingir esse objetivo, as vendas de produto, quer sejam físicos ou intelectuais, continuam a ser a maior fonte de rendimento mesmo num mundo em constante mudança. Desta forma, é importante para os agentes de decisão das empresas conseguirem controlar os valores das suas vendas, para assim perceberem a "saúde" da sua empresa e a definirem os próximos passos. Uma das técnicas de suporte à tomada de decisão que os agentes de decisão têm ao seu dispor é a previsão de vendas. As vendas podem ser vistas como uma série temporal em que se agregam os valores e números de itens vendidos num determinado período de tempo. Desta forma, é possível utilizar técnicas de *Data Mining* com capacidade para processar séries temporais de modo a realizar previsão de vendas. Usualmente, para cada empresa, é necessário adaptar os algoritmos de *Data Mining* à sua realidade de forma a ser possível a previsão de vendas, levando assim a processos morosos de adaptação de algoritmos e que acarretam elevados custos. O objetivo deste projeto passa, por utilizar apenas um algoritmo de *Data Mining* que, sem parametrização alguma, consiga obter a melhor previsão de vendas possível para quatro empresas com diferentes características. Assim, serão explorados os resultados que os modelos ARIMA, ARTXP, ARXTP + ARIMA, NN e SVM conseguem obter na realização de previsões tendo em conta as características das empresas. Também será considerada a combinação da utilização de variáveis internas e externas às empresas, juntamente com os seus valores mensais de vendas. No final, são apresentados resultados que permitem identificar o algoritmo de SVM como um algoritmo com boa capacidade de generalização no qual é possível obter resultados com baixa margem de erro na análise global às quatro previsões.

Palavras-chave: *Data Mining*, Previsão de Vendas, Séries Temporais, Variáveis Internas, Variáveis Externas, EPAM, SVM, ARIMA, NN, ARTXP.

TIME SERIES EXPLORATION IN SALES FORECAST PROCESSES

ABSTRACT

For most of the world enterprises, the final goal is profit. With a vast variety of ways to accomplish this goal, no matter if it is a sale of a physical product or an intellectual one, one of the best ways to achieve it is through sales. This way, is important for managers to know how enterprise sales are going for decision making purposes. One of the most common techniques used by decision makers is sales forecast. Sales can be seen as a time series of sales value which happen s at the same instant of time. This let us use Data Mining techniques with capability to process time series to generate sales forecasts. Usually for each enterprise it is needed to create a Data Mining model adapted to those enterprise characteristics, leading to long processes of algorithms adaptation and big resources waste. This thesis focuses on searching for an algorithm with capability to generalize and make good sales forecast over different enterprise without any model adaptation. Different Data Mining models will be explored mainly ARIMA, ERXTP, ARIMA + ARTP, NN and SVM models to find which one can obtain better forecast performance (smaller error value). Will also be considered the combination of internal and external variables along with sales values to help at the forecast process. At the end results will be shown proving that the SVM algorithm have a good generalization capability which allow him to get results with a low error of forecasting at sales forecast for all the enterprises.

Keywords: Data Mining, Sales Forecast, Time Series, Internal Variables, External Variables, MAPE, SVM, ARIMA, NN, ARTXP

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

CAPÍTULO 1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

As novas tecnologias desempenham um papel fulcral na evolução e desenvolvimento de diversas áreas, não sendo exceção o mundo dos negócios. A informatização dos negócios passa a ser assim inevitável, sendo usual as empresas terem praticamente todos os processos de negócio informatizados. Uma das ferramentas normalmente utilizadas pelas empresas são os sistemas de ERP (*Enterprise Resource Planning*) que permitem gerir de forma integrada os diversos processos das empresas. Estas ferramentas têm associado a si um aumento acentuado da quantidade de dados na posse das empresas (Bendoly, 2002) que, aliado ao aparecimento de novos paradigmas como o *cloud computing* e *software as service* (Martson et al., 2009), implica não só o aumento da quantidade de dados mas também uma maior variação das suas fontes. Segundo dados apontados pela Gartner (2012), é possível verificar que todas as tendências tecnológicas apontam para um aumento de necessidade de recurso a sistemas de armazenamento de dados cada vez mais sofisticados. Desta forma, será necessária a evolução dos sistemas de armazenamento de dados convencionais uma vez que as empresas necessitarão cada vez mais de melhores sistemas de armazenamento e tratamento de dados em grande escala (Tay e Shen, 2002). Os sistemas de DW (*Data Warehouse*) são já amplamente usados por empresas que necessitam de integrar dados de diferentes fontes de dados e que necessitam de rápido acesso a informação histórica. Tal como referido por Bose e Mahapatra (2001), com a criação destes sistemas as empresas passaram a possuir uma ferramenta que permite organizar e guardar grandes volumes de dados, para que possam ser facilmente analisados.

Com o aparecimento destes conceitos, um facto tornou-se claro: *"We are data rich, but information poor"* (Han et al., 2012). Embora a informação armazenada esteja hoje em níveis de armazenamento muito elevados, as empresas apenas utilizam a informação já conhecida, desperdiçando deste modo, muita informação que se encontra armazenada mas que as empresas não sabem da sua existência ou como a utilizar. Urge portanto, a necessidade de utilizar mecanismos que permitam às empresas tirar partido

das grandes volumes de dados que possuem, para se poderem posicionar de forma favorável em relação à sua concorrência. Parafraseando Bose e Mahapatra (2001), é necessário identificar formas inovadoras de capturar e aumentar o *share* de mercado enquanto se reduzem custos.

Aliado à necessidade de evoluir a forma como as tecnologias da informação são aplicadas na análise e no armazenamento de dados, surge o aparecimento da área de *Business Inteligence* (BI) (Connolly e Begg, 2010). Os sistemas de DW são apenas um dos sistemas integrantes de BI. Para além de sistemas de armazenamento de grande volume de dados e de visualização dos mesmos, o BI é também caracterizado por fornecer um conjunto de soluções que permitam identificar padrões nos dados, ou seja, sistemas capazes de retirar informação desconhecida dos dados conhecidos. As ferramentas que permitem essa extração de conhecimento são apresentadas por Fayyad como *Knowledge Discovery* in *Database* (KDD), sendo *Data Mining* (DM) o processo de extração de conhecimento (Fayyad et al., 1996ª).

Existem métodos e algoritmos muito distintos em DM. Enquanto alguns surgem com a evolução das tecnologias, outros aparecem com a junção de métodos já conhecidos. As técnicas de DM são amplamente usadas em áreas como a saúde (Hu et al., 2010), a segurança (Damle e Yalcin, 2006), a genética (Ponomarenko et al., 2002), entre outras, nas quais alguns dos métodos mais utilizados são as *Decision Trees, Rules, Classification*, *Clustering*, *Support Vectore Machines*, *Neural Networks* e *Genetic Algorithm*.

Como referido anteriormente, as empresas procuram constantemente meios que permitam adquirir vantagem competitiva sobre a sua concorrência. De forma indireta, um bom sistema de suporte à decisão irá colocar os gestores de uma empresa em vantagem em relação aos seus concorrentes, pois permitirá realizar tomadas de decisão mais céleres e precisas. Dentro das ferramentas que auxiliam a tomada de decisões, as previsões são um conjunto de ferramentas bastante importante para os agentes de decisão empresarial. A potencialidade das previsões é patente no artigo de Tsaih et al. (1998). Segundo os autores, com base numa boa previsão os investidores podem, com uma pequena quantia de capital, retirar um lucro muito elevado de um determinado investimento.

Embora as previsões possam ser usadas para diversos fins, um dos seus objetivos, passa por prever valores futuros de vendas. Saber qual o lucro expectável num período

futuro é algo bastante aliciante, porém, modelar sistemas de previsões de vendas não é algo simples. Vivendo num mundo tão complexo como o atual, as vendas são influenciadas por muitos fatores, tanto internos como externos. Desta forma, é necessário considerar modelos que atentem não só ao valor histórico de vendas, mas também ao valor das variáveis que rodeiam o negócio. Outro fator característico das vendas é a necessidade de enquadramento temporal. Este facto transforma estes problemas em problemas de séries temporais, pois a sua ordem no tempo é relevante. As séries temporais são sequências de dados ordenados cronologicamente, nos quais um determinado valor é em termos temporais, posterior ao seu antecedente e anterior ao seu sucessor na série (FU, 2011). O facto de as séries temporais se caracterizarem por serem facilmente interpretáveis torna-as muito atrativas para serem utilizadas em meios empresariais, visto que qualquer gestor poderá facilmente concluir elações sobre o que estas representam (Agon e Esling, 2012). Considerando que as vendas se encontram representadas sob a forma de séries temporais, surge a hipótese de aplicar algoritmos de DM para poder retirar informação dos dados sob a forma de previsão. Entre estes algoritmos perfilam-se alguns como *Support Vector Machines* (SVM), *Neural Networks* (NN) e *Autoregressive Trees* (ART), suscitando interesse no estudo destes algoritmos por forma a perceber como podem ser utilizados e como se comportam quando aplicados com fins preditivos.

1.2 Motivação e Objetivos

Se um indivíduo for questionado sobre a possibilidade de prever os resultados atirando um dado de seis faces sobre a mesa, a sua resposta será provavelmente negativa. A verdade é que é possível prever qual o resultado de atirar o dado sobre a mesa. Para tal seria necessário conhecer a sua composição, o seu estado inicial, a sua velocidade de lançamento, as condições atmosféricas, os ângulos de incidência sobre a mesa e as características da superfície da mesma. Os cálculos físicos iriam dar a conhecer qual o lado que ficaria voltado para cima. Desta forma, é possível perceber que a capacidade de previsão de um acontecimento depende do conhecimento de todas as variáveis envolvidas, o que torna o estudo de métodos de previsão algo fascinante, uma vez que nem sempre conhecemos todas as variàveis. Considerando a abundância de dados que existem no meio empresarial, é essencial transformar os dados em conhecimento para que os agentes de decisão sejam munidos de bases sólidas de informação para maximizar a

qualidade dos seus processos de tomada de decisão. A utilização de métodos de previsão que recorram a determinadas variáveis de negócio pode trazer aos agentes de decisão a informação que estes necessitam para tomar uma determinada decisão. A possibilidade de munir os agentes de decisão de tal informação, torna o desafio de desenvolver métodos de previsão de vendas, num desafio bastante aliciante, pois caso este desafio seja ultrapassado um resultado positivo irá ter impacto em diversas empresas. Para além das motivações aplicacionais, existiu uma grande motivação académica para o desenvolvimento deste projeto, pois o desenvolvimento do mesmo permite aquisição e evolução de conhecimentos nas áreas de DM bem como o desenvolvimento de conhecimentos de BI de forma abrangente.

De forma genérica os produtos disponibilizados nos mercados de bens e serviços têm como foco ou grandes massas de consumidores ou nichos específicos de mercado. Tipicamente os bens e serviços orientados a massas possuem custos de produção menores e consequentemente o seu valor de aquisição é menor. Por sua vez os produtos orientados a nichos de mercado, pelo facto de serem artigos especializados, possuem um valor acrescentado superior, aumentando o seu valor final de aquisição. Os serviços de previsão de vendas encontram-se usualmente neste grupo, uma vez que na sua maioria são desenvolvidos à medida do cliente, tornando-se serviços altamente especializados.

Desta forma, este projeto teve como objetivo a implementação de algoritmos de previsão que procuram o oposto, identificar métodos que permitam ser aplicados a diferentes empresas para que o custo de posse para o consumidor final seja o menor possível, permitindo assim que este tipo de métodos esteja disponível a um maior número de empresas. Este objetivo apenas pode ser atingido comparando o desempenho de diversos algoritmos em relação a um conjunto de variáveis transversais. Assim torna-se importante verificar qual o resultado que cada algoritmo obtém para diferente combinação de variáveis transversais. Esta combinação de variáveis implicou analisar o desempenho de cada algoritmo perante a utilização de variáveis internas e externas às empresas bem como a utilização do histórico de vendas. A avaliação de desempenho de cada algoritmo perante as diversas combinações de variáveis será realizada com base no menor erro absoluto médio uma vez que no final será comparado para cada algoritmo os resultados dos últimos doze meses de dados conhecidos.

Durante o trabalho prático desenvolvido ao longo deste projeto, este objetivo principal tomou a forma de dois tipos de objetivos, os objetivos de negócio e os objetivos de DM.

O trabalho prático do presente projeto foi elaborado em simultâneo com a realização de um projeto de estágio na empresa Primavera BSS. Nesse projeto o objetivo centrou-se em encontrar um algoritmo que permitisse um bom *trade off* entre custos de implementação e qualidade da previsão. Este objetivo prendeu-se com a intenção de utilizar o mesmo algoritmo para empresas com diferentes tipos de características. Desta forma, do ponto de vista de negócio foi importante verificar se este *trade off* poderia ser ou não obtido perante dados reais de diferentes empresas. Por sua vez, os objetivos de DM prenderam-se com a vontade de desenvolver conhecimento na área de DM. Neste sentido, desenvolveram-se diferentes algoritmos de DM bem como a identificação e tratamento de diferentes tipos de variáveis e a escolha de métodos adequados à comparação de resultados.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Para além do presente capítulo, este documento encontra-se organizado em mais 6 capítulos, nomeadamente:

- **Capítulo 2 – Previsão de Vendas** – Neste capítulo estão apresentados os diversos conceitos utilizados ao longo deste projeto. Numa primeira parte são dadas a conhecer as características das vendas para ser possível compreender melhor a sua natureza. De seguida, o tema "previsão" é apresentado e discutido para que seja possível através da sua história e características compreender melhor qual o tipo de técnicas de previsão a utilizar. As séries temporais e a utilização de DM na previsão de vendas sob series temporais são as secções que terminam este capítulo, apresentando informação para identificar as características das séries temporais e a forma como as técnicas de DM podem ser aplicadas sobre as mesmas.
- **Capítulo 3 – *Data Mining* e Técnicas de Previsão de Vendas** –É abordado o DM de forma geral, discutindo alguns dos seus aspetos mais pertinentes tais como metodologias existentes para desenvolver projetos de DM, quais as técnicas e os algoritmos de DM que podem ser desenvolvidos e utilizados. Por fim,

apresentam-se os algoritmos que serão utilizados na realização dos trabalhos deste projeto.

- **Capítulo 4 – Análise e Preparação dos Dados** –É descrito o processo de análise e de preparação de dados que foi realizado, a análise de negócio levada a cabo, bem como a apresentação de como estudar as quatro empresas utilizadas como casos de estudo. Além disso, apresentam-se as ferramentas de DM, os conjuntos de dados utilizados e a forma como foram preparados para a realização do trabalho de mineração pretendido.
- **Capítulo 5 – Modelação dos Modelos de Previsão** – A modelação dos modelos de previsão é descrita neste capítulo em que é apresentado o processo de modelação dos algoritmos de mineração, a avaliação dos modelos, as métricas de desempenho definidas e a seleção de atributos que foi realizada, para que fosse possível definir como os modelos seriam modelados para a obtenção das previsões pretendidas.
- **Capítulo 6 – Apresentação e Discussão de Resultados** – Com todos os modelos concebidos já aplicados, neste capítulo faz-se uma apresentação e discussão detalhada dos resultados alcançados.
- **Capítulo 7 – Conclusões e Trabalho Futuro** – Neste capítulo são apresentadas as conclusões sobre todo o trabalho realizado e retiradas as respetivas ilações dos resultados obtidos. São ainda apontadas algumas linhas de orientação para trabalho futuro por forma a dar seguimento ao trabalho desenvolvido neste projeto.

CAPÍTULO 2

PREVISÃO DE VENDAS

2.1 Características das Vendas

De uma forma geral, uma venda é o ato de vender um artigo ou bem em troca de dinheiro ou outra compensação. As vendas são algo fulcral no funcionamento das empresas. O processo de vendas é muitas vezes cuidadosamente analisado para ser possível identificar quais os pontos a melhorar, para que seja possível aumentar, ou pelo menos manter, o nível de vendas das empresas. A econometria é um conjunto de métodos que visam encontrar relações económicas em dados económicos, o que torna as vendas, alvo de estudos econométricos para melhor perceber quais os fatores que as influenciam (Allen e Fildes, 2001).

Existem diversos fatores que podem influenciar uma venda, quer sejam estes internos às empresas ou externos, que condicionam toda a conjuntura económica em que tanto o vendedor como o comprador se encontram. Quando os fatores se relacionam com a empresa em si e o mercado em que esta atua são os fatores microeconómicos que estão em causa, sendo a microeconomia a área da economia que estuda como é que decisões e comportamentos dos mercados em que a empresa atua podem influenciar a procura e oferta dos bens e dos serviços. A macroeconomia contrasta com a microeconomia por não tratar de fatores locais mas sim de fatores globais, focando-se principalmente na causa-efeito das flutuações dos lucros nacionais e no crescimento económico a longo termo. Como fatores internos apresentam-se como exemplo, o número de vendedores, a capacidade de inovação, as encomendas e a capacidade de produção. Estes são os fatores que dependem diretamente das empresas, estando diretamente ligados ao volume das vendas realizadas. Além disso, influenciam direta ou indiretamente a quantidade e a qualidade de produtos que as empresas podem disponibilizar aos consumidores dos seus bens ou serviços, bem como a capacidade de apresentar um maior valor acrescentado ao consumidor final, levando-o a adquirir o produto da empresa em detrimento de um produto da concorrência. Fatores como o clima ou alterações socioeconómicas podem influenciar fortemente as vendas das empresas, sendo por isso necessário considerar fatores externos, quando se avaliam os aumentos e as diminuições das vendas. As grandes crises económicas são exemplos importantes, pois estas crises fazem com que diversos

fatores externos às empresas, como alterações no PIB (Produto Interno Bruto), se traduzam em ciclos económicos negativos fazendo diminuir as vendas. Estes fatores que influenciam as vendas indiretamente são denominados de fatores externos.

2.2 PREVISÃO

As decisões são tomadas com base no conhecimento atual sobre determinadas variáveis com o intuito de atingir uma dada meta no futuro. Usualmente procede-se a um planeamento dos objetivos para que se possa definir um caminho para atingir os mesmos. Uma previsão permitirá tomar algumas decisões em detrimento de outras, possibilitando adaptar melhor o planeamento das atividades com o intuito de atingir os objetivos pretendidos. Embora seja impossível prever o futuro de forma completamente assertiva, as técnicas de previsão procuram minimizar ao máximo os erros de previsão.

Geweke e Whiteman (2006) reportam uma definição simplista da previsão como a utilização de informação sob forma de modelos formais, dados e até mesmo experiência pessoal para fazer afirmações sobre o curso provável de eventos futuros. Os agentes de decisão que recorrem a previsões vão desde uma determinada pessoa que decide levar um guarda-chuva, pois a previsão meteorológica prevê chuva, ao acionista que decide investir numa empresa, sendo que as previsões económicas são de crescimento para um dado segmento de mercado. A previsão é assim, algo bastante ambicionado por parte dos agentes de decisão ligados às empresas. Com previsões de vendas, as empresas passam a poder otimizar decisões como por exemplo encomendas, investimentos e compras. Esta previsão é também bastante importante hoje em dia pois, tal como indicado por Ostrow (2011) num *whitepaper* do AberdeenGroup, um exemplo de empresas que sofrem uma enorme pressão para entregar previsões, são as empresas de vendas que são bastante pressionadas tanto pelos *stakholders* internos quer externos para entregar previsões relativas a vendas o mais assertivas possíveis, uma vez que estas vão melhorar a capacidade de previsão das próprias empresas a longo prazo.

Para compreender de forma plena a utilidade das previsões e as suas aplicações é necessário compreender em que consistem os métodos preditivos e como os mesmos são utlizados. Desde o primórdio dos tempos que o homem procura prever o futuro para poder planear quais tarefas deverá executar. Os primeiros métodos de previsão utilizados foram os métodos qualitativos, sendo um exemplo a utilização dos mesmos pelos nossos

antepassados quando estes tentavam perceber como se iriam desenvolver as plantações ao longo do tempo. Eles utilizavam a informação que dispunham como as condições atmosféricas, relacionando-as com a época do ano em que se encontravam aliando estes factos a conhecimentos de anos anteriores, de forma a prever qual o desenvolvimento das plantações ao longo do ano. Com o desenvolver da ciência, as previsões também começaram a evoluir deixando de ser apenas qualitativas e passando a ser também quantitativas. Os estudos quantitativos são estudos que se baseiam em valores quantitativos e não qualitativos. Desta forma, um exemplo da utilização de métodos qualitativos é a utilização por parte dos primeiros físicos, quando estes começaram a estudar as relações matemáticas que levaram a compreender o comportamento de um corpo, não estando mais do que a estabelecer relações matemáticas que consigam descrever o movimento do corpo utilizando relações quantitativas para prever acontecimentos futuros.

No livro Philosophiae Naturalis Principia Mathematica (Newton, 1687) Sir Isaac Newton apresentou as três leis de Newton através das quais é possível descrever o comportamento de um corpo, sendo assim possível prever o movimento de um corpo quando as condições iniciais são conhecidas. As fórmulas matemáticas são assim utilizadas em diversos campos da ciência, como a física, a mecânica ou a eletrónica para estabelecer relações entre dados conhecidos, por forma a ser obtido um determinado resultado, sendo assim possível "prever", com certeza, o resultado. Na utilização destas fórmulas matemáticas usualmente todas as variáveis são conhecidas e são estáticas, contudo, por vezes os sistemas são demasiado complexos ou as variáveis têm valores incertos, o que torna impossível obter resultados precisos. Assim sendo, foram desenvolvidos modelos matemáticos que, com alguma incerteza associada ao resultado obtido, conseguem estabelecer relações entre variáveis com o intuito de prever o resultado do modelo. Nestes casos, os modelos de regressão são usualmente utilizados para estabelecer relações entre variáveis e prever o possível resultado. Sendo as ciências económicas um campo que lida com sistemas muito complexos, com variáveis que se encontram em constante mudança, estas viram a potencialidade da previsão de resultados melhorar com a utilização de modelos de regressão, sendo o modelo MARS (*Multivariate Adaptive Regression Splines*) um dos modelos utilizado para fins de previsão, apresentado por Friedman (1991). Este modelo é utilizado, por exemplo, por Lu et al. (2012) para fazer previsão de vendas de computadores por retalhistas. A utilização de modelos de

previsão em contexto económico, e mais propriamente no contexto de vendas tem vindo a aumentar com o aumento da capacidade dos computadores, pois com a evolução das tecnologias cada vez é mais fácil aceder às variáveis e processar modelos mais complexos, aumentando a credibilidade das previsões.

Atualmente verifica-se a existência de uma vasta panóplia de aplicações para realizar previsões. Como tal, é necessário identificar os diferentes tipos de previsões para que seja possível escolher o melhor modelo a aplicar em cada caso. Como apresentado no parágrafo anterior, as previsões podem derivar de métodos qualitativos e quantitativos. No que diz respeito aos métodos quantitativos, estes podem ser divididos em duas classes, os métodos não causais e os métodos causais. Os métodos não causais assentam nos valores históricos da variável a prever, sendo utilizado usualmente métodos de regressão simples. Entre os vários métodos não causais, um dos mais utilizados é o método ARIMA (*AutoRegressive Integrated Moving Average*), que é uma generalização do modelo ARMA (*AutoRegressive Moving Average*), que se baseia nas médias móveis do valor histórico da variável, bem como na autoregressão das previsões em relação aos valores passados. Este modelo foi apresentado por Whittle (1951). Porém, este modelo só foi popularizado com a sua sistematização por George Box e Gwilym Jenkins, sendo por isso muitas vezes apelidado de Box-Jenkins (Box et al., 1994). O modelo ARIMA acrescenta aos processos autorregressivos e aos processos de médias móveis do modelo ARMA, um processo de integração que permite transformar séries temporais em séries estacionárias caso estas não o sejam.

Devido ao facto de muitas das variáveis a prever serem influenciadas por outras variáveis, é necessário recorrer a métodos que não se baseiem apenas nos valores históricos da variável a prever, mas também em valores históricos de outras variáveis sendo este o caso das variáveis causais. Em Mohammad e Nishida (2010) é feita uma breve introdução sobre causalidade sendo possível resumir uma relação causal como "se *x* então y", o que leva à interpretação que "se *x* acontecer, então y também vai acontecer". Os métodos causais são métodos que procuram relacionar a variável a prever com outras variáveis que possam explicar o seu comportamento. Estes recorrem habitualmente a técnicas de regressão múltipla. Neste sentido, as técnicas de DM tomam uma posição relevante na escolha de técnicas, pois através dos seus diversos algoritmos, permitem a regressão de variáveis influenciando o valor de previsão da variável a prever. Assim é necessário analisar o tipo de variáveis a considerar. As variáveis a utilizar podem tomar

três tipos de comportamento em relação à variável a prever. Estas podem ter um comportamento causal, em que uma alteração nesta variável auxiliar irá provocar uma alteração na variável a prever, ou apresentar um comportamento indiferenciado, no qual uma alteração da variável auxiliar não está relacionada com o comportamento da variável a prever. Por fim, podem ainda ter um comportamento consequente, em que a variável auxiliar varia após alteração da variável a prever. Desta forma, para efeitos de aprimoramento de previsão, é desejável que sejam consideradas como variáveis auxiliares variáveis que tenham comportamentos causais em relação à variável a prever.

2.3 Séries Temporais

A capacidade de armazenar dados relativamente a uma variável sem qualquer contexto temporal pode não ser suficiente para um dado agente de decisão. Nas mais diversas áreas como a economia, para saber datas de vendas, ou na saúde, para saber o valor de batimentos cardíacos num determinado instante de tempo, é imprescindível atribuir um contexto temporal a uma dada variável. Desta forma, as séries temporais representam conjuntos de dados ordenados cronologicamente (Esling e Agon, 2012). Matematicamente, esta relação pode ser expressa da seguinte forma:

$$T = (t_1, \dots, t_n), t_i \in \mathbb{R}$$

tendo cada variável *n* associada a si um instante de tempo *t*, em que *t* é cronologicamente precedente a *t-1* e antecedente a *t+1*.

A grande utilização deste formato de armazenamento e apresentação de dados, remonta já ao final do século XX, uma vez que entre 1974 e 1989 uma amostra de 4000 gráficos de 15 jornais de todo o mundo, apresentava 75% dos seus gráficos em forma de séries temporais (Ratanamahatana et al., 2005). Associado ao grande potencial que a representação de dados através de séries temporais possui, associam-se diversas dificuldades na representação e manuseamento deste formato de dados. Segundo Yang e Wu (2006) um dos problemas mais desafiantes na investigação em DM é a mineração de dados sequenciais e de dados de séries temporais, uma vez que frequentemente os dados que são guardados sob o formato de séries temporais possuem demasiado ruído, levantando barreiras na interpretação correta de dados. De forma geral, o ruído associado às séries temporais é obtido devido ao uso errado de agentes de agregação de informação, sendo obtidas por vezes informação a mais e outras a menos, comprometendo análises

sob as séries temporais. Este problema é patente principalmente quando se pretendem realizar previsões sob séries temporais, pois torna-se fulcral escolher os agentes preditivos corretos para de facto ser garantido o menor ruído possível nas séries temporais a analisar. Esling e Agon (2012) salientaram que o facto das séries temporais terem associadas a si uma grande dimensionalidade e apresentarem alguma dificuldade na definição de medidas de similaridade baseadas na perceção humana, levanta três problemas:

- a apresentação de dados, uma vez que será necessário compreender como deverão ser as formas/contornos apresentadas pelas séries temporais para que estas não percam as suas características essenciais;
- as medidas de similaridade, uma vez que é necessário possuírem boas medidas de similaridade para que se consiga perceber se duas séries são ou não idênticas, mesmo que não o sejam matematicamente;
- os métodos de indexação, uma vez que as séries temporais muitas das vezes representam grandes volumes de dados, é importante perceber como estas séries podem ser indexadas para que seja possível lançar *querys* sobre as séries de forma mais rápida e eficaz.

A identificação destes desafios deve-se à necessidade de encontrar soluções para que seja possível tirar partido de todo o potencial das séries temporais uma vez que estas são utilizadas para diversas finalidades.

Do ponto de vista de aplicação de técnicas de DM sobre séries temporais, estas são essencialmente alvo de análise com base em técnicas de classificação, agregação, sumarização, segmentação, indexação, deteção de anomalias ou previsão (Keogh e Kasetty, 2002; Esling e Agon, 2012; Ratanamahatana et al., 2005; Laxman e Sastry, 2006).

Através da classificação pretende-se que seja possível identificar a que grupo pertence uma dada série temporal, comparando a mesma com outras séries temporais já definidas e divididas em grupo. O algoritmo é assim treinado criando grupos perante diferentes séries temporais para ser possível identificar a que grupo pertence a nova série. Esta comparação pode ser realizada em relação a várias séries temporais dentro de um grupo ou então por questões de desempenho, podendo ser escolhido apenas um representante para cada grupo, e a comparação ser feita apenas perante esse representante (Geurts,

2001), (Keogh e Pazzani, 1998). É usual esta técnica de DM sob séries temporais ser utilizada para reconhecer linguagem gestual, transformando as sequências de imagens em séries temporais e identificando, assim, a que grupo ou palavra a nova série temporal pertence.

A agregação é uma técnica similar à classificação uma vez que também agrupa as séries temporais quanto à sua característica. Porém, ao contrário das técnicas de classificação, nas técnicas de agregação não existe uma fase de aprendizagem na qual são criados os grupos de comparação. No caso da agregação, os grupos são criados conforme as características dos dados inseridos. A técnica de agregação pode ser aplicada segundo duas perspetivas, uma de agregação de séries temporais completas em que se agregam as séries pelas suas semelhanças (Xiong e Yeung, 2002) e outra de agregação de subséries, sendo a agregação feita conforme certas características que permitam determinar sequências de características diferentes numa série temporal (Kalpakis,Gada e Puttagunta, 2001). Para fins de deteção de similaridade em séries económicas é usual recorrer a técnicas de agregação para que seja possível identificar a título de exemplo, séries financeiras que se movam com as mesmas tendências na bolsa. Por vezes, as séries temporais são demasiado grandes para apresentar. Para esses casos sugere-se a utilização de técnicas de sumarização sob séries temporais. Com as técnicas de sumarização pretende-se reduzir o número de dados das séries temporais sem retirar os seus componentes essenciais, tornando-se possível representar gráficos que não caberiam numa folha mantendo a mesma representação visual relativamente ao original.

Uma das técnicas de DM mais utilizadas com séries temporais, são as técnicas de indexação ou de *querying* por conteúdo. Estas técnicas permitem, perante uma *query* lançada contra a base de dados, detetar que série temporal mais se assemelha à série temporal utilizada como *query*. Este processo pode ser realizado a dois níveis, ao nível total, no qual a *query* lançada apenas procura séries temporais completas que se assemelhem à série temporal utilizada com *query*, ou a nível parcial, em que é pesquisado em todas as séries temporais disponíveis se alguma subsequência se aproxima da série temporal utilizada como *query*. É assim possível, a título de exemplo, procurar qual o ritmo cardíaco que mais se aproxima do pretendido num eletrocardiograma.

A deteção de anomalias é uma técnica de DM que consiste na análise de séries temporais com o intuito de perceber que subsequências não apresentam um comportamento considerado normal. Usualmente, este processo de deteção é realizado

fornecendo ao sistema uma série temporal em que se encontrem identificadas algumas sequências de dados que seriam caracterizadas como anómalas, sendo para tal necessário um treino prévio do algoritmo. Este tipo de técnicas é amplamente usado no mundo empresarial pelas empresas para detetarem possíveis casos ou situações de fraude (Ngai et al., 2011).

Sendo as séries temporais um formato de armazenamento de dados que permite estabelecer relações temporais entre as suas variáveis, elas tornam-se por excelência alvo de técnicas de previsão. Considerando que o objetivo da previsão é estimar um possível valor futuro, as previsões sob séries temporais, focam-se em estimar para uma dada série temporal T quais serão os valores de "T + *n"*. Como as séries temporais possuem informação dos valores de uma dada variável ao longo do tempo, a utilização de métodos regressivos para a realização de previsões está facilitada, uma vez que facilmente se consegue utilizar os valores existentes de forma ordenada em processos de extrapolação através de métodos de regressão para valores futuros. Uma das áreas em que as técnicas de previsão são utilizadas abundantemente, é a área dos negócios. Um exemplo da sua utilização é a capacidade de prever as vendas de um dado artigo, para meses futuros, com base no seu histórico que se encontra em forma de série temporal.

2.4 DATA MINING NA PREVISÃO DE VENDAS SOB SÉRIES TEMPORAIS

2.4.1 DATA MINING SOB SÉRIES TEMPORAIS EM ÁREAS FINANCEIRAS

Devido às vantagens que as técnicas de DM trazem aos ambientes de negócio, a sua utilização neste domínio é bastante proveitosa tanto para gestores como para analistas de negócio. Bose e Mahapatra (2001) apresentam um artigo em que revêm a utilização de técnicas de DM nos negócios, baseando o seu estudo na revisão de artigos publicados em diversos jornais científicos que possuíssem o termo DM quer no seu título quer nas suas *keywords.* Na tabela 2.1 está apresentada a aplicação de técnicas de DM distribuídas por área de negócio. Como é percetível através da sua análise, a maior aplicação destas técnicas centra-se nas áreas financeiras com 28.3%, sendo patente desta forma o forte interesse das técnicas de DM.

Tabela 2.1. Distribuição de aplicação por área funcional. Tabela adaptada de (Bose e Mahapatra, 2001)

Functional Area	Number of applications	Percentage
Finance	17	28,3
Marketing	12	20,0
Web analysis	9	15,0
Telecom	7	11,7
Others	15	25,0
Total	60	100

Tabela 2.2. Distribuição da aplicação por área funcional e categoria de problema. Tabela adaptada de (Bose e Mahapatra, 2001)

	Finance	Marketing	Web analysis	Telecom	Other	Total	Percentage
Classification	2	3	5	2	8	20	31,7
Prediction	10	3		2	4	19	30,3
Association		6	6			12	19,0
Detection	5		1	3	3	12	19,0
Total	17	12	12	7	15	63	100

Aumentando o nível de detalhe da análise, é possível através da tabela 2.2 identificar que a grande maioria dos problemas nas áreas das finanças são os problemas de previsão. Tomando em consideração o grande interesse na utilização de DM para obtenção de informação com fins preditivos e considerando o grande valor apresentado por dados em forma de séries temporais, torna-se, assim, bastante motivante utilizar técnicas de DM na exploração de previsões sob séries temporais em áreas financeiras.

As técnicas de DM com fins preditivos podem ser divididas como tarefas de classificação e tarefas de regressão. As tarefas de classificação focam-se em estabelecer relações entre os dados, classificando estas relações para que a ocorrência de um evento possa levar à previsão de outro caso tenha sido classificado uma associação entre os dois eventos. É então estabelecida uma relação "se A ocorre, então B também ocorre", podendo assim ser previsto que se A ocorreu então B irá ocorrer (Dhanabhakyam e Punithavalli, 2011).

O *Market Basket Analysis* é um bom exemplo da utilização de técnicas de previsão, já que com este método é possível com base no histórico de compras anteriores e associações já conhecidas, identificar quais os produtos que serão adicionados a um carro de supermercado, com base nos produtos que já se encontram no mesmo. Contudo, as técnicas de DM com fins preditivos que tenham por base tarefas de classificação encontram-se para lá do objetivo deste projeto. Todavia, para os interessados, em

(Dhanabhakyam e Punithavalli, 2011) está apresentada uma revisão de literatura bastante interessante acerca desta problemática.

As técnicas de regressão procuram relacionar modelos matemáticos e estatísticos para que seja possível, com base em valores passados, prever valores futuros. Existem diversos trabalhos na área financeira sobre utilização de técnicas de DM para previsão de vendas sob séries temporais. Grande parte do trabalho realizado em áreas financeiras recai sobre previsões para bolsas financeiras. De Oliveira et al. (2013) apresentaram um estudo no qual propõem a utilização de NN na previsão dos valores da bolsa da empresa Petrobas, que foram denominadas de PETR4 e que são negociadas na bolsa de valores de São Paulo. Neste estudo, foi apresentada uma abordagem utilizando NN recorrendo a vários KPI (*Key Performance Indicators*) e a outros indicadores macroeconómicos para auxiliar o modelo na previsão dos valores futuros na bolsa. O trabalho foi dividido em quatro fases: estudo do domínio, pré-seleção e recolha de amostras, pré-processamento e modelação e previsão. Em cada secção foram justificadas as opções tomadas até se obter o modelo final de NN, que assentou num modelo *feed forward multilayer perceptron* recorrendo a um nível de *input layer*, um de *hidden layer* e um de *output layer*. No final, os resultados obtidos foram considerados como positivos, uma vez que foi atingido um MAPE (*Mean Absolute Percentage Error*) de 5,45%, demonstrando assim que, neste caso, a utilização de NN ao invés de métodos tradicionais, apresentaram resultados bastante favoráveis à futura utilização de NN.

Kim (2003) apresentou também uma abordagem para a realização de previsões sobre ações, usando SVM e comparando o seu desempenho perante a capacidade preditiva de mecanismos baseados em *back-propagation neural networks* e *case-based reasoning*. Neste estudo foi analisada a alteração diária das ações no KOSPI (*Korean composite stock price índex*) recorrendo-se a várias fórmulas matemáticas que permitiram obter informação mais precisas sobre as variações das ações. O autor concluiu que o modelo SVM proposto tem um desempenho superior aos outros modelos utilizados, advertindo porém para a sensibilidade das funções de *kernel* do algoritmo SVM que podem influenciar fortemente o resultado final.

Nos estudos anteriores, o objetivo de estudo foi a análise da alteração positiva ou negativa dos valores das ações, porém Chen e Du (2009) apresentam uma abordagem a outro elemento de interesse nas bolsas de valor, a análise das dificuldades financeiras. Muitas vezes as empresas conseguem ocultar a sua dificuldade financeira, levando

acionistas a comprar ações das mesmas, quando estas não se encontram saudáveis do ponto de vista financeiro. O artigo apresenta um modelo baseado em NN que permite prever quando uma empresa se irá encontrar em dificuldades financeiras, para que os investidores possam ter outro elemento de ponderação na compra (ou não) de ações, antes de existir uma variação dos valores na bolsa da empresa. O estudo analisa empresas que violaram leis do TSEC (*Taiwan Stock Exchange Corporation)* operando fora das condições exigidas pela bolsa, comparando-as com empresas que permaneceram na bolsa, e recorrendo também a outras variáveis auxiliares consideradas como causais em relação ao aparecimento de *stress* financeiro. São previstas pelos autores situações de *stress* financeiro para que seja possível obter sinais de alarme o mais cedo possívelconcluindo-se que é possível prever as situações de *stress* financeiro e que algoritmos baseados em NN apresentam melhor resultado do que os algoritmos de agregação.

Também na área financeira, mas orientado ao mercado de conversões monetárias se pode analisar a implementação de um modelo híbrido apresentado por Khashei e Bijari (2010). Este estudo, reporta um modelo híbrido em que recorreram a um modelo ARIMA para pré-processar os dados semanais das taxas de conversão entre a libra britânica e o dólar americano, aplicando à *posteriori* um modelo baseado em NN. Os resultados obtidos com este modelo são superiores à utilização de um modelo que recorre apenas a NN para realizar a previsão de um período subjacente em relação aos dados conhecidos, sendo atribuído pelo autor o facto da utilização de modelos que se contradizem bastante, permitir uma menor variação quando generalizado, levando a uma redução de incertezas do modelo. Por forma a melhorar a qualidade de previsão recorre-se muitas vezes a variáveis macroeconómicas que permitam perceber como se irão mover as condições exteriores às empresas em causa para prever os efeitos inesperados, como por exemplo, a recessão que se iniciou em 2008.

2.4.2 Previsão de vendas através de técnicas de Data Mining

Os casos apresentados anteriormente, enquadraram-se numa perspetiva de previsão sob séries temporais na área financeira, não sendo focado o tema de previsão de vendas. As previsões de vendas são úteis de diferentes pontos de vista, uma vez que podem ajudar a tomar decisões antecipadamente para alterar planos de negócio, visto que as previsões podem não ir ao encontro do expectável, ou podem simplesmente ser utilizadas para perceber de forma indireta como será o consumo de um dado artigo e assim, perceber

quais as quantidades a produzir para evitar excedentes ou para evitar ruturas de *stock*. Quer seja por questões meramente financeiras, quer por razões mais associadas a processos logísticos ou por questões meramente ambientais e morais, as previsões de vendas desempenham um papel importante no quotidiano dos agentes de decisão. Estas encontram-se aplicadas nas mais diversas indústrias, sendo exemplos as indústrias alimentares, automóvel, de confeção e de vestuário e até mesmo a imprensa. Embora todos estes tipos de empresas possam ser alvos dos mesmos algoritmos, existem determinadas características associadas a cada uma que transformam cada caso num caso único.

Na indústria alimentar, Doganis e os seus colaboradores apresentam um estudo direcionado à previsão de vendas para alimentos com tempo de prateleira de curta duração no qual recorreram a NN (Doganis et al., 2006). Os autores dão alguma relevância ao facto de considerarem que a escolha das variáveis corretas, dentro de uma panóplia de variáveis candidatas, é algo fulcral para a realização de previsões uma vez que defendem que a utilização de todas as variáveis candidatas pode prejudicar a previsão. Desta forma, apresentam um algoritmo que denomina por GA-RBF (Genetic Algorithm – Radial Basis Function) o qual é constituído por dois algoritmos distintos: o algoritmo GA, que tem como objetivo codificar as diversas variáveis candidatas como se de cromossomas se tratasse, tendo em conta a inclusão ou não destas no sistema, e o algoritmo NNRBF (*Neural Networks with Radial Basis Function*) que irá ser aplicado sobre as diversas variáveis, tendo como objetivo a previsão com menor MSEP (*Mean Squared Prediction Error*) com vista à definição do melhor conjunto de variáveis a utilizar e, consequente, do modelo de NNRBF.

Dando também bastante importância à necessidade de determinar quais as melhores variáveis auxiliares a utilizar de um conjunto de variáveis candidatas, Lu et al. (2012) recorreram a um algoritmo baseado em MARS (*Multivariante Adaptive Regression Splines*) para identificar quais as melhores variáveis a utilizar, comparando depois se se obteve um melhor resultado recorrendo apenas a este algoritmo ou aplicando-o juntamente com outros. O algoritmo MARS funciona da forma "*divide and conquer*" atribuindo diferentes equações de regressão a diferentes áreas dos grupos de dados de treino. Desta forma, o algoritmo tem a capacidade de aplicar diferentes equações de regressão perante a utilização de diferentes variáveis, sendo aplicado neste caso para avaliar variáveis referentes à média móvel dos últimos períodos. Desta forma, os autores

comparam a aplicação do algoritmo MARS com os algoritmos SVR, BPN, ELM, CMACNN, ARIMA e MLR, bem como com os algoritmos MARS-SVR, MARS-BPN, MARS-ELM e MARS-CMACNN, concluindo que o algoritmo MARS se encontra sempre associado a melhores resultados na previsão de vendas de computadores por parte de retalhistas. Recorrendo a técnicas de regressão, Yu, Qi e Zhao (2013) apresentaram um algoritmo baseado em *Support Vector Regression* (SVR) para realizar previsões de vendas de revistas e jornais em diversas lojas. Devido à existência de diversas variáveis externas como a existência de diferentes tipos de lojas (livrarias, papelarias, etc.) e emprego, educação, idade, sexo e ordenado dos frequentadores das mesmas. O artigo refere que o algoritmo baseado em SVR apresenta resultados bastante positivos quando comparados com resultados de outros algoritmos. Numa outra área de estudo, mais precisamente na indústria automóvel, foram também realizados outros estudos que apresentam a utilidade da aplicação de técnicas de DM na previsão de vendas, sendo um desses estudos, apresentado em (Hülsman et al, 2012) discutindo e analisando de forma genérica diversos algoritmos de DM na previsão de vendas no mercado automóvel alemão e americano. Recorreu-se a diferentes variáveis auxiliares para os dois mercados, sendo no caso do mercado alemão utilizadas as variáveis DAX (*Deutscher Aktienindex*) e IFO (*Institute for Economic Research*), sendo a primeira variável um indicador associado ao desenvolvimento das 30 empresas que mais vendem na bolsa de valores de Frankfurt e o segundo respetivo a um indicador causal lançado todos os meses pelo *German Institute for Economic Research*.

No caso do mercado Americano são utilizados como variáveis registos de novos carros, PIB, salário, taxa de desemprego, taxas de juros, preço médio de bens de consumo, preço médio de gasolina, índice de consumo privado, variações médias mensais *Dow Jones* e por fim, a variação mensal dos índices BCI (*Business Confidence Index*). O objetivo do estudo foi verificar qual o desempenho de diversos algoritmos usualmente utilizados, sendo estes OLS (*Ordinary Least Squares*), QR (*Quantile Regression*), SVM, DT (*Decision Trees*), KNN (*k-Nearest Neighbor*) e RF (*Random Forest*). Desta forma, pretendeu-se verificar qual a capacidade preditiva de cada um dos algoritmos quando aplicados a valores reais e perante a utilização de diversas variáveis. O algoritmo SVM atingiu bons resultados, ao contrário dos algoritmos OLS e QR que, devido à sua linearidade, não se revelaram muito adequados para estas tarefas de DM. Entre os algoritmos que apresentaram melhores resultados estão os algoritmos DT, KNN e RF,

conseguindo o algoritmo DT apresentar os melhores resultados entre todos numa avaliação global. Um mercado bastante interessante para análise do ponto de vista da previsão de vendas baseadas em métodos de DM é o mercado têxtil, segmentado para a fabricação de vestuário. O fabrico têxtil para vestuário é uma área de produção bastante complexa devido ao grande nível de agregação que os diversos grupos de têxteis podem atingir, sendo esta complexidade demonstrada na figura 2.1. Através da sua análise pode-se verificar que uma simples peça de roupa pode ser agrupada em até oito categorias, sendo cada categoria composta por diversas variantes. Para além da existência de diversos artigos distintos, existem outras características deste mercado que tornam as previsões em tarefas bastante complexas, como é o caso do conceito abstrato de "moda". O facto de um determinado artigo "estar ou não na moda" condiciona bastante as vendas, tanto porque pode ditar o sucesso de um determinado artigo, como inviabiliza muitas vezes a sua utilização na mesma época do ano seguinte caracterizando estes artigos, como artigos de tempo de vida curto (Choi, Hui e Yu, 2014). Outros fatores que condicionam a compra de roupa são o caso do clima e do calendário, uma vez que épocas festivas podem proporcionar a compra de nova roupa e o clima pode condicionar o tipo de roupa a comprar. Este mercado não se encontra porém à margem de fatores externos como as condições macroeconómicas ou a existência de fortes concorrentes, o que, juntamente com as suas características próprias apresentadas anteriormente, transformam de facto este mercado num sistema bastante complexo, onde existem diversas variáveis causais praticamente independentes entre si que influenciam o valor final das vendas (Choi, Hui e Yu, 2014).

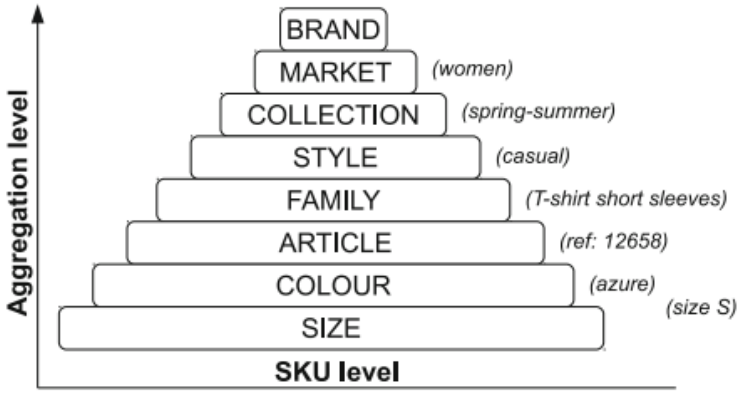


Figura 1.1 Exemplo de agregação de informação por tipologia de produtos – figura extraída de (Choi et al., 2014).

Thomassey e Fiordaliso (2006) apresentaram um trabalho intitulado "*A hybrid sales forecasting system based on clustering and decision trees*" no qual apresentaram um mecanismo de previsão de vendas recorrendo a dois algoritmos distintos: numa primeira fase, um algoritmo de *clustering*, *k-means,* para agregar os itens conforme a sua semelhança em termos de perfis de vendas permitindo assim aplicar o algoritmo de árvores de decisão sobre conjuntos já bem definidos; depois, um algoritmo de árvore de decisão, o C4.5, que irá ser aplicado aos diversos perfis de vendas já separados, derivando pelas relações entre critérios definidos, identificando assim séries temporais de vendas associadas a artigos que tenham obtido o mesmo tipo de derivação, por entre as árvores de decisão. As séries temporais de vendas identificadas como semelhantes no final da aplicação do algoritmo C4.5 são utilizadas como protótipos e considerados como séries de previsão das vendas. Mais tarde, Thomassey (2010) apresentou um artigo no qual expos uma pesquisa mais desenvolvida em torno deste mercado, apresentando de forma bastante descritiva as diversas variáveis que o influenciam, bem como as ferramentas usadas por parte da indústria com a finalidade de realizar previsões. Neste estudo, o autor realça que a escolha de algoritmos a utilizar varia com a finalidade da previsão. Isto é, se o objetivo for o de realizar previsões com um horizonte previsional pequeno, um a doze meses. No entanto, os melhores algoritmos a utilizar são algoritmos baseados em NN, se o horizonte de previsão for superior a um ano. Além disso, o autor identifica algoritmos que sejam baseados em sazonalidade e que se baseiem em interferências *fuzzy* por parte de variáveis explanatórias, como a melhor opção. Contudo, o autor ressalva que estas escolhas só se aplicam na existência de histórico de dados. Caso os dados sejam escaços, o que acontece principalmente se a previsão for avaliada ao nível do artigo, é defendida como melhor opção pelo autor a utilização de algoritmos baseados em algoritmos de agregação e de classificação. Dentro do mesmo domínio, Yu, Choi e Hui (2011) apresentaram um algoritmo baseado em redes neuronais *feedforward* apenas com um nível de nodos ocultos, denominado por ELM (*Extreme Learning Machine*). Este algoritmo permite de forma bastante rápida identificar os valores ótimos em termos de quantidade de séries de dados a utilizar para treino, bem como número de nodos ocultos. No caso de estudo, o autor considera as variáveis cor, tamanho e preço, como fatores que influenciam as vendas dos artigos, sendo os dados referentes às vendas de três meses registados num sistema de POS (*Point Of Sales*) e sendo utilizados 60% dos dados para treino, 20% para validação e os restantes 20% para testar a capacidade de previsão. Recorrendo a um algoritmo baseado em redes neuronais, foram realizados diversos testes

de forma sistemática por forma a serem identificados rapidamente os parâmetros a otimizar do algoritmo, obtendo depois uma previsão de qualidade considerável. Um último exemplo da aplicação de técnicas de mineração de dados em sistemas de previsão de vendas para o mercado de retalho de vestuário foi apresentado por Ni e Fan (2011), que apresentaram um modelo previsional de duas fases. Os autores recorrem a um algoritmo de árvores autorregressivas para aplicar técnicas de regressão ao longo dos nodos da árvore obtendo no final a previsão de vendas de um dado artigo. Neste estudo, as previsões foram divididas em dois grupos, um em que a previsão foi feita para um conjunto de lojas e um segundo na qual a previsão é feita ao nível do artigo. Desta forma, é possível o modelo identificar quais as variáveis a utilizar para auxiliar a previsão de vendas. Numa primeira fase o modelo realiza previsões a longo termo, utilizando para tal apenas dados históricos de vendas. A segunda fase do modelo permitiu realizar previsões a curto termo, adicionando informação em tempo real aos dados históricos, para que a previsão pudesse atingir um melhor nível de previsão nos primeiros períodos em análise. Segundo os autores é assim obtida uma previsão com maior qualidade, quer a curto quer a longo termo, sendo destacado que deverão ser estudadas quais as melhores variáveis a utilizar em trabalhos futuros, bem como a utilização de outros algoritmos que não as árvores autorregressivas.

Esta apresentação de bibliografia, focada mais especificamente no mercado do vestuário, teve apenas como objetivo demonstrar que, mesmo dentro de um mesmo domínio, existem inúmeras formas de abordar a previsão de vendas e algoritmos que podem ser aplicados. Assim, é expectável que tenham sido identificadas diversas técnicas utilizadas em previsão de vendas acerca da utilização de séries temporais, bem como retiradas ilações quanto ao esforço necessário para serem obtidas previsões de elevada qualidade e o quão diferente podem ser os resultados das previsões conforme uma maior ou menor exposição a determinadas variáveis causais.

2.5 CONCLUSÃO

Nos dias que correm os meios tecnológicos tomam um papel fulcral nos ambientes empresariais. Verifiquemos, pois, os exemplos dos *marketters* que recorrem a ferramentas capazes de realizar análise de perfis dos clientes, dos recursos humanos que conseguem com a utilização de CMS (*Content Management System*) chegar a todos os colaboradores,

e mesmo dos responsáveis de produção que conseguem recorrer a algoritmos capazes de otimizar processos de produção a um nível impensável em pleno século XX. Os sistemas de informação permitem um controlo fulcral em termos de competitividade e respetiva prosperidade das empresas. As vendas não fogem assim a esta revolução dos sistemas de informação. Embora a utilização de ferramentas de BI pelos agentes de decisão seja cada vez mais frequente na melhoria e na análise da informação, bem como em processos de toma de decisão sobre as vendas das suas empresas, a verdade é que, com toda a evolução tecnológica e constante evolução dos mercados, esta é uma área em constante crescimento. Com o aumento da capacidade de armazenamento e de processamento de dados por parte das empresas, as técnicas de DM começam a ganhar cada vez uma maior importância. Estas permitem às empresas procurar informação que não era conhecida previamente nos seus dados. Embora estas técnicas possam ser utilizadas para procurar relações entre vendas, permitindo melhor caracterizar os mercados nas quais as empresas se posicionam, elas podem também ser utilizadas para realizar previsões de vendas. A previsão de vendas é um tema bastante aliciante para agentes de decisão que trabalhem em áreas logísticas ou de gestão cujo trabalho seja direta ou indiretamente influenciado por vendas, pois uma previsão sólida das suas vendas futuras irá transmitir-se num excelente dado de suporte a decisões que poderão levar a uma vantagem competitiva tremenda. A maior problemática que as técnicas de previsão de vendas com base em técnicas de mineração de dados enfrentam são as diversas variáveis que se encontram associadas às vendas de um produto. Desta forma, foram apresentados diversos trabalhos neste campo para que fosse possível perceber não só quais as técnicas mais utilizadas para realizar previsões de vendas mas também a abordagem que é usualmente feita quando se lida com diferentes mercados.

CAPÍTULO 3

DATA MINING E TÉCNICAS PARA PREVISÃO DE VENDAS

3.1 DATA MINING

A informatização do mundo que nos rodeia implica um aumento exponencial da quantidade de dados gerados, quer estes sejam provenientes do aumento de utilizadores na rede para uso privado, quer sejam provenientes de processos de negócio cada vez mais complexos e com maior poder de captação de dados. Este elevado número de dados leva a que muitas vezes exista informação valiosa que não é utilizada, uma vez que a sua existência simplesmente não é conhecida ou não se conhece a sua utilidade. Desta forma, torna-se necessário o desenvolvimento de mecanismos que permitam recolher informação a partir de grandes conjuntos de dados que muitas vezes não se encontram diretamente relacionados entre eles. Para este fim, foram desenvolvidas as técnicas de DM que permitem "minerar" dados que se encontram em bruto. Diferentes autores definem "*Data Mining*" obviamente de formas diferentes. Turban e seus colaboradores definem DM como sendo um processo que recorre a técnicas estatísticas, matemáticas, inteligência artificial e de *machine-learning* para extrair e identificar informação a partir de grandes volumes de dados (Turban et al., 2007). Por sua vez Fayyad, juntamente com Piatetsky-Shapior e Smyth apresentam o DM como sendo algo integrante de um processo maior, o processo de KDD, sendo que neste ponto de vista o processo de DM diz apenas respeito à extração de conhecimento, considerando o pré-processamento de dados, e a interpretação de resultados outras fases pertencentes ao processo de KDD. Porém, independentemente das interpretações existentes, um ponto é comum: DM diz respeito à extracção de conhecimento a partir de dados que resultem em nova informação para o utilizador. Desta forma, as técnicas de DM perfilam-se como boas técnicas para realizar previsões de vendas com base em séries temporais, extraindo informação sobre o formato de previsões a partir de históricos de vendas e de outras variáveis causais.

3.2 METODOLOGIAS DE DATA MINING

Devido à complexidade inerente à implementação de técnicas de DM é fulcral sustentar a implementação destas técnicas em metodologias que permitam garantir o sucesso e viabilidade dos projetos. Usualmente, estes projetos são de média ou grande duração, e

necessitam de um nível considerável de conhecimento acerca do negócio. Além disso, os projetos de DM alocam muitos recursos das empresas, quer seja a nível monetário quer seja a nível humano. As diversas metodologias de DM permitem desenvolver os projetos de forma sustentada tentando evitar ao máximo a sua inviabilização em fases mais maduras provocada por limitações funcionais ou tecnológicas.

A primeira metodologia de implementação de projetos de DM surge nos primórdios do DM (Fayyad et al., 1996ª). Embora os autores não considerem DM como uma técnica que possa ser utilizada isoladamente, estes denominam a metodologia que implementa processos de DM como o processo de KDD. Este considera nove tarefas essenciais, nomeadamente:

1. Analisar o domínio em que o projeto vai ser implementado, bem como perceber quais os objetivos do cliente para com a implementação do processo de KDD.
2. Criar e identificar o conjunto de dados que vão ser analisados com o intuito de identificar nova informação.
3. Realizar o pré-processamento e a limpeza dos dados. As fontes de dados nem sempre se encontram prontas a ser utilizadas em processos de DM. Por isso, este passo é importante para remover possíveis ruídos que os dados contenham, bem como identificar e definir o que fazer perante a falta de dados ou a existência de dados que variem ao longo do tempo.
4. Uma vez que são utilizados grandes volumes de dados, neste passo é realizada a redução e a projeção de certos conjuntos de dados, uma vez que a redução do número de variáveis utilizadas e a identificação de dados que não variam, podem ajudar a diminuir o volume de dados em análise melhorando assim toda a performance dos processos sequentes.
5. Nesta fase do processo de KDD são alinhados os objetivos do cliente com as diversas técnicas de DM, identificando-se quais as melhores técnicas a aplicar com o intuito de satisfazer os objetivos do cliente.
6. Depois de identificadas as técnicas de DM a aplicar, estas serão utilizadas com o intuito de permitir a seleção do(s) algoritmo(s) e dos métodos para identificação de dados a utilizar. É tida em conta não só a capacidade de processamento de dados, mas também outros aspetos como a preferência do cliente por obter um algoritmo mais orientado à previsão de dados ou a facilidade de análise dos mesmos.

7. É nesta etapa que o processo de DM propriamente dito tem efeito. Nesta etapa são aplicados os algoritmos e são obtidos os novos dados. A análise de dados é posterior a esta fase, não estando ainda os dados prontos a ser entregues ao cliente.
8. Os dados obtidos através do processo 7 são analisados e é tomada a decisão se os mesmos vão ao encontro do que o cliente pretende ou se deve ser refeita alguma das iterações anteriores.
9. Nesta última etapa, os resultados obtidos são documentados e, se necessário, são introduzidos em outros sistemas, para que seja possível entregar a nova informação "descoberta" num formato interpretável para o cliente.

Todos estes passos encontram-se representados na figura 3.1, que representa a visão global do processo de KDD para descoberta de conhecimento (Fayyad et al., 1996ª).

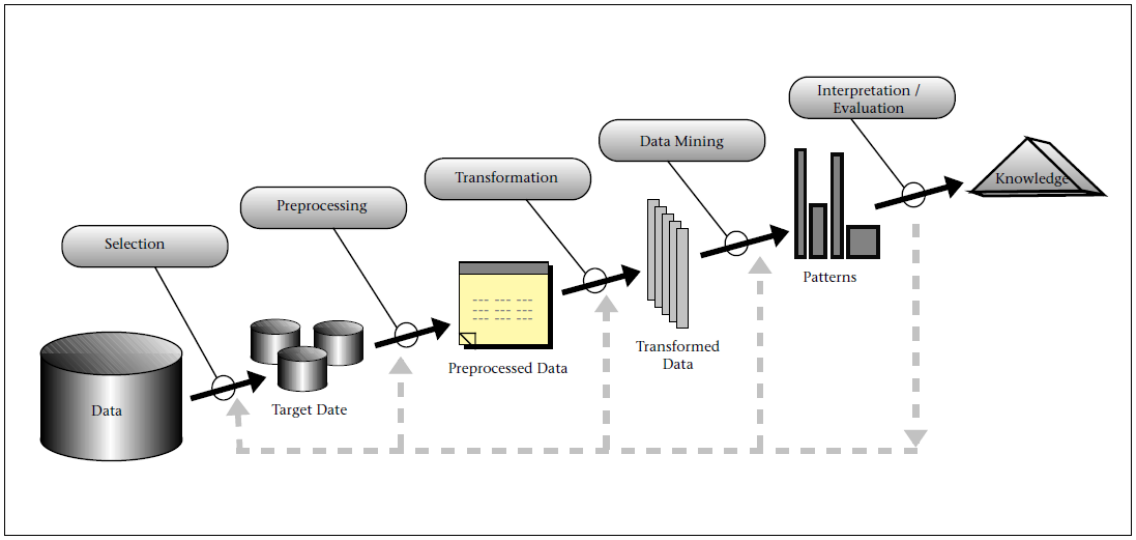


Figura 3.1. Revisão dos diversos processos que compõem o KDD – adaptado de Fayyad et al. (1996).

Embora a metodologia KDD seja amplamente utilizada nos meios académicos, a comunidade empresarial recorre frequentemente a uma outra metodologia, a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) (Chapman et al., 1999). Esta metodologia partiu da iniciativa das empresas DaimlerChrysler, SPSS e NCR em desenvolver uma metodologia que fosse transversal a toda a indústria, independente de qualquer plataforma e que permitisse, de forma rápida e prática, apresentar as mais-valias da implementação de projetos de DM na indústria. Um consórcio foi realizado para que

fosse possível discutir e analisar diversas opiniões e pontos de vista diferentes, obtendo grande adesão por parte de diversas entidades globais demonstrando uma grande vontade por parte da indústria mundial em estabelecer padrões para a implementação de projetos de DM. Desta forma, em meados de 1999 surgiu a versão 1.0 do CRISP-DM que como metodologia pretendeu descrever as principais etapas de um projeto, desde as tarefas que constituem cada fase, às relações que as diversas fases têm entre si. Como modelo de processos, foi pretendido possibilitar uma análise de todo o processo de DM, desde a análise do negócio até à entrega do produto final. Os processos da CRISP-DM decompõem-se nas seguintes fases e etapas (Chapman et al.,1999):

1. Aprendizagem do negócio.
    - Identificar objetivos de negócio.
    - Verificar capacidade de acesso a recursos.
    - Determinar objetivos da mineração de dados.
    - Produzir planificação do projeto.
2. Análise dos dados.
    - Recolha de dados iniciais.
    - Descrição dos dados.
    - Exploração dos dados.
    - Verificação da qualidade de dados.
3. Preparação dos dados.
    - Seleção de dados.
    - Limpeza de dados.
    - Construção de dados.
    - Integração de dados.
    - Formatação dos dados.
4. Modelação.
    - Seleção da técnica de modelação.
    - Gerar esquema de testes.
    - Construir modelo.
    - Consultar modelo.
5. Avaliação.
    - Avaliar resultados.
    - Rever processos.

- Determinas próxima tarefa.

6. Distribuição.
    - Planear a distribuição.
    - Planear a manutenção e a monitorização.
    - Produzir os relatórios finais.
    - Rever o projeto.

As duas primeiras fases de ambos os métodos são bastante idênticas, uma vez que dizem respeito à recolha de informação relativa ao negócio, quer em termos de dados, quer em termos de contexto empresarial. Ao contrário do método KDD, que utiliza as fases 3 e 4, no método CRISP-DM, o tratamento de dados é realizado numa só fase. Na fase 3 do método CRISP-DM é realizado o processo necessário para selecionar, limpar e construir as fontes de dados para a posterior aplicação dos algoritmos. Na fase 4 do método CRISP-DM realizam-se os processos equivalentes a duas das fases do método KDD, as fases 5 e 6 já apresentadas anteriormente. Na primeira fase do método CRISP-DM, é realizada uma análise sobre os objetivos da mineração de dados que permitirão suportar a escolha de algoritmos a aplicar posteriormente na fase 4. É nesta fase que os modelos são escolhidos, projetados e implementados, sendo apenas analisados os resultados na fase seguinte. Na fase 5 deste método são analisados os resultados obtidos e é revisto todo o processo para que seja possível, no final, decidir-se se é necessário repetir alguma fase do método ou não, sendo estes passos realizados na fase 7 e 8 do método KDD. A distribuição, a manutenção e a monitorização do(s) modelo(s) são realizados na fase 6, terminando a última fase do modelo CRISP-DM com a produção de relatórios finais e a revisão do projeto.

Não sendo a sequência de fases de implementação algo rígido no CRISP-DM, a sequência mais comumente aplicada é a que está apresentada figura 3.2. Este um método mais focado na indústria, sendo por excelência, o método mais aconselhado para a aplicação de técnicas de mineração de dados com fins industriais.

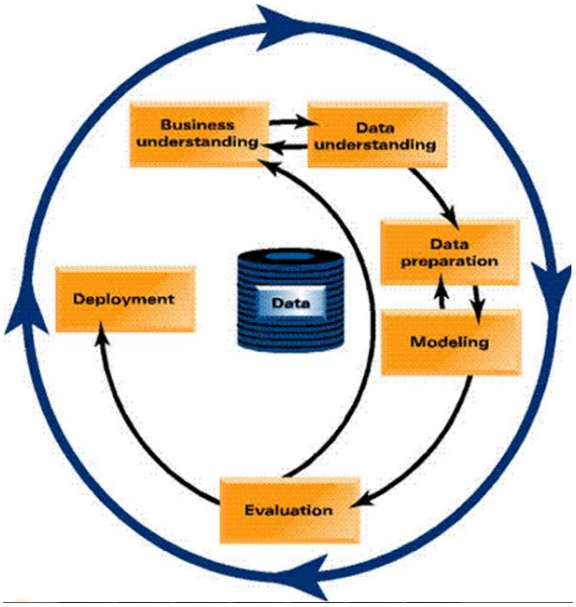


Figura 3.2. Ciclo de vida de um projeto desenvolvido de acordo com a metodologia CRISP-DM – extraído de Chapman et al. (1999).

3.3 ALGORITMOS DE DATA MINING

Para que se compreenda melhor as tarefas a realizar pelos algoritmos de DM, vamos agrupá-los em dois tipos: descritivos e preditivos. Os modelos descritivos podem ter como objetivo a segmentação ou a sumarização dos dados. Por sua vez, os modelos preditivos podem ter como objetivo a regressão ou a classificação dos dados por forma a determinar um valor não conhecido. Isto acontece porque embora estas sejam as duas grandes categorias em que os algoritmos de DM se enquadrem devido à grande variedade de algoritmos existentes hoje em dia, a linha que separa as duas categorias é muito ténue, sendo usual recorrer a algoritmos preditivos que tomam uma função de análise tão descritiva, que, no final, apresentam comportamentos característicos de métodos descritivos.

Os modelos descritivos são modelos cujo principal objetivo passa por caracterizar e descrever determinadas características dos dados que permitam ao utilizador relacionar as diversas características existentes e que permitam retirar elações de como os dados se relacionam entre si. Dentro dos modelos descritivos, os mais comumente utilizados são os métodos de *clustering* e os métodos de *sumarisation* (Fayyad et al., 1996ª) .

Os métodos de *clustering* são métodos que procuram agrupar os dados de acordo com as suas características. Desta forma, perante as métricas de agrupamento definidas, é possível determinar os grupos de dados existentes sem que para isso seja necessário treinar o algoritmo com grupos pré-definidos. A capacidade do algoritmo de identificar grupos com propriedades similares deve-se ao facto de estes algoritmos procurarem maximizar as similaridades intraclasses e minimizar as similaridades intercalasses. Um exemplo de utilização deste tipo de algoritmos é a criação de grupos sobre o perfil dos clientes, para poder fornecer às unidades de marketing os diferentes perfis de compradores e tornar as técnicas de marketing mais efetivas. Os grupos criados por *clustering* não necessitam de ser exclusivos, existindo a possibilidade de serem únicos ou de partilharem elementos (Figura 3.3).

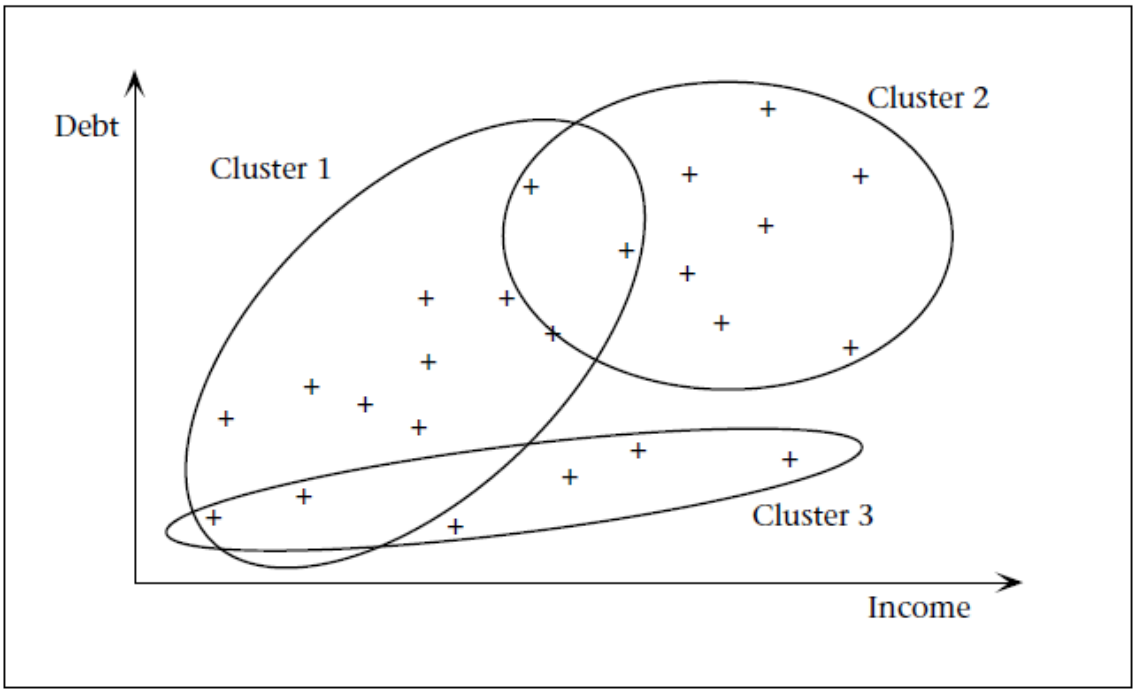


Figura 3.3. Exemplo de Clustering – extraído de (Fayyad et al., 1996ª)

Os modelos de *summarisation* procuram identificar características dos dados que permitam descrever um dado grupo. Isto é, ao invés da segmentação em que são criados grupos a partir das características, a sumarização procura identificar quais as características que os dados têm em comum para que possam ser resumidos. Um exemplo de utilização de técnicas de *summarisation* são os cubos OLAP nos quais os dados são agregados e sumariados perante certas características facilitando a visualização dos mesmos – ex: compras do ano 2013. Quando o objetivo passa por utilizar os dados atuais

para prever valores futuros é necessário recorrer a modelos preditivos, sendo os modelos mais conhecidos os modelos regressão (*regressio*n) e classificação (*classification).*

Quando aplicados, os algoritmos de classificação possuem duas fases distintas. Num primeira fase, estes algoritmos realizam a chamada "fase de treino" na qual o algoritmo é testado contra um conjunto de dados da amostra para que consiga estabelecer correlações e, por conseguinte, derivar as regras de classificação necessária para obter a melhor classificação possível. Numa segunda fase, o modelo matemático definido na fase de testes é aplicado ao restante conjunto de dados para realizar a classificação Um cenário em que a classificação é bastante útil é no momento de cedência de crédito por parte das entidades bancárias. Quando existe um novo cliente, a entidade bancária tem de perceber de acordo com o seu perfil se esta deverá ou não conceder crédito. Desta forma e com base em casos passados, o algoritmo de classificação compara o novo cliente com dados históricos e "percebe" com base na regra derivada se o cliente tem um perfil mais próximo dos clientes cumpridores ou incumpridores. Exemplos da implementação de modelos de classificação são as árvores de decisão e as redes neuronais, sendo estas abordadas detalhadamente nas próximas secções, uma vez que no caso das árvores de decisão é possível ao longo dos seus ramos fazer derivações exclusivas, acabando por identificar, no nodo final, a que grupo pertence o novo elemento. No caso das redes neuronais, quando utilizadas em tarefas de classificação, são atribuídas unidades de processamento a cada neurónio, estando estes interligados e possuindo um determinado peso atribuído. De seguida, apresenta-se um pequeno exemplo de como um modelo de classificação poderá separar os dados de pessoas que poderão obter um empréstimo daquelas que não o conseguirão, de acordo com o seu crédito e o seu ordenado (Figura 3.4).

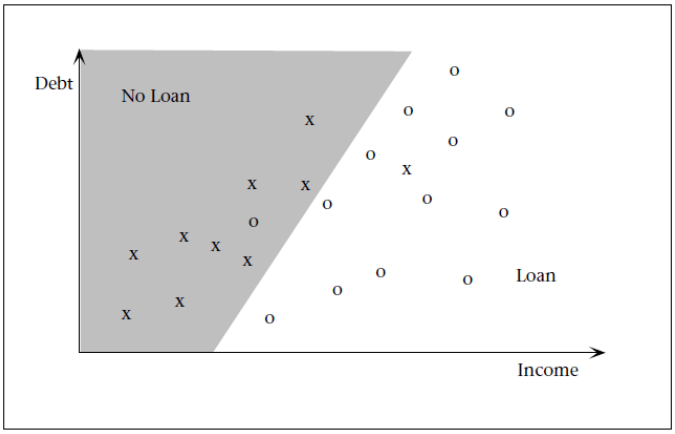


Figura 3.4. Um simples classificador linear de limites para a amostra de dados de empréstimos – extraído de (Fayyad et al., 1996[a]).

Enquanto os modelos de classificação fornecem previsões fortemente ligadas a dados nominais, devido à sua natureza descritiva na abordagem aos problemas, os modelos de regressão permitem o estabelecimento de modelos numerais para prever os valores futuros das variáveis. Por sua vez, os modelos descritivos são mais utilizados para descrever qual o nome de uma potencial classe, enquanto que os modelos de regressão são mais utilizados para identificar situações de valores nulos ou de falta nos dados. Tal como os algoritmos de classificação, os algoritmos de regressão também passam por uma fase de treino na qual se criam fórmulas matemáticas para correlacionar as diferentes variáveis do problema, com o intuito de criar a "regressão" que melhor descreve o potencial comportamento futuro da variável. A previsão de temperaturas é um exemplo de utilização de modelos de regressão, no qual se recorre a dados passados como valores de temperaturas de anos anteriores, estado atmosférico etc., para se estabelecer uma correlação entre os dados e, depois, com base na regressão prever as temperaturas futuras (Almonacid, et al., 2013)

Na implementação de projetos de DM o conhecimento do problema a abordar toma um papel preponderante na escolha de algoritmos. Conhecendo as vicissitudes do problema em questão e as características das diversas abordagens existentes a problemas de DM, é possível escolher quais os algoritmos que melhor se adequam a um dado problema. Contudo, não menos importante é identificar as limitações dos algoritmos de DM e o tipo de dificuldades que estes projetos têm a si associadas. Mesmo estando num mundo rico em dados em que as técnicas de DM trazem valor acrescentado aos negócios, a verdade é que estas possuem limitações que podem inviabilizar a sua implementação. Em 1996, Fayyad, e seus colaboradores apresentaram alguns dos problemas encontrados no desenvolvimento de projetos de DM. De referir:

- **Bases de dados de grandes dimensões**. Inerente à sua própria existência, os projetos de mineração de dados terão sempre associados a si grandes volumes de dados. Desta forma, torna-se essencial desenvolver-se algoritmos capazes de lidar com o crescimento do tamanho das bases de dados ou recorrer-se a técnicas de amostragem e de resumo de dados que permitam diminuir o volume de dados a utilizar.
- **Grande dimensionalidade dos dados**. O elevado espaço de armazenamento ocupado pelas bases de dados encontra-se também ligada à grande dimensionalidade dos mesmos. Um incremento do número de atributos dos dados leva a que o número de variáveis e de combinações entre as mesmas aumente de forma exponencial. Desta forma, a complexidade dos problemas a analisar é cada vez maior, o que implica que se recorra a algoritmos que consigam encontrar de forma célere quais as variáveis que realmente interessam para o problema e, também evitar que este aumento de atributos aumente também o número de descobertas sem qualquer relevância para o problema em causa.
- **Sobre ajustamento**. Este fenómeno diz respeito a uma aproximação demasiado grande entre o modelo de dados obtido pelos algoritmos e os modelos de dados reais, que aparece frequentemente quando a quantidade de dados a analisar é reduzida ou se utilizam grandes percentagens de dados para a construção de modelos de DM.
- **Avaliação de pontos estatisticamente significativos**. Quando um algoritmo de DM procura entre vários modelos quais os modelos estatisticamente mais significativos, por vezes o algoritmo diminui o seu desempenho simplesmente por entrar em ciclo na tentativa de decidir entre dois modelos similares qual é o estatisticamente mais significativo.
- **Dados e conhecimento em mudança**. Sendo os projetos de DM aplicados quase sempre em ambientes ricos em dados não estacionários é usual que tanto os dados como as métricas sofram alterações, o que pode tornar inválidas as conclusões atuais. Porém, estas alterações podem também abrir portas a novas descobertas, quer em relação aos modelos de implementação já aplicados, quer no sentido de abrir a possibilidade de, através de outros modelos, estudar qual o impacto das alterações realizadas.

- **Dados em falta e dados com ruido**. Uma vez que os modelos aplicados são baseados em relações entre dados já existentes é essencial verificar a qualidade dos dados, pois a existência de dados com ruido ou de dados em falta erroneamente identificados, podem levar a conclusões completamente desfasadas da realidade.
- **Compreensão dos parâmetros**. Numa área em que se interage tão profundamente com o utilizador é essencial permitir uma boa interpretação dos parâmetros, para que desta maneira seja possível que todos os intervenientes do projeto consigam perceber e interpretar os dados da mesma forma.
- **Interação do utilizador e conhecimento prático**. Considerando a pluralidade de métodos de DM bem como as suas bases teóricas, é normal que nem todos os modelos possuam a mesma simplicidade. É assim importante conhecer o relacionamento anterior dos utilizadores com as técnicas de DM selecionadas, para que estes possam aproveitar todo o potencial do modelo utilizado.
- **Integração com outros sistemas**. Os modelos de mineração de dados apenas realizam a descoberta de conhecimento. Por isso, é usual que estes sistemas estejam interligados a mais dois sistemas, nomeadamente: um sistema de dados que permita recolher todos os dados necessários e um sistema de apresentação que permita mais facilmente apresentar o novo conhecimento adquirido, facilitando a sua assimilação.

Existem outros algoritmos que permitem implementar lógicas de regressão para poder realizar previsões, sendo alguns desses algoritmos os SVR (*Support Vector Machines for Regression*), as NN (*Neural Networks*), as ART' (*Auto Regressive Trees*) e os ARIMA (*Auto Regressive Moving Average*). De seguida serão apresentados individualmente e será explicado de forma sucinta o funcionamento de cada um destes algoritmos.

3.3.1 Support Vector Machines for Regression

As SVR são uma adaptação de SVM's para problemas de regressão. Assim, é necessário compreender o que são os SVM. Os SVM são algoritmos baseados na teoria de aprendizagem estatística, a qual caracteriza as propriedades que permite às máquinas de aprendizagem generalizar os seus resultados para diferentes casos (Smola e Scholkopf, 2002). A generalização é a capacidade que um algoritmo tem de apresentar bons

resultados perante conjuntos de dados diferentes dos conjuntos de dados de treino. Em 1979, Vapnik apresentou um SVM com a capacidade de encontrar um híper plano ótimo que consegue separar uma classe positiva de uma classe negativa maximizando as suas margens. Platt apresentou no seu trabalho uma imagem que ilustra de forma simples esta separação (Figura 3.5).

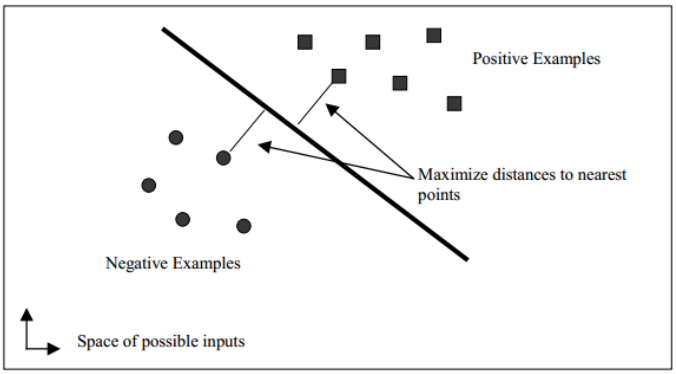


Figura 3.5. A linear Support Vector Machine (Um SVM Linear) – extraído de (Platt, 1998)

Matematicamente os SVR são bastante próximos dos SVM, sendo apresentadas de seguida as fórmulas matemáticas que explicam o funcionamento de um SVR. Assumamos, então, que $(x_1, y_1) \dots (x_k, y_k)$ com, $x_i \in \mathbb{R}^n$ e $y_i \in \mathbb{R}$, representa as variáveis de entradas e as variáveis de saída respetivamente. O objetivo da função de regressão é encontrar uma função linear de aproximação como a seguinte:

$$f(x) = (\omega \dots x) + b \qquad (1)$$

com $\omega, x \in \mathbb{R}^n, b \in \mathbb{R}$ e $(\omega \dots x)$ $os\ produtos\ escalares\ em\ \mathbb{R}^n$

Nesta função, ω representa um vetor normal para o híper plano e x é a variável de entrada. Após treino, o y correspondente pode ser obtido através de $f(x)$, para os x que se encontrem fora dos dados de treino (Peixian et al., 2011). Em (1) é pretendido obter o menor ω possível, o que pode ser conseguido minimizando a norma $||\omega||^2$, obtendo assim o seguinte problema de otimização e respetivas restrições:

$$\text{Minimizar } \frac{1}{2}||\omega||^2 = \frac{1}{2}(\omega.\omega) \qquad (2)$$

$$\text{Sujeito a: } \begin{cases} y_i - (\omega.x_i + b) \leq \mathcal{E} \\ (\omega.x_i + b) - y_i \leq \mathcal{E} \end{cases} \text{ i=1,2,..,l} \qquad (3)$$

em que $\mathcal{E}$ representa a precisão aceitável em termos de erro por parte do algoritmo.

Uma vez que este problema de otimização pode originar uma procura por um mínimo global sem solução ou pode ser pretendida a aceitação de erros para evitar o sobre ajustamento do modelo, foram introduzidas as variáveis de folga ($\xi\ e\ \xi^*$) na equação dos SVR para replicar o comportamento das margens suaves dos SVM, originando assim o seguinte problema:

Minimizar $\frac{1}{2}||\omega||^2 + C\ \sum_{i=1}^{l}(\xi + \xi^*)$ (4)

Sujeito a: $\begin{cases} y_i - (\omega.x_i) - b \leq \mathcal{E} + \xi_i \\ (\omega.x_i) + b - y_i \leq \mathcal{E} + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$ (5)

no qual C representa a penalização atribuída aos erros que se apresentam para além do intervalo de valores com limite $\mathcal{E}$. O intervalo de valores entre -$\mathcal{E}$ e +$\mathcal{E}$ representa uma área em forma de tubo que identifica a área na qual não existe penalização de erros, ou seja, segundo a função *$\mathcal{E}$-insensitive* apenas os valores do tubo são penalizados de forma linear.

$$|\xi|_{\epsilon} := \begin{cases} 0\ , se\ |\xi| \leq \mathcal{E} \\ |\xi| - \mathcal{E}, caso\ contrário \end{cases} \quad (6)$$

A adição de alguns multiplicadores de *Lagrange* não negativos (α_i, α_i^*) permite passar o problema da sua forma primal para uma forma dual passando o problema a ser descrito pela seguinte maximização:

Maximizar $\begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i.x_j) \\ -\epsilon \sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases}$ (7)

Sujeito a $\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)$=0 e $\alpha_i, \alpha_i^* \in [0, C[$

Desta forma ω será substituído por $\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\, x_i$ originando a função.

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\omega.x_i) + b \quad (8)$$

A computação de *b* será então realizada com base nas condições *KKT*, Karush-Kuhn-Tucker (Karush, 1939; Kuhn e Tucker 1951). Estas condições garantem que quando a solução é atingida o produto entre as variáveis duais e as restrições tem de ser anulado.

$\alpha_i(\mathcal{E}+\xi_i - y_i + (\omega.x_i) + b)=0$ (10)

$\alpha_i^*(\mathcal{E}+\xi_i^* + y_i - (\omega.x_i) - b)=0$

e

$(C - \alpha_i)\xi_i = 0$ (11)

$(C - \alpha_i^*)\xi_i^* = 0$

Desta forma é garantido que todos os valores dentro do tubo terão $(\alpha_i - \alpha_i^*) = 0$, logo x_i será anulado pois segundo os multiplicadores de *Lagrange* apenas fora do tubo $(\alpha_i - \alpha_i^*)$ podem ser diferentes de 0, permitindo assim concluir que apenas os valores não descartados estarão presentes no problema de otimização sendo estes os conhecidos vetores de suporte.

Outra informação mais detalhada, sobre a implementação de SVR pode ser consultada em Smola e Scholkopf em que é possível obter informação sobre as diversas funções de *kernel* bem como sobre as diferentes funções de custo. De destacar o modelo SMO (*Sequential Minimal Optimization*), apresentado por Platt em 1998 o qual separa os diversos problemas de PQ (*Programação Quadrática*) dos SVM, em problemas PQ mais pequenos. Os modelos mais pequenos são resolvidos analiticamente, o que evita realizar ciclos de processamento internos para resolver os problemas de PQ segundo o autor. Este modelo permite assim que os algoritmos de SVM e, consequentemente, de SVR, sejam processados mais rapidamente tornando os mesmos uma opção viável para utilização em problemas de DM.

3.3.2 Neural Networks

As *Neural Networks* são sistemas que procuram simular o processo de raciocínio daquele que é o mais fascinante agente de tomada de decisão, o cérebro humano. Em Zahedi (1991), tabela 3.1 é possível encontrar uma introdução simplista dos componentes do cérebro humano e dos constituintes das redes neuronais, sendo eles os seguintes:

Tabela 3.1. Comparação de características entre cérebro humano e Neural Networks. Adaptado de Zahedi (1991).

Cérebro Humano	**Neural Networks**
Neurónios: Unidade simples do cérebro capaz de realizar processos simples.	Camadas: As estruturas mais comuns de *Neural Networks* possuem uma camada de dados de entrada, uma camada de dados de saída e possíveis camadas ocultas adicionais.
Dendrites: Ramos dos neurónios que formam as redes neuronais.	Nodos: Cada camada possui diversos nodos.
Ligações sinápticas: As ligações pelas quais passam os sinais elétricos que ligam os neurónios e as dendrites.	Ligações: Os nodos das diversas camadas encontram-se ligados entre si apresentando comummente uma topologia de ligação "camadas de alimentação adjuvante", em que as ligações seguem o sentido da camada de entrada, para a camada oculta, e desta para a camada de saída.
Axónio: O longo ramo de neurónios que leva o sinal de saída (químico ou elétrico) das dendrites para os neurónios.	Pesos: Cada ligação desde o nodo i até ao nodo j possui um peso w_{ij}, em que o valor do peso varia entre 0 e 1 (-1 e 1 em algumas variantes).
Ativação e desativação de neurónios: Estímulos químicos ou elétricos provenientes quer de estímulos externos, quer de estímulos de outros neurónios que possuem um efeito negativo ou positivo sobre o neurónio alvo.	Valores de entrada nas redes de nodos: O valor de entrada da rede para um nodo i pode ser obtido através da fórmula $net_i = \sum w_{ji}\, activ_i + exitinput_i\,,$ $\forall j\ nodosligados\ a\ i;$ (12) Em que $activ_i$ representa os valores de saída do nodo j e $exitinput_i$ é o valor total de valores de entrada externos/de ambiente sobre o nodo i. A explicação matemática desta soma pode ser matematicamente explicada pelo produto interno dos vetores $w_i = (w_{1i}, w_{2i}, \dots, w_{ni})$, (13) $ACTIV = (activ_1, active_2, \dots, active_n)$, (14) Ignorando a influência dos estímulos externos. Assim sendo o nodo com w_i mais próximo de ACTIV será o nodo mais estimulado.
Atividade dos Neurónios: Uma vez que o estímulo atinge um certo patamar, o neurónio liberta um sinal elétrico ou químico sendo a força da atividade do neurónio medida com base na frequência de sinais por segundo libertados.	Valores de saída dos nodos: os nodos libertam os valores de saída quando atingem um determinado limite T de tal forma que, $active_i = \begin{cases} 0\ se\ nte_i \le T, \\ 1\ caso\ o\ valor\ de\ saida\ seja\ discreto, \\ active_i - T caso\ o\ valor\ de\ saida\ seja\ continuo. \end{cases}$ (15) Um nodo pode contudo possuir um valor de saída contínuo se utilizar funções de transferência como a função logística $active_i = \frac{1}{1+exp^{\frac{-net_i}{T}}}$. (16)
Paralelismo Massivo: Um neurónio tem até 10,000 ligações sinápticas. Estima-se que o cérebro humano possui cerca de 10^{11}neurónios, o que torna o processamento do cérebro humano algo altamente paralelo.	

As NN são classificadas conforme o seu método de aprendizagem, podendo assim ser divididas em aprendizagem em tempo real, aprendizagem não supervisionada e aprendizagem supervisionada. A aprendizagem em tempo real é caracterizada pelos valores de saída dos nodos da camada de saída de uma NN irem para os nodos da camada de entrada, para que seja possível voltar a realizar o processamento dos dados. Por sua vez, a aprendizagem não supervisionada caracteriza uma aprendizagem no qual nenhum

objetivo é estabelecido para que seja possível reduzir o erro e por isso o processamento do algoritmo é realizado com base num padrão predefinido e capaz de desenvolver resultados com base nesse padrão. Por fim, a aprendizagem supervisionada é caracterizada pela utilização de um conjunto de dados de treino que irão permitir afinar o modelo para reduzir o seu erro. Uma vez descobertos os parâmetros pretendidos, a aprendizagem supervisionada permite que através da característica de generalização das NN, estas obtenham resultados com menor nível de erro.

Um exemplo entre os diversos algoritmos de aprendizagem supervisionada das redes neuronais é o algoritmo de *Backpropagation* (*backward propagation of errors*). Este algoritmo foi desenvolvido inicialmente por Werbos (Werbos 1974; Rumelhart, Hinton, e Williams 1986) e baseia-se na correção de erros a partir do erro detetado no nodo de saída. Os valores de ativação dos nodos são obtidos pela equação 15, gerando um valor de saída com um erro associado. Desta forma o erro necessita de ser minimizado e corrigido (se possível) sendo então utilizada a fórmula:

$$w_i^{new} - w_i^{old} = \beta\ error_i f'(net_i) ACTIVE \quad (17)$$

Se a função de ativação for logística tal como a função apresentada em (16) então:

$f'(net)_i$ tomará o valor:

$$f'(net_i) = net_i(1 - net_i) \quad (18)$$

As camadas escondidas possuem erros associados à propagação do erro da camada de *output* definido pelo $f'(net_k)$ do nodo *k*, como:

$$error_j = \sum_{\forall\ k\ ligado\ a\ j} f'(net_k) w_{jk}^{new} error_k \quad (19)$$

Um tipo de NN que normalmente utiliza o algoritmo de *Backpropagation* como função de aprendizagem supervisionada é o algoritmo *Multi-layer perceptron* (perceptron multi-camada). Esta é uma NN de camadas de alimentação- adjuvante estando os nodos ligados através de pesos, formando a soma dos mesmos o sinal de saída que é modificado por uma simples função não-linear de ativação. Algumas das vantagens deste algoritmo são a sua capacidade de modelar funções complexas, ignorar parâmetros de entrada irrelevantes e barulho – este algoritmo encontra-se descrito detalhadamente em (Gardner e Dorling, 1998).

3.3.3 AUTOREGRESSIVE TREE MODELS

Meek, Chickering e Heckerman (2002) incitados pela procura de modelos que pudessem aprender facilmente através dos dados, suportassem previsões assertivas e fossem fáceis de interpretar desenvolveram um modelo de previsões ART. Este modelo de árvores autorregressivas assenta nos princípios das árvores de decisão em que um determinado nodo i possuidor de uma função f (x) vai devolver o valor de entrada para um dos $j_1, j_2, \dots, j_n$ nodos subjacentes. Contudo, este modelo procura adaptar as estruturas das árvores introduzindo funções de regressão nas suas folhas que permitam realizar previsões sobre séries temporais.

Um modelo autorregressivo é definido através da seguinte fórmula:

$$f\left(y_t \middle| y_{t-p}, \dots, y_{t-1}, \theta\right) = N\left(m + \sum_{j-1}^{p} b_j y_{t-j}, \sigma^2\right) \quad (20)$$

Onde N (μ, σ^2) é uma distribuição normal com média μ e variação σ^2 e $\theta = \left(m, b_1, \dots, b_p, \sigma^2\right)$ são os parâmetros do modelo.

Desta forma, uma árvore autorregressiva é um modelo em que os limites são definidos por uma árvore de decisão e as suas folhas são compostas por funções lineares autorregressivas como em (20). Considerando que cada nodo possui um booleano caso não seja uma folha final, a árvore autorregressiva pode ser definida pelo modelo:

$$f\left(y_t \middle| y_{t-p}, \dots, y_{t-1}, \theta\right) = \prod_{i-1}^{L} f_i\left(y_t \middle| y_{t-p}, \dots, y_{t-1}, \theta\right)^{\emptyset_i} = \prod_{i-1}^{L} N\left(m_i + \sum_{j-1}^{p} b_{ij} y_{t-j}, \sigma_i^2\right)^{\emptyset_i} \quad (21)$$

Onde L é o número de folhas, $\theta = (\theta_1, \dots, \theta_L)$ e $\theta_i = \left(m_i, b_{i1}, \dots, b_{ip}, \sigma_i^2\right)$ são os parâmetros do modelo para a regressão linear na folha l_i, i = 1,…,L.

Desta forma, um modelo de árvore autorregressivo consegue modelar relações não lineares ao contrário dos modelos regressivos que apenas conseguem modelar relações lineares (Figura 3.6), sendo um exemplo de uma árvore autorregressiva apresentado na figura (Figura 3.7).

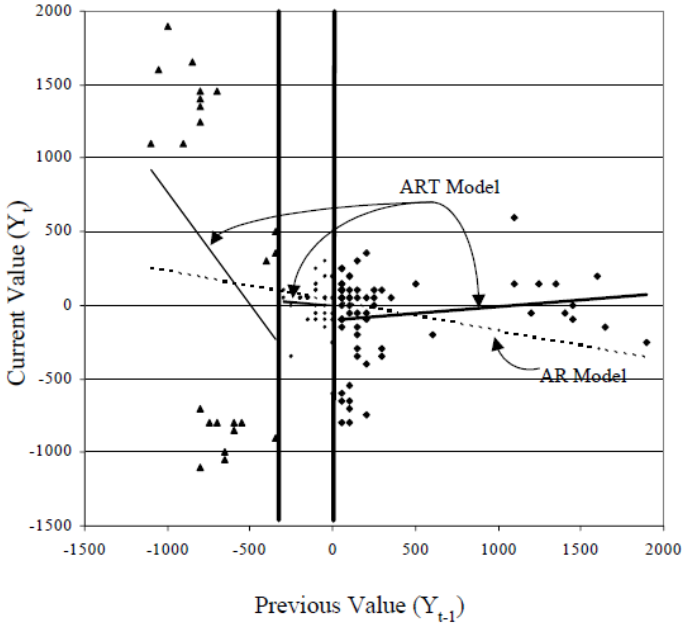


Figura 3.6. Gráfico de dispersão de dados de series temporais provenientes de modelos AR(1) e ART(1), extraído de (Meek et al., 2002)

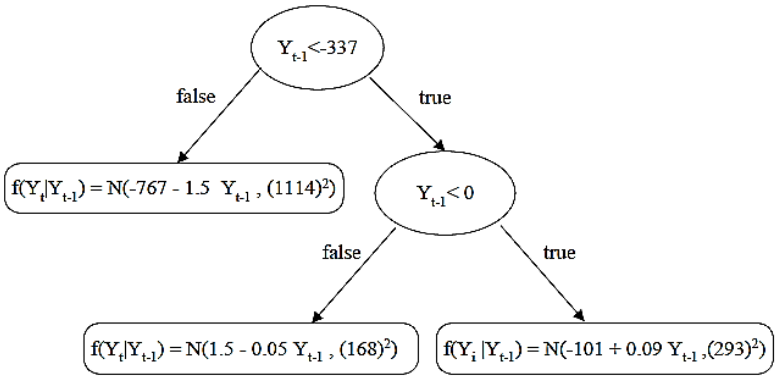


Figura 3.7. Exemplo de árvore autoregressiva, extraído de (Meek et al. 2002)

Meek apresenta no artigo Autoregressive Tee Models for Time-Series Analysis (Meek et al., 2002) como pode este modelo de árvores ser usado para fins de previsão sobre séries temporais.

3.3.4 ARIMA

O modelo ARIMA , também conhecido como modelo Box-Jenkins, foi apresentado por George Box e por Gwilym Jenkins em 1976, sendo um modelo que procura de forma matemática e através de processos autorregressivos, processos de médias móveis e a utilização de um elemento de integração, realizar previsões futuras de séries temporais. Este modelo é apresentado como ARIMA (p,d,q) em que p, d e q representam, respetivamente, a ordem do processo autorregressivo (AR), o grau de diferenciação envolvido (I) e a ordem dos processos de média móvel. Os processos autorregressivos (AR) podem ser descritos pela equação:

$$X_t = c + f_1 X_{t-1} + f_2 X_{t-2} + \cdots + f_p X_{t-p} + e_t \quad (22)$$

em que *t* corresponde ao período temporal atual, *c* corresponde a uma constante, *p* corresponde à ordem do processo autorregressivo, f_p corresponde à previsão para ordem de regressão *p*, X_t corresponde ao valor da série no período *t*, e e_t corresponde ao erro do período *t*. Através do elemento I, elemento de integração, é possível determinar se os modelos observados são modelados diretamente ou não. Ou seja, se I=0 então a modelação é referente à observação atual, se I=1 a modelação é referente a observações consecutivas e assim sucessivamente. Por sua vez, os processos de MA (*Moving Average*) são descritos pela equação:

$$X_t = c - q_1 e_{t-1} + -q_2 e_{t-2} - \cdots - q_q e_{t-q} + e_t \quad (23)$$

em que q corresponde à ordem do processo de média móvel, *c* corresponde a uma constante, *t* corresponde ao período temporal atual, q_q corresponde à MA de ordem 1 q, e e_t corresponde ao erro do período *t*. Um modelo ARIMA (1,0,1) pode então ser descrito da seguinte forma:

$$X_t = c + f_1 X_{t-1} - q_1 e_{t-1} + e_t \quad (24)$$

Uma particularidade dos modelos ARIMA é que estes apenas podem ser implementados em séries temporais estacionárias, por isso é necessário proceder a uma transformação prévia dos dados caso estes não sejam estacionários, procurando eliminar a tendência e a sazonalidade das séries temporais em causa.

3.4 CONCLUSÃO

Embora os dados das empresas sejam cada vez mais abundantes é necessário saber retirar informação útil a partir deles para que a sua abundância tenha algum significado. No processo de DM do KDD, diversas técnicas são aplicadas sobre dados para que seja possível extrair a partir deles nova informação. Embora este processo diga respeito apenas à extração em si de nova informação, sem ter em conta todo o tratamento necessário para tornar esta percetível aos agentes envolvidos nos projetos, nos dias de hoje o termo DM é regularmente utilizado para descrever o processo de mineração de dados, desde a preparação dos dados, passando pela extração do conhecimento e terminando na apresentação do novo conhecimento (Fayyad et al. 1996ª).

Todas as técnicas de extração de conhecimento revelaram-se bastante úteis ao longo do tempo. Assim, o mercado empresarial ganhou interesse sobre elas dada a sua potencialidade para aumentar a competitividade do negócio e potenciar os lucros. Neste sentido, as empresas DaimlerChrysler, SPSS e NCR tomaram a iniciativa de criar um consórcio no qual diversas entidades experientes na área tivessem a oportunidade de dar o seu contributo, para delinear uma metodologia transversal às diversas empresas. Desta forma surgiu a metodologia CRISP-DM que com o contributo de diversas entidades procura conseguir responder de forma mais rápida e flexível às necessidades do mercado. Esta metodologia, procura assim ser, para além de um guia das diversas etapas pelas quais os projetos de DM devem passar, um documento de suporte para consultar, a qualquer altura do ciclo de vida dos projetos de DM, sempre que for necessário.

Para ser implementado um projeto de DM com o intuito de realizar previsões de vendas sobre séries temporais é necessário em primeiro lugar identificar os diversos tipos de algoritmos existentes e quais se podem adaptar ao problema em questão. Uma vez que previsões e vendas sobre séries temporais procuram estabelecer relações causais entre acontecimentos separados cronologicamente e representados por valores contínuos. Tal implica que os algoritmos e as técnicas de DM a utilizar devam respeitar essas premissas. O algoritmo ARIMA procura estabelecer uma relação autorregressiva juntamente com o valor da média móvel dos valores em análise. Por sua vez, as árvores autorregressivas possuem algoritmos de regressão nas suas folhas e procuram através de algoritmos de pontuação criar uma iteração ao longo dos seus ramos, que permita prever qual o próximo valor da sequência. Por seu lado, as *NN* procuram estabelecer relações entre os diversos parâmetros de entrada utilizando funções de regressão nos seus nodos transportando os

parâmetros de entrada para uma camada escondida e calculando um resultado final numa camada de saída. Por fim, as SVM procuram através de uma modelação próxima de modelos de aprendizagem estatística identificar quais os pontos do hiperplano que permitem maximizar as margens entre classes distintas identificando assim quais os pontos que serão utilizados nos cálculos de autorrepressão.

CAPÍTULO 4

ANÁLISE E PREPARAÇÃO DE DADOS

4.1 ANÁLISE DE NEGÓCIO

O objetivo do DM é a descoberta de informação desconhecida a partir de dados já conhecidos. Este processo pode não ser relevante se não existir previamente um correto conhecimento do negócio. Quando as técnicas de DM são utilizadas com fins preditivos, não conhecer o negócio em causa pode levar a que sejam utilizados dados errados como valores de entrada, o que pode conduzir a conclusões erradas ou à não interpretação correta dos dados finais. Desta forma, é importante identificar e perceber o negócio por traz dos algoritmos aplicados e apresentados ao longo da redação deste projeto.

A base do trabalho prático deste projeto surge a partir de um projeto de estágio curricular realizado na empresa Primavera BSS, que teve como objetivo desenvolver um componente de *software* que demonstrasse ser possível implementar um algoritmo de previsão de vendas transversal a diferentes tipos de negócios. Nesse estágio foram utilizados dados de quatro empresas com características distintas, que foram fornecidos pela Primavera BSS. As fontes destes dados não podem ser referidas por questões de confidencialidade. Como tal, essas empresas serão referidas neste projeto por A, B, C e D. As empresas são distintas no que diz respeito ao seu ramo de negócio, ciclo de vendas e volume de negócio. De forma mais detalhada, na tabela 4.1 são apresentadas as empresas. De seguida, nas figuras 4..1, 4.2, 4.3 e 4.4, respetivamente, apresentamos os valores de vendas mensais dessas empresas.

Tabela 4.1. Caracterização de empresas estudadas.

Empresa	Área de Negócio	Volume de Negócio	Sazonalidade	Comentários
A	Agência de viagens	10.000.001€ e 50.000.000€	Sim	Tal como espectado o negócio das agências de viagens é um negócio extremamente sazonal uma vez que as viagens são sempre vendidas em maior número durante as férias.
B	Comércio por grosso de máquinas-ferramentas	500.000€ e 2.000.000€	Sim	Embora não seja um negócio ao qual seja atribuído sazonalidade de forma tão "direta" como uma agência e viagens, a análise dos dados de vendas desta empresa demonstraram que estes também são sazonais.
C	Atividades das sociedades não gestoras e participações sociais não financiadas	10.000.000€ e 50.000.000€	Não	Esta empresa tem uma atividade internacional fazendo exportações para todo o mundo. Desta forma o negócio é realizado de forma totalmente alheia a ciclos.
D	Comércio por grosso de equipamentos eletrónicos e de telecomunicação	2.000.001€ a 10.000.000€	Não	Tal como a empresa C, também a empresa D não apresenta sazonalidade pois tem um negócio que não é influenciável por particularidades do ano.

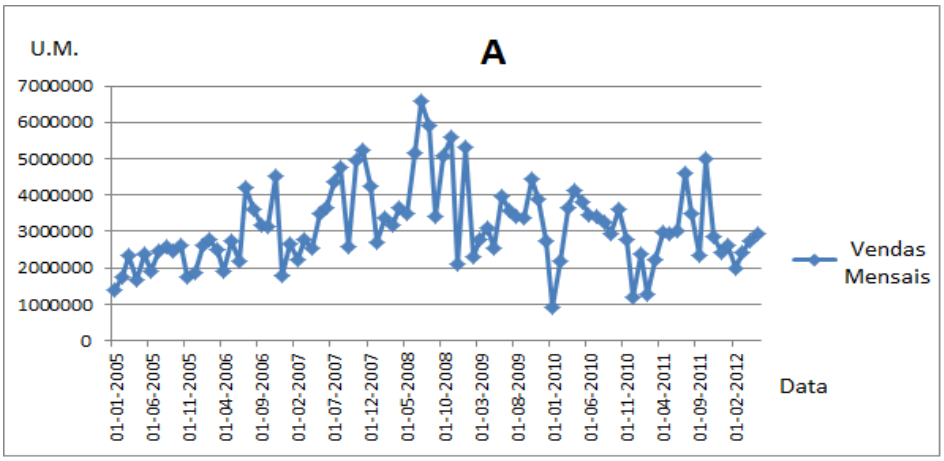


Figura 4.1 Vendas mensais da empresa A – 01-01-2005 e 01-02-2012.

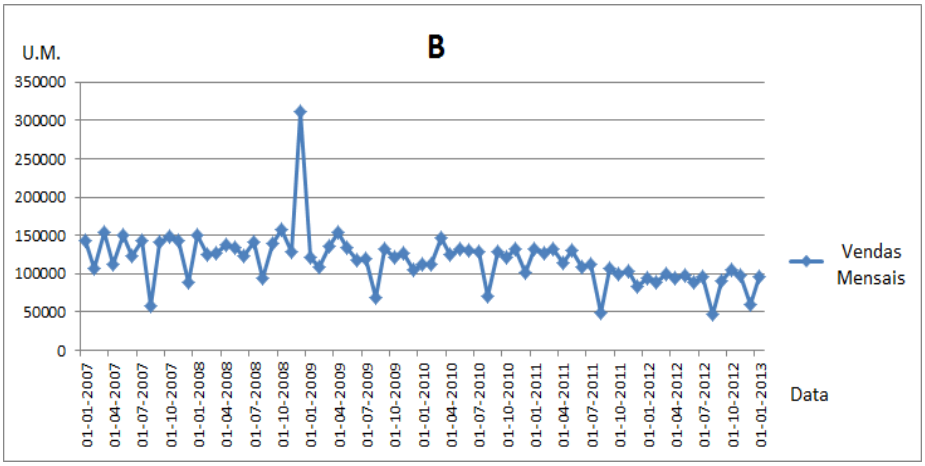


Figura 4.2. Vendas mensais da empresa B – 01-01-2007 e 01-01-2013.

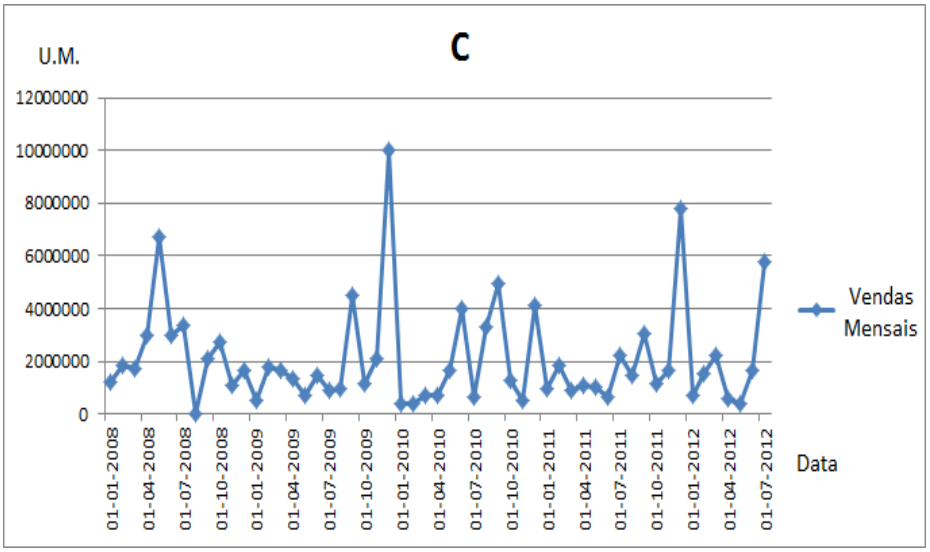


Figura 4.3. Vendas mensais da empresa C – 01-01-2008 e 01-07-2012.

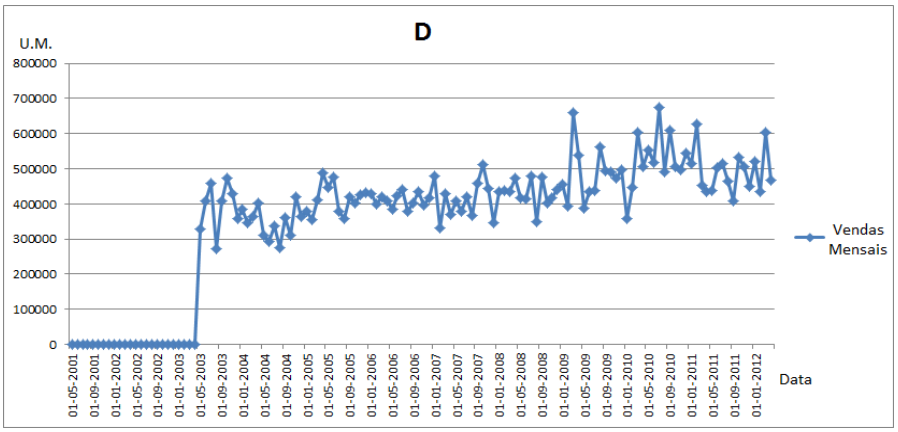


Figura 4.4. Vendas mensais da empresa D – 01-05-2001 e 01-01-2012.

Para além dos dados das vendas foram utilizadas também algumas variáveis internas e externas. Sabendo que existem fatores que se encontram correlacionados com o aumento e diminuição das vendas, mantendo o princípio de manter o modelo o mais simples possível para diminuir futuros custos de implementação, as variáveis internas utilizadas são provenientes dos sistemas ERP da Primavera BSS. Para ser possível determinar quais as variáveis internas a utilizar, foram questionados dez elementos de cargos superiores da Primavera BSS. Embora exercendo o seu cargo em áreas como o marketing, logística, vendas e gestão, todos os elementos possuíam conhecimentos de vendas por experiência que foram acumulando e por formação académica. Entre as diferentes variáveis internas propostas foram selecionadas quatro variáveis transversais a todas as empresas, sem que estas fossem demasiado específicas a uma determinada área de negócio. Assim sendo, as quatro variáveis internas utilizadas são: o número de novos *itens* por mês, o número de novos clientes, o número de vendedores e o número de empresas/sucursais. No que diz respeito às variáveis externas foram escolhidas variáveis externas que tivessem um impacto na economia de forma global. Assim sendo, as variáveis recolhidas foram o produto interno bruto, a procura interna, o consumo privado, as exportações, o consumo de cimento, o investimento de empresas estrangeiras e o índice de confiança do consumidor.

4.1.1 Objetivos de negócio

Após ter sido apresentado o negócio sobre o qual a implementação do projeto de DM se desenrolou, procedeu-se à delineação dos objetivos de negócio. Através dos objetivos de negócio pretendeu-se estabelecer diretrizes que permitissem guiar todo o trabalho prático

do projeto e que permitissem também definir se os objetivos propostos do projeto são alcançados ou não.

O trabalho desenvolvido teve em consideração, o seguinte objetivo de negócio:

> *Determinar qual o melhor algoritmo para realizar previsões de vendas sob séries temporais que possa ser aplicado de forma genérica a diferentes tipos de empresa considerando o equilibrio entre custo de implementação e qualidade da previsão.*

Este objetivo encontra-se fortemente ligado ao projeto que deu origem ao trabalho prático que sustenta este projeto de mestrado, uma vez que a empresa tinha como intuito encontrar um algoritmo que beneficiasse o baixo custo de implementação, face à qualidade final da previsão. Embora seja este o principal objetivo, um segundo objetivo surgiu devido às características do problema em causa. Considerando que as vendas são fortemente influenciadas por fatores externos e internos às empresas, surgiu como segundo objetivo de negócio, determinar se a utilização de variáveis internas e externas auxiliam no aumento de qualidade da previsão. Assim sendo, estes foram os objetivos de negócio que serviram como mote para o estabelecimento de objetivos de DM bem como para a implementação e desenvolvimento dos algoritmos.

4.1.2 Objetivos do Data Mining

Com os objetivos de negócio definidos, surgiu a necessidade de estabelecer como deveria ser conduzida a escolha e implementação dos algoritmos de DM. Desta forma, a necessidade de estabelecer objetivos de DM que permitam de forma ponderada estabelecer como os dados devem ser preparados, quais os algoritmos de DM a implementar e como os implementar, qual o melhor método de análise dos resultados e como se devem tirar elações de qual o melhor algoritmo tornam-se questões a responder. Neste sentido, os objetivos de DM para este projeto são os seguintes:

- Identificar, limpar e preparar as variáveis internas e externas a utilizar.
- Estudar a implementação de algoritmos SVM, NN, ART e ARIMA em trabalhos de previsões de vendas sobre séries temporais.
- Implementar os algoritmos SVM, NN, ART e ARIMA considerando quatro configurações diferentes de valores de entrada, apenas vendas; vendas e variáveis internas; vendas e variáveis externas; vendas e variáveis internas e externas.

- Analisar métodos de comparação de resultados de previsão para classificar qual o algoritmo com melhor desempenho.
- Comparar resultados e apresentar qual o melhor algoritmo e respetiva configuração para realizar a previsão de vendas perante os objetivos de negócio estipulados.

4.1.3 CRITÉRIO DE SUCESSO

Perante os objetivos de negócio e de DM apresentados, foi considerado como espectável obter resultados que permitissem responder aos objetivos propostos.

Considerando os objetivos anteriormente propostos, o trabalho desenvolvido apresenta como critério de sucesso os seguintes pontos:

- Identificar pelo menos um algoritmo que obtenha um erro médio inferior a 20%.
- Identificar qual a combinação de variáveis que melhor se perfilam para auxiliar a previsão de vendas.

4.2 FERRAMENTAS DE *DATA MINING*

O trabalho prático realizado para produzir os resultados apresentados neste projeto teve por base duas ferramentas de DM distintas, SSAS (SQL Server Analysis Services 2012) e Rapid Miner. Derivado do facto de a Primavera BSS ser parceira de negócio da Microsoft, os algoritmos NN, ART e ARIMA foram implementados através do *software* SSAS. Desta forma a Microsoft disponibiliza aos seus utilizadores uma ferramenta intuitiva capaz de construir e entregar bases de dados analíticas para poderem ser utilizadas em sistemas de suporte à decisão. Além disso, a Microsoft procura apresentar aos seus utilizadores uma ferramenta que se integra facilmente com as outras ferramentas da empresa, nomeadamente com o MS SQL Server e com o MS Excel, e que fosse ao mesmo tempo intuitiva para que os utilizadores conseguissem sem um grande nível de conhecimento técnico tirar partido desta ferramenta. Neste sentido, a componente orientada ao DM do SSAS possui alguns algoritmos pré-definidos que permitem através de um interface gráfico intuitivo chegar rapidamente a resultados. Uma vez que todos os sistemas de dados da Primavera BSS assentam em tecnologia Microsoft, a escolha de ferramenta de DM recaiu sobre o sistema SSAS. Embora o SSAS disponha de diversos algoritmos, o sistema apresenta um algoritmo que considera como o mais adequado para

cada grupo de tarefa, existindo assim um algoritmo associado a tarefas de associação, um para segmentação e um para redes neuronais, acabando por limitar as possibilidades de testar diversos algoritmos para a mesma tarefa. Como o algoritmo de SVM é amplamente referido na literatura relacionada com previsão (Yu, X., Qi, Z., Zhao, Y., 2013; Trafalis, T. B., Ince, H., 2000; Lu, C., 2014) e havendo a intenção de utilizar o mesmo para comparar resultados com os resultados obtidos pelos outros três algoritmos, procedeu-se à utilização de outra ferramenta de DM que permitisse implementar o algoritmo de SVM sem ter de desenvolver o mesmo no SSAS.

Perante a necessidade de implementar um algoritmo de SVM, foi necessário encontrar um *software* livre que tivesse uma grande comunidade de utilizadores. Sendo proprietário de uma empresa com o mesmo nome, o RapidMiner, um *software opensource* que conta com uma vasta comunidade de utilizadores. Ao longo da sua existência, esta comunidade contribui com diversas bibliotecas de algoritmos de DM, o que disponibilizou um grande leque de algoritmos a aplicar. Não se ficando apenas pelas tarefas de mineração de dados, o RapidMiner disponibiliza também uma panóplia de funcionalidades bastante grande, que inclui, por exemplo, meios para a extração de dados a partir de diversas fontes (como bases de dados e folhas de cálculo), para a integração de linguagens (como scripts em R), para fazer a limpeza e tratamento de dados, para gerar gráficos de resultados, entre outros.

4.3 Análise de Dados

Nesta secção serão analisados os dados disponíveis, identificadas as suas fontes e as suas características. Este passo é crucial em todo o projeto, pois é aqui que se identifica se os dados necessitam de ser tratados ou não. Uma má análise dos dados pode levar a que estes sejam utilizados de forma incorreta o que pode originar resultados imprevisíveis.

4.3.1 Recolha de dados

Os dados utilizados neste projeto tiveram origem em duas fontes diferentes, sendo as variáveis internas provenientes do ERP Primavera de cada empresa e as variáveis externas do Banco de Portugal. Quanto aos dados provenientes do ERP Primavera, no sistema de planeamento e gestão de recursos é guardada informação sobre cada venda efetuada, cada novo vendedor contratado, cada novo produto criado e cada sucursal existente. Posteriormente, estes dados são agrupados e enviados para um DW (Data Warehouse) existente num servidor centralizado que agrega os valores ao longo do tempo. Por sua vez, o Banco de Portugal possui informação sobre todos os indicadores

que afetam o país, tendo por isso disponível informação sobre as variáveis externas utilizadas. Estes dados são então introduzidos diretamente numa tabela de factos do DW para que seja possível registar a sua evolução ao longo do tempo. Desta forma, os dados são centralizados num único DW armazenando os dados nas tabelas de factos e nas dimensões de variáveis internas e externas como apresentado no esquema da figura 4.5.

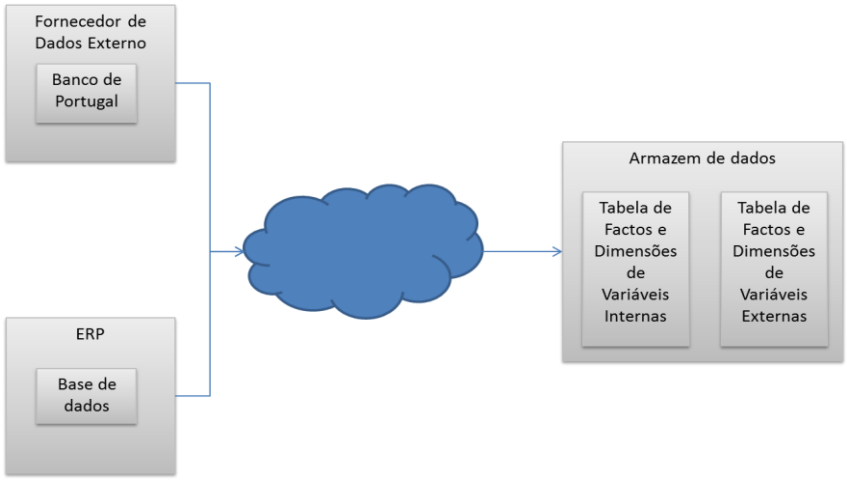


Figura 4.5. Arquitetura de sistema de armazenamento de dados para alimentação dos algoritmos de DM.

As dimensões permitiram contextualizar os factos que representam informação sobre o valor de uma determinada variável para um determinado instante de tempo, contextualizando também em função da empresa no que diz respeito às variáveis internas.

4.3.2 Características da fonte de dados

A fonte de dados será um DW, sendo possível aceder às diversas dimensões e tabelas de factos que este possui. Todavia, só nos interessam as tabelas de factos e as dimensões relacionadas com as previsões de vendas. Para contextualizar a tabela de factos de vendas é necessário fornecer a esta dados das variáveis internas, das variáveis externas, das vendas (produtos a serem vendidos) e das empresas. Desta forma será possível realizar uma análise ao nível da granulidade "previsão de vendas mensais", uma vez que, para cada empresa e para cada conjunto de variáveis, será possível consultar o seu valor num determinado mês (Figura 4.6). Embora algumas destas dimensões não possuam um elevado número de registos, a junção de todos os dados culmina com um conjunto de dados com um tamanho aproximado de dois milhões de registos numa *view* sobre as previsões das quatro empresas.

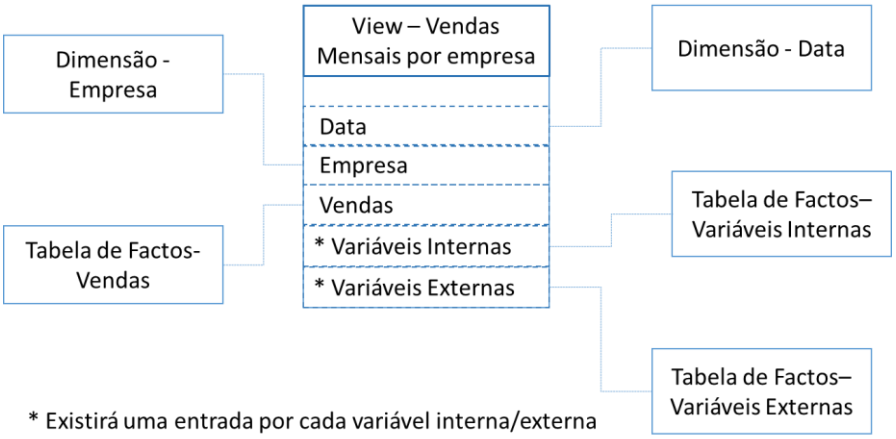


Figura 4.6. Ilustração do esquema da tabela de factos de vendas mensais.

4.3.3 Atributos e Enriquecimento de Dados

Uma vez que todos os atributos em análise são quantitativos, o que vai ao encontro do tipo de dados necessário para a utilização de modelos de previsão com algoritmos regressivos, e o objetivo era realizar a previsão a partir de variáveis internas já existentes ou de varáveis externas provenientes de um provedor externo de dados, não existiu a necessidade de definir novos atributos que pudessem auxiliar a previsão. Desta forma, não foi necessário proceder a qualquer enriquecimento de dados, primeiro porque os dados existentes apresentavam a informação necessária no formato necessário e, em segundo, porque um enriquecimento de dados poderia implicar o aumento de complexidade do sistema para implementações futuras.

4.3.4 Consistência, Integridade e Poluição de Dados

Sendo as variáveis internas provenientes de um sistema de ERP a sua consistência, integridade e possível poluição dos dados é assegurada e verificada de forma geral. Embora existam valores de vendas que se distanciam dos restantes, como é o caso das vendas de Dezembro de 2008 da empresa B e algumas vendas da empresa C, estes valores fora do padrão, *outliers*, não sofreram qualquer tipo de tratamento, tendo sido utilizados nas previsões. Considerando que serão realizadas previsões a doze meses, a exclusão destes valores, que são valores reais e não erros de leitura, poderá implicar a não inclusão de um sinal de mudança do volume de negócio das empresas. Quanto às variáveis externas, cujos valores provêm do Banco de Portugal, após análise verificou-se que estes valores não possuíam qualquer *outlier*, embora apresentem grandes variações nos últimos

anos, refletindo a crise económica internacional que se fez sentir a partir de 2007. Assim, todos os valores encontravam-se dentro do domínio em questão, sendo íntegros e não possuindo ruido.

4.3.5 Valores Nulos

As variáveis internas e as vendas não apresentaram valores nulos. Contudo, algumas variáveis externas são trimestrais, o que implica a existência de valores nulos nos períodos que não coincidem com meses de final de trimestre. Os valores nulos são um dos maiores problemas na aplicação de técnicas de DM, pois a sua existência limita a interpretação os dados e é necessário ter um grande conhecimento de negócio para ser possível interpretar a existência deste tipo de dados corretamente.

4.3.6 Outras características das fontes de dados

De destacar que, os valores iniciais de vendas da empresa D são 0. Foi investigado o porquê destes valores e foi detetado que estes valores se devem ao facto de se ter verificado uma migração de sistemas. Ou seja, a migração dos dados implicou a migração das vendas sem o seu valor, sendo apenas utilizado de forma correta a partir de Abril de 2003. As vendas desta empresa foram, assim, consideradas apenas a partir do quarto mês de 2003, embora existam registos de vendas nas fontes de dados anteriores a essa data.

4.4 O Conjunto de dados de trabalho

Uma vez que os dados utilizados para previsão se encontram agrupados em tabelas de factos e dimensões criadas maioritariamente para a realização de previsões, e como estes de uma forma geral se encontravam consistentes, íntegros e não poluídos, o conjunto de dados de trabalho utilizado é próximo do conjunto de dados iniciais. Para ser possível centralizar todos os dados necessários, foi utilizada uma *View* sobre a tabela de factos de variáveis externas, a tabela de factos das variáveis internas e a tabela de factos de vendas. Na tabela 4.2 podemos ver os dados de trabalho provenientes dessa *View* – *vwSalesForecastTS*.

Tabela 4.2. Descrição das variáveis utilizadas no armazenamento dos dados do processo de DM.

Nome	Fonte	Tipo	Descrição	Observações
Empresa	Dimensão de empresas	Nominal	Nome da empresa	Este atributo é utilizado para distinguir a empresa à qual dizem respeito os valores.
Ano/Mês	Tabela de Factos de Vendas	Numérico	Data de observação	Com este atributo o mês e o ano são registados para que seja possível manter a granulidade das previsões e contextualizar as observações em termos temporais.
Vendas Totais	Tabela de Factos de Vendas	Numérico	Valor de vendas totais mensais	Através deste atributo é registado o total de vendas líquidas faturadas para um determinado mês.
PD	Tabela de Factos de Variáveis Externas	Numérico	Procura Doméstica	Este atributo permite registar a evolução trimestral[1] da procura doméstica.
CP	Tabela de Factos de Variáveis Externas	Numérico	Consumo Privado	Será possível através deste atributo registar trimestralmente qual a variação do consumo privado.
PIB	Tabela de Factos de Variáveis Externas	Numérico	Produto Interno Bruto	Com este atributo é registado trimestralmente qual a variação do PIB.
CC	Tabela de Factos de Variáveis Externas	Numérico	Consumo de cimento (Cimpor + Secil)	Será possível através deste atributo registar qual o consumo mensal de cimento a nível nacional.
IEE	Tabela de Factos de Variáveis Externas	Numérico	Investimento de empresas estrangeiras	Este atributo permite registar mensalmente a variação do investimento de empresas estrangeiras em Portugal.
ICC	Tabela de Factos de Variáveis Externas	Numérico	Índice de confiança do consumidor	É possível com este atributo registar a evolução mensal do índice de confiança do consumidor.
Número de Sucursais	Tabela de Factos de Variáveis Internas	Numérico	Número e sucursais da empresa	Através deste atributo é possível registar se a empresa está ou não em expansão registando o número de sucursais existentes ao longo do tempo.
Número e clientes	Tabela de Factos de Variáveis Internas	Numérico	Número de clientes que consumiram serviços/produtos no presente mês.	Com este atributo será registado o número de clientes que foram servidos em um determinado mês.
Número de vendedores	Tabela de Factos de Variáveis Internas	Numérico	Número de funcionários a efetuarem vendas e/ou prestarem serviços	Através deste atributo é possível registar o número de funcionários que efetuaram vendas/Serviços no presente mês.

Considerando que o algoritmo a selecionar deverá ser suficientemente generalista para ser aplicado de forma satisfatória a empresas com diferentes perfis, os atributos serão sempre utilizados em conjunto, ou seja, serão utilizadas todas as variáveis externas e todas as variáveis internas ou vice-versa. Esta decisão prende-se com o facto de as variáveis que não influenciam uma empresa, poderem, obviamente influenciar uma segunda empresa. Assim sendo, não foi considerado qualquer tipo de relacionamento

[1] Valores trimestrais serão apresentados em três valores mensais para garantir que todos os dados se encontram no mesmo nível de granulidade. Desta forma o valor trimestral é uma média dos restantes valores diminuindo o bias introduzido.

entre os atributos. A variável de previsão escolhida foi a variável "Vendas Totais". As variáveis "Empresa" e "Ano/Mês" serão utilizadas para fins de contextualização dos dados, enquanto as restantes variáveis serão utilizadas para auxiliar a previsão.

4.5 PREPARAÇÃO DOS DADOS

Antes de serem utilizados na aplicação dos modelos de previsão, os dados necessitam de ser preparados para garantir o seu processamento pelos algoritmos selecionados. A preparação dos dados varia com o tipo de algoritmo a utilizar. Considerando que as previsões serão realizadas através de algoritmos de regressão, os dados terão de ser numéricos e não poderão conter valores nulos. Devido à diferença entre escalas dos valores em questão, todas as variáveis serão normalizadas para melhorar o desempenho do algoritmo. Embora seja comum nesta fase proceder à remoção de atributos redundantes, dado que os dados provêm de uma *View* sobre um DW, neste processo não serão removidos quaisquer atributos, uma vez que nenhum deles é redundante - informação mais detalhada sobre os diversos processos de preparação de dados e as técnicas mais comummente utilizadas para realização das mesmas pode ser encontrada em *Data Preparation for Data Mining* de Dorian Pyle (1999).

4.5.1 CONVERSÃO DE DADOS NOMINAIS PARA NUMERAIS

Uma vez que o objetivo era o de aplicar algoritmos de regressão que retornassem valores numerais, é usual transformar as variáveis nominais para que estas possam ser utilizadas como variáveis nos processos de regressão. Embora possam ser aplicadas técnicas como o *Scoring Method*, método que consiste em atribuir pontuações aos possíveis valores das variáveis nominais, não foi necessário neste projeto transformar qualquer atributo nominal, uma vez que todos os atributos eram numéricos.

4.5.2 NORMALIZAÇÃO DE DADOS

Existindo diferentes variáveis provenientes de contextos distintos é natural que as variáveis apresentem valores de grandeza díspares, que podem conduzir a interpretações incorretas por parte dos algoritmos ou a problemas de desempenho. Tanto o algoritmo de NN como o algoritmo de SVM, necessitam de ter os dados normalizados para poderem ser implementados. Embora nem todos os algoritmos necessitem de receber variáveis

normalizadas, este processo pode beneficiar todos os algoritmos tal como referido por Pyle (1999). Tanto na normalização dos dados para o algoritmo de NN como para o algoritmo de SVM, o método utilizado foi o método de transformação *z*, também conhecido por normalização estatística ou pontuação normal (Jayalakshmi e Santhakumaran, 2011) . Este método é baseado na seguinte fórmula:

$$z = \frac{x - \mu}{\sigma}$$

Em que *x* representa os valores dos atributos, aos quais é subtraída a média dos atributos, e se divide a diferença pelo desvio padrão.

4.5.3 Tratamento de Valores Nulos

A existência de valores nulos torna-se bastante penalizadora para os algoritmos de DM, podendo inclusive limitar a execução de alguns algoritmos. Desta forma, aquando da existência de valores nulos é necessário tomar uma decisão que possa permitir a ultrapassagem deste problema. Numa primeira análise existem duas abordagens possíveis, removem-se as variáveis ou se substituem os valores nulos. Uma vez que todas as variáveis são importantes neste problema, a sua remoção nunca foi considerada. Para a substituição de valores nulos existem diversas técnicas tal como reportado em Luengo *et al* (2011). Tal como referido, nenhuma técnica de substituição é a melhor para todos os casos, por isso, com o intuito de manter o modelo o mais simples possível foi adotada uma abordagem através do método *Mean Mode Imputation* (Luengo, García e Herrera, 2011). Segundo este método, o valor em falta será substituído pelo valor médio da variável. No projeto, esta técnica apenas é aplicada nas variáveis PD, CP e PIB., pois estas variáveis como são trimestrais apresentam valores nulos para os dois meses seguintes a cada medição.

4.6 Conclusão

Nesta secção apresentamos os princípios sobre os quais os dados foram obtidos após a análise do negócio e quais as ferramentas a utilizar para aplicar os algoritmos de DM sobre os mesmos. De seguida, são apresentados os dados recolhidos, os dados que serão utilizados para teste, e por fim, quais as transformações necessárias para que os dados estejam prontos a ser utilizados nos modelos de DM. No que diz respeito à análise de

negócio, foram apresentadas quatro empresas das quais foram retirados os valores para o processo de previsão. Foram apresentados os principais objetivos do projeto e realçado o facto de todo o projeto ter surgido de um estágio curricular desenvolvido na Primavera BSS que culminou no estudo de algoritmos de DM para a previsão de vendas líquidas totais mensais das empresas.

Parte do trabalho prático deste projeto foi desenvolvido em paralelo com um estágio na empresa Primavera BSS. Devido a uma parceria tecnológica entre a Primavera BSS e a Microsoft. Esta última influenciou a escolha da principal ferramenta de DM que utilizámos neste trabalho: o SSAS. Contudo, devido à não existência do algoritmo de SVM como algoritmo nativo, foi necessário escolher uma outra ferramenta para que se pudesse aplicar tal algoritmo. Esta escolha recaiu sobre o RapidMiner por ser um *software opensource* com bastantes librarias disponibilizadas *online* e com uma boa comunidade a qual permite uma boa integração e se apresenta disponível para tirar eventuais dúvidas. O conjunto de dados a analisar podem provir de quatro tabelas distintas:

1. a tabela de dimensão de empresas possui a informação proveniente do ERP Primavera sobre as empresas utilizadas;
2. a tabela de factos possui informação total de vendas mensal de cada empresa sendo estes dados provenientes de uma tabela de factos com base nos registos de vendas do ERP da Primavera BSS;
3. a tabela de factos das variáveis internas sendo esta povoada com informação também proveniente do sistema ERP da Primavera BSS;
4. a tabela de factos de variáveis externa é a única que não tem por base o sistema de dados da PrimaveraBss mas sim um conjunto de indicadores proveniente do Banco de Portugal.

Após a análise dos dados verificou-se que a única fonte de dados a apresentar valores nulos é a tabela de factos das variáveis externas. Sendo os indicadores do consumo privado e do produto interno bruto trimestrais, estes indicadores possuem valores nulos de dois em dois meses. Com a análise das fontes de dados concluída foi procedido à sua preparação. Como foram detetados valores nulos foi necessário fazer o seu tratamento. O método escolhido foi a utilização de valores médios para os meses em que não existem valores conhecidos. A média tem por base os dois valores conhecidos temporalmente mais próximos. Foi ainda necessário recorrer à normalização de dados

devido às características específicas do algoritmo de SVM e de NN. A normalização das variáveis de previsão foi realizada com auxílio do método *z-transformation*, permitindo que ambos os algoritmos, NN e SVM, fossem utilizados sem qualquer tipo de restrição. Desta forma os dados ficaram preparados.

CAPÍTULO 5

DESENVOLVIMENTO DOS MODELOS DE PREVISÃO

5.1 AVALIAÇÃO DOS MODELOS E MÉTRICAS DE DESEMPENHO

Uma vez que foram utilizados vários algoritmos no desenvolvimento do trabalho prático deste projeto, a sua modelação variou conforme as necessidades. Porém para que fosse possível comparar o resultado entre todos os algoritmos procurou-se utilizar os mesmos princípios na obtenção de variáveis a utilizar bem como método de avaliação de modelos e aplicação de métricas de desempenho. A obtenção de resultados que satisfaçam os objetivos propostos não pode ser considerada como válida se não forem aplicados métodos de avaliação de desempenho transversais a todos os algoritmos. Em trabalhos de DM usualmente são utilizados como modelos de avaliação de desempenho técnicas como *Cross-Valdiation*, *Bootstrap* ou *Holdout*.

5.1.1 CROSS-VALIDATION

O método *Cross-Validation* consiste em dividir *x* exemplos de dados por *y* partições para que o número de exemplos seja sempre aproximadamente *x/y*, garantindo que o número de elementos se mantem constante por amostra, sendo também garantido que cada elemento não pode ser repetido por amostra. Seguidamente, serão realizados *y* treinos em que uma das partições é utilizada para testes e os restantes dados utilizados para a aprendizagem do modelo. Embora este método permita garantir que todos os dados são utilizados para testes, o facto de as partições serem criadas de forma aleatória pode comprometer um dos princípios dos dados em estudo, que é o facto de estes serem séries temporais. A escolha de elementos aleatórios pode levar à introdução de *bias* no que diz respeito a medições com diferentes espaçamentos temporais ao invés do espaçamento mensal entre medições, bem como mitigação de determinadas características sazonais caso os dados as possuam.

5.1.2 BOOTSRAP

O método de *Bootsrap* é similar ao método de *Cross-Validation* no que diz respeito à escolha de grupos de teste e de treino de algoritmo com a exceção de que os exemplos escolhidos para as amostragens se podem repetir em grupos de amostragem diferentes. Tal como o método *cross-validation* o problema inerente à possibilidade de comprometer as séries temporais é mantido.

5.1.3 HOLDOUT

Ao contrário dos dois métodos apresentados anteriormente, o método de *Holdout* permite escolher uma percentagem da amostragem dos dados a ser usada para treino e consequente para a criação do modelo, sendo os restantes elementos utilizados para testes do modelo. Embora este método seja usualmente mais utilizado para grandes volumes de dados, este é o único modelo que garante que uma determinada sequência de dados com o mesmo espaçamento temporal e que perfaçam ciclos anuais de vendas sejam utilizados. Desta forma foi diminuida a possibilidade de retirar características sazonais ao modelo evitando a possível introdução de bias no sistema. Assim, para validar o modelo é realizada uma análise com base no método de *Holdout*, que ao invés de ser escolhida uma percentagem de dados para validação é definido como intervalo temporal o último ano de dados conhecidos. Deste modo, a validação será sempre correta uma vez que irá garantir que é validado um ano completo de dados, não correndo o risco de não incorporar um período de sazonalidade. A comparação do desempenho dos diferentes algoritmos é realizada com base no MAPE (sigla em inglês mais comumente usada no meio cientifico para Erro Percentual Absoluto Médio). O MAPE é calculado com base na seguinte fórmula:

$$E = \frac{1}{n}\sum_{t=1}^{n} \frac{[x_t - w_t]}{w_t}$$

Assim, torna-se possível calcular qual a variação média do erro ao longo dos diversos instantes de tempo. Considerando que esta fórmula é adequada para analisar o erro em termos de previsões de séries temporais (uma vez que permite analisar o erro ao longo do tempo), foi também considerado que esta apresenta também bons resultados em termos de fiabilidade, validade da construção, sensibilidade e relacionamento com tomadas de decisão. Em análise ao trabalho apresentado por Amstrong em 1992, sobre a utilização de medidas de erro para métodos de previsão, foi concluído que o MAPE era a medida de

erro que melhor se enquadrava nas necessidades deste projeto, validando assim a escolha deste método.

5.2 SELEÇÃO DE ATRIBUTOS

Em relação às variáveis, uma vez que o grupo de quatro variáveis internas e externas são sempre usadas como um todo, será sempre avaliado se todas as variáveis internas ou todas as variáveis externas deverão ser usadas ou não. Ao longo dos desenvolvimentos práticos realizados foram testados os algoritmos selecionados, considerando o envolvimento de variáveis internas e externas, apenas variáveis internas, apenas variáveis externas ou até mesmo a não utilização de variáveis. Desta forma, o método de seleção de atributos utilizado foi o método *Wrapper*, o qual possibilita a análise da utilização ou não de determinados atributos, conforme o desempenho do algoritmo. Com este modelo, para cada um dos grupos de dados em estudo foi aplicado o algoritmo, variando a utilização dos conjuntos de variáveis tal como planeado. No final, os diversos resultados foram analisados para perceber se, de facto, a inclusão ou não de determinadas variáveis permitiu obter melhores resultados. Esta análise será apresentada e discutida junto com os restantes resultados.

5.3 ALGORITMOS DE DATA MINING

Como referido anteriormente, este estudo teve por base um conjunto de algoritmos nomeadamente NN, SVM, ART, ARIMA e ART + ARIMA. Os algoritmos NN, ART, ARIMA e ART + ARIMA foram implementados utilizando-se a ferramenta SSAS. Sendo algoritmos distintos, os parâmetros otimizados em cada um dos algoritmos foram diferentes. No que diz respeito ao algoritmo de NN este foi parametrizado e analisado conforme a utilização de 30% ou 60% de dados para criação do modelo bem como a utilização de 1, 4 ou 8 níveis de nodos ocultos. Os algoritmos de ARIMA, ARTXP e ARIMA + ARTXP foram modelados a partir do mesmo modelo do SSAS. Neste modelo foram alteradas os parâmetros Forecast *Method*, HMC (*Historic Model Count*) e HMG*(Historic Model Gap*). O parâmetro *Forecast Method* permite escolher qual dos três modelos se pretende aplicar. Por sua vez os parâmetros HMC, quantidade de modelos históricos a ser construídos, e HMG, intervalo de tempo entre dois modelos consecutivos, foram alternados durante os testes para perceber como estas variações poderiam

influenciar cada um dos modelos No algoritmo NN, os parâmetros que foram tidos em consideração foram os parâmetros *Hidden Node Ratio, Hidden Node Percentage* e *Holdout Seed*. O valor de *Holdout Seed* foi construído com base no nome do modelo, a função utiliza a palavra que representa o nome do modelo para gerar um valor numérico por forma a garantir que este não variava ao longo dos testes. Os outros valores foram alterados procurando identificar qual o impacto que a utilização de maiores ou menores quantidades de dados para treino e modelação dos modelos (variar *Hidden Node Ratio*) poderia ter no resultado das previsões bem como a utilização de um maior ou menor número de nodos ocultos (variação do parâmetro *Hidden Node Ratio)*. Contrariamente aos modelos anteriores, o modelo de SVM foi implementado com base no *software* RapidMiner. Devido às suas características, a otimização de parâmetros a utilizar baseou-se na escolha do tipo de *kernel* e de SVM a utilizar num primeiro ponto. No que diz respeito ao tipo de algoritmo de SVM a utilizar, foi escolhido o algoritmo ε-SVR por este permitir modelar um algoritmo de SVM para regressão clássico tal como apresentado anteriormente. O tipo de *kernel* escolhido foi um *kernel* do tipo RBF (*Radial Basis Function*) pois a sua capacidade de realizar mapeamentos não lineares torna-se ideal para detetar a relação entre os diversos atributos. Em alguns testes com um dos modelos de dados foi também detetado que este *kernel* apresentava melhores resultados do que *kernel* polinomiais ou sigmoids. Considerando o *kernel* e o algoritmo escolhido, foram então definidos como parâmetros variáveis o valor de *gamma*, *C* e ε em que gamma corresponde ao valor da função de *kernel*, C o custo associado ao aumento de erros e ε é o valor que representa a tolerância ao critério de término do algoritmo. Foi também detetado durante testes preliminares que o único parâmetro a apresenta uma influência significativa nas previsões foi o parâmetro ε, o qual influencia diretamente o número de iterações que o algoritmo vai realizar até achar o seu modelo ótimo, independentemente de este ser um ótimo local ou global. Para o modelo de SVM foi também considerada a utilização de janelas de series temporais para que fosse possível modelar o mesmo a séries temporais, tendo sido variado o HMC e o HMG.

5.4 O PROCESSO DE ITERAÇÕES

A implementação dos diversos modelos teve por base uma série de iterações que permitiram fazer o seu teste perante diferentes combinações de variáveis e parâmetros. A sistematização das iterações permitiu também diminuir possíveis erros e garantir que

todas as combinações foram testadas. Desta forma para cada um dos modelos foram realizadas várias iterações que permitissem em primeiro lugar, escolher a combinação de variáveis a utilizar. Depois de escolhida a combinação de variáveis, foi escolhida a combinação de parâmetros. Aplicou-se, de seguida, o modelo. Quando todas as combinações de parâmetros forem testadas, foi realizado o teste com outra combinação de variáveis, sendo feito assim, sucessivamente, até se garantir que todas as combinações possíveis foram utilizadas. Na figura 5.1 podemos observar o modelo de iterações adotado.

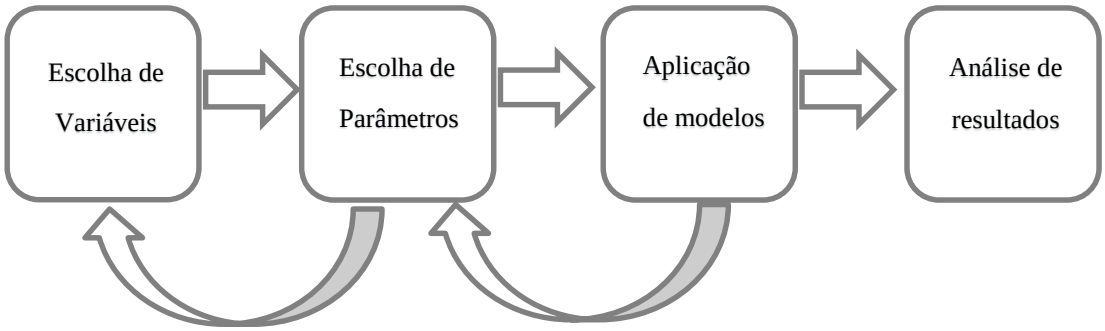


Figura 5.1. Sequência de iterações adotada na aplicação dos modelos.

No decorrer das aplicações as variáveis utilizadas em cada um dos modelos foram aquelas que estão apresentadas na tabela 5.1.

Tabela 5.1. Valores utilizados como variáveis e parâmetros.

Algoritmo	Variáveis: Valores	Parâmetros
ARIMA	HMG: 1,2,6,12 HMC: 1,2,6,12	Variáveis Internas + Vendas; Variáveis Externas mais vendas; Variáveis Internas + Variáveis Externas + Vendas; Vendas
ARTXP	Historical Model Gap: 1,2,6,12 Historic Model Count: 1,2,6,12	
ARIMA + ARTXP	Historical Model Gap: 1,2,6,12 Historic Model Count: 1,2,6,12	
NN	Hidden Node Ratio: 1,4,8 Holdout Percentage: 30%,60%	
SVM	E: 0.001,1 HMC: 2,6,12 HMG: 1,2,6	

A única exceção à regra de iterações adotada foi verificada com o modelo ARIMA, uma vez que este apenas realiza as seguintes combinações de pares de (*Historical Model Gap, Historic Model Count*): (1,12), (12,1), (2,6), (6,2). Deste modo, é expectável fazer-se o teste do modelo perante a utilização de uma janela temporal de doze meses, doze janelas temporais de um mês, duas janelas temporais de seis meses, ou seis janelas temporais de dois meses.

5.5 CONCLUSÃO

A modelação do problema é o culminar da recolha de informação que será utilizada para implementar os modelos de DM escolhidos. Depois de ser conhecido o âmbito do trabalho, recolhida informação sobre quais os algoritmos a utilizar e quais as ferramentas a utilizar na sua implementação, é, então, primordial delinear um modelo de implementação. Numa primeira fase é definido como serão avaliados os modelos e qual a métrica de desempenho a utilizar. Uma vez que são estabelecidos objetivos, este passo torna-se importante para que seja possível definir qual a melhor maneira de avaliar os diferentes modelos, introduzindo o menor *bias* possível na comparação de desempenho. Desta forma, foi delineado que o melhor modo de comparar os modelos seria com base no MAPE pois seria assim possível considerar o erro ao longo das séries temporais. Para validação dos modelos foi definido que seriam utilizados os últimos 12 meses de dados permitindo assim avaliar a qualidade da previsão ao longo de um ano completo. Devido às considerações já apresentadas, a escolha dos atributos assumem um papel global, ou seja, não é analisado a utilização ou não de uma variável mas sim o impacto da utilização de grupos de variáveis. Assim sendo, as combinações de atributos utilizados tem por base as vendas e a utilização de variáveis internas ou externas. Por fim, para que seja possível garantir que a aplicação dos modelos é executada de forma a cobrir todos os casos é realizado o planeamento das diversas iterações. Num primeiro ponto é escolhido o algoritmo a utilizar. Depois, para esse algoritmo, definem-se quais os parâmetros a utilizar executando-o para todas as combinações de variáveis. Quando todas as combinações de variáveis já foram utilizadas, é então seleccionado o próximo parâmetro, garantindo que são executados todos os modelos para todas as combinações de parâmetros e variáveis.

CAPÍTULO 6

APRESENTAÇÃO E DISCUSSÃO DE RESULTADOS

6.1 ANÁLISE DE RESULTADOS

A execução dos modelos de DM tem como resultado um conjunto de previsões para os últimos doze meses de dados conhecidos para cada empresa em estudo. Assim sendo, de seguida, serão apresentados para cada um dos algoritmos os resultados obtidos para todas as firmas, para que seja possível numa primeria fase analisar o desempenho individual de cada um dos algoritmos e depois, comparar os resultados entre todos os algoritmos.

6.1.1 ARIMA

O primeiro modelo a ser executado foi o modelo ARIMA. Devido às suas características, este algoritmo não permite a utilização de variáveis que auxiliem a previsão de vendas pelo que não foram desta forma testadas iterações com variação de variáveis internas ou externas. Embora esta limitação seja inegável, a sua implementação enquanto algoritmo de baixo custo é deveras aliciante para um estudo que se foca em encontrar o algoritmo com melhor desempenho e que apresente um baixo esforço de implementação. Tendo isto em consideração, foram realizadas quatro iterações para cada empresa sendo alternado entre a combinação de valores de HMG e HMC. A tabela 6.1 apresenta o MAPE para os últimos doze meses de valores conhecidos para cada combinação de variáveis.

Tabela 6.1. MAPE obtido para o modelo ARIMA.

Modelo	**Combinação de parâmetros**	**MAPE por empresa**				**MAPE total**
		A	B	C	D	
ARIMA	HMC:1;HMG:12	17,52	44,86	143,90	16,98	55,81
	HMC:12;HMG:1	29,71	69,72	126,07	15,42	60,23
	HMC:2;HMG:6	30,44	63,82	108,42	40,14	60,70
	HMC:6;HMG:2	47,07	99,47	102,06	23,62	68,05
Legenda	HMC: *Historical Model Count*					
	HMG: *Historical Model Gap*					

Como é possível analisar na tabela 6.1, a combinação de parâmetros que corresponde a um melhor desempenho é a utilização de um HMG de doze e um HMC de um com

MAPE de 55,8%. Embora este modelo apresente um MAPE elevado, não deixa de ser verdade que, de todos os modelos aplicados, este é o mais simplista. O menor MAPE é contudo obtido com o modelo que possui um HMG de um e um HMC de doze com o valor de 15,4% para a empresa D.

6.1.2 ARTXP

Ao contrário do modelo anterior, o modelo ARTXP permite que sejam combinadas as variáveis internas e externas com as vendas por forma a criar correlações que permitam auxiliar os modelos na execução dos algoritmos para se obter as previsões de vendas. Desta forma, serão utilizados em todos as variações apresentadas deste modelo em diante o MAPE obtido, não só com a variação dos parâmetros, mas também com a utilização (ou não) de variáveis auxiliares à previsão. De seguida, na tabela 6.2, encontram-se os MAPE obtidos para o modelo de ARTXP

Tabela 6.2. MAPE obtido para o modelo ARTXP.

Modelo	**Combinação de parâmetros**	**Combinação de variáveis**	**MAPE por empresa**				**MAPE total**
			A	B	C	D	
ARTXP	HMC:1;HMG:12	Vendas	23,73	87,51	70,80	10,40	48,11
		Vendas + Variáveis internas	23,98	82,69	70,80	9,90	46,85
		Vendas + Variáveis exernas	73,81	61,66	186,21	10,07	82,94
		Todas as variáveis	15,73	45,28	143,54	13,94	54,62
	HMC:12;HMG:1	Vendas	23,10	33,47	85,96	10,49	38,26
		Vendas + Variáveis internas	27,21	37,27	162,28	10,25	59,25
		Vendas + Variáveis exernas	20,34	39,70	90,80	10,49	40,33
		Todas as variáveis	22,64	48,93	84,19	10,51	41,57
	HMC:2;HMG:6	Vendas	30,91	34,04	95,79	10,39	42,78
		Vendas + Variáveis internas	24,23	50,13	424,23	9,95	127,14
		Vendas + Variáveis exernas	34,54	43,40	95,58	10,12	45,91
		Todas as variáveis	22,45	51,38	86,72	24,68	46,31
	HMC:6;HMG:2	Vendas	27,22	31,14	81,53	10,03	37,48
		Vendas + Variáveis internas	27,80	26,64	185,66	9,85	62,49
		Vendas + Variáveis exernas	24,73	41,77	98,75	10,01	43,82
		Todas as variáveis	32,60	70,38	84,64	15,51	50,78

Legenda	HC: *Historical Count*
	HG: *Historical Gap*

Observando os resultados obtidos é possível concluir que tanto o MAPE mais baixo para uma empresa em específico como para o conjunto de todas as empresas é obtido através da combinação dos parâmetros HMC:6 e HMG:2. Considerando que o segundo melhor resultado, este resultado é obtido com base nos parâmetros HMC:12 e HMG:1, o que indica que o algoritmo ARTXP consegue obter melhores resultados quando generalizado para empresas de diferentes características se forem considerados HMC's superiores aos HMG's. Porém, este algoritmo apresenta um melhor MAPE para uma combinação de parâmetros que apresentou o pior resultado aquando da sua aplicação a todas as empresas com o algoritmo ARIMA. Assim, será interessante verificar nos próximos resultados, quais os parâmetros e combinações de variáveis que permitem obter melhores resultados quando só dois algoritmos são utilizados em conjunto.

6.1.3 ARIMA + ARTXP

Com base nos modelos ARIMA e ARTXP, o modelo ARTXP + ARIMA permite verificar quais os melhores resultados quando se utiliza uma combinação dos dois algoritmos para realizar as previsões e assim verificar o MAPE das previsões. Considerando as mesmas variações de parâmetros e de variáveis utilizadas no modelo ARTXP, os resultados obtidos podem ser observados na tabela 6.3.

Tabela 6.3. MAPE obtido para o modelo ARIMA + ARTXP.

Modelo	Combinação de parâmetros	Combinação de variáveis	MAPE por empresa				MAPE total
			A	B	C	D	
Misto	HMC:1;HMG:12	Vendas	16,46	28,69	123,11	14,40	45,67
		Vendas + Variáveis internas	16,05	43,02	123,11	14,24	49,11
		Vendas + Variáveis exernas	15,92	45,25	143,62	14,32	54,78
		Todas as variáveis	15,73	45,28	143,54	13,94	54,62
	HMC:12;HMG:1	Vendas	24,16	44,93	98,80	10,41	44,57
		Vendas + Variáveis internas	26,11	49,30	136,96	10,59	55,74
		Vendas + Variáveis exernas	19,24	54,71	93,01	10,63	44,40
		Todas as variáveis	22,64	48,93	84,19	10,51	41,57
	HMC:2;HMG:6	Vendas	29,46	44,06	91,06	25,30	47,47
		Vendas + Variáveis internas	26,63	49,13	218,90	25,16	79,96
		Vendas + Variáveis exernas	19,28	53,77	95,66	25,25	48,49
		Todas as variáveis	22,45	51,38	86,72	24,68	46,31
	HMC:6;HMG:2	Vendas	36,71	62,19	87,40	15,05	50,34
		Vendas + Variáveis internas	36,88	61,06	137,80	14,95	62,67
		Vendas + Variáveis exernas	27,84	72,49	93,39	15,03	52,19
		Todas as variáveis	32,60	70,38	84,64	15,51	50,78

Legenda:	HC: *Historical Count*
	HG: *Historical Gap*

Analisando os resultados apresentados na tabela 6.3, verifica-se que o melhor MAPE obtido para todas as empresas foi de 41,57% e o melhor MAPE global de 10,41%. Ambos os resultados têm em comum o facto de apresentarem melhores valores perante uma combinação dos parâmetros HMC:12 e HMG:1. Apresentando um melhor resultado global para uma combinação de parâmetros HMC:12 e HMG:1, este modelo não apresenta, porém, resultados que vão ao encontro dos dois modelos anteriores. Contudo a combinação de parâmetros que origina o segundo melhor resultado de cada um dos dois modelos anteriores coincide com a combinação de parâmetros que apresenta o melhor resultado do presente modelo. Avaliando os três modelos aplicados é possível também identificar que a empresa que apresenta melhores resultados é a empresa D e a que apresenta piores resultados é a empresa C. Esta constatação leva a pensar que as características da empresa D a tornam uma empresa com características que beneficiam a previsão, ao contrário da empresa C, ou são os próprios modelos que se adaptam muito melhor às características da empresa D, em oposto à empresa C.

6.1.4 NN

De seguida procedeu-se à análise dos resultados do algoritmo NN de forma a perceber como é que a utilização de um algoritmo que assenta em bases teóricas diferentes pode trazer melhores resultados globais e obter um "melhor resultado" e um "pior resultado" individual para empresas diferentes.

As NN são por si só bastante "maleáveis" devido à sua grande capacidade de aprendizagem. Como tal, será interessante analisar os próximos resultados de forma a perceber se a utilização de um algoritmo com forte capacidade de generalização nos trás vantagens na realização de previsões de vendas, quando aplicado em empresas de diferentes características. O algoritmo de NN utilizado apresenta, porém, uma característica que limita as combinações de variáveis a utilizar, que é o facto de não possuir a capacidade de realizar previsões apenas com uma variável. Desta forma, não é possível realizar previsões utilizando apenas os valores das vendas. Logo todas as iterações utilizam as vendas como elemento da previsão. De seguida, na tabela 6.4 podemos observar a variação do número de HiddenNodes e a combinação de variáveis perante a utilização de um *Holdout* de 30%, enquanto na tabela 6.5 podemos observar a mesma variação de parâmetros e variáveis perante a utilização de um Holdout de 60%.

Tabela 6.4. MAPE obtido para o modelo NN com 30% de Holdout.

Modelo	**Combinação de parâmetros**	**Combinação de variáveis**	**MAPE por empresa**				**MAPE total**
			A	B	C	D	
NN H:30%	HN:1	Vendas + Variáveis internas	38,03	42,06	152,23	15,11	61,86
		Vendas + Variáveis exernas	32,77	55,84	82,19	17,01	46,95
		Todas as variáveis	37,69	30,30	86,95	12,89	41,96
	HN:4	Vendas + Variáveis internas	26,78	21,92	86,80	19,91	38,85
		Vendas + Variáveis exernas	37,04	60,83	101,58	17,46	54,23
		Todas as variáveis	27,16	22,73	106,93	13,69	42,63
	HN:8	Vendas + Variáveis internas	22,90	24,03	109,18	22,59	44,67
		Vendas + Variáveis exernas	136,53	171,75	72,33	95,34	118,99
		Todas as variáveis	32,57	22,95	96,19	12,74	41,11

Legenda: H: *Holdout*

HN: *HiddenNode*

Tabela 6.5. MAPE obtido para o modelo NN com 60% de Holdout.

Modelo	Combinação de parâmetros	Combinação de variáveis	MAPE por empresa				MAPE total
			A	B	C	D	
NN H:60%	HN:1	Vendas + Variáveis internas	26,38	34,65	122,81	14,10	49,48
		Vendas + Variáveis exernas	38,15	55,14	109,95	14,38	54,41
		Todas as variáveis	47,37	27,42	93,60	11,33	44,93
	HN:4	Vendas + Variáveis internas	33,11	30,71	75,16	17,61	39,15
		Vendas + Variáveis exernas	37,80	55,36	92,42	17,48	50,77
		Todas as variáveis	32,18	24,15	101,37	14,80	43,12
	HN:8	Vendas + Variáveis internas	25,59	27,99	114,62	19,93	47,03
		Vendas + Variáveis exernas	41,65	53,99	100,28	17,44	53,34
		Todas as variáveis	33,57	24,04	95,38	15,50	42,12
Legenda:	H: *Holdout*						
	HN: *HiddenNode*						

Os resultados apresentados na tabela 6.4 permitem verificar que 4 níveis de *hidde nnodes* permitem atingir um MAPE mais baixo (38,85%), se utilizadas variáveis internas juntamente com as vendas. O melhor desempenho em termos individuais é em relação à empresa D, que tem como parâmetros 8 *hidden nodes* e utiliza todas as variáveis obtendo um MAPE de 12,74%. Naquilo que diz respeito à tabela 6.5 é possível constatar que o melhor resultado global (39,15%) é também obtido perante a utilização de 4 *hidden nodes* e o conjunto de vendas e variáveis internas. Embora a melhor previsão individual seja também para a empresa D, esta é obtida com base em apenas um nível de *hidden nodes* e considerando tanto variáveis internas como externas para auxiliar a previsão sendo o seu MAPE de 11,33%. É deveras curioso constatar que apenas uma diferença de 0,15% impedem que a utilização dos mesmos parâmetros e variáveis produza o melhor resultado de previsão para apenas uma empresa, pois o segundo melhor resultado do modelo baseado em NN, com Holdout de 30% é de 12,89%, para a utilização de apenas um nível de *hidden nodes* e todas as variáveis.

6.1.5 SVM

As iterações com base no algoritmo de SVM são apresentadas conforme o valor de ε, pois é este valor que vai influenciar o processamento realizado até se encontrar um modelo ótimo local. Em cada uma das tabelas são apresentados os resultados segundo as diversas combinações de variáveis considerando três combinações de parâmetros, nomeadamente HC e HG, 12:1,6:2 e 2:6, respetivamente. De seguida, apresentam-se as tabelas 6.6 e 6.7. Nelas é possível verificar o desempenho do algoritmo de SVM ao longo de diversas iterações para as diversas combinações de parâmetros e de variáveis.

Tabela 6.6. MAPE obtido para o modelo SVM COM ε = 0.0001.

Modelo	**Combinação de parâmetros**	**Combinação de variáveis**	**MAPE por empresa**				**MAPE total**
			A	B	C	D	
SVM ε=0.0001	HC:12;HG:1	Vendas	9,89	11,34	6,71	0,47	10,18
		Vendas + Variáveis internas	27,82	73,87	67,48	12,76	42,82
		Vendas + Variáveis exernas	9,76	11,27	6,78	2,11	7,48
		Todas as variáveis	9,82	11,28	6,78	2,11	7,09
	HC:6;HG:2	Vendas	9,88	11,37	6,75	0,48	10,19
		Vendas + Variáveis internas	27,84	73,84	67,49	12,76	42,82
		Vendas + Variáveis exernas	9,76	11,23	6,72	2,11	7,46
		Todas as variáveis	9,82	11,28	6,78	2,11	7,09
	HC:2;HG:6	Vendas	9,88	11,37	6,75	0,48	10,19
		Vendas + Variáveis internas	27,82	73,87	67,48	12,76	42,82
		Vendas + Variáveis exernas	9,76	11,27	6,78	2,11	7,48
		Todas as variáveis	9,82	11,28	6,78	2,11	6,97

Legenda: HC: *Historical Count*

HG: *Historical Gap*

Tabela 6.7. MAPE obtido para o modelo SVM COM ε = 1.

Modelo	Combinação de parâmetros	Combinação de variáveis	MAPE por empresa				MAPE total
			A	B	C	D	
SVM ε=1	HC:12;HG:1	Vendas	19,62	61,49	39,69	9,52	32,58
		Vendas + Variáveis internas	22,51	85,54	84,91	12,56	51,38
		Vendas + Variáveis exernas	19,62	61,49	39,42	8,76	32,32
		Todas as variáveis	19,62	61,49	39,42	8,76	32,32
	HC:6;HG:2	Vendas	19,62	61,49	39,69	9,52	32,58
		Vendas + Variáveis internas	22,51	85,54	84,91	12,56	51,38
		Vendas + Variáveis externas	19,62	61,49	39,42	8,76	32,32
		Todas as variáveis	19,62	61,49	39,42	8,76	32,32
	HC:2;HG:6	Vendas	19,62	61,49	39,69	9,52	32,58
		Vendas + Variáveis internas	22,51	85,54	84,91	12,56	51,38
		Vendas + Variáveis externas	19,62	61,49	39,42	8,76	32,32
		Todas as variáveis	19,62	61,49	39,42	8,76	32,32
Legenda:	HC: *Historical Count*						
	HG: *Historical Gap*						

Com base nos resultados apresentados na tabela 6.6 (MAPE obtido para o modelo SVM COM ε = 0.0001) é possível constatar que os melhores resultados obtidos a nível individual e a nível geral se encontram em extremos opostos. Enquanto o MAPE mais baixo a nível individual (0,47%) é obtido para um HC e HG de 12 e 1, respetivamente, e sem auxílio de variáveis internas ou externas, o MAPE mais baixo para todas as empresas (6,97%) é obtido para um HC de 2 e um HC de 6, considerando tanto variáveis internas como variáveis externas para auxiliar a previsão. Por sua vez, a tabela 6.7 demonstra que sempre que foram realizadas iterações com todas as variáveis ou com variáveis externas o resultado culmina com o mesmo MAPE de 8,76% e 32,32% para uma avaliação individual e global respetivamente. Este resultado indica assim que para um ε=1 existe uma grande probabilidade de que as variáveis externas tenham um impacto preponderante na obtenção de resultados.

6.2 ANÁLISE GLOBAL

Partindo dos resultados obtidos para cada algoritmo, é possível realizar uma comparação de todos os algoritmos a um nível individual e a um nível global. Uma comparação de todos os algoritmos em termos individuais permitirá identificar qual o algoritmo que tem melhor capacidade de previsão para determinadas características das empresas. Por sua vez, a avaliação global permitirá também observar o resultado que é o principal foco deste estudo: qual o algoritmo que de forma global consegue ter um melhor desempenho de previsão quando aplicado a empresas com diferentes características. Para realizar esta comparação será apresentada uma tabela contendo em cada uma das suas linhas pelo menos um "melhor resultado individual" (ou global) relativamente a cada algoritmo. O melhor resultado do algoritmo em relação à empresa será identificado com um fundo verde, sendo apenas apresentadas as combinações de variáveis e parâmetros que levaram a esse resultado. A tabela 6.8 terá, então, tantas linhas quantas necessárias para representar todos os melhores resultados de cada algoritmo, apresentando mais do que uma linha, se necessário, caso exista mais do que um "melhor resultado" para um dado algoritmo.

Tabela 6.8. Melhor MAPE obtido para todas as empresas.

Modelo	Combinação de parâmetros	Combinação de variáveis	MAPE por empresa				MAPE TOTAL
			A	B	C	D	
ARIMA	HC:1,HG:12	-	17,52	44,86	-	-	55,81
ARIMA	HC:12,HG:1	-	-	-	-	15,42	-
ARIMA	HC:6,HG:2	-	-	-	102,60	-	-
ARTXP	HC:1,HG:12	Todas as variáveis	15,73	-	-	-	-
ARTXP	HC:6,HG:2	Vendas + variáveis Internas	-	26,64	-	9,85	-
ARTXP	HC:1,HC12	Vendas	-	-	70,80	-	-
ARTXP	HC:1,HG12	Vendas + Variáveis Internas	-	-	70,80	-	-
ARTXP	HC:6,HG:2	Vendas	-	-	-	-	37,48
MISTO	HC:1,HG:12	Todas as variáveis	15,73	-	-	-	-
MISTO	HC1,HG:12	Vendas	-	28,69	-	-	-
MISTO	HC:12,HG:1	Todas as variáveis	-	-	84,19	-	41,57
MISTO	HC:12,HG:1	Vendas	-	-	-	10,41	-
NN	H:30%,HN:8	Vendas + Variáveis Internas	22,90	-	-	-	-
NN	H:30%,HN:4	Vendas + Variáveis Internas	-	21,92	-	-	38,85
NN	H:30%,HN:8	Vendas + Variáveis externas	-	-	72,23	-	-
NN	H:30%,HN:8	Todas as variáveis	-	-	-	12,74	-
NN	H:60%,HN:8	Vendas + variáveis internas	25,59	-	-	-	-
NN	H:60%,HN:4	Todas as variáveis	-	24,15	-	-	-

NN	H:60%,HN:4	Vendas + variáveis internas	-	-	75,16	-	39,15
NN	H:60%,HN:1	Todas as variáveis	-	-	-	11,33	-
SVM	ε=0.0001,HC:12,HG:1	Vendas + Variáveis Externas	9,76	-	-	-	-
SVM	ε=0.0001,HC:6,HG:2	Vendas + Variáveis Externas	9,76	11,23	-	-	-
SVM	ε=0.0001,HC:2,HG:6	Vendas + Variáveis Externas	9,76	-	-	-	-
SVM	ε=0.0001,HC:12,HG:1	Vendas	-	-	6,71	0,47	-
SVM	ε=0.0001,HC:2,HG:6	Todas as variáveis	-	-	-	-	6,97
SVM	ε=1,HC:12,HG:1	Vendas	19,62	61,49	-	-	-
SVM	ε=1,HC:12,HG:1	Vendas + Variáveis externas	19,62	61,49	39,42	8,76	32,32
SVM	ε=1,HC:12,HG:1	Todas as variáveis	19,62	61,49	39,42	8,76	32,32
SVM	ε=1,HC:6,HG:2	Vendas	19,62	61,49	-	-	-
SVM	ε=1,HC:6,HG:2	Vendas + Variáveis externas	19,62	61,49	39,42	8,76	32,32
SVM	ε=1,HC:6,HG:2	Todas as variáveis	19,62	61,49	39,42	8,76	32,32
SVM	ε=1,HC:2,HG:6	Vendas	19,62	61,49	-	-	-
SVM	ε=1,HC:2,HG:6	Vendas + Variáveis externas	19,62	61,49	39,42	8,76	32,32
SVM	ε=1,HC:2,HG:6	Todas as variáveis	19,62	61,49	39,42	8,76	32,32

Legenda: HC: Historical Count

HG: Historical Gap

HO: Holdout

HN: HiddenNode

Analisando os resultados obtidos, é possível constatar que os melhores resultados foram obtidos unanimemente a partir do algoritmo de SVM com ε=0.0001. Contudo, se por um lado se verifica uma unanimidade quanto ao algoritmo que obtém os melhores resultados, o mesmo não acontece quanto à combinação dos parâmetros HC/HG e das respetivas variáveis. No que diz respeito aos parâmetros HC e HG, tanto a combinação HC:2/HG:6 como a combinação HC:6/HG:2 apresentam dois resultados cada uma delas com o MAPE mais baixo. De salientar que o melhor MAPE obtido a nível global é conseguido para um HC de 2 e um HG de 6. A combinação que apresenta um maior número de resultados com menor MAPE é a combinação HC:12 e HG:1 com três resultados. Analisando as variáveis utilizadas é percetível que entre todos os melhores resultados obtidos, nenhum utiliza apenas as variáveis internas para auxiliar a previsão. Este facto sugere dois cenários possíveis: 1) as variáveis internas utilizadas não são as mais adequadas perante as empresas analisadas, ou 2) as empresas analisadas possuem características que as tornam menos suscetíveis às influências das variáveis internas utilizadas. Das restantes combinações de variáveis utilizadas, tanto a utilização de vendas e variáveis externas como apenas a de vendas para realizar as previsões apresentam duas empresas cada com os melhores resultados de previsão. Contudo o melhor resultado geral é obtido para a utilização de todas as variáveis (vendas, variáveis externas e variáveis internas) em conjunto.

Considerando os resultados obtidos após a aplicação dos diversos algoritmos de DM com as restrições e parametrizações já apresentadas, os melhores resultados alcançados foram de 9,76% de MAPE para a empresa A, 11,23% de MAPE para a empresa B, 6,71% de MAPE para a empresa C, 0,47% de MAPE para a empresa D e 6,97% de MAPE médio para todas as empresas. Na tabela 6.9 e na Figura 6 é possível ver os valores previstos para os últimos 12 meses para cada uma das empresas, sendo apresentado para cada uma delas a melhor previsão e a melhor previsão global encontrando-se todos os dados em unidades monetárias (UM).

Tabela 6.9. Melhor previsão local e previsão global para os últimos 12 meses da empresa A.

Empresa	Data	Valor Real	Melhor Previsão Local	Melhor Previsão Global
A	31-05-2011	11689623,49	12408348,56	12415866,55
	30-06-2011	12068702,89	12408348,56	12415866,55
	31-07-2011	21401285,13	18981666,42	18989184,41
	31-08-2011	14928188,37	14011323,97	14005648,19
	30-09-2011	8262120,278	9179174,441	9186177,567
	31-10-2011	23770228,36	18981666,42	18989184,41
	30-11-2011	11190402,28	12107607,83	12116280,85
	31-12-2011	8770727,809	9687635,283	9695739,282
	31-01-2012	9762207,005	10678931,84	10685538,65
	29-02-2012	6056127,112	6972939,78	6979362,866
	31-03-2012	8779527,113	9696402,803	9703584,403
	30-04-2012	10624912,56	11541794,7	11548675,02
	31-05-2012	11595830,9	12408348,56	12415866,55

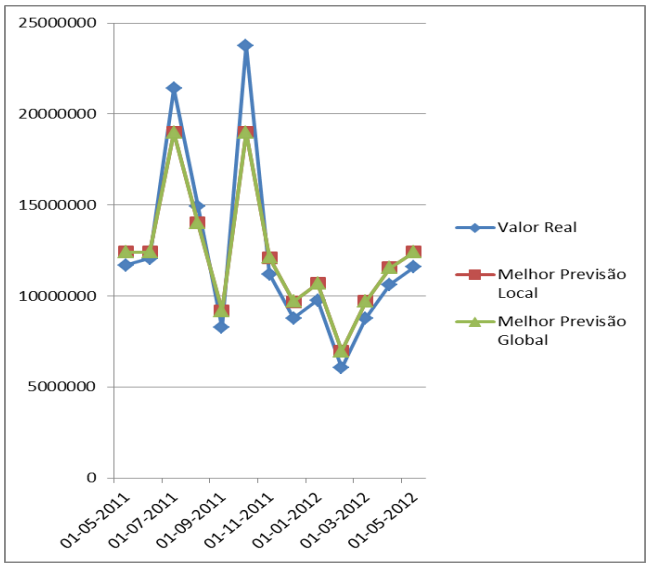


Figura 6.1. Representação gráfica da melhor previsão local e previsão global para os últimos 12 meses da empresa A.

Através da análise do gráfico é possível constatar que mesmo tendo vários pontos em que os valores previstos e os valores reais são ligeiramente diferentes, a variação dos

mesmos foi sempre seguida pelo algoritmo de previsão. Foi também possível constatar que o algoritmo não consegue acompanhar as mudanças mais repentinas de valores mesmo que consiga identificar que de facto iria existir uma mudança de valores como no caso do mês 7 e 10 de 2011 (consultar tabela 6.9).

Tabela 6.10. Melhor previsão local e previsão global para os últimos 12 meses da empresa B.

Empresa	Data	Valor Real	Melhor Previsão Local	Melhor Previsão Global
B	31-01-2012	280339	280347,1163	280439,7753
	29-02-2012	245885	245894,1036	246082,8631
	31-03-2012	308941	308948,1193	309094,8341
	30-04-2012	284945	284954,2149	285013,6538
	31-05-2012	306229	306235,9532	306436,2417
	30-06-2012	252803	252810,7213	252864,3343
	31-07-2012	288154	288161,2451	288219,9367
	31-08-2012	46507,6	102246,9384	101896,5207
	30-09-2012	257846	257855,0623	258047,4203
	31-10-2012	346639	346647,9501	346727,2168
	30-11-2012	306912	306918,7617	307145,0513
	31-12-2012	80167,2	102246,9384	101896,5207
	31-01-2013	289742	289751,3205	289788,8719

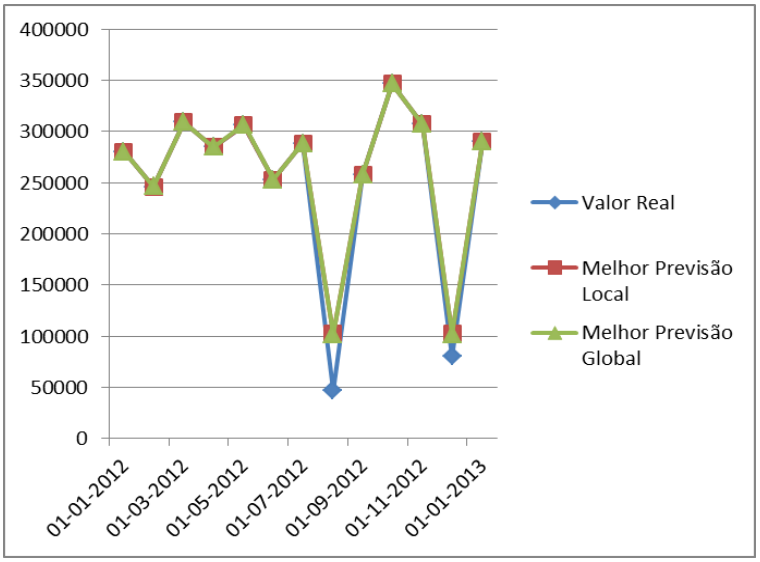


Figura 6.2. Representação gráfica da melhor previsão local e previsão global para os últimos 12 meses da empresa B.

Como nos resultados para a empresa A, também nos resultados para a empresa B foi possível identificar os meses para os quais se verificaram maiores discrepâncias foram os meses 8 e 12 de 2013 (consultar tabela 6.10), sendo que em todos os outros meses foi possível obter uma previsão muito próxima da realidade.

Tabela 6.11. Melhor previsão local e previsão global para os últimos 12 meses da empresa C.

Empresa	Data	Valor Real	Melhor Previsão Local	Melhor Previsão Global
C	31-07-2011	753453	753429,5653	753156,4987
	31-08-2011	496103	496124,8844	496274,0744
	30-09-2011	1018190	1018163,591	1017895,965
	31-10-2011	368667	368706,9294	368827,1344
	30-11-2011	561178	561205,3816	561447,5863
	31-12-2011	2631085	1261794,737	1260719,186
	31-01-2012	221560	221574,4246	221707,6439
	29-02-2012	518012	518046,3778	518212,1114
	31-03-2012	742385	742355,6328	742099,7825
	30-04-2012	196170	196212,1013	196520,5962
	31-05-2012	117365	117368,8114	117700,8788
	30-06-2012	558725	558747,6905	559063,7432
	31-07-2012	1946296	1261794,737	1260719,186

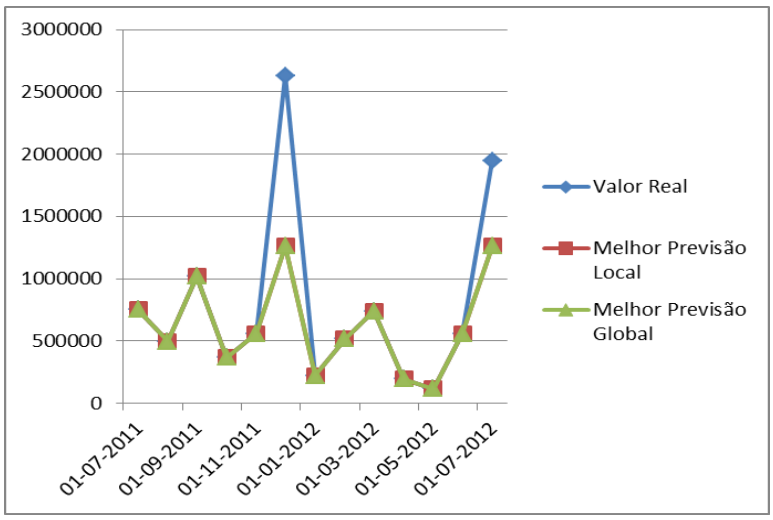


Figura 6.3. Representação gráfica da melhor previsão local e previsão global para os últimos 12 meses da empresa C.

Não fugindo às características já identificadas nas empresas anteriores, os pontos em que os valores previstos mais se afastaram dos valores reais foram os meses 12 de 2011 e 7 de 2012 (consultar tabela 6.11)

Tabela 6.12. Melhor previsão local e previsão global para os últimos 12 meses da empresa D.

Empresa	Data	Valor Real	Melhor Previsão Local	Melhor Previsão Global
D	30-04-2011	2176988	2145530,195	2145424,043
	31-05-2011	2189187	2157728,845	2157597,933
	30-06-2011	2508668	2477240,328	2477439,057
	31-07-2011	2568566	2537101,865	2536804,673
	31-08-2011	2322411	2290974,065	2291163,283
	30-09-2011	2045515	2014078,065	2014267,283
	31-10-2011	2667305	2635877,328	2635716,563
	30-11-2011	2534121	2502665,598	2502758,015
	31-12-2011	2243684	2212251,882	2212446,265
	31-01-2012	2607796	2576337,695	2576575,165
	29-02-2012	2173512	2142075,615	2141925,941
	31-03-2012	3017802	2676335,111	2676354,093
	30-04-2012	2337810	2306346,065	2306437,045

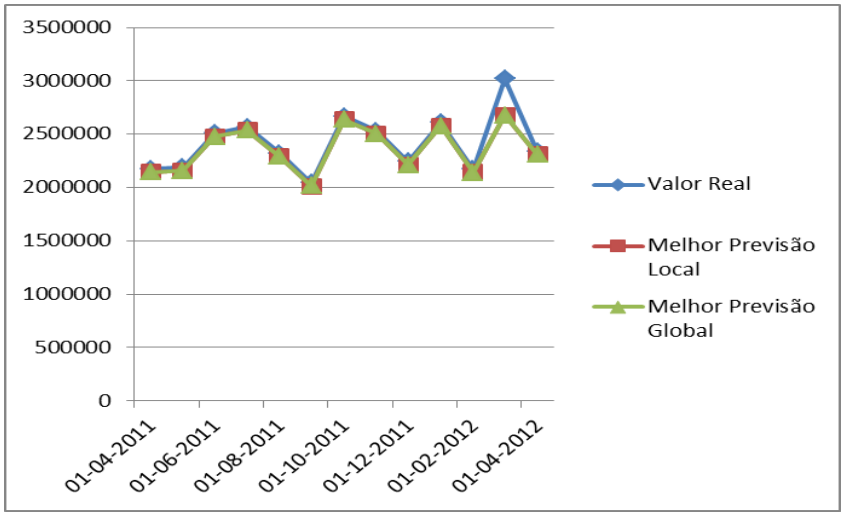


Figura 6.4. Representação gráfica da melhor previsão local e previsão global para os últimos 12 meses da empresa D.

De todas as empresas, a empresa D é a empresa que obteve um MAPE local e global mais baixo. É possível constatar esse facto atrvés de uma breve análise da figura 6.4 onde é patente que apenas no mês 3 de 2012 existiu um desalinhamento significativo entre os valores previstos e os valores reais, não sendo contudo, um desvio tão acentuado como o observado nas outras empresas.

Através dos gráficos apresentados, é possível rapidamente identificar que a diferença entre os resultados do melhor algoritmo local e global é pequena sendo apenas possível identificar nas tabelas. É também possível verificar que mesmo não tendo sempre um resultado de previsão certeiro, a previsão segue a tendência dos valores reais.

CAPITULO 7

CONCLUSÕES E TRABALHO FUTURO

7.1 ALGORITMOS DE DM

O presente projeto avaliou o desempenho de diferentes algoritmos de DM tendo em conta a sua capacidade de generalização e realização em processos de previsão sobre vendas de empresas de diferentes tipologias sem otimizar a sua parametrização segundo os parâmetros de qualquer empresa, em particular, com o objetivo de baixar "custos" de implementação. O algoritmo de ARIMA apresentou-se um pouco limitado noq eu diz respeito à combinação de variáveis de análise a utilizar, uma vez que apenas utiliza os dados de vendas para realizar as previsões. O mesmo não se verificou no algoritmo de ARTXP, permitindo assim que a combinação dos algoritmos ARTXP e ARIMA também não apresentassem a limitação de apenas realizar previsão sobre dados de vendas. Entre os três algoritmos aquele que apresentou a melhor capacidade para realizar previsões para as empresas em questão foi o algoritmo ARTXP. Este desempenho deve-se à sua capacidade de realizar cálculos de correlação entre as variáveis com base nas suas funções de regressão presentes nos nodos da árvore. Quanto ao algoritmo de NN, este apresentou um desempenho satisfatório em termos de generalização, uma vez que os dois MAPES mais altos entre os MAPES obtidos são mais baixos do que os mesmos resultados quando comparados com os algoritmos ARIMA, ARTXP e ARIMA+ARTXP. Porém, na realização das previsões para as empresas A e B, os dois MAPES de valores mais reduzidos obtidos possuem valores mais elevados em relação aos mesmos MAPES das mesmas empresas, quando se utilizou o algoritmo ARTXP. Esta circunstância levou a que o algoritmo ARTXP apresentasse um resultado global melhor do que o algoritmo NN apresentou. Sendo os algoritmos NN bastante complexos e podendo os seus resultados apresentar avriações consideráveistendo em conta o tipo de *kernel* utilizado, seria interessante em trabalhos futuros perceber quais os tipos de algoritmos NN que melhor se adequarão à realização da previsão de vendas sob séries temporais, de forma transversal a empresas com diferentes atividades comerciais.

Perante os resultados obtidos, o algoritmo que apresentou melhor capacidade de generalização e adaptação a diferentes empresas foi o algoritmo de SVM. Este resultado espelha a capacidade que este algoritmo tem na procura de pontos ótimos com base em

teorias matemáticas de vetores de suporte, conferindo-lhe assim, uma grande capacidade para obter pontos ótimos, facilitando a generalização do algoritmo e permitindo uma rápida aprendizagem com novos dados que eventualmente alimentem o modelo. São estas características que permitem que este algoritmo tenha obtido o melhor desempenho. Este algoritmo consegue assim obter o MAPE mais baixo, quer para cada empresa individualmente quer para a avaliação global do MAPE para todas as empresas.

7.2 As Variáveis de Análise

Não sendo o principal alvo de estudo deste projeto, analisámos também a influência da inclusão (ou não) de variáveis internas e externas às empresas como auxilio aos processos de previsão realizados. Com base na tabela 6.8 do capítulo anterior (melhor MAPE obtido para todas as empresas) é possível analisar o número de vezes que cada combinação de variáveis permitiu obter um melhor MAPE, quer a nível individual de cada empresa, quer a nível global – os valores obtidos podem ser consultados na tabela 7.1.

Tabela 7.1. Número de MAPES mais baixos por combinação de variáveis.

Combinação de variáveis	Número de melhores resultados entre todos os algoritmos	Número de melhores resultados do melhor algoritmo
Vendas	12	2
Vendas + Variáveis Internas	9	-
Vendas + Variáveis Externas	20	4
Todas as variáveis	23	1

A tabela 7.1 permite-nos verificar que o maior número de melhores resultados entre todos os algoritmos coincide com a utilização de todas as variáveis seguido de muito perto da utilização dos valores de vendas e das variáveis externas para obtenção dos valores de previsão. Assim, é possível concluir que as variáveis externas possuem um grande impacto nas vendas das empresas, tendo as variáveis internas uma menor influência, uma vez que a diferença de resultados é apenas de 20 (vendas + variáveis externas) contra uma diferença de 23 envolvendo todas as variáveis. Em relação ao algoritmo que atingiu todos os melhores resultados, o resultado com menor MAPE, para todas as empresas, tem por base a utilização de todas as variáveis. O facto de o estudo ser

feito sobre empresas de diferentes características e existirem resultados que apontam para a utilização de vendas e variáveis internas, isso justifica o facto do melhor resultado global ter por base a utilização de todas as variáveis. Contudo, seria interessante em trabalhos futuros analisar a influência de variáveis quer internas quer externas das empresas. A existência de poucos resultados que contemplem a utilização de vendas e variáveis internas pode ser devido à utilização de variáveis internas que não influenciem da mesma forma todas as empresas. Assim sendo, seria interessante em estudos futuros analisar, também, as diversas variáveis internas das empresas envolvidas e estudar qual a sua correlação com as vendas e como estas têm influências que podem implicar o aumento ou a diminuição das mesmas. O mesmo se poderá aplicar às variáveis externas, pois um estudo mais aprofundado em relação a um maior número de empresas de diferentes características pode revelar resultados diferentes em relação à influência deste tipo de variáveis na realização de um qualquer processo de previsão de vendas.

7.3 PREVISÃO DE VENDAS

O principal objetivo deste projeto era encontrar um algoritmo que permitisse obter a melhor previsão de vendas para diferentes empresas sem adaptar os algoritmos a nenhuma das empresas em específico. Este resultado foi medido em termos de MAPE, tendo sido considerado que um resultado inferior a 20% seria um bom resultado. Analisando a tabela 7.1 - melhor MAPE obtido para todas as empresas -, é possível constatar que o melhor resultado obtido é de 6,97%. Este resultado permite concluir que o principal objetivo foi alcançado uma vez que foi possível realizar uma previsão para todas as empresas com um valor global abaixo de 20%, sendo a maior de todas de 11,28% proveniente das previsões da empresa B. O avanço dos algoritmos de DM permite que algoritmos com grande capacidade de generalização, como os algoritmos baseados em SVM, sejam agora usados para tarefas nas quais a adaptação a diferentes perfis de dados seja essencial. Estudos mais aprofundados, incluindo um maior número de empresas e tipos de SVM, permite investigar se existem outros algoritmos baseados em SVM que apresentem uma capacidade de generalização ainda mais assertiva.

7.4 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

No que diz respeito ao *software* usado, ambos os produtos conseguiram apresentar resultados razoáveis de previsão. O SSAS apresentou-se como um produto de *software* mais simples de utilizar, com modelos de DM já predefinidos para cada tipo de aplicação. Porém, esta simplicidade acarreta a desvantagem de os algoritmos serem menos costumizáveis. Por sua vez, o RapidMiner foi bastante simples de utilizar, embora tendo uma curva de aprendizagem ligeiramente superior ao SSAS. A vantagem deste *software* apresentou-se ao nível da customização disponível. O RapidMiner apresenta uma vasta panóplia de soluções para preparação, análise e tratamento de dados, para além de tarefas de DM obviamente. Desta forma, para trabalhos futuros considero que o RapidMiner se pode perfilar como uma melhor ferramenta. Seria ainda interessante verificar se os resultados obtidos neste projeto se verificam em situações com um maior número de empresas, bem como se os resultados aqui alcançados se verificam quando se utilizar dados de variáveis externas e internas diferentes daqueles aqui reportados.

REFERÊNCIAS

Pyle, D. (1999) Data Preparation for Data Mining [Online] *Morgan Kaufman Publishers, Inc.* Available from: http://hfs1.duytan.edu.vn/upload/ebooks/3836.pdf [Accessed: 19 May 2015]

Allen, P. G. Fildes, R. (2001) "Econometric forecasting", Chapter 11 in Armstrong, J.S.(ed), *Principles of Forecasting: A Handbook for Researchers and Practicioners*, Klwuer Academic Press, Norwell, MA

Armstrong, J. S. (1992), Error Measurement for Generalizing About Forecasting Merhods: Empirical Comparsions

Almonacid, F. et al., Generation of ambient temperatures hourly time series for some Spanish locations by artificial neural networks.

Beeg, C. E., Connoly, T. M. (2009) Database Systems: A pratical approach to design, implementation and management. Pearson, 5 edition (March 6, 2009)

Bendoly, E. (2002) Theory and support for process frameworks of knowledge discovery and data mining from ERP sytem. Information & Management. (2003) 639-647

Box, G. E. P. e Jenkins, G. M. (1989) Time series analysis: forecasting and control

Bose, I. e Mahapatra, R. K. (2001) Business data mining - a machine learning pespective. Information & Management. (2001) 211 – 225

Chapman, P. et al. (2000) CRISP-DM 1.0

Chen, W. e Du, Y. (2009) Using neural networks and data mining techniques for the financial distress prediction model. Exper Systems with Applications 36 (2009) 4075 - 4086

Choi, T., Hui, C. e Yu.Y. (2014) Intelligent Fashion Forecasting Systems: Models and Applications; Chapter 2: Sales Forecasting Apparel and Fashion Industry: A Review,2014

Damle, C., Yalcin, A. (2006) Flood prediction using Time Series Data Mining Journal of Hydrology (2007) 333, 305 - 316

Dhanabhakyam, M. e Punithavalli, M. (2011) A survey on data mining for market basket analysis. Global Journal of Computer Science and Technology

Doganies, P. (2006) Time Series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. Journal of Food Engineering 75 (2) 196-204

Esling, P. e Agon, C. (2012). Time-Series data mining. *ACM Comput. Surv.* 45, 1, Article 12 (November 2012), 34 pages.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996) From data mining to knowledge discovery in datbases American Association for Artificial Intelligence

Friedman, J.H. (1991) Multivariate Adaptive Regression Splines. The annals of statistics

Fu, T. (2011) A review on time series data mining Engineering Applications of Artificial Inteligence

Gardner, S. e Dorling, M. (1998) Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. Atmospheric Environment, 32(14–15),2627–2636.

Gartner, (2012) Gartner Identifies the Top 10 Strategic Technology Trends for 2013. [Online] Available From: http://www.gartner.com/newsroom/id/2209615 [Accessed: 20 de Março de 2014]

Geurts, P. (2001) Pattern extraction for time series classification

Geweke, J. e Whiteman, C. (2006) Bayesian Forecasting Handbook of Economic Forecasting, Volume I, Ed por Elliott, G., Granger, C.W.J. e Timmerman, A

Han, J; Kamber, M; Pei, J (2012) Data mining - Concepts and techniques

Hu, X., et al. (2010) A data minin framework for time series estimation Journal of Biomedical Informatics

Hülsman, M. et al. (2012) General sales forecast models for automobile markets and their analysis. Transactions on Machine Learning and Data Mining Vol.5, No 2 (2012) 65-86

Jayalakshmi, T. e Santhakumaran, Dr. A. (2011) Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*

Kalapakis, K., Gada, D. e Puttagunta, V. (2001) Distance Measures for Effective Clustering of ARIMA Rime-Series

Karush, W. (1939). Minima of Functions of Several Variables with Inequalities as Side Constraints

Keogh, E. J. e Oazzani, M. J. (1998) Scaling up Dynamic Time Warping to Massive Dataset

Khashei, M. e Bijari, M. (2010) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing 11 (2011) 2664 - 2675*

Kim, K. (2003) Financial time series forecasting using support vector machines Neurocomputing 55(2003) 307 – 319

Kuhn, H. W., Tucker, A. W., (1951). Nonlinear programming. Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press. pp. 481–492. MR 47303

Laxzman, S. e Sastry, P. S. (2006) A survey of temporal data mining

Lu, C. (2014) Sales Forecasting of computer producs based on variable selection scheme and supportverctor regression. *Neurocomputing*

Luengo, J., García, S. e Herrera, F. (2011) On the choice of the best imputation methods ofr missing values considering three groups of classification methods.

Martson, S. L. et al. (2011) Cloud computing - The business prespective

Meek, C., Chickering, D. M. e Heckerman, D. (2002) Autoregressive tree models for time-series analysis

Mohammad, Y. e Nishida, T. (2010) Mining causal relationships in multidimensional time series, In: Szczerbicki, E. e Nguyen, N. T. (eds) *Smart information and knowledge management: advances, challenges, and critical issues*. Springer pp.309-338

Newton, I. (1687) Philosophiae Naturalis Principa Mathematica

Ngai, E. W. T. et al. (2011) The Application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems,* Februery 2011, Pages 559-569

Ni, Y. e F. Fan (2011). A two-stage dynamic sales forecasting model for the fashion retail. Expert Systems with

Applications 38(3): 1529-1536

de Oliveira, F. A., Nobre, C. N., Zárate, L. E., (2013) Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index - Case study of PETR4, Petrobras, Brazil. Expert System with Applications 12/2013; 40(18):7596-7606. DOI: 10.1016/j.eswa.2013.06.071

Ostrow, P. (2011) Sales Performance Management 2012 - How the best-in-class optimize the front line to grow the botom line

Ponomarenko, J., et al. (2002) Mining DNA sequences to predict sites which mutations cause genetic diseases. Knowledge - Based Systems 15 (2002) 225 – 233

Pyle, D. (1999) Data Preparation for Data Mining [Online] Morgan Kaufman Publishers, Inc. Available from: http://hfs1.duytan.edu.vn/upload/ebooks/3836.pdf [Accessed: 19 May 2015]

Ratanamahatana, C. A. Et al. (2005) Mining time series data Data Mining and Knowledge Discovery Handbook pp 1069 – 1103

Rumelhar, D., Hinton, D. e Williams, G. (1986) Learning internal representations by error propagation

Smola, A. e Scholkopf, B. (2004) A tutorial on support vector regression. Statistics and Computing 14: 199-222

Tay, F. E. H. e Shen, L. (2002) Economic and financial prediction using rough set model. European Journal of Operational Research 141(2002) 641 – 659

Thomassey, S. e Fiordaliso, A. (2006) A hybrid sales forecasting system based on clustering and decision trees. Journal Decision Support Systems Volume 42 Issue 1, October 2006, Pages 408 – 421

Thomassey, S. (2010) Sales forecasts in clothing industry: the key success factor of the supply chain management. International Journal of Production Economics. 128 (2), 470 - 483

Trafalis, T. B., Ince, H.(200) Support Vector Machine for Regression and Applications to Financial ForecastingTsaih, R., Hsu, Y. e Lai, C. C. (1998) Forecasting S & P 500

stock index futures with a hybrid AI system Decision Suport Systems 23 (1998) 161 - 174

Turban Et Al. (2006) Decision Support and Business Inteligence Systems (8th Edition)

Werbos, P. (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences

Whittle, P. (1951), Hypothesis Testing in Time Series Analysis

Xiong, Y., Yeung, D. Y. (2002) Mixtures of ARMA models for model-based time series clustering. Proceeding - IEEE International Conference on Data Mining,ICDM, 2002, p 717 - 720

Yang, Q. e Wu, X. (2006) 10 Challenging Problems In Data Mining Research. International Journal of Information Technology & Decision Making

Yu, Y. Choi, T. e Hui, C. (2011) Na intelligence fast sales forecasting model for fashion products. Expert Systems with Applications. Volume 38, issue 6, June 2011, Pages 7373 - 7379

Yu, X., Qi, Z. e Zhao, Y.(2013) Support Vector Regression for Newspaper/Magazine Sales For\ecasting. *First International Conference on Information Technology and Quantitative Management*

Zahedi, F. (1991) Na introduction to neural networks and a comparison with artificial intelligence and expert systems