NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Aplicação Data Mining para Análise e Previsão das Estratégias de *Pricing* em Companhias Aéreas

Estudo de Caso: Registos das Tarifas da Rota SSA-LIS

Pedro Artur Alves Rita

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

APLICAÇÃO *DATA MINING* PARA ANÁLISE E PREVISÃO DAS ESTRATÉGIAS DE *PRICING* EM COMPANHIAS AÉREAS

Estudo de Caso: Tarifas da Rota SSA-LIS

Por: Pedro Artur Alves Rita

Estudo de Caso: Registos das Tarifas da Rota SSA-LIS

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação com Especialização em Gestão do Conhecimento e Business Intelligence

Orientador: Mauro Castelli

Maio de 2018

DEDICATÓRIA

Ao meu Amantíssimo Poder Superior

"Ninguém escapa ao sonho de voar, de ultrapassar os limites do espaço onde nasceu, de ver novos lugares e novas gentes. Mas saber ver em cada coisa, em cada pessoa, aquele algo que a define como especial, um objeto singular, um amigo é fundamental. Navegar é preciso, reconhecer o valor das coisas e das pessoas, é mais preciso ainda!"

Antoine de Saint-Exupéry

AGRADECIMENTOS

A conclusão de um projeto de Mestrado implica uma grande jornada. É o resultado de conhecimentos e vivências adquiridas ao longo de uma vida. É produto, também, de conselhos, exemplos e orientações de docentes e colegas. Não é uma realização cujo mérito seja de uma pessoa só.

A todos aqueles que, direta ou indiretamente possibilitaram a realização deste trabalho, expresso aqui a minha gratidão e o meu bem hajam.

Quero, por isso, agradecer em primeiro lugar a todo o Departamento de Gestão de Informação da Universidade Nova de Lisboa, em especial ao corpo docente do Mestrado em Gestão de Informação, com especialização em Gestão do Conhecimento e Business Intelligence, pelos conhecimentos que me transmitiram, e pelo exemplo de competência e excelência.

Agradeço também ao meu orientador de projeto Professor Mauro Castelli, pedra angular deste trabalho, pela sua orientação sempre sábia, generosa e paciente. Por me ter apoiado, motivado e principalmente por me ter colocado no caminho certo para terminar este trabalho para que pudesse seguir com a minha vida profissional.

À empresa onde trabalho, "TAP-PORTUGAL – Departamento de *Pricing* and Demand", por ter viabilizado a concretização deste projeto, me ter transmitido conhecimento sobre a área e por me ter autorizado utilizar um caso de implementação. Agradeço em especial à minha grande amiga e colega, Ana Almeida, por ter cooperado na obtenção de informações, que em muito enriqueceram, este trabalho.

Ao meu "ginásio mental", em especial à Dra. Margarida Cordo, a minha profunda gratidão pelo seu apoio. A disponibilidade com que ouviu as minhas angústias e incertezas, o incentivo que me deu nos momentos de maior cansaço e desalento face ao que ainda faltava fazer. Por me ter permitido descomprimir após os dias de trabalho e ajudar-me a ter força para trabalhar de noite e em gozo de férias.

A todos os meus Amigos por me terem dado força quando ela me faltou e por me terem animado quando o cansaço me atacou. Sem a sua ajuda não teria sido possível chegar ao termo desta caminhada e manter o equilíbrio emocional indispensável para poder continuar a trabalhar.

Agradeço em particular, à minha grande amiga Benvinda, pela amizade, paciência e dedicação. Pelo amor e apoio incondicional, com que todas as vezes me incentivou a sair da minha toca isolada para fazer pausas de conversas e risadas animadas. Por nunca ter desistido de mim quando eu próprio já em nada acreditava. Uma verdadeira força da natureza.

Por último, agradeço à minha família por toda a ajuda que me deram, pela força e por acreditarem em mim.

Em especial aos meus queridos pais, que tanto AMO, pelo incentivo permanente e apoio inexcedível. Sem a sua ajuda incondicional, nunca na vida conseguiria chegar a este patamar de felicidade e realização pessoal, nem nunca seria a pessoa hoje sou! Olhar para eles, sentir e ver o seu orgulho estampado no rosto, será a minha maior conquista! Gratidão… Gratidão… Gratidão…

RESUMO

O foco deste trabalho é estudar a aplicabilidade e pertinência do uso de árvores de decisão como modelo preditivo do *pricing* para uma companhia aérea. Para isso, são utilizados como amostra dados da rota Salvador-Lisboa da TAP. As variáveis identificadas por este tipo de modelo preditivo como mais determinantes para o preço pago pelos bilhetes nesta rota foram o momento de compra (medido em número de dias antes da partida do voo), o ponto de venda do bilhete e o número de dias da estadia no destino. Este *outcome* corresponde com o conhecimento empírico de negócio da gestão de rotas da TAP. Foi também efetuada uma análise de clusterização dos clientes da rota de forma a descobrir alguns padrões comportamentais que suportem a tomada de decisão das métricas a utilizar.

Podemos aferir que as técnicas de *Machine Learning* e *Data Mining* utilizadas neste projeto podem servir de suporte na obtenção de melhores resultados, numa lógica complementar aos modelos matemáticos existentes, que têm como objetivo a maximização de receita. Estas técnicas permitem descrever com maior riqueza de informação o comportamento esperado da procura. Com a leitura que estas técnicas nos apresentaram para a presente análise da rota Salvador-Lisboa, conseguimos nomear quais das 11 métricas/dimensões atualmente utilizadas (com a metodologia "*Bayesian Forecasting"* – Guilhotina) são mais importante e vão dar à companhia maiores benefícios.

Assim, podemos manipular e dar maior ênfase no nosso trabalho diário, àquelas dimensões que, segundo os resultados obtidos pelo SAS, têm maior preponderância na definição da procura. Métricas/dimensões essas que são utilizadas numa base diária pelo analista com a função de *Pricing & Demand*.

Acreditamos que a leitura e análise dos resultados dos modelos utilizados neste trabalho sejam uma mais-valia e suporte fundamental na tomada de decisão das nossas ações diárias, com a finalidade comum de obtenção de receita máxima e adequação da oferta às necessidades dos clientes.

PALAVRAS-CHAVE

Data Mining, *Pricing*, Airlines, Business Intelligence, Yield, Decision Support Systems

ABSTRACT

The main purpose of this work is to study the applicability and relevance of using decision trees as a predictive model for an airline *pricing* department. For this, a dataset containing the information related to the TAP Salvador-Lisbon route was considered. The variables identified by this type of predictive model as the most determinant for the price paid for the tickets, were the moment of purchase (measured as the number of days before departure of being flight), the point of sale and the number of days of the ticket issued. This outcome corresponds to TAP's empirical business knowledge from the analyst. A clustering analysis of the route customers was also carried out to discover some behavioral patterns that support the decision making of the dimensions to be used.

We can verify that the Machine Learning and Data Mining techniques could be used to gain a better knowledge of the business of the company, by complementing the existing mathematical models. These techniques allow us to describe the expected behavior of demand with a greater wealth of information, leading to a maximization of the revenue. With the information gained by using these techniques on the Salvador-Lisbon route, we were able to name which of the 11 metrics/dimensions currently used (with the methodology "Bayesian Forecasting" - Guillotine) are more important and will give the company greater benefits.

Thus, one can manipulate and give greater emphasis, in daily activities, to those dimensions that, according to the results obtained by SAS, have a greater preponderance in the definition of demand. These metrics/dimensions are the ones used by the analysts operating in the *Pricing* & Demand area.

We believe that the analysis of the results obtained by the models used in this work provides an added value and a fundamental support for the daily decision-making activities. This allows to maximize the revenue and to increase the matching between the adequacy of the offer and the needs of the clients.

KEYWORDS

Data Mining, *Pricing*, Airlines, Business Intelligence, Yield, Decision Support Systems

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

TAP – Transportes Aéreos Portugueses

SSA – São Salvador

LIS - Lisboa

HUB – Placa Giratória

DW – *Data Warehouse*

LCC – *Low Cost Carriers*

APEX - *Advance Purchase Excursion Fares*

SAS – *Enterprise Miner*

BI – *Business Intelligence*

ERP – *Enterprise Resource Planning*

DM – *Data Mining*

ODIF - *Origin & Destination, Itinerary, Fare class*

OD - *Origin & Destination*

POS – *Point-of-sale*

PROS – *Revenue & Profit Optimization*

EMSR - *Expected Marginal Seat Revenue*

RASK - *Revenue per Available Seat-Kilometer*

DSS - *Decision Support Systems*

WTP - *Willingness to Pay*

ETL - *Extract, Transform and Loading*

CRM - *Customer relationship management*

CCC - *Cubic Clustering Criterion*

VFR - Visiting Friends and Relatives

KDD - *Knowledge Discovery from Databases*

WTA - *World Travel Awards*

ZED - *Zonal Employee Discount*

1. INTRODUÇÃO

A indústria da qual as companhias aéreas fazem parte insere-se num ambiente turbulento e complexo. Ao contrário de outras indústrias, a aviação está sujeita a rápidas alterações no que diz respeito às expetativas dos clientes, movimentos da concorrência, desenvolvimentos do fornecedor, regulamentações governamentais e dinâmicas dos funcionários (Riwo-Abudho et al, 2013).

De acordo com Lawrence *et. al (2003),* para uma companhia aérea, a prática da otimização de receitas através do controlo da disponibilidade e preço de lugares num voo é, habitualmente, designada por gestão de receitas. Sistemas sofisticados de gestão de receitas estão já em utilização no seio de todas as grandes companhias aéreas, e são vistos por muitos com uma componente crítica na infraestrutura tecnológica de uma companhia.

1.1 ENQUADRAMENTO

No contexto do atual mercado empresarial, as exigências ao nível da competitividade, levam à necessidade de uma melhor e mais rigorosa Gestão da Informação. Neste sentido, vingam as empresas que conseguem implementar estratégias que permitam reunir não apenas o máximo de informação possível, mas também informação com maior qualidade, para responder às questões fulcrais das várias áreas de negócio, por forma a otimizar e suportar as tomadas de decisão.

Dados, os avanços nas tecnologias de informação, existe crescente necessidade de transformar grandes volumes de dados, previamente armazenados, em informação, e essa em conhecimento útil. É nessa perspetiva que o *Business Intelligence* tem um papel fundamental porque vai permitir agregar um conjunto vasto de tecnologias. Em particular, as técnicas de *Data Mining,* ou seja, aquelas que nos permitem descobrir informação previamente desconhecida, para a construção de modelos preditivos partindo de um conjunto de hipóteses e pressupostos, que nos irão guiar na tarefa da definição do modelo a aplicar. No caso das ferramentas deste tipo de exploração, o objetivo é "deixar os dados falar", no sentido de criar condições para que os "dados se possam expressar", e extrair-se padrões e tendências que possam responder às questões previamente formuladas.

O *Data Mining* e a consequente modelação preditiva podem ser processos relevantes para um sistema de gestão de receitas, permitindo prever, com antecedência, quanto uma determinada empresa poderá faturar, com base em informação histórica contida nas bases de dados. Deste modo, os gestores poderão estar sempre prevenidos contra emergências e tomar decisões com menor risco, o que é uma mais-valia significativa na conjuntura atual das companhias aéreas.

A motivação para a escolha deste tema, prende-se com o fato do autor, em contexto profissional se deparar com várias questões, que poderão ser auxiliadas na área de *Business Intelligence* e aplicação de técnicas de *Data Mining,* associadas à *previsão* de *Pricing*, como é o que se pretende elaborar neste projeto em específico.

1.2 OBJETIVOS

O principal objetivo do presente projeto é desenvolver um modelo preditivo, com base em dados de *Pricing* de uma rota aérea, por forma a identificar padrões relevantes para apoiar futuramente na gestão e implementação de preços para essa mesma rota. Para o efeito, pretende-se extrair e

analisar um conjunto de informação recolhida ao longo de uma década na plataforma de gestão da informação da Transportadora Aérea Portuguesa (TAP), o sistema PROS. As principais variáveis a serem trabalhadas serão a segmentação de clientes e as variáveis determinantes do *pricing* para a rota Salvador-Lisboa. E com isto fazer uma reflexão, a uma pequena escala, sobre o valor da gestão de informação e *Data Mining* na TAP, através deste modelo preditivo para o apoio à decisão na área de *Pricing Demand*.

Pretende-se, igualmente, mostrar que a aplicação de técnicas de *Data Mining (Software SAS)* numa empresa de aviação, como é a TAP, que engloba várias áreas de negócio. O que torna a gestão da informação muito complexa e específica. Neste sentido, também se pretende avaliar a viabilidade e possíveis dificuldades ao introduzir um modelo preditivo num nicho tão restrito, como é uma rota aérea.

Objectivos específicos deste projecto são:

- Avaliar a complementariedade do *software SAS* utilizado neste metodologia para apoio à decisão face aos resultados finais;

- Perceber em que medida esta metodologia apresenta um contributo científico, que sirva de informação de base em projectos de *Pricing*, em termos de definição de tarifas para uma rota áerea em específico, podendo eventualmente ser aplicada a outras.

- Analisar a importância da integração da informação extraída do sistema PROS, e respectiva selecção de variáveis para a modelação preditiva, cujo objectivo final é a atribuição de tarifas numa rota aérea, que representa uma pequena amostra da complexidade dos sistemas de informação e apoio à decisão da TAP.

1.3 ESTRUTURA

Com base nos objetivos descritos no ponto anterior, este projeto está estruturado da seguinte forma:

- **Capítulo 2** é feita uma **revisão bibliográfica**, com levantamento do quadro teórico de referência na área de *Business Intelligence,* desde as suas origens até aos dias de hoje, bem como a aplicação de técnicas de *Data Mining* e as suas diversas vertentes e aplicações;
- Segue-se, no **capítulo 3**, apresentação da **metodologia** a ser aplicada neste projeto, começando pela metodologia geral, recolha e seleção de dados, descrição geral dos procedimentos metodológicos a serem efetuados no *software SAS.*
- No **Capítulo 4**, é feita a representação gráfica dos **resultados** gerados após a implementação e processamento da metodologia no *software SAS,* seguindo-se uma análise e discussão dos resultados obtidos, tirando as ilações do trabalho no seu todo e tecendo algumas considerações mais relevantes em termos de *pricing*; fiabilidade da metodologia implementada e perspetiva de aplicações futuras em outras rotas em termos de previsão de preços.
- E finalmente nos dois capítulos **(5 e 6)** é feita uma análise exploratória dos dados de saída, mais favoráveis, à segmentação de tarifas e do respetivo valor acrescentado para a rota em estudo. E neste sentido, pretende-se tirar ilações e recomendações sobre a viabilidade desta metodologia para aplicações futuras em outras rotas.

1.4 CARACTERIZAÇÃO DA COMPANHIA AÉREA TAP

A TAP Portugal é a maior companhia aérea portuguesa, a operar desde **1945**, tem os seus *hubs* em **Lisboa** e no **Porto,** garantindo ligações aos quatro continentes (Africa, Europa, América do Norte e América do Sul) para mais de 76 destinos, espalhados por 29 países. Os seus segmentos de tráfego com maior peso são o "*Leisure*" o "Étnico" (emigrantes) e tráfego *Corporate*. Todos estes segmentos caracterizam-se por uma elevada sensibilidade ao preço, apresentando um nível de fidelização bastante elevado.

Prosseguindo uma orientação estratégica cuja prioridade é a **satisfação das expetativas dos Clientes**, a TAP procura continuamente proporcionar as melhores e mais fáceis soluções para as suas viagens, agregando cada vez mais valor aos produtos que oferece.

Com esse objetivo, a Empresa estabelece também as melhores parcerias, em terra e no ar, disponibilizando assim um número alargado de destinos servidos em ***code-share*** com companhias suas congéneres, além de um diversificado conjunto de **vantagens e benefícios** associados.

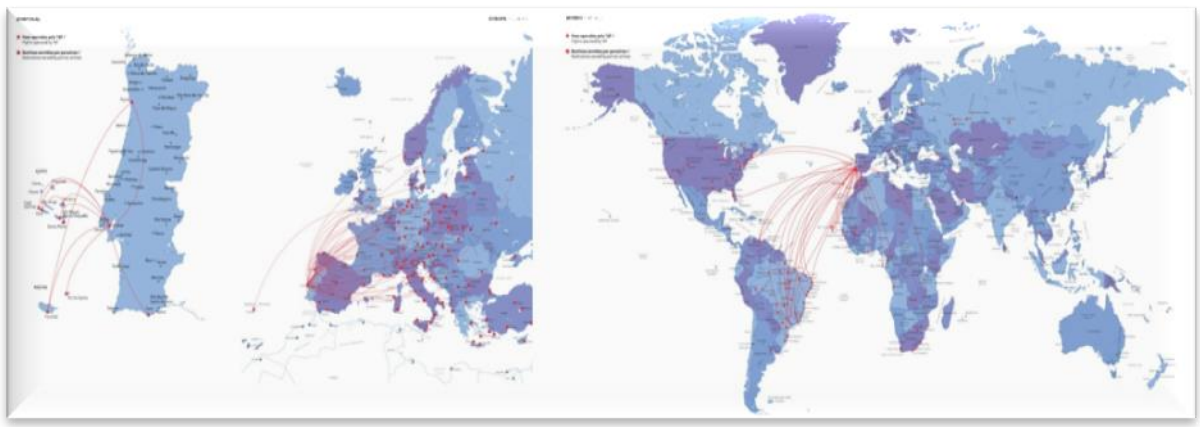


Figura 1 - Representação gráfica de todos os destinos TAP incluindo rotas em code-share. Fonte: SRS Analyser

A TAP viaja para sete cidades distintas em Portugal e para vários países da Europa, tais como: Espanha (nove cidades), Itália (quatro cidades), França (cinco cidades), Reino Unido (duas cidades), Suíça (duas cidades), Benelux (duas cidades), Alemanha (cinco cidades), Dinamarca (uma cidade), Noruega (uma cidade), Suécia (uma cidade), Helsínquia (uma cidade) República Checa (uma cidade), Hungria (uma cidade) e Bucareste (uma cidade) e Rússia (uma cidade). Viaja ainda para, mais dez países em África: Magreb (quatro cidades), Cabo Verde (quatro cidades), Senegal (uma cidade), Guiné-Bissau (uma cidade), São Tomé e Príncipe (uma cidade), Angola (uma cidade), Moçambique (uma cidade), Costa do Marfim (uma cidade), Gana (uma cidade) e Togo (uma cidade). Ainda voa para os Estados Unidos da América e Canada (quatro cidades); para Venezuela (uma cidade) e Brasil (10 cidades). Perfazendo um total de setenta e sete cidades em todo o mundo, aos quais se juntarmos as cem companhias aéreas parceiras da TAP temos um total de 1330 destinos possíveis, em 192 países.

Figura 2 - Rotas TAP Portugal. Fonte: SRS Analyser

A TAP de momento conta na sua frota com os aviões mais modernos e fiáveis, adquiridos á gigante construtora europeia, Airbus.

A frota da TAP é constituída por: vinte e um Airbus A319, dezanove unidades do Airbus A320, três do Airbus A321 e dezasseis do Airbus A330 (o mais recente avião da TAP, que se encontra muito bem equipado e por isso, é muito utilizado para voos de longo curso), e ainda conta com quatro unidades do Airbus A340, conforme o ilustrado na Figura abaixo.

Figura 3 **-** *Frota TAP Portugal. (Fonte: http://www.flytap.pt)*

A TAP Portugal é a **companhia aérea Portuguesa líder** de mercado e membro da ***Star Alliance*** desde 14 de março de 2005. Foi reconhecida pela UNESCO e pela *International Union of Geological Sciences* com a atribuição do **Prémio Planeta Terra IYPE 2010**, na categoria **"Produto Sustentável Mais Inovador"**.

Em 2012, a Manutenção & Engenharia TAP ganhou o prémio ***Silver*** de publicidade, na categoria ***Airline Contract Maintenance***, da prestigiada revista ***Air Transport World***.

Eleita em quatro anos consecutivos (2009, 2010, 2011 e 2012) a **Companhia Aérea Líder Mundial para a América do Sul**, pelos ***World Travel Awards*** (WTA), viu assim reafirmada a sua liderança naquele mercado, em forte crescimento. A prestigiada "***Global Traveler***" dos EUA elege-a como a **Melhor Companhia Aérea da Europa** em 2011 e 2012.

E em dezembro de 2012, a TAP é distinguida como **Líder Mundial para África** pelo 2º ano consecutivo.

Em 2013, o vídeo de segurança da TAP vence **prémio *Passenger Choice Awards*** na APEX para melhor *inflight video*, assim como vence o galardão de Ouro na categoria Comunicação Institucional dos **Prémios Meios & Publicidade**.

1.4.1 Política de Preços das Companhias Aéreas Versus Outros Sectores Empresariais

Na tabela 1 apresentam-se as diferenças entre a política de preços de uma companhia aérea e outros sectores empresariais, para melhor entendimento deste tipo de negócio.

Tabela 1 - Descrição comparativa da Política de Preços das companhias aéreas versus outros sectores.

Companhias Aéreas	Outros Sectores Empresariais
• **Vendem serviços, e um lugar vazio é uma perda.** • **A oferta é limitada à capacidade do avião.** • **Os custos operacionais fixos são muito elevados.** • **Cada viagem, para cada origem /destino (O&D) é um diferente mercado e potencialmente com preços diferenciados.**	• **Podem-se armazenar. Não se vende hoje, vende-se "amanhã".** • **A capacidade de produção é variável. Se a procura aumenta, a produção aumenta em conformidade.** • **O custo variável é importante e o custo marginal é um *input* fundamental para a definição do preço.** • **Existe um leque de vários produtos com preços diferenciados.**

1.4.2 A evolução das estruturas tarifárias da TAP

Nas figuras 4, 5, 6 e 7 apresenta-se a evolução das estruturas tarifárias da TAP nos últimos anos.

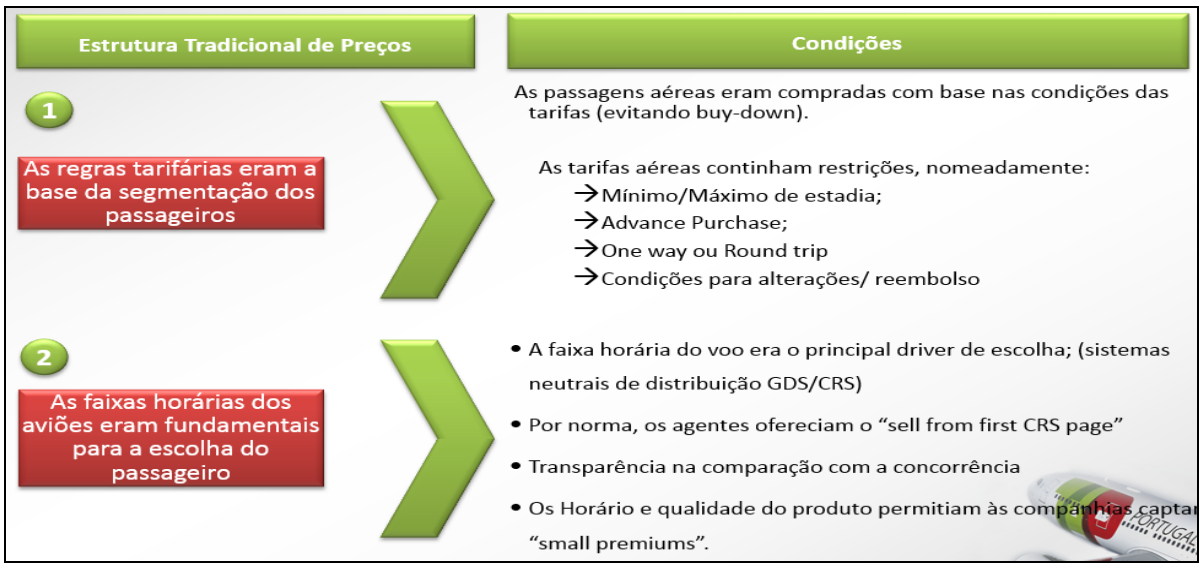


Figura 4 - Representação da evolução das estruturas tarifárias da TAP (Estrutura tradicional de Preços).

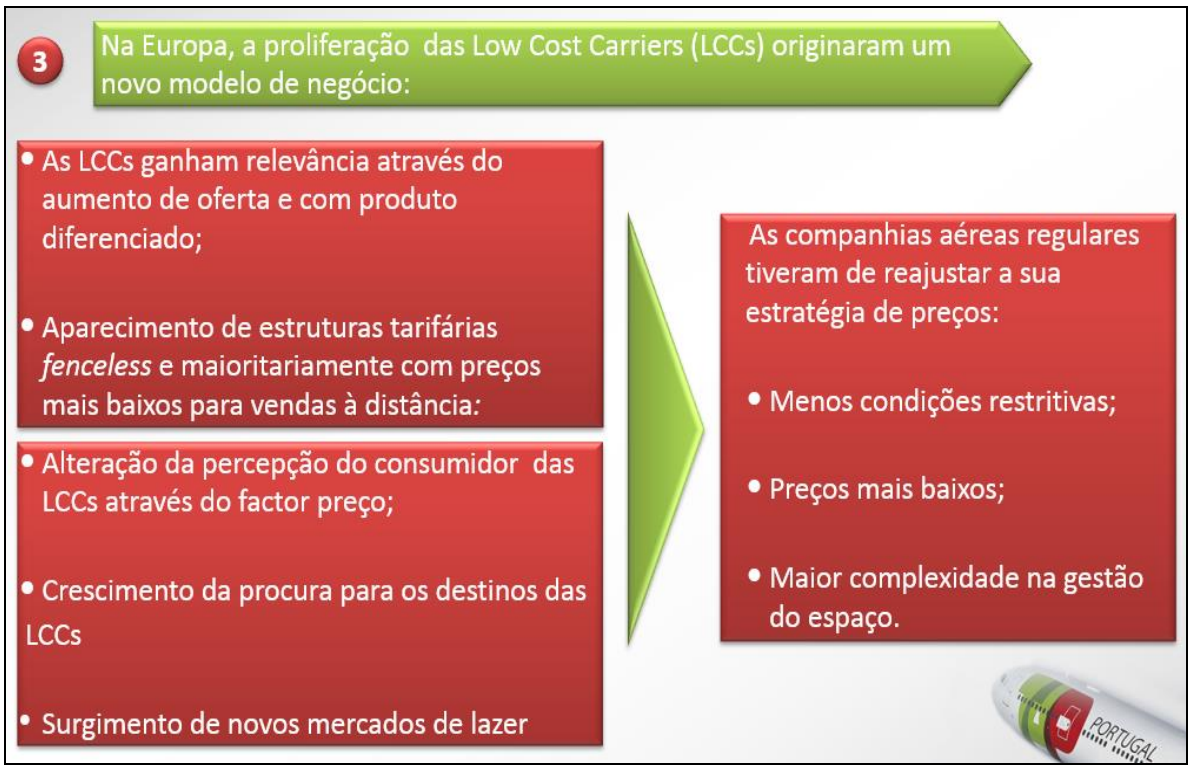


Figura 5 - *Representação da evolução das estruturas tarifárias da TAP (Proliferação das Low Cost).*

Figura 6 *- Representação da evolução das estruturas tarifárias da TAP (Mudanças no mercado).*

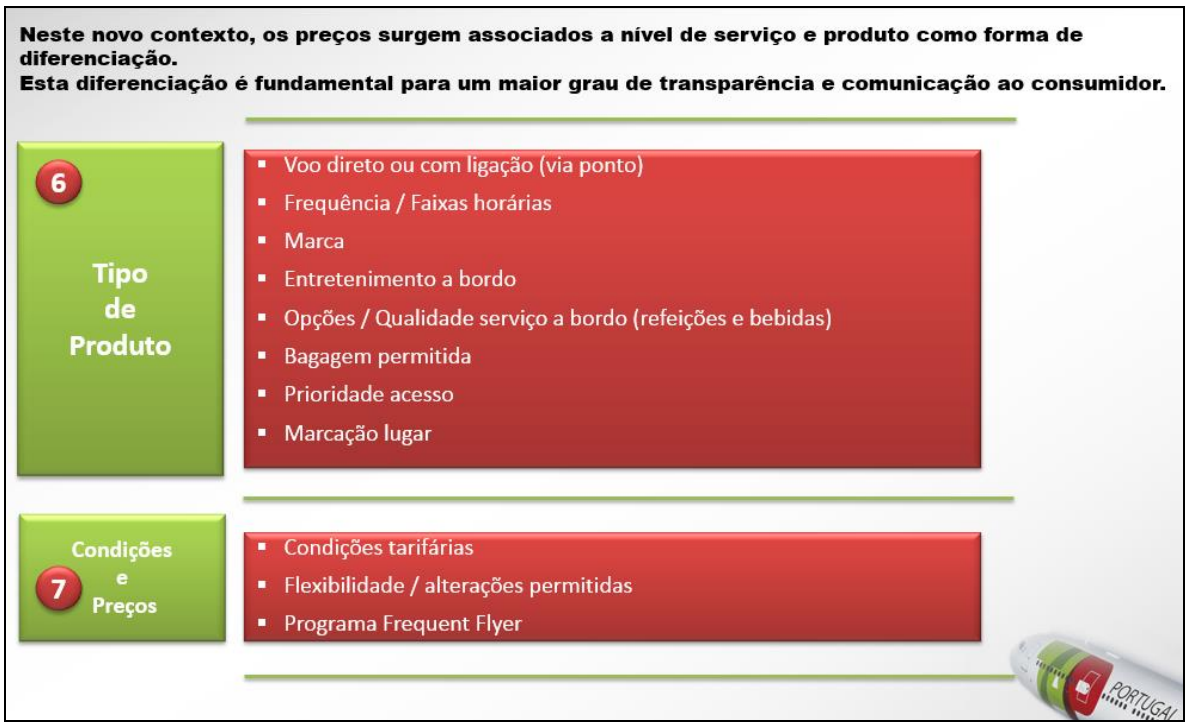


Figura 7 *- Representação da evolução das estruturas tarifárias da TAP (Necessidade de diferenciação dos produtos).*

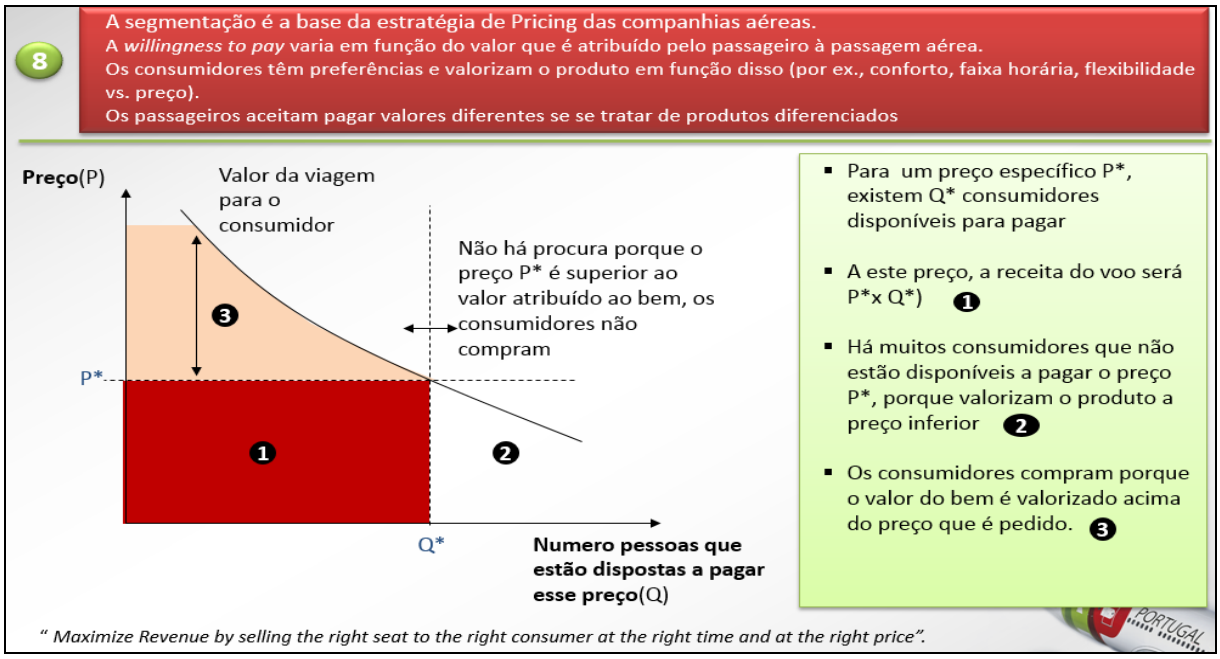


Figura 8 - Representação da evolução das estruturas tarifárias da TAP (Segmentação através da estratégia de Pricing).

1.4.3 Definição de *Pricing* e sua Principal Função na Aviação

Define-se estratégia de *Pricing* (curto a médio prazo para os mercados) como o objetivo para maximizar a receita, respeitando os objetivos comerciais e estratégicos da empresa, considerando o produto TAP, a concorrência, e as necessidades de mercado.

Esta estratégia de *pricing* na aviação em geral e na TAP em particular visa:

- A definição da estrutura de preços públicos e confidenciais que otimizem a gestão da receita em função da segmentação para a rede TAP;
- A definição da estrutura de preços para mercados *beyond* através de acordos de parceria com outras companhias (*code-share e interline*) por forma a ampliar a rede TAP e possibilitar vendas adicionais;
- A definição e implementação de ações promocionais globais, ações pontuais pró-ativas e reação da concorrência em função do produto TAP e necessidades da empresa;
- A análise e negociação de propostas de *pricing* dos mercados (público e confidencial) com a área Vendas;
- A monitorização dos preços TAP versus concorrência para avaliação de ações a tomar a nível pontual ou estratégico;
- Reuniões regulares dentro da equipa de *Pricing e Revenue Management* sobre questões específicas relativas ao comportamento dos mercados/rotas e novas estratégias de *pricing*;
- A definição, implementação e atualização das condições das tarifas, mais precisamente os níveis de tarifa e condições de segmentação (MIN/MAX estadia, taxas de penalidades remarcação e reembolso, *surcharges, stopovers*, etc);

- A Análise e estudos pontuais das linhas na ótica de suporte à decisão e implementação de novas ações de *pricing;*
- A avaliação mensal dos resultados das rotas e elaboração de relatórios com ações a tomar.

1.4.4. Os Cinco Elementos Fundamentais para a Maximização da Receita

Na figura 9 apresenta-se uma síntese dos cinco elementos para a maximização da receita no sector da Aviação.

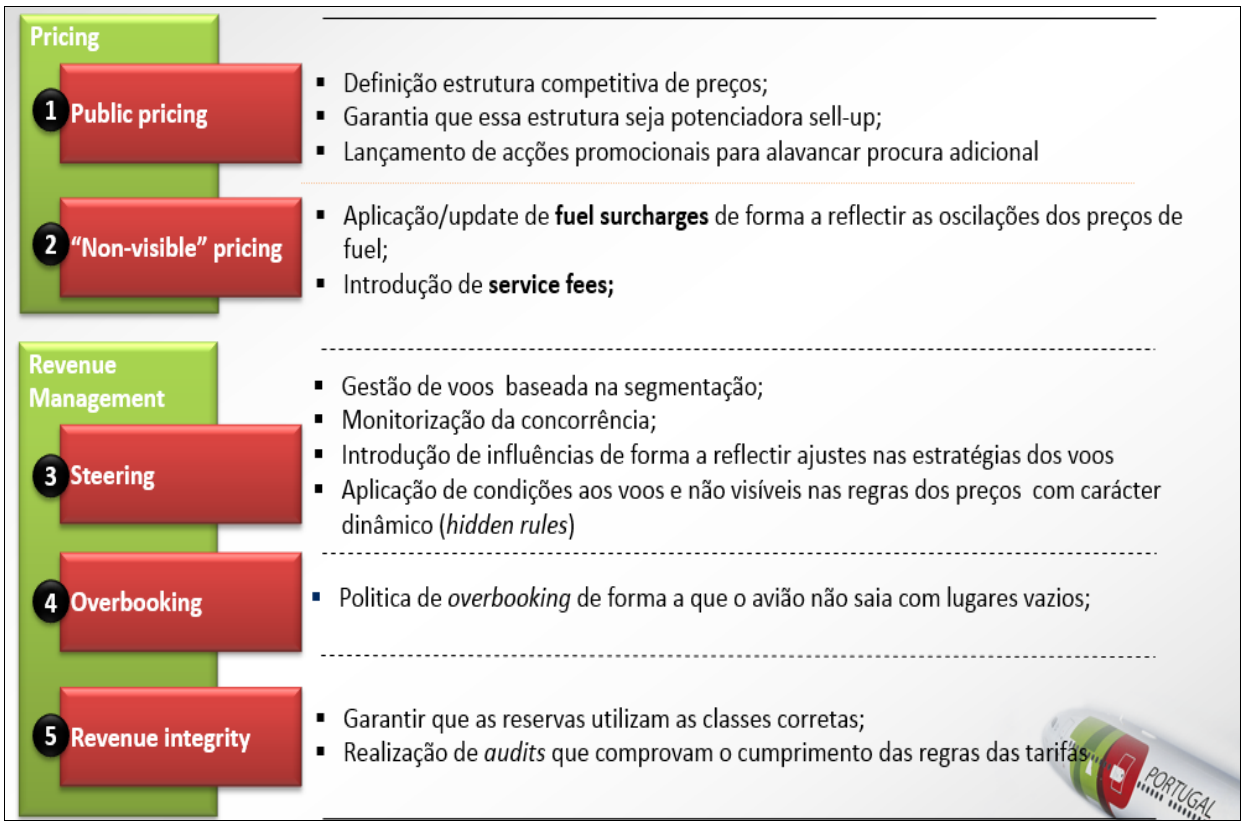


Figura 9 - *Representação da evolução das estruturas tarifárias da TAP (Elementos de Maximização de Receitas).*

1.4.5 Definição de Willingness To Pay e seus Objetivos

Willingness To Pay (WTP) define-se como a disponibilidade que um passageiro tem para adquirir uma viagem por determinado valor.

A gestão da disponibilidade de classes no *Revenue Management* tradicional assenta na segmentação básica do tráfego baseada no conceito de *willingness to pay dos* consumidores, com o objetivo de:

- Por um lado, impedir que os passageiros ***product oriented***, dispostos a pagar tarifas mais altas, comprem tarifas abaixo da sua *WTP*.

Exemplo: Tráfego que normalmente viaja por motivos profissionais/negócios estará dispostos a pagar mais, em troca de maior conforto em terra e a bordo e de menores restrições tarifárias

Resumindo: o *Revenue Management tem como finalidade*: "Não Vender Hoje um lugar que pode ser vendido mais tarde, mais caro".

- Por outro lado, estimular a procura do tráfego de lazer, mais sensível ao preço, ou seja, **P*rice Oriented.***

Maximização da Receita

- A maximização da receita através da aplicação de técnicas de *Revenue Management* pressupõe, portanto, a oferta de produtos tarifários diferenciados, que permitam uma segmentação efetiva da procura já que, estando a assumir que os passageiros têm necessidades diferentes, e também que estarão *dispostos a pagar* diferentes preços por diferentes tipos de serviços.

- A gestão ótima do inventário de cada voo/data traduz-se, na prática, na determinação do número de lugares a disponibilizar para cada diferente nível de preço (classe de reserva), procurando sempre vender cada lugar ao preço o mais alto possível.

Receita Marginal

- Sendo o transporte aéreo uma indústria em que o peso dos custos fixos é relativamente elevado, quando comparado com o peso dos custos variáveis, a obtenção de receita marginal pode ter um contributo muito elevado para o lucro final.

- O enfoque de ação do *Revenue Management* é maximizar a receita marginal esperada por lugar[1] e a aplicabilidade do *Revenue Management* à gestão de espaço em transporte aéreo. Que passou a revelar-se fundamental para a rentabilidade do negócio a partir do momento em que a sua expansão e desregulamentação, iniciada nos E.U.A. nos anos 70-80, o transformou num meio de transporte de massas, a operar num ambiente altamente competitivo. Com o aumento da oferta e da concorrência, o fator "preço" passou a ter um papel preponderante na tomada de decisão de compra de uma percentagem elevada do tráfego.

Gestão Ótima do Inventário

A ótica de abordagem correta à gestão do inventário pressupõe que exista (e se mantenha de forma consistente) um enfoque equilibrado entre os níveis de:

- ***YIELD*** ou receita por passageiro

- ***Load-Factor*** ou taxa de ocupação

De forma, a que o resultado dessa gestão se traduza efetivamente na obtenção de RASK - *Revenue per Available Seat-Kilometer* mais elevados, este é um método de medição de receitas comummente utilizado pelas companhias aéreas, que se traduz em receitas por lugares disponíveis por quilómetro.

$$RASK\ (€cent) = \frac{Passengers\ Flown\ Revenue}{ASK} \times 100 \qquad (1)$$

[1] *EMSR - Expected Marginal Seat Revenue resultado da multiplicação do preço pela probabilidade de a procura aparecer, a esse preço, para cada lugar, num voo)*

Para a determinação do número de lugares a atribuir a cada classe num voo (o *fare mix* ótimo) são considerados:

- **A CAPACIDADE** – oferta de lugares
- Os níveis de **PREÇO**S existentes
- A **PROCURA** prevista para cada um desses níveis de preço

O Grande Desafio ao *Revenue Management* Tradicional é:

- O aparecimento e rápida expansão das companhias *LOW COST*, observado primeiro nos E.U.A. e poucos anos depois na Europa, veio evidentemente abanar profundamente todos os princípios em que se baseava o *Revenue Management* tradicional.
- A adoção de estruturas tarifárias simplificadas (ou *fenceless)*, onde apenas o nível de preço, varia e não segmentam a procura através de regras tarifárias. Neste sentido, os riscos de diluição de receita associados à adoção progressiva e generalizada deste tipo de estruturas tarifárias pelas companhias tradicionais, obrigaram a ajustes nos sistemas de *Revenue Management*, que garantissem um controlo mais eficaz desses riscos.

1.5 CARACTERIZAÇÃO DA ROTA SALVADOR- LISBOA (SSA-LIS)

Tal como tem sido evocada diversas vezes no cancioneiro brasileiro…***Salvador Terra musa inspiradora de poetas***…

"***Tudo, tudo na Bahia faz a gente querer bem / A Bahia tem um jeito / Que nenhuma terra tem***" escreveu um dos maiores compositores baianos, Dorival Caymmi em 1941. E é justamente a música o que melhor traduz este estado brasileiro cuja capital, Salvador, foi também a primeira capital do Brasil. Mais de 500 anos depois, é ali que sentimos, como em nenhum outro lugar, a força da mistura dos povos africanos, portugueses e índios. Terra de João Gilberto, Gilberto Gil, Caetano Veloso e outros grandes génios musicais, os sons dos variadíssimos ritmos ecoam por toda a parte. Desta polifonia distingue-se um instrumento muito particular: o berimbau, que, entre outras coisas, marca o tempo da capoeira, arte marcial, música, dança & desporto: a capoeira é uma mistura de tudo isto. Levada para o Brasil por escravos africanos – acredita-se que do sul de Angola, no séc. XVII -, foi proibida durante todo o período colonial e ate meados do séc. XX. Património Imaterial da humanidade pela UNESCO desde 2014 é um dos marcos da cultura brasileira exportada com orgulho para o resto do mundo. O movimento básico da capoeira chama se Ginga e o supermanequim, também *personal trainer,* Mauro Lopes gingou no frio outono lisboeta durante várias horas.

1.5.1 Histórico

A TAP opera a rota Salvador-Lisboa há mais de 20 anos. Inicialmente operava-se em voo circular com outros pontos do Nordeste (REC, FOR, NAT) ou até com o Rio de Janeiro.

O voo non-stop diário para a cidade de Salvador da Bahia já surgiu depois do ano 2005. Até há pouco tempo Salvador era o grande centro (*hub)* do Nordeste, ou seja, todas as companhias domésticas operavam este aeroporto ligando destinos do interior do Brasil (essencialmente Nordeste) a Rio de Janeiro e São Paulo. Haviam operações internacionais quer para a Argentina, quer para a Europa.

Recentemente o destino Recife ultrapassou o destino Salvador, que é agora o aeroporto do Nordeste com maior tráfego aéreo, e consequentemente com mais ligações internacionais.

Também o destino Fortaleza ganhou preponderância face ao destino Salvador com o anúncio do lançamento das operações das companhias aéreas *Aire France* e da *KLM* (*Royal Dutch Airlines*).

1.5.2 Tipo de Tráfego

Em termos de caracterização do tipo tráfego da rota Salvador, trata-se de uma rota essencialmente de lazer, existindo também uma componente muito interessante de *Visiting Friends and Relatives* (VFR).

Relativamente ao Tráfego Internacional**,** para além da TAP operam, a *Air Europa* com três frequências semanais para Madrid e as linhas Aéreas Argentinas com voos para Buenos Aires e para Córdoba (Argentina).

1.5.3 Mercados Dominantes

O Brasil é o mercado dominante com mais de 50% das vendas. Segue-se Portugal, com cerca de 12% das vendas desta rota, as restantes percentagens de vendas distribuem-se por Itália, França e Alemanha, que são os principais mercados de sextas, ou seja, com o direito de transportar passageiros e carga, através do território do estado de nacionalidade da aeronave, entre o território de um terceiro estado (ponto aquém) e o território do outro estado contratante.

1.5.4 Origens e Destinos Dominantes.

33% Dos passageiros da rota fazem apenas o percurso entre Lisboa e Salvador. Os restantes são passageiros que fazem ligação entre Lisboa e outros destinos, sendo que Paris é o primeiro Destino em ligação e seguem-se Roma, Milão, Porto, Barcelona, Zurique, Madrid e Londres.

1.5.5 A Importância da Rota Salvador-Lisboa na TAP

As Companhias aéreas através da verificação dos destinos preferenciais dos clientes fazem um estudo de mercado de rotas aéreas (com ou sem escala de viagem), cujo objetivo é definir de uma forma sustentável a sua inclusão e/ou exclusão.

No seguimento destes estudos de mercado, a rota SSA-LIS, é considerada uma das rotas com maior procura de passageiros a nível europeu.

Após a introdução de uma nova hierarquia na rede TAP, implementada através do sistema PROS/O&D, levou a adoção de uma metodologia única de classes, comum a toda a sua rede. Sendo que esta mudança para O&D, obrigou a ajustamentos nos pacotes, e a rota SSA-LIS foi a primeira a efetuar esta conversão, no mercado brasileiro.

2. REVISÃO BIBLIOGRÁFICA

2.1 GESTÃO DA INFORMAÇÃO

A gestão da informação é um processo que consiste nas atividades de recolha, identificação, classificação, processamento, armazenamento e disseminação da informação, independentemente do formato ou meio em que se encontra, seja via documentos físicos ou digitais (Robertson, 2005). A sua finalidade é fazer chegar as informações às pessoas que delas necessitam para decidir oportunamente, no momento certo, de forma eficaz e eficiente. Neste sentido, requer-se competências específicas por parte dos gestores e/ou responsáveis pela gestão da informação.

Para Robertson (2005) a gestão da informação é um termo genérico englobando sistemas e processos dentro de uma organização para criar e usar informações empresariais, sendo muito mais do que apenas Tecnologias da Informação, pressupondo práticas negociais, gestão de recursos humanos, cultural e organizacional. Este autor enumera então, dez princípios para garantir que a gestão da informação tenha sucesso, e são eles:

1. Reconhecer a complexidade;
2. Concentrar-se sobre a implementação;
3. Oferecer benefícios visíveis;
4. Hierarquizar objetivos de acordo com necessidades;
5. Estabelecer um percurso de muitas etapas;
6. Providenciar uma liderança forte;
7. Mitigar os riscos;
8. Comunicar de modo amplo;
9. Proporcionar uma experiência contínua;
10. Escolher cuidadosamente o projeto de arranque.

Por outro lado, importa introduzir o conceito de gestão de conhecimento ao conceito de gestão da informação. Pois, segundo Khandekar e Sharma, (2006) e Prieto e Revilla (2006), a gestão do conhecimento e a aprendizagem organizacional estão intimamente relacionadas e são encaradas como uma estratégia para as organizações se manterem atualizadas, face às turbulências e exigências do mercado competitivo, elevando, assim, o desempenho organizacional.

Atualmente é ainda vulgar que em muitas organizações, seja utilizado um conjunto de ferramentas de planeamento, sem com isso obter o benefício de um enquadramento que seja a base da estratégia da organização. Neste sentido, pretende aclarar no presente projeto, alguns conceitos e recomendações relacionadas com a gestão da informação e gestão do conhecimento, por forma a otimizar os processos de apoio à decisão de uma forma sustentada, para a obtenção de melhores resultados no que respeita à adoção da melhor metodologia a seguir.

Para um melhor entendimento, importa resumir as diferenças práticas entre os conceitos de gestão da informação e gestão do conhecimento, que são elas:

A **Gestão da Informação** é um conjunto de estratégias que visa identificar as necessidades informacionais, mapear os fluxos formais de informação nos diferentes ambientes da organização, assim como sua coleta, filtragem, análise, organização, armazenagem e disseminação, objetivando apoiar o desenvolvimento das atividades cotidianas e a tomada de decisão no ambiente corporativo

A **Gestão do Conhecimento** é um conjunto de estratégias para criar, adquirir, compartilhar e utilizar ativos de conhecimento, bem como estabelecer fluxos que garantam a informação necessária no tempo e formato adequados, a fim de auxiliar na geração de ideias, solução de problemas e tomada de decisão.

Na literatura observa-se que algumas correntes fundem os dois modelos de gestão, ou ainda, confundem um modelo com o outro. Por esse motivo, é muito comum, em diferentes segmentos económicos, empresários falarem que fazem gestão do conhecimento nas suas empresas, quando na realidade o que fazem é gestão da informação.

No entanto, algumas correntes definem muito claramente o papel de cada um destes modelos de gestão. Sem dúvida nenhuma, as duas gestões convergem para o fato de que pretendem apoiar/subsidiar as atividades desenvolvidas no dia-a-dia, e a tomada de decisão na organização.

Para isso, focam fluxos informacionais diferenciados. A **gestão da informação** apoia-se nos **fluxos formais** (conhecimento explícito) e a **gestão do conhecimento** nos **fluxos informais** (conhecimento tácito).

A gestão da informação **trabalha no âmbito do registado**, não importando o tipo de suporte: papel, disquete, CD-ROM, Internet, Intranet, fita, DVD, etc., constituindo-se por **ativos informacionais tangíveis**.

A gestão do conhecimento **trabalha no âmbito do não registado**: reuniões, eventos, construção individual de conhecimento, valores, crenças e comportamento organizacional, experiências práticas, educação corporativa, conhecimento do mundo etc., constituindo-se por **ativos intelectuais (intangíveis).**

Na tabela 2, apresenta-se o foco que cada um dos modelos de gestão (Informação e conhecimento) em relação às suas **atividades de base**, **objeto** e **âmbito da gestão**.

Tabela 2 - Gestão vs. conhecimento

GESTAO DA INFORMAÇÃO	GESTAO DO CONHECIMENTO
ÂMBITO Fluxos formais	**ÂMBITO** Fluxos informais
OBJETO Conhecimento explícito	**OBJETO** Conhecimento tácito
ATIVIDADES BASE	**ATIVIDADES BASE**
- Identificar demandas necessidades de informação - Mapear e reconhecer fluxos formais - Desenvolver a cultura organizacional positiva em relação ao compartilhamento/ socialização de informação - Proporcionar a comunicação informacional de forma eficiente, utilizando tecnologias de informação e comunicação - Prospetivar e monitorizar informações - Coletar, selecionar e filtrar informações - Tratar, analisar, organizar, armazenar informações, utilizando tecnologias de informação e comunicação - Desenvolver sistemas corporativos de diferentes naturezas, visando o compartilhamento e uso de informação - Elaborar produtos e serviços informacionais - Fixar normas e padrões de sistematização da informação - Retroalimentar o ciclo	- Identificar demandas necessidades de conhecimento - Mapear e reconhecer fluxos informais - Desenvolver a cultura organizacional positiva em relação ao compartilhamento/socialização de conhecimento - Proporcionar a comunicação informacional de forma eficiente, utilizando tecnologias de informação e comunicação - Criar espaços criativos dentro da corporação - Desenvolver competências e habilidades voltadas ao negócio da organização - Criar mecanismos de captação de conhecimento, gerado por diferentes pessoas da organização - Desenvolver sistemas corporativos de diferentes naturezas, visando o compartilhamento e uso de conhecimento - Fixar normas e padrões de sistematização de conhecimento - Retroalimentar o ciclo

Para o processo de inteligência competitiva organizacional esses dois modelos de gestão são essenciais para o seu funcionamento. Por esse motivo, tanto a gestão da informação quanto a gestão do conhecimento se fazem necessárias para sua efetividade corporativa.

2.2. OUTROS CONCEITOS ASSOCIADOS À GESTÃO DA INFORMAÇÃO - DADOS, CONHECIMENTO, INFORMAÇÃO

2.2.1. Dados

Os dados podem ser definidos como um conjunto de factos discretos e objetivos sobre os acontecimentos. São pontos no espaço e no tempo, sem referência ao tempo e ao espaço. Os dados apenas descrevem parte do sucedido, não proporcionando nenhum juízo de valor ou interpretação (Serrano e Fialho, 2004).

Os dados são itens referentes a uma descrição primária de objetos, eventos, atividades e transações que são gravados, classificados e armazenados, mas não chegam a ser organizados de forma a transmitir algum significado específico (Turban, McLean e Wetherbe, 2004).

Quando nos deparamos com dados, atribuímos-lhes algum significado. É este atribuir de sentido, esta contextualização e compreensão dos dados à luz do que cada um sabe, que constitui a gestão da informação, o processo de planeamento, organização, direção e controlo da informação, aos níveis estratégico, tácito e operacional.

A informação e conhecimento são a mesma coisa?

Pode parecer um detalhe semântico, mas faz muita diferença na gestão organizacional. Hoje em dia qualquer pessoa tem acesso abundante de Informação sobre qualquer assunto na internet. No entanto, a informação pela informação não tem valor algum. É a capacidade de filtrar, de articular e de aplicar essas informações de forma a dar uma solução – desde a mais simples até a mais sofisticada – que faz a diferença. É a capacidade de realizar, de executar e de criar que geram real valor para a organização. E, para isso, precisamos do conhecimento.

2.2.2. Conhecimento

O conhecimento passa pela capacidade de reter e assimilar informação, pelas experiências, pela visão do mundo e até pelos valores pessoais. A mesma informação pode ser interpretada de maneira totalmente diferente dependendo de quem tem acesso a ela. Tudo depende de como cada um vai transformar a informação que recebe. Alguns profissionais fazem gestão da informação pensando que estão fazendo gestão do conhecimento. Organizar documentos e extrair relatórios a partir de bases de dados são atividades importantes, mas gestão do conhecimento vai muito para além do que é gerir o que está explícito.

O conhecimento é essencialmente intangível e depende da inteligência humana para se manifestar. A própria inteligência artificial é um produto da inteligência humana.

A capacidade de inferência, de dedução, de formulação de novas hipóteses a partir da interpretação de um conjunto de dados, de informações e da perceção da realidade é que fazem com que os profissionais (e as organizações) evoluam continuamente.

Gerar e replicar informações é mais natural. Gerar novos conhecimentos obriga a uma certa dose de ousadia. Precisamos exercitar os músculos da aprendizagem mais do que armazenar um volume enorme de informações perecíveis.

2.2.3 Valor da Informação

Dos vários recursos de uma organização (financeiros, humanos ou logísticos), a informação é provavelmente o mais valioso de todos, porque faz uma interligação, e descreve, os recursos físicos e o meio onde se encontram.

O valor da informação só é importante quando considerados os objetivos da organização, uma vez que esta é determinada pelo utilizador, nas suas ações e decisões, e depende do contexto em que é utilizada na tomada das decisões finais.

A informação só é valorizada com base nas decisões eficazes, não tendo qualquer valor se esta não tiver qualquer utilidade para a tomada de decisões, no presente ou no futuro.

2.2.4 Importância da Informação

A importância da informação é universalmente aceite, e cada vez mais defendida como um dos recursos mais importante dentro das organizações, em que a sua gestão e aproveitamento têm uma influência direta no sucesso empresarial.

A importância da informação nas organizações assume três vertentes: **o Recurso, Ativo e a Mercadoria**

• A informação é entendida como **Recurso**, quando serve como forma de recolha de dados, respetivo tratamento, de modo a dar satisfação às exigências pretendidas;

• A informação como **ativo** verifica-se quando a organização consegue rentabilizar os recursos existentes de modo tornar-se mais competitiva;

• A informação como **Mercadoria** encontra-se quando as organizações podem vendê-la, sob forma de jornais, revistas e outras publicações (Gordon & Gordon,1999).

2.2.5 Relação entre Dados, Conhecimento e Informação

Ainda que com significados distintos, os dados e a informação relacionam-se devido à necessidade constante das organizações, em captarem, identificarem e analisarem os dados, para que se possa obter informação útil.

Contudo, e segundo Davis (1974) existe uma utilidade diferente no conceito de "informação", na medida em que, o que é para um utilizador poderá ser diferente para outro, tal como um produto acabado de uma secção de fabrico poderá ser matéria-prima para a secção seguinte.

Por outro lado, Boisot & Canals (2004) defendem que a informação é uma extração de dados que, modificando as distribuições de probabilidade relevantes, tem capacidade para realizar um trabalho útil na base de um agente do conhecimento. Ainda segundo estes autores, e como se pode verificar na representação da figura 10, os agentes operam dois tipos de filtros ao converterem estímulos recebidos em informação. Apenas os estímulos que passam **pelos filtros preceptivos** são registados como dados e os **filtros conceptuais** extraem informação com base nos dados registados. Ambos os filtros são "sintonizados" pelas expectativas dos agentes cognitivos e afetivos, sendo moldados de acordo com os conhecimentos ao longo da vida, no sentido de atuar seletivamente tanto nos estímulos como nos dados.

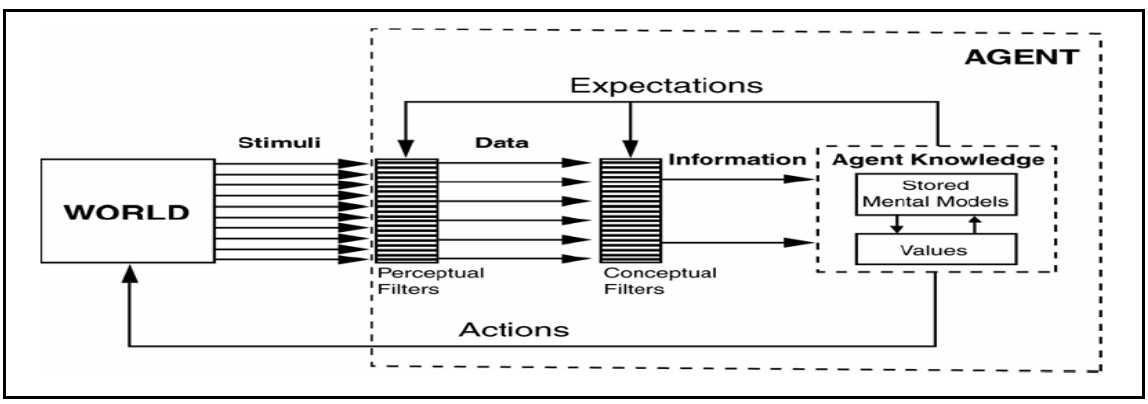


Figura 10 – *Relacionamento entre dados, informação e conhecimento (Boisot & Canals, 2004)*

Em complemento aos conceitos descritos por Boisot & Canals (2004), Tuomi (1999) identifica mais dois tipos de conhecimento, a **inteligência e a sabedoria**.

Como é possível verificar na Figura 11, em que, segundo Tuomi (1999) é ilustrada a visão convencional da hierarquia do conhecimento, cujos dados são descritos como simples factos isolados, que em determinado contexto e combinados com uma estrutura única, dão lugar a informação. E inteligência surge assim, na fase em que a mente humana usa este conhecimento para escolher entre alternativas. Por fim, quando os valores e os comportamentos culturais são como diretrizes no comportamento humano, pode-se dizer que este comportamento se baseia na sabedoria.

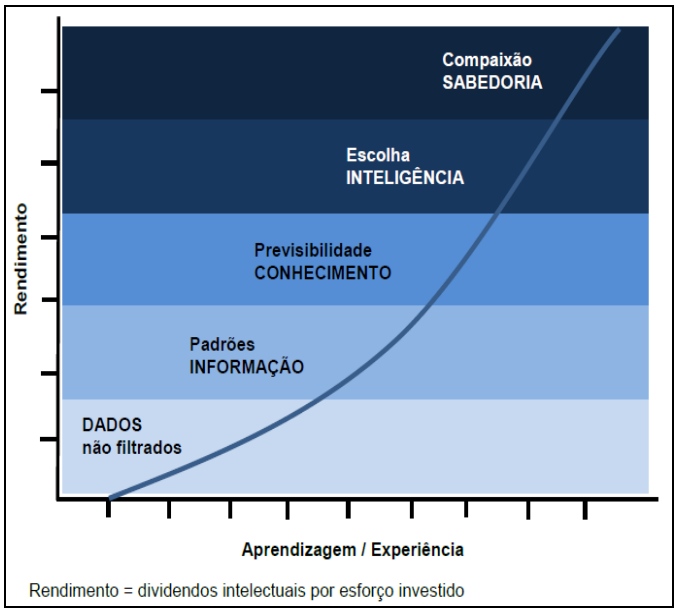
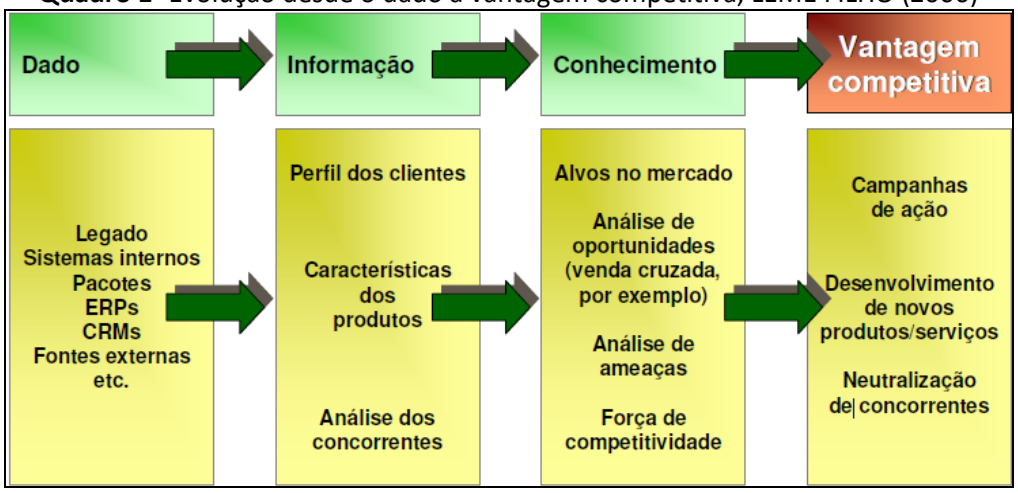


Figura 11 **-** *A visão convencional da hierarquia do conhecimento, adaptado de (Tuomi, 1999).*

2.2.6 Vantagem Competitiva da Utilização Sustentável dos Dados, Informação e Conhecimento

LEME FILHO (2006) apresenta um quadro que demonstra a evolução do dado até a obtenção de vantagem competitiva

Quadro 1- Evolução desde o dado à vantagem competitiva, LEME FILHO (2006)



As fontes podem estar disponíveis para as empresas a partir de seus próprios sistemas computacionais (sistemas internos, ERPs, CRMs) ou a partir de fontes externas.

Esses dados trabalhados convertem-se em informação, que oferecem às empresas um entendimento sobre sua atuação no mercado: perfil de consumo de seus clientes, produtos mais ou menos rentáveis, comparação de preços, prazos e participação de mercado perante os concorrentes, estudos de sazonalidade, entre outros.

A informação, em seguida, deve ser transformada em conhecimento. Para PONCHIROLLI (2005), conhecimento é informação internalizada pela pesquisa, estudo ou experiência que tem valor para a organização. NONAKA e TAKEUCHI (1997) consideram o conhecimento como um processo humano dinâmico de justificar a crença pessoal com relação à verdade. São dois entendimentos complementares, pois trata-se de uma dimensão poderosa de análise que, democratizada, potencializa o apoio às estratégias tornando-as mais assertivas, evitando "matar moscas com balas de canhão".

Nesse momento, cruzando informações de perfis de consumo de clientes e características de participação e rentabilidade de produtos, é possível traçar alvos específicos no mercado. Vendas cruzadas podem ser estimuladas a partir das informações de necessidades dos clientes, aproveitando assim oportunidades em primeira mão. E compreendendo a atuação da concorrência, pode-se mitigar com maior eficácia os efeitos das ameaças.

Finalmente, o conhecimento adquirido resulta em vantagem competitiva, nomeadamente com campanhas de marketing mais direcionadas, novos produtos ou serviços podem ser desenvolvidos com mais clareza e chegar público-alvo que se pretende atingir (dentro do contexto demográfico e de poder aquisitivo), estudos de correlação podem indicar quais fatores influenciam diretamente variáveis quantitativas, ajudando gestores a direcionar os seus esforços para as causas, e não mais para os efeitos (por exemplo, dentro do universo de cartões de crédito, qual a importância de um gestor saber quais variáveis influenciam mais os atrasos de pagamento?).

Tem-se a impressão de que, de acordo com a análise previamente feita, o conhecimento derivado da informação e dos dados, é um bem precioso para as organizações. De fato, há a máxima de que 'informação é poder'. Neste sentido, PONCHIROLLI (2005) afirma:

(...) estamos no limiar de uma nova era, na qual o conhecimento é reconhecido como o principal ativo das organizações é a chave para uma vantagem competitiva sustentável.

Até então fortemente caracterizada pelos bens tangíveis, como o capital financeiro e as estruturas físicas, a fonte de riqueza e competitividade passa a ser, agora, o próprio conhecimento. Sociedade do conhecimento; era do conhecimento; era do capital intelectual; sociedade pós-capitalista são algumas denominações para esta nova época.

No entanto, o autor adverte que acréscimo de informação tende a confundir em vez de esclarecer. Pois a quantidade e disponibilidade da informação cresce exponencialmente, confundindo as pessoas e dificultando, a gestão do conhecimento.

Sabe-se que tanto a escassez quanto o excesso podem ser prejudiciais, ou seja, perder-se em uma quantidade incontrolável de informações é tão nocivo como não as possuís. Temos de aprender a eliminar, em vez de acumular informações. Adotar a máxima "quanto menos melhor".

PEREIRA (1997) segue a mesma linha de advertência quanto ao excesso de informações, alegando que a informação está inserida no contexto da linguagem. Desse modo, informação é uma mensagem que contém novidades. Quando a mensagem não contém novidades (traz apenas assuntos ou dados conhecidos), é chamada de redundância. Com isso, sugere a criação de um indicador denominado 'índice de informação':

Seguindo a revisão da literatura existente, buscar a tratar as informações certas (adequadas aos utilizadores e ao momento) é fator crítico de sucesso para gerar conhecimento e vantagem competitiva.

2.3. BUSINESS INTELLIGENCE

Se conhece o inimigo e a si mesmo, a vitória será inquestionável.

Se conhece o Terreno e o Tempo a sua vitória será total.

Sun Tzu

O conceito *Business Intelligence* (BI) presta-se a muitas interpretações e tem evoluído ao longo do tempo. Contrariamente ao que se possa supor, não é um conceito novo. Ele foi concebido originalmente por Hans Peter Luhn, investigador da IBM, em outubro de 1958, num artigo da *IBM Journal of Research and Development,* que definia BI como a capacidade em conhecer as relações entre os factos apresentados de forma a atingir um objetivo desejado (Luhn, 1958). Segundo Gartner (2013), BI é o conjunto de boas praticas e uma alavanca tecnológica que possibilita a visão de gestão a partir das aplicações e dados das empresas.

A Business Intelligence (BI) tem um papel decisivo na criação de vantagens competitivas em qualquer organização (Evelson, Karel *et al.* ,2010).

Para avaliar a maturidade de BI é indispensável ter noção de quais são as principais ferramentas duma plataforma. A construção de uma plataforma ideal é algo complexo, pois são várias as componentes que têm de ser consideradas, incluindo a integração de dados, limpeza, modelação, armazenamento, criação de métricas, relatórios, *queries* entre muitos outros, com combinações e abordagens infinitas de forma a torna-las uteis e significativas.

Zhang *et al.* (2011) definem *Business Intelligence* como o uso da tecnologia de *Data Warehouse (DW)* para armazenar e gerir dados operacionais e, através de diversas ferramentas de análise estatística e técnicas de *Data Mining*, analisar esses mesmos dados, de forma a providenciar uma variedade de relatórios analíticos que, por sua vez, podem oferecer informação relevante no processo de apoio à decisão.

O conceito de *Data Warehouse* tem diversas definições consoante os autores. Mencionando em específico Gartner (2013), que define que o DW é uma arquitetura de armazenamento desenhada para guardar os dados extraídos dos sistemas transacionais e de outras fontes externas. O armazém (*warehouse*) agrega os dados sumarizando-os de forma a adequar aos relatórios e análise de dados para as necessidades de negócio pré-definidas. Este autor define, ainda, cinco componentes do DW, que são eles: sistema fontes de dados de produção, extração de dados e transformação, o sistema de gestão de base de dados do DW, a administração do DW e as ferramentas de BI.

Por outro lado, e segundo TDWI (2013), o *Data Warehouse*, ou mais concretamente, o processo de *Data Warehousing* incorpora os repositórios de dados e os modelos conceptuais, lógicos e físicos para suportar os objetivos de negócio e as necessidades dos utilizadores finais.

O DW é a base para o sucesso de um programa de BI. A construção de um DW requer a correspondência entre dados das fontes e dos destinos, e a captura dos detalhes da transformação dos dados em metadados. O DW providencia uma única e abrangente fonte da situação atual e histórica. As técnicas e ferramentas de DW incluem as plataformas, arquiteturas, estruturas, escalabilidade, serviços e segurança e o próprio DW como um serviço (TDWI 2013).

Outros autores de referência como Inmon e Kimball (2002), enfatizam e utilizam conceitos semelhantes de DW afirmando que, o DW ultrapassou os teóricos que queriam colocar todos os dados numa única base de dados, "sobrevivendo" ao desastre das *dot.com* originado pelos capitalistas de visão curta.

Para Inmon (2002), um *Data Warehouse* é um repositório de dados orientados para temas, integrado, independente do tempo e não volátil que suporta os processos de tomada de decisão. Mas para Kimball (2002), DW tem uma definição mais lata, definindo que este, é o conjunto de dados passíveis de serem consultados e tem os seguintes objetivos: simplificar o acesso à informação; apresentar a informação de uma forma consistente; ser flexível, adaptável e resistente às mudanças; ser seguro; ser a base para melhoria da tomada de decisão e ser aceite pelos utilizadores (Kimball e Ross, 2002).

Ross (2002) define e liga o DW ao BI, afirmando que a missão do DW é providenciar a informação de negócio, consistente e harmonizada, baseada nos dados operacionais, de suporte à decisão e externos para todas as unidades de negócio.

Para atingir este fim, os dados devem ser analisados, compreendidos, transformados e disponibilizados. Portanto, a administração do DW deve coordenar e supervisionar o desenvolvimento, gestão e manutenção de todo o ambiente do DW (Moss e Adelman 2000).

Para Inmon, a história do DW está ligada à evolução dos sistemas de informação de suporte à decisão (DSS - *Decision Support Systems*), e é um repositório organizado de dados, separado do sistema operacional e preparado para ser consultado de uma forma simples e intuitiva.

Por outro lado, Kimbal utiliza a metáfora do restaurante para descrever um sistema de DW. A comparação é bastante prática, porque com uma imagem do mundo real, é possível compreender rapidamente as várias áreas do DW.

No restaurante existe a zona da cozinha (back room), normalmente escondida e não acessível aos clientes, que prepara os pratos que serão servidos. Os ingredientes chegam do exterior (sistemas fonte), são preparados e transformados (ETL) nos pratos que irão ser servidos na sala de refeições. Na sala de refeições (front room), os clientes (utilizadores) escolhem os pratos através dos menus. Os clientes nunca (ou raramente) entram na cozinha. Qualquer pedido é feito sempre na sala de refeições. Por vezes, os clientes pedem algumas alterações nos pratos constantes no menu. Por vezes os pedidos são aceites, por vezes não, por falta de matéria-prima ou por necessitar de muito tempo.

Para Kimball e Ross (2002), os objetivos do DW são:

- Facilitar e simplificar o acesso a informação da organização;
- Dar consistência à informação;
- Ser adaptável e imune às mudanças das necessidades de negócio;
- Proteger a informação;
- Ser a base para a tomada de decisão.

As abordagens de Inmon e Kimball são diferentes, não só na estrutura, mas também na metodologia. Diversos autores têm-se dedicado a descobrir as diferenças e similaridades entre as duas abordagens que por vezes são diametralmente opostas.

Breslin sintetiza as diferenças e os aspetos específicos de cada abordagem, referindo que Inmon e Kimball propõem abordagens e perspetivas diferentes, por vezes totalmente opostas, dependendo dos modelos propostos. Um modelo de um DW corporativo só é possível com uma abordagem *top-down*. A aproximação de Kimball é por áreas temáticas. Por esse motivo o sub-conjunto de dados (*Data Marts) temáticos* constitui o *Data Warehouse, enquanto* Inmon privilegia o DW corporativo.

O desenvolvimento do DW é mais demorado seguindo a metodologia de Inmon. O modelo de Kimball necessita de pequenas equipas enquanto o modelo de Inmon só é possível com equipas maiores de especialistas. Tal repercute-se no investimento inicial e em termos financeiros a aproximação de Kimball é menos onerosa no início, uma vez que o esforço é dirigido apenas à construção de um *Data Mining*, em oposição à construção do DW corporativo, defendido por Inmon, onde o esforço financeiro é maior.

Sendo o objetivo principal do BI permitir a fácil interpretação de dados para auxiliar a gestão de qualquer negócio, e ao mesmo tempo identificar novas oportunidades com vista a implementar uma estratégia efetiva baseada nos dados. Neste sentido, pretende promover negócios com vantagem competitiva no mercado, conferindo uma melhor estabilidade a longo prazo.

O DW é o conceito base para montagem de um sistema de dados utilizados em BI, onde a corporação pode unificar todos os seus sistemas para ter uma base única para extração de relatórios, em que os dados serão posteriormente analisados através de *Data Mining* (Mineração de Dados) que também podem ser aplicadas a essa base de dados.

O principal elemento do BI é o Data Warehouse (DW): um grande banco de dados onde são armazenadas informações sobre transações da empresa, dados externos e donde se pode efetuar consultas analíticas. O DW é definido por Inmon (1997) como "um conjunto de dados baseado em assuntos, integrado, não-volátil e variável em relação ao tempo, de apoio às decisões da gestão". A integração do DW ao BI pode ser melhor explanada a partir da figura 12.

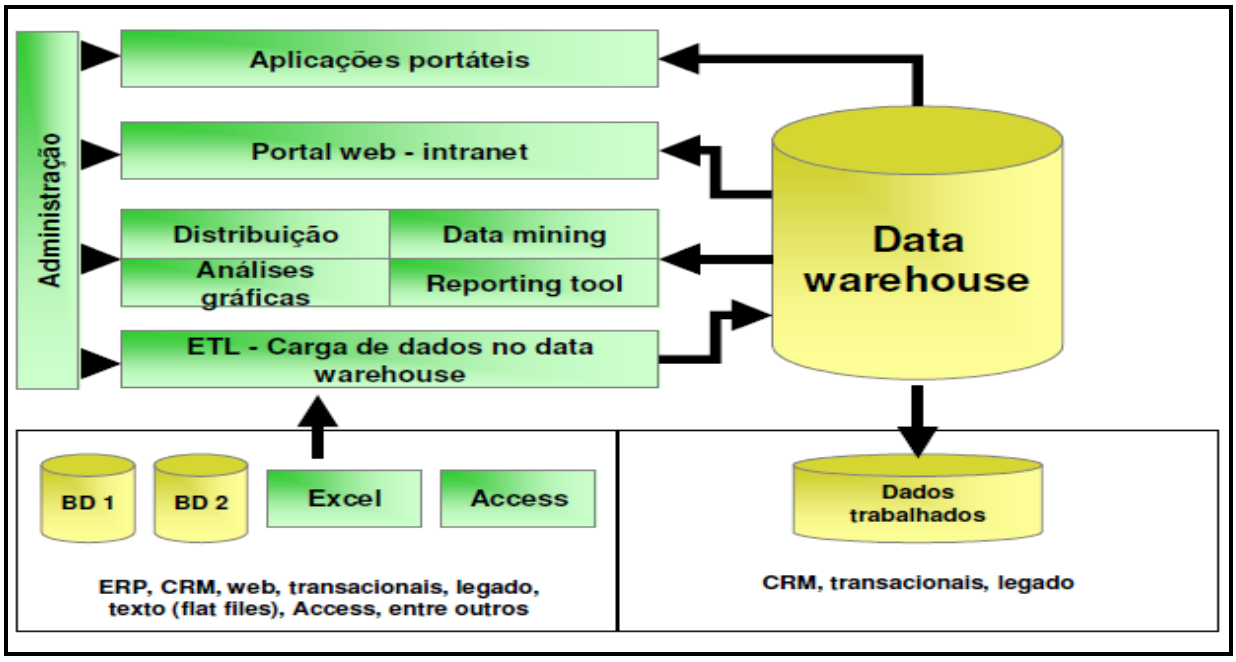


Figura 12 - *Modelo esquemático do ambiente tecnológico de Bussiness Intelingence (Fonte: Leme Filho, 2006)*

1. Sistemas fonte – São os sistemas nucleares necessários para sustentar o negócio. Compreende os vários sistemas de informação, como os ERP, sistemas externos e os próprios ODS. Kimball considera que os ODS são apenas estruturas temporárias. Para efeitos de consulta existem os *reporting* ODS que servem para consultas — *off-line* aos sistemas operacionais.

2. ETL (*Extract, Transform and Loading*) – zona que corresponde à transformação dos dados, desde a sua extração dos sistemas fonte da empresa, passando pela limpeza de erros, torná-los conformes, até ao seu carregamento no repositório central. Os sistemas fontes podem ser externos à empresa.

3. Área de apresentação dos dados – é a zona onde os dados são guardados de forma a permitir as análises multidimensionais.

4. Exploração dos dados (BI) – esta é a zona onde se faz a exploração dos dados pelos utilizadores. Essa exploração vai desde consultas (ad-hoc), relatórios, (*dashboards)*, análises (*what-if)* até ao *Data Mining*.

5. Metadados – é uma zona transversal a todo o DW e consiste em toda a informação que define as estruturas, operações, conteúdo do DW e está dividida em: metadados técnicos, de negócio e de processo.

6. Infraestrutura e segurança – é a estrutura que suporta e protege o DW. Inclui toda a infraestrutura física (servidores, discos, comunicações). A segurança desempenha um papel fundamental uma vez que procura conciliar a facilidade de consulta e acesso aos dados com a privacidade e confidencialidade dos mesmos.

Fatores críticos de sucesso no desenvolvimento de Projetos de Business Intelligence/Data Warehouse

No seguimento da revisão de conceitos de DW/BI, será importante salientar alguns fatores que são relevantes para o sucesso, e os cuidados a ter quando se desenvolvem projetos de DW/BI.

Um estudo da McKinsey[2] de 2012 em colaboração com a Universidade de Oxford revela que metade dos projetos de TI excede o orçamento previsto (McKinsey, 2012).

Um outro estudo[3] publicado no IEEE Computer Society, com base em diversos relatórios, revela que uma percentagem significativa de projetos não chega ao seu término e a percentagem dos projetos cancelados e falhados foi aproximadamente de 34% em 2005 e 26% em 2007 (Emam e Koru 2008).

[2] Inquérito realizado em 2012 em colaboração com a Universidade de Oxford com base em 5400 projetos de IT.

[3] Inquérito realizado em 2005 (236 respostas) e em 2007 (156 respostas) via WEB (www.cutter.com) dirigido aos clientes da Cutter Consortium.

Este último estudo, efetuou um resumo de outros relatórios com percentagens ainda superiores de insucesso, baseados em relatórios do Standish Group, que são muito mais penalizadores. Apesar de alguma reserva de Glass (2006) em relação à metodologia que é implementada no Chaos Report do Standish Group estes números constituem uma importante medida na indústria de TI (Dominguez 2009).

Segundo Moss e Atre (2003), os projetos de Tecnologias da Informação (TI), em geral, e para os projetos de BI em particular, cerca de 60% dos projetos falham por vários motivos, tais como:

- Mau planeamento;
- Má gestão de projeto;
- Incapacidade de ir ao encontro dos requisitos de negócio;
- Má qualidade.

Porquê falham as empresas? Porquê falham os projetos?

São questões que têm sido estudadas ao longo do tempo. Vários autores e estudos abordaram o tema de sucesso ou insucesso dos projetos de TI, englobando também projetos de DW e BI.
Rockart (1979), num estudo do *Massachusetts Institute of Technology*, propôs uma aproximação por fatores críticos de sucesso. Nessa aproximação, e em termos de gestão, define fatores críticos de sucesso (FCS) como, o número limitado de áreas que, se asseguradas, constituem para qualquer negócio, o desempenho competitivo de sucesso para a organização.
Parece existir um paradoxo dentro das organizações, sugerindo um obstáculo à coleta, tratamento, uso e democratização de informações para a formulação de estratégias. Embora as informações sejam consideradas importantes no contexto corporativo, como apoio à boa tomada de decisão, o fator tempo continua a ser uma barreira ao planeamento de ações estratégicas.
Sabe-se que muitas organizações investem em *softwares* de BI, mas nem sempre os dados armazenados e os relatórios produzidos são úteis o suficiente para auxiliar o processo de entendimento deste tipo de ambiente e apoiar nas ações a tomar.
No seguimento destas considerações, poderá concluir-se que o processo de definição de estratégias apoiadas em soluções de BI, é claramente viável e vantajoso para as organizações, contudo, estas devem agir de forma pró-ativa, no sentido de melhor planearem as suas decisões que criem cenários que vão de encontro aos seus valores, missão e objetivos de negócio.

"O Business Intelligence como apoio à formulação de estratégia"

Trajano Leme Filho
Centro Universitário Nove de Julho – UNINOVE

Depois desta breve revisão de conceitos e considerações feitas por vários autores, será importante salientar, para concluir, que o Business Intelligence (*BI*) enquanto processo de coleta, organização, análise, compartilha e monitorização de informações oferece um suporte à gestão de negócios, com o auxilio do *Data Warehouse* (*armazém de dados*), que enquanto sistema computacional, permite um armazenamento das informações (em grande escala) relativas às atividades de uma organização, em forma de bases de dados consolidada. O desenho da base de dados favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão. A exploração de grandes volumes de dados nas organizações pode ser apoiada

por diversas técnicas de *Data Mining (*Prospeção de Dados), que exploraram os dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detetar relacionamentos sistemáticos entre variáveis, detetando assim novos subconjuntos de dados.
O DW, enquanto depósito de dados, projetado especialmente para organizar os dados de tal forma que facilite e viabilize o acesso a informações, o que não é possível no modelo tradicional de armazenamento de dados.
Quando um sistema é construído, o objetivo, é facilitar a coleta e armazenamento de dados no dia-a-dia das organizações, porém o modelo tradicional usado privilegia a gravação e leitura, sem se preocupar com a geração de informações e conhecimento. Por outro lado, o BI são técnicas usadas em conjunto com o DW para analisar os dados. Neste sentido, poderá colocar-se a questão: as ferramentas de BI podem funcionar sem DW? Sendo autossuficientes, usando metodologia própria para organizar e analisar os dados sem DW? De certa forma, sim, mas deve-se ter em conta um detalhe importante, que é: todas as ferramentas de BI, quando não usam um DW usam uma metodologia própria para organizar e analisar os dados, e muitas vezes, usam o DW juntamente com essa metodologia. Assim sendo, se os *softwares* de BI usam metodologias próprias para coletar, organizar e analisar os dados, então está a ser criado um DW, pois usando um DW externo ou sua própria metodologia, as ferramentas de BI, nada mais são do que técnicas automatizadas para geração de informações. No seguimento, de todo este processo, as técnicas de *Data Mining*, vem ajudar a "refinar" os dados e descobrir informações e gerar um conhecimento relacionado com estas informações onde essas técnicas foram aplicadas. Como existem diversas técnicas, conhecidas como algoritmos, o *Data Mining* é sem dúvida o mais complexo, exigindo um conhecimento elevado de quem faz uso, tanto na preparação dos dados, quanto na interpretação das informações.
Em resumo, dados geram informação, informações geram conhecimento, logo, DW armazena os dados de tal forma a facilitar a geração de informações. *Business Intelligence* são as técnicas usadas na geração e análise dessas informações, e *Data Mining* são as técnicas usadas para a descoberta de padrões e tendências, que possam vir a apoiar os processos de decisão nas organizações. Dada a importância desta última técnica de refinamento dos dados, no ponto seguinte será abordado o tema *Data Mining* com mais detalhe.

2.3.1. Breves considerações sobre análise preditiva

A análise preditiva consiste no uso de dados, algoritmos estatísticos e técnicas de *Machine Learning* para identificar a probabilidade de resultados futuros com base em dados históricos.
O objetivo é ir além da estatística descritiva e dos relatórios sobre o que aconteceu para fornecer uma melhor avaliação sobre **o que vai acontecer no futuro**. O resultado final é a simplificação da tomada de decisão e a geração de novos insights que levem a melhores ações.
Os modelos preditivos utilizam os resultados conhecidos para desenvolver (ou treinar) um modelo que possa ser usado para prever valores para dados diferentes ou novos. Os resultados da modelação em previsões que representam a probabilidade da variável-alvo com base na importância estimada a partir de um conjunto de variáveis de entrada. Isso é diferente dos modelos descritivos, que ajudam a entender o que aconteceu, ou dos modelos de diagnóstico, que ajudam a entender as principais relações e a determinar, porquê algo aconteceu.
Cada vez mais organizações estão a voltar-se para a análise preditiva visando aumentar seu lucro e a sua vantagem competitiva. **E os principais motivos são:**

- Crescentes volumes e tipos de dados e mais interesse na utilização de dados para produzir informações valiosas.
- Computadores mais rápidos, mais baratos e *software* mais fáceis de usar.
- Agravamento das condições econômicas e uma necessidade de diferenciação competitiva.
- Com o *software* interativo e fácil de usar, tornando-se cada vez mais predominante a análise preditiva, que não é mais, apenas o domínio de matemáticos e estatísticos.

Os analistas e especialistas de negócios também estão a usar essas tecnologias, para:

- Identificar tendências;
- Entender os clientes;
- Melhorar o desempenho dos negócios;
- Promover a tomada de decisão estratégica;
- Prever o comportamento.

Algumas das aplicações mais comuns das análises preditivas incluem:

Deteção de fraude e segurança – A análise preditiva pode ajudar a pôr um fim às perdas ocorridas por atividades fraudulentas antes que elas ocorram. Ao combinar vários métodos de deteção, tais como: regras empresariais, deteção de anomalias, análises preditivas, *link analytics*, etc.

Marketing – O uso da análise preditiva pode ajudar a entender melhor os clientes. A maioria das organizações modernas usa a análise de dados para determinar as respostas ou compras dos clientes, bem como para promover oportunidades de vendas cruzadas. Os modelos preditivos ajudam as empresas a atrair, reter e desenvolver os clientes mais rentáveis e maximizar seus gastos com o marketing.

Operações – O *analytics* desempenha um papel importante nas operações para muitas organizações, permitindo que elas operem sem problemas e de forma eficiente. Muitas empresas utilizam modelos preditivos para:

- Prever o stock e gerir os recursos das empresas;
- Companhias aéreas usam a análise preditiva para decidir quantos bilhetes devem ser vendidos por cada preço, para um voo.
- Hotéis tentam prever o número de hóspedes esperado em qualquer noite para ajustar os preços para maximizar a ocupação e aumentar a receita.
- A análise preditiva de dados também é usada em recursos humanos, manutenção de ativos, no governo e ciências da vida e saúde.

Risco – Um dos exemplos mais conhecidos de análise preditiva é a pontuação de crédito. As pontuações de crédito são usadas de modo onipresente para avaliar a probabilidade de incapacidade financeira de um comprador para as compras que vão desde casas até carros e seguros.

Para fazer uma análise preditiva, tem de se ter em consideração alguns pontos importantes:

1 Para começar a análise preditiva é necessário **ter um problema para resolver**. O que saber sobre o futuro com base no passado? O que entender e prever? Considerar o que será feito com as previsões. Que decisões serão conduzidas pelos insights? Que medidas serão tomadas?

2 **Existirem dados**. No mundo de hoje, isso significa dados de muitas fontes. Os seus sistemas transacionais, os dados coletados por sensores, informações de terceiros, notas de call-centers, registros na web, etc.

3 Ter um data wrangler, ou alguém com experiência em gestão de dados, para limpar e deixar os dados preparados para a análise.

4 Preparar os dados para um exercício de modelação preditiva também exige alguém que entende tanto os dados quanto o problema da empresa.

5 Definir uma meta é essencial para entender como poder interpretar o resultado (a preparação de dados é considerada um dos aspetos mais demorados do processo de análise).

6 Depois disso, a construção do modelo preditivo começa. Com *software* cada vez mais fácil disponível no mercado, é possível desenvolver modelos analíticos, de preferência com um analista de dados que possa refinar seus modelos e chegar ao melhor desempenho.

7 Implementar modelos, significa colocar os modelos em produção, a trabalhar os dados selecionados, para se obterem os resultados.

8 A modelação preditiva exige uma abordagem em equipa pessoas que entendam do problema empresarial a ser resolvido, saibam como preparar os dados para análise e que possam construir e refinar os modelos e garantir que a organização tem uma infraestrutura certa de *analytics* para construir e implantar o modelo adequado à realidade empresarial.

2.4. ANÁLISE PREDITIVA COM *DATA MINING*

As técnicas de Data Mining surgem, hoje em dia, como uma ferramenta importante e crucial para o sucesso de um negócio. O considerável volume de dados que atualmente se encontra disponível, por si só, não traz valor acrescentado. No entanto, as ferramentas de *Data Mining*, capazes de transformar dados e mais dados em conhecimento, vêm colmatar esta lacuna, constituindo, assim, um trunfo que ninguém quer perder.

A prospeção de Dados é uma das formas a utilizar, e enquanto processo, visa organizar os dados, e encontrar aqueles que a computação consegue dar significado por forma a lidarmos com o volume crescente de dados que têm sido gerados e escolher somente os mais importantes. E a partir disso, objetivar relações relevantes entre eles e reconhecer padrões de comportamento.

Podemos constatar que a quantidade de dados existente no mundo, não para de aumentar (I. H. Witten, Frank, & Hall, 2011). Estima-se que mais de 90% da totalidade do conhecimento que temos hoje começou a ser adquirido por volta de 1950 (Nisbet, Elder, & Miner, 2009). Um fator crítico de sucesso das empresas é a sua capacidade de tomar partido de toda a informação disponível. Este

desafio torna-se mais difícil com o constante aumento do volume de informação, tanto interno como externo às empresas uma vez que quanto maior for a quantidade de informação disponível, menor será a proporção de dados que o ser humano consegue analisar (Angelis, Polzonetti, & Re, n.d.; I. H. Witten et al., 2011).

A informação dispersa pelo volume de dados disponível poderá ser decisiva no sucesso de um negócio e uma mais-valia aquando da tomada de decisão. Torna-se assim indispensável encontrar a melhor forma de extrair toda a informação que se encontra camuflada numa base de dados. As teorias e ferramentas capazes de auxiliar os humanos na extração de informação útil dos grandes volumes de dados disponíveis são a base da descoberta de conhecimento em bases de dados (Lavalle, Hopkins, Lesser, Shockley, & Kruschwitz, 2010).

Para colocar estes conceitos em prática, são usados *software* que trabalham em conjunto com cientistas da informação e profissionais de gestão. Esses programas usam de artifícios como inteligência artificial, estatística e aprendizagem de máquina (*Machine Learning*) para analisar os dados brutos e produzir informações que podem ser usadas para conhecer melhor os clientes e gerar novos indicadores para a empresa.

2.4.1. *Data Mining* / Knowledge Discovery from Databases

Data Mining é uma área relativamente recente que começou a ser desenvolvida nos anos 90 e que ganhou identidade própria nos primeiros anos do século XXI (Nisbet et al., 2009). Alguns autores defendem KDD e *Data Mining* como sinónimo (Kononenko & Matjaz, 2007). No entanto, e tal como defende Fayyad, *Data Mining* é uma etapa específica do processo KDD.

Existem várias definições de DM que dependem da visão de diferentes autores, enumerando de seguida algumas correntes de pensamento:

- Um método direcionado para a descoberta de mensagens escondidas, tais como tendências, padrões e relações existentes nos dados" (Hsu & Ho, 2012).
- A extração de informação implícita, anteriormente desconhecida e potencialmente útil dos dados" (I. H. Witten et al., 2011);
- A aplicação de algoritmos específicos para a extração de padrões dos dados" (Fayyad et al., 1996);
- Utilizado para descobrir padrões e relações nos dados, com ênfase em grandes bases de dados" (Friedman, 1997);

No fundo, o processo de DM consiste na atribuição de significado aos dados e na resultante extração de conhecimento. As ferramentas de DM permitem às organizações tomar decisões fundamentadas e eficientes, uma vez que preveem tendências e acontecimentos através da leitura de padrões encobertos pelas bases de dados (Silltow, 2006).

Data Mining consiste assim na junção de várias áreas de interesse já bastante cimentadas, tais como a análise de dados tradicional, a inteligência artificial e a aprendizagem automática (Nisbet et al., 2009).

2.4.2. Relação entre o *Data Mining* e o Big Data

Geralmente, a prospeção de dados é feita com amostragens menores, o que limita a quantidade de resultados que ela pode oferecer. Quanto à prospeção do Big Data é um processo similar ao que é feito em *Data Mining*, mas numa escala maior em termos de quantidade e tipo de dados. A prospeção de dados é mais usada com dados mais estruturados, como folhas de cálculo, bancos de dados relacionais e dimensionais. Sendo as escalas e os tipos de dados diferentes, os períodos de análise e os seus resultados também diferem. Enquanto o *Data Mining*, se refere a um processo mais pontual, que gera relatórios para responder a questões específicas, o Big Data é uma análise feita duma forma contínua por períodos maiores. Por esse motivo, o Big Data pode ser usado para fazer previsões e indicar caminhos para mudanças estratégicas na forma de gestão.

O termo Big Data está desde logo associado ao volume de dados. Porém, grandes quantidades de dados são apenas um dos aspetos deste conceito.

Uma possível definição de Big Data é referida por Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh e Byers da consultora Mckinsey (2011) que define Big Data como o conjunto de dados, cujo tamanho vai para além da capacidade das ferramentas típicas de bases de dados no que respeita à captura, armazenamento, gestão e análise dos dados.

Por outro lado, Bernard Marr (2013) define este conceito de acordo com a habilidade das pessoas em recolher e analisar um vasto volume de dados que estamos a gerar no mundo. Segundo Mazhelis, as características principais do Big Data estão associadas a um termo específico, da autoria de Doug Laney (2001), os "*3Vs*", que congregam as palavras Volume, Velocidade e Variedade.

2.4.3. Descoberta do Conhecimento em Base de Dados

A descoberta de conhecimento é frequentemente usada como parte integrante da sigla KDD (Knowledge Discovery from Databases), neste caso, aplicando-se para bases de dados trajetórias. KDD é o processo de nível superior na obtenção de fatos através da prospeção de dados e destilação destas informações sobre conhecimento ou ideias sobre o minimundo descrito pelos dados. Este geralmente requer uma inteligência a nível humano para orientar o processo e interpretar o resultado baseado em conhecimentos pré-existentes (Miller et al., 2001). O processo KDD vai procurar qualquer padrão arbitrário de um banco de dados; em vez disso, a prospeção de dados busca apenas aqueles que são interessantes. Esses padrões são válidos (um padrão generalizável, e não simplesmente uma anomalia de dados), romance (inesperado), útil (relevante) e compreensível (pode ser interpretado e destilada no conhecimento) (Fayyad et al., 1996). O processo KDD geralmente envolve as seguintes etapas principais agrupadas em categorias de atividade maiores (Fayyad et al., 1996; Miller et al. 2001; Qi et al. 2003), que também serão seguidas neste projeto, tal como se pode ver na Figura 13.

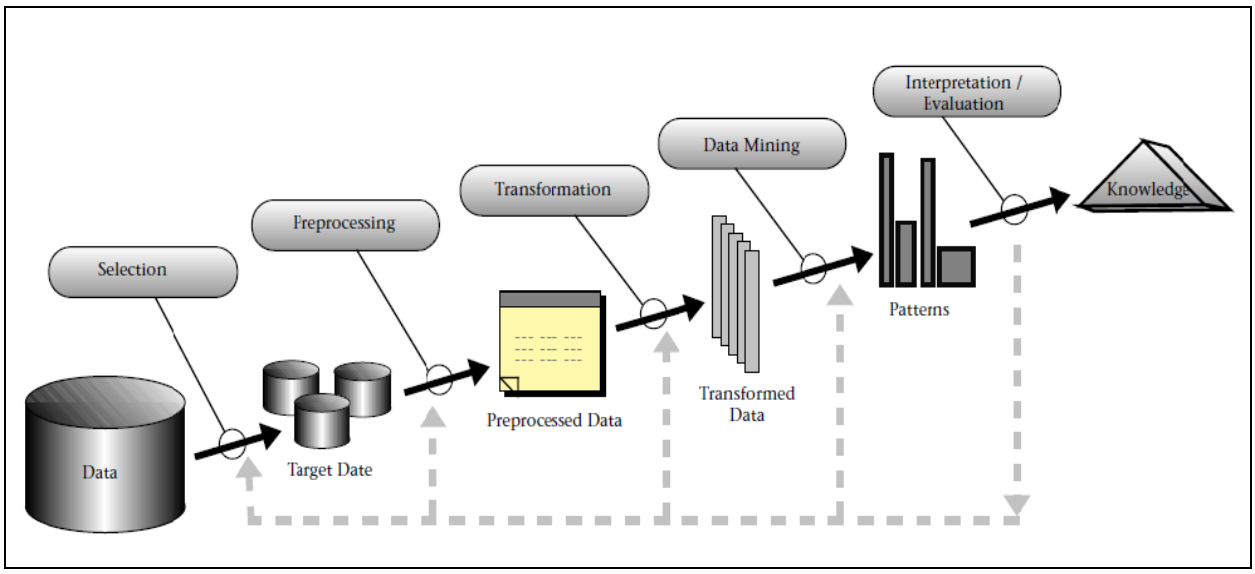


Figura 13 - *Processo KDD (adaptado de Fayyad et al., 1996)*

Estas etapas, de uma forma resumida, consistem (1) na seleção ou segmentação de um subconjunto de dados relevantes para um objetivo em concreto, (2) na eliminação de informação desnecessária e na consistência do formato dos dados, (3) na transformação dos dados em dados adequados e úteis para a etapa de *Data Mining*, (4) na extração de padrões dos dados e (5) na conversão dos padrões obtidos em conhecimento (Pujari, 2001).

São as três primeiras etapas do processo KDD que garantem a qualidade dos resultados obtidos nas duas últimas etapas da figura 13 (Fayyad et al., 1996).

2.4.4. *Data Mining* como ferramenta de apoio a decisão na Aviação

As companhias aéreas estão persistentemente a procura de aprimorar as suas atividades de tomada de decisão para melhorar os processos de negócio e criar vantagem competitiva. Cada dia elas recolhem e armazenam grandes quantidades de dados que podem ser analisados para reduzir custos, aumentar as receitas, melhorar eficiências e prever tendências futuras e comportamentos dos passageiros. A prospeção de dados, que é a extração automatizada de informações preditivas de grandes bancos de dados, que ajudam a ligar grandes volumes de dados heterogêneos e permitem que as companhias os analisem de diversas perspetivas.

A prospeção de dados utiliza algoritmos matemáticos sofisticados de forma automática e faz uma análise sistemática duma grande quantidade de dados para encontrar relacionamentos e avaliar a probabilidade de eventos futuros.

Com base nas consultas abertas dos utilizadores, o *software* de prospeção de dados facilita a descoberta de conhecimento, processo que analisa relações e padrões em dados de transações armazenadas. Assim, o primeiro passo no processo da prospeção de dados é a recolha de informações e dados (geralmente através do uso de uma base de dados). No entanto, a recolha de dados não é suficiente, os utilizadores das empresas precisam de localizar esses dados e aperfeiçoa-los para posterior utilização.

De seguida, a organização precisa desenvolver um modelo para conhecer outras situações e aplicá-lo noutros casos. Como modelo, que é, usa um algoritmo para atuar em conjunto com os dados, para que, os utilizadores finais possam executar consultas para determinar possíveis relacionamentos e definir uma solução para um problema que permita às organizações analisar os dados em diferentes perspetivas, classificá-los e usar essa informação para prever futuras tendências e comportamentos, e com isto, diminuir os custos, aumentar as receitas e melhorar os processos de *pricing*.

Além disso, a prospeção de dados reduz as consultas demoradas e permite que a organização tome decisões de uma forma mais expedita. As empresas podem aproveitar estas técnicas para melhorar fidelização de clientes através da segmentação de mercado, entenda o que seus concorrentes estão a fazer, prever as vendas, monitorizar o desempenho do negócio e detetar fraude, desperdício e abuso (Anderson-Lehman *et al*., 2004).

O case study da Continental Airlines vem demonstrar isto mesmo, que apesar do enorme investimento feito para o suporte de prospeção de dados nas atividades de *Data Warehouse*, o retorno dos benefícios quantificáveis resultantes deste investimento, podem ser ainda maiores.

2.4.4.1. A perspetiva empresarial em geral e nas companhias aéreas

O *Data Mining* responde a problemas empresariais que, num passado recente, constituíam constrangimentos importantes, na medida em que exigiam demasiado tempo no seu tratamento. Assim, estas ferramentas exploram bases de dados em busca de "*padrões escondidos*", encontrando informação de carácter preditivo, que os peritos podem não encontrar unicamente porque se encontra fora das suas expectativas.
A atualidade é fortemente marcada pelas condições financeiras difíceis em que as empresas operam. A verdade é que o controlo das despesas e a redução de investimento constitui a preocupação mais importante dos gestores. Neste contexto, a necessidade de simplificar e racionalizar processos, cortar nas atividades supérfluas e restringir o investimento ao desenvolvimento de projetos indispensáveis está no topo da agenda de todos os gestores, quer sejam públicos ou privados. Paralelamente, nunca como hoje a necessidade de inovação foi tão indispensável e urgente. A inovação nos produtos, mas também a inovação nos processos, nas práticas de gestão, nos canais de comercialização, etc.
Todas as organizações modernas possuem enormes quantidades de dados, que são recolhidos de forma automática e que promovem o aparecimento de mega base de dados. Estas bases de dados constituem a imagem digital da atividade empresarial e continuarão a crescer de forma muito significativa no futuro. Apesar de constituírem uma importante oportunidade de aprendizagem e compreensão da atividade, este recurso tem sido muito pouco explorado. Como descrevem Bisson, et. al. 2010 no texto seguinte:

"Although the volume of data created is expected to increase fivefold over the next five years, best-guess estimates suggest that less than 10 percent of the information created is meaningfully organized or deployed. That number will only shrink as the rate of information production goes up.".

Esta situação foi caricaturada numa capa de um número da revista *The Economist* dedicada ao dilúvio de dados (*The data deluge*) como se pode observar na Figura 14.

Figura 14 – Capa da revista The Economist de 27 de fevereiro de 2010 sobre o "diluvio de dados"

A verdade é que esta abundância de dados continuará a ser um subproduto inesgotável da economia do conhecimento, sendo que os melhores e mais aptos serão capazes de alavancar o crescimento com base neste "novo" recurso. Neste novo milénio a informação constituirá o principal "*driver*" dos aumentos de produtividade e da inovação, e informação é algo de que ninguém se pode queixar de não ter.

Esta ideia, de utilizar os dados para criar valor, aumentando a produtividade e promovendo a inovação, é de tal forma apelativa que nos últimos anos inúmeras publicações, mais ou menos técnicas, têm vindo a destacar esta como uma das tendências mais importantes, tanto em termos computacionais, como em termos de impacto na gestão.

Uma das áreas com maior contribuição para a integração do *Data Mining* no contexto empresarial tem sido o Marketing. De facto, o Marketing constituiu um dos "*early adopters*" desta tecnologia, em particular na tentativa de alavancar o conhecimento sobre o cliente como fator de crescimento empresarial.

Há alguns anos as organizações compreenderam que, na maior parte das indústrias, é mais dispendioso recrutar novos clientes do que manter e aprofundar a relação com os clientes existentes. Como é óbvio os clientes só se manterão como tal, caso estejam satisfeitos com a empresa e a relação que com ela mantêm. A preocupação central das organizações passou a ser o conceito de satisfação do cliente.

A partir desta observação nasceu o conceito de CRM (*customer relationship management*) que consiste na ideia de aprofundar o conhecimento sobre o cliente e por essa via ser capaz de adequar os serviços/produtos oferecidos, desenvolvendo uma relação de maior confiança, geradora de maior satisfação para o cliente e mutuamente compensadora.

2.4.5. *Data Mining* e Estatística

Um dos aspetos mais distintivos do *Data Mining* com a Estatística relaciona-se com a dimensão dos dados, quer em termos de dimensionalidade quer em termos de tamanho. A dimensionalidade retracta a "largura" da base de dados, ou seja, o número de variáveis existentes e suscetíveis de serem utilizadas nos modelos.

Tamanho refere-se à "profundidade" da base de dados, ao número de registos. Em qualquer um destes critérios a diferença entre os conjuntos de dados normalmente tratados na Estatística e no *Data Mining* é abissal.

Parte significativa, dos desenvolvimentos teóricos da Estatística, fizeram-se em torno da seguinte preocupação: ***"qual o número mínimo de elementos que tenho que observar por forma a poder retirar conclusões fiáveis sobre o comportamento da população?"*** Esta é, na maior parte das circunstâncias, uma questão irrelevante no contexto do *Data Mining*, uma vez que a própria população se encontra disponível.

Assim, todas as questões relacionadas com representatividade ou significância estatística passam a ter muito pouca importância, ou significado. Na maior parte dos processos de exploração de bases de dados sabemos de antemão quais os resultados que procuramos. Podendo ficar surpreendidos com os resultados, o facto é que sabíamos à partida que eles existiam e que poderiam ser analisados. Este tipo de interrogação das bases de dados, típico da Estatística, pressupõe que avancemos com hipóteses sobre a natureza do nosso problema, o que no contexto empresarial está na maior parte das vezes relacionado com o comportamento dos clientes. Uma das singularidades do *Data Mining* consiste no facto de procurar informação que o utilizador desconhece existir, o que se poderá traduzir, com propriedade, na exploração dos dados. A descoberta de relações entre variáveis e determinados comportamentos não intuitivos constitui uma das maiores promessas desta nova tecnologia. Esta "*procura automática de novidades*" tem sido um dos aspetos mais enfatizados por todos os que procuram divulgar e promover esta nova "*disciplina*".

Um aspeto verdadeiramente importante, desta característica singular, relaciona-se com o facto de sendo padrões não intuitivos, poderemos esperar que possuam um enorme potencial para se tornarem a fonte de decisões empresariais inovadoras, com eventual impacto na criação de vantagens competitivas.

Obviamente, isto não acontecerá sempre, nem de forma contínua, no entanto, quando acontece pode produzir resultados verdadeiramente surpreendentes.

Um outro aspeto a ter em conta no *Data Mining*, e que de certo modo também constitui novidade em relação a processos anteriores, relaciona-se com a necessidade de compreender o porquê dos resultados. Tende a existir a ideia de que o *Data Mining* constitui algo de sobrenatural que descobre coisas importantes, mas que a forma como o faz está para além da nossa compreensão. Esta imagem, muitas vezes promovida pelos próprios divulgadores da tecnologia, *está* longe de ser correta. Apesar de existirem algumas ferramentas em que a compreensão do modelo subjacente pode ser difícil, como por exemplo as redes neuronais, isto não quer dizer que todas as ferramentas sofram do mesmo problema.

Existe uma forma mais exigente de utilização do *Data Mining*, que passa pela compreensão, por parte do utilizador dos mecanismos envolvidos e que por esse motivo estará em condições de proceder a escolhas na especificação do modelo.

Existe uma certa dificuldade na distinção entre *Data Mining* e a análise estatística, porque basicamente existe similaridade entre ambas, e também pelo fato de que este procedimento de análise ser, geralmente, utilizado conjuntamente com os métodos estatísticos. Entretanto, essa

dificuldade pode ser diluída se as técnicas de *Data Mining* forem diferenciadas, ou pelo menos entendidas como uma adaptação das técnicas estatísticas tradicionais, visando a análise de enormes bancos de dados.
O termo *Data Mining* parece não ser novo para muitos estatísticos e econometristas, e tem sido utilizado para descrever o processo de pesquisa num conjunto de dados na espectativa de identificar comportamentos ou características comuns.
Data Dredging, Data Snooping e Fishing podem ser vistos como sinónimos de *Data Mining*, e têm sido utilizados para nomear a extração de estruturas suspeitas e identificar padrões em conjuntos de dados (Hand, 1998 e Potts, 1998).
Apesar de *Data Mining* e análise estatística terem o mesmo objetivo, a construção de modelos parcimoniosos e compreensíveis, que incorporem as dependências entre as descrições de uma determinada situação e os resultados destas descrições, neste sentido, *Data Mining* e a análise estatística representam dois procedimentos diferentes para análise de dados.
Enquanto a análise estatística tem como base um procedimento hipotético-dedutivo, *Data Mining* é, além disso, um processo indutivo (Hand, 1998).
Assim existe uma forma mais exigente de utilização do *Data Mining*, que passa pela compreensão, por parte do utilizador dos mecanismos estatísticos envolvidos e que por esse motivo estará em condições de proceder a escolhas na especificação do modelo.

2.4.6. Modelação *Data Mining*

Neste ponto iremos abordar o conjunto de tarefas, normalmente, desenvolvidas no âmbito do *Data Mining*. Apesar da diversidade de aplicações podemos caracterizar as tarefas típicas em 2 grandes conjuntos:
Modelação descritiva – onde o objetivo consiste em obter descrições sumárias dos dados e aumentar o conhecimento e compreensão do analista sobre a base de dados
Modelação preditiva – onde o objetivo consiste em "aprender" um critério de decisão que nos permita classificar exemplos novos e desconhecidos.
Estes dois grandes conjuntos englobam todas as tarefas de *Data Mining*, mas abrangem tarefas/métodos bastante diferentes. A modelação descritiva, em particular, engloba métodos que vão da análise de *clusters* à visualização/resumo, passando pelas regras de associação ou *link analysis*. Já a modelação preditiva pode ser subdividida em 2 grandes tipos de tarefa: classificação e regressão.
Globalmente poderemos subdividir os dois grandes tipos de modelação em:

- Classificação
- Regressão
- Clustering
- Visualização/Resumo
- Regras de Associação ou Link analysis

No caso da **classificação**, a tarefa consiste, basicamente, em analisar as características de um novo elemento e associá-lo a uma, de entre um conjunto de classes pré-definidas. Esta é, provavelmente, a tarefa mais comum que encontramos no *Data Mining*. Constitui um imperativo humano, na medida em que o nosso processo de aprendizagem pode ser visto, em larga medida, como o desenvolvimento de um modelo de classificação do real. Por forma a compreender o mundo,

demasiado complexo, estamos constantemente a classificar, é desta forma que aprendemos a distinguir entre laranjas e limões, entre animais e pessoas, entre carros e motas, etc.
No caso do *Data Mining* estes elementos, geralmente, correspondem a registos de uma base de dados nos quais temos um campo em branco que necessita de ser preenchido com um código. Assim, procede-se à classificação de novos exemplos em classes com base num conjunto de treino (a maior parte das vezes utilizando também um conjunto de validação) com exemplos já classificados. A tarefa consiste em construir um modelo que possa ser aplicado a dados não-classificados, por forma a permitir a sua classificação. O aspeto geral do processo de classificação é apresentado na Figura 15, onde se pode verificar a distinção entre a parte do processo que se relaciona com a aprendizagem e a classificação de novos exemplos propriamente dita.

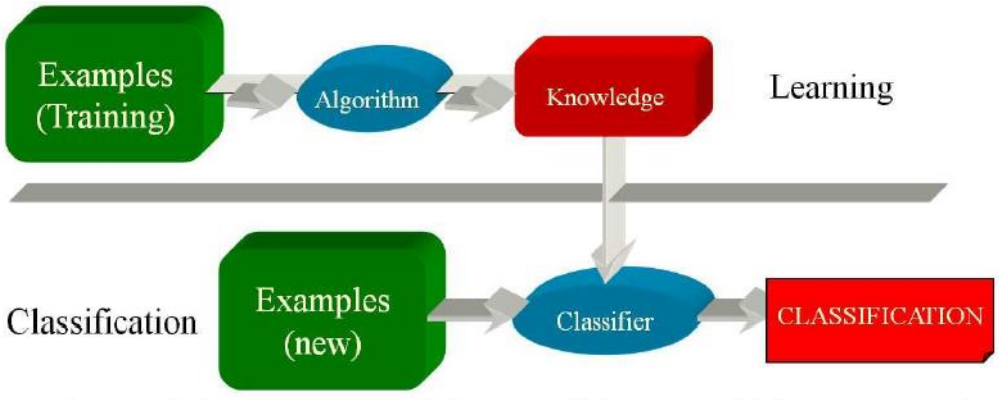


Figura 15 - *Visão geral do processo de modelação preditiva que se inicia com um conjunto de dados (exemplos) pré-classificados onde através de um algoritmo (p.e. regressão, rede neuronal ou árvore de decisão) é extraído conhecimento que é posteriormente aplicado à classificação de novos elementos. (Bação, 2016).*

Para efetuar modelação preditiva, existem várias ferramentas de *Business Intelligence* disponíveis no mercado, contudo, no presente projeto, por questões académicas, em que o *software* SAS, foi a ferramenta eleita, para a realização de vários trabalhos ao longo do mestrado. Neste sentido, no presente projeto, as várias técnicas de *Data Mining* serão realizadas através do SAS. Para o efeito, no ponto seguinte, será feita uma breve revisão, sobre este programa e as suas principais funcionalidades, que posteriormente serão analisadas e selecionadas para serem feitas as prospeções dos dados deste projeto.

3. METODOLOGIA E PROCESSAMENTO

No presente trabalho pretende-se tratar dados relativos à rota Salvador-Lisboa da TAP, extraídos do *software* PROS, que constituem a base dados utilizada. Esta base de dados será tratada, para que possa ser introduzida no *software SAS Enterprise Miner*. O objetivo é fazer uma análise comparativa, o que permitirá investigar a fiabilidade desta alternativa de BI.

O sistema tem um mecanismo de previsão com base nas reservas históricas. Tipicamente é com recurso a essas previsões que se determina a melhor combinação para cada mercado e rota, respondendo às necessidades dos passageiros. Quanto mais essa previsão refletir os padrões reais, mais facilmente o analista garante que está a oferecer ao passageiro/mercado a melhor opção com o maior retorno em receita. Os analistas de *forecast* têm também de ter um conhecimento profundo do mercado em análise para aplicar as alterações necessárias à previsão feita pelos sistemas.

Esta análise comparativa tem como objetivo fazer a pergunta de partida, que é entender como os passageiros da TAP se comportam relativamente à compra de tarifas disponíveis para as Rotas que atravessam o segmento Salvador - Lisboa.

Este é o primeiro passo em busca da vantagem competitiva, no que respeita ao enfoque da gestão O&D/POS (Origem/Destino por Ponto de Venda). Com base nesta premissa, pretendemos identificar qual o intervalo ótimo de dias antes da partida para que as tarifas (disponibilidade diferenciada em Real Time por O&D/POS) possibilitem para a rede e planeamento maiores receitas nas vendas. Isto é, o propósito é prever com a maior acuidade e consistência a procura através do *ODIF* (Origin & Destination, Itineray, fare class), que baseia as suas previsões de vendas em faixas horárias de voos e dados históricos.

Aproveitando esta amostragem, pretende-se ainda fazer uma breve análise preditiva para otimização das receitas da rota Salvador-Lisboa.

3.1. PROCEDIMENTOS METODOLÓGICOS

Para aferir a questão de partida deste projeto, a prospeção dos dados será elaborada no *software SAS Enterprise Miner*. Esta exploração irá basear-se no processo SEMMA que compreende as seguintes etapas:

- *Sample* (Amostrar)
- *Explore* (Explorar)
- *Modify* (Modificar)
- *Model* (Modelar)
- *Assess* (Avaliar)

Estas etapas encontram-se descritas em pormenor na *literature review*, secção 2.4.

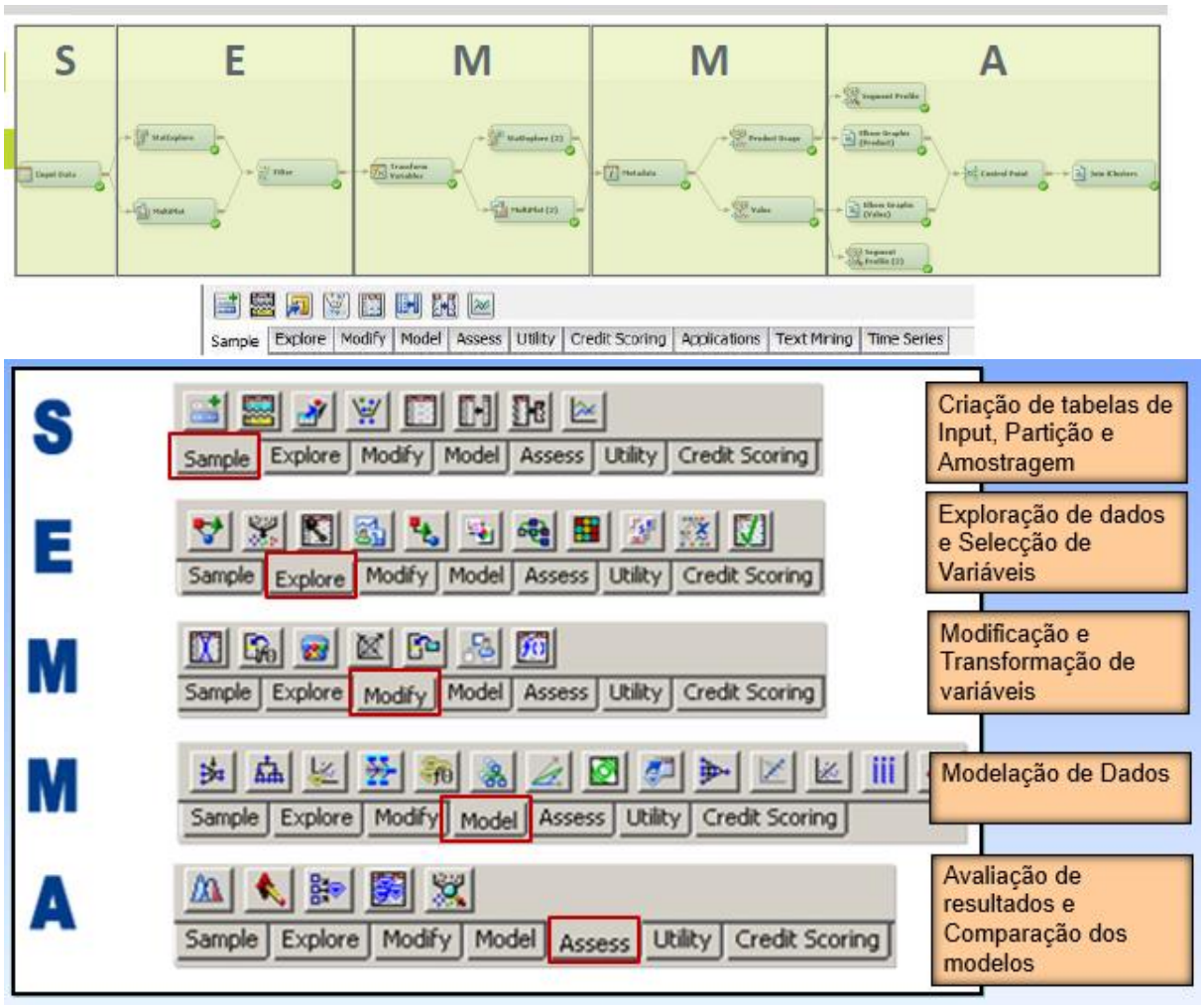


Figura 16 - *Metodologia SEMMA*

O processo é conduzido por um fluxograma, que pode ser modificado e gravado. Este é desenhado para que o analista do negócio, que tem poucos ou nenhuns conhecimentos de *Data Mining*, o possa utilizar para descobrir informação face a novos dados.

O *Enterprise Miner* contém um conjunto de tarefas de análise que podem ser combinadas de modo a criar e comparar múltiplos modelos. Para além destas existem tarefas para preparação dos dados, nomeadamente para deteção de pontos isolados, transformação de variáveis, amostragem e partição dos dados em conjuntos de treino, teste e de validação. As suas tarefas de visualização mais avançadas permitem uma análise rápida e fácil dos dados e informação obtidos.

3.2. DADOS

Os dados originais são formados por 20 variáveis que contém informações socioeconômicas, demográficas e de "consumo" de 46.763 registos.

3.2.1 Classificação de variáveis

Na corrente secção apresenta-se uma listagem das variáveis, bem como uma descrição e informação quanto ao tipo de variável.

Tabela 3 - Variáveis da Base de Dados da Rota Salvador-Lisboa

Variável	Descrição	Tipo de variável
PNR	*Passenger Name Record* - Referência da reserva	Nominal
Marketing_Airl_Code	Código repartido da companhia aérea em que o passageiro viajou	Nominal
FLT	Número do voo para os segmentos LIS-SSA-LIS	Nominal
Flight_Year	Ano da viagem	Nominal
Flight_Month	Mês da viagem	Nominal
DOW	*"Day of the week":* Dia da semana da viagem	Nominal
POS	*"Point of sale":* Ponto de venda na compra da viagem	Nominal
Segment	Segmento/Percurso da viagem	Nominal
ORIG	Ponto de Origem da viagem	Nominal
DEST	Ponto de Destino da viagem	Nominal
Days_to_Flight	Número de dias antes da partida em que foi efetuada a compra do bilhete	Interval
StayDuration_Days	Número de dias em que o passageiro regressa a origem	Interval
Cabin	Cabine onde foi efetuada a viagem	Nominal
RBD	Subclasse da reserva que corresponde a um valor monetário	Nominal
RBD_VALUE	Peso atribuído a cada RBD	Interval
Pax_BKD	Número total de passageiros que reservaram no mesmo PNR	Interval
Owner	Canal de vendas	Nominal
Sales_Year	Ano em que efetuada a compra da viagem	Nominal
Sales_Month	Mês em que efetuada a compra da viagem	Nominal
KO	Quartil da tarifa mais baixa em determinada viagem	Ordinal
KY	Quartil da tarifa mais alta em determinada viagem	Ordinal

3.2.2 Outliers

No corrente trabalho a base de dados que nos foi fornecida já se apresentava "limpa", pelo que não foi necessário tratar os *outliers*.

3.2.3 Missing values

Os *missing values* foram utilizados para a decisão de escolha dos novos ramos da árvore, dado que a opção escolhida no Painel de Controlo do SAS Miner foi *"use in search"* para as 4 árvores de decisão produzidas, conforme se verá mais à frente. Assim, no caso de observações com *missing values*, um dado *missing value* é utilizado como uma observação válida, ao invés de, por exemplo, ser alocado ao nó com maior número de observações.

3.2.4 Data Partition

Os dados foram partidos em três grupos: treino (60%), validação (20%) e teste (20%). Esta divisão permite a construção de uma árvore de decisão, como se verá mais à frente.

Os dados de treino permitem ao algoritmo aprender e mapear as regras de decisão da árvore, treinando o modelo. Segue-se uma segunda fase, dividida em duas partes: validação e teste. O grupo de validação serve sobretudo para avaliar os modelos criados durante a fase de treino e selecionar a abordagem com melhor performance[4]. Finalmente, o grupo de teste permite estimar a precisão da abordagem selecionada em data desconhecida (novas observações), depois de já selecionado o algoritmo preferido.

3.2.5 Variáveis escolhidas para o modelo preditivo

A variável target escolhida foi a *RBD_Value*, que nos dá informação sobre o preço de compra associado a cada classe RBD. Trata-se de uma variável contínua (numérica), pelo que as previsões obtidas com as árvores de decisão construídas são os valores médios dentro de cada um dos subgrupos associados aos nós da árvore.

Na árvore de decisão 1 as variáveis preditivas foram escolhidas de forma automática pelo algoritmo do *software SAS Enterprise Miner*. Nas restantes árvores de decisão as variáveis preditivas foram escolhidas tendo como base o conhecimento empírico do negócio da TAP na rota Salvador-Lisboa. Coincidentemente, as variáveis preditivas das árvores 1 e 2/3 são as mesmas. Isto resultou de uma convergência entre os *outcomes* do modelo preditivo e os procedimentos vigentes da TAP na sua gestão de receitas.

[4] No caso do corrente trabalho, a performance é medida recorrendo aos erros dos quadrados médios, já que a variável target (*RDB_Value)* é uma variável numérica, conforme pode ser constatado na tabela 1. Esta questão será aflorada na secção 3.4.2.

3.3 CLUSTERIZAÇÃO

De forma a preparar a construção de um modelo preditivo através de árvores de decisão – permitindo um melhor conhecimento sobre os dados – procedemos à segmentação dos clientes através de uma análise de cluster. A análise de cluster é uma técnica exploratória de análise multivariada de dados que permite classificar um conjunto de categorias em grupos homogéneos, observando as similaridades ou dissimilaridades entre elas. Podem ser utilizados métodos hierárquicos, que obrigam ao cálculo de uma matriz de semelhança/distâncias ou os não-hierárquicos que se aplicam diretamente sobre os dados originais e que partem de uma repartição inicial dos indivíduos por um número de grupos pré-definido.

Utilizando o programa SAS Miner, é possível definir um número fixo de *clusters* que se quer atingir ou deixar o SAS encontrar o número de *clusters* ótimo (escolher "*Automatic*" no campo "*Selection Criterion*"). Esta seleção automática processa-se da seguinte forma:

- Define inicialmente um grande número de *clusters seeds* preliminares, sendo as observações alocadas à *seed* mais próxima. As médias destes *clusters seeds* são calculadas.
- Um algoritmo hierárquico é utilizado para aglomerar e consolidar os *clusters* preliminares. É calculado o *Cubic Clustering Criterion* a cada passo desta consolidação.
- Este indicador (CCC) permite a escolha do númerode *clusters*. É escolhido o menor número de *clusters* que obedeça aos seguintes critérios: a) o número de *clusters* deverá ser superior ao mínimo ("Minimum") indicado na secção "*Selection Criterium*"; b) O número de *clusters* apresenta valores do CCC superiores ao "*CCC Cutoff*" selecionado na secção "Selection *Criterium*); c) o número de *clusters* é inferior ao "*Final Maximum*" indicado na mesma secção; d) existe um máximo local no número de *clusters*[5].

O método de seleção do número de *clusters* utilizado no âmbito do presente estudo obedeceu a um método hierárquico, que incorpora também algumas características do método *k-means*[6]. Trata-se do método de Ward.

3.3.1 Método de Ward

Este método não utiliza as distâncias entre *clusters* para os combinar, procura antes juntar *clusters* para que a variabilidade dentro de cada cluster aumente o menos possível.
Este método apresenta algumas limitações, nomeadamente:

- Agrega *clusters* com poucas observações;
- Minimiza a variância dentro de cada cluster, pelo que tende a produzir *clusters* homogéneos e uma hierarquia simétrica;
- Tende a encontrar *clusters* de tamanho semelhante e forma aproximadamente esférica;
- Tem uma performance fraca no que diz respeito à aglomeração de *clusters* de forma alongada.

[5] Caso estas condições não sejam respeitadas, o *SAS enterprise Miner* irá escolher como número de *clusters* ótimo o primeiro máximo local.

[6] Podendo ser visto como o análogo hierárquico do método *k-means.*

3.3.2 Seleção do número de *clusters*

A seleção do número de *clusters* foi efetuada através da análise do *Cubic Clustering Criterion.* Conforme se pode observar na figura 17, o máximo local corresponde a 9 *clusters*.

Assim, a clusterização efetuada resulta em 9 *clusters*, conforme se verá na apresentação de resultados.

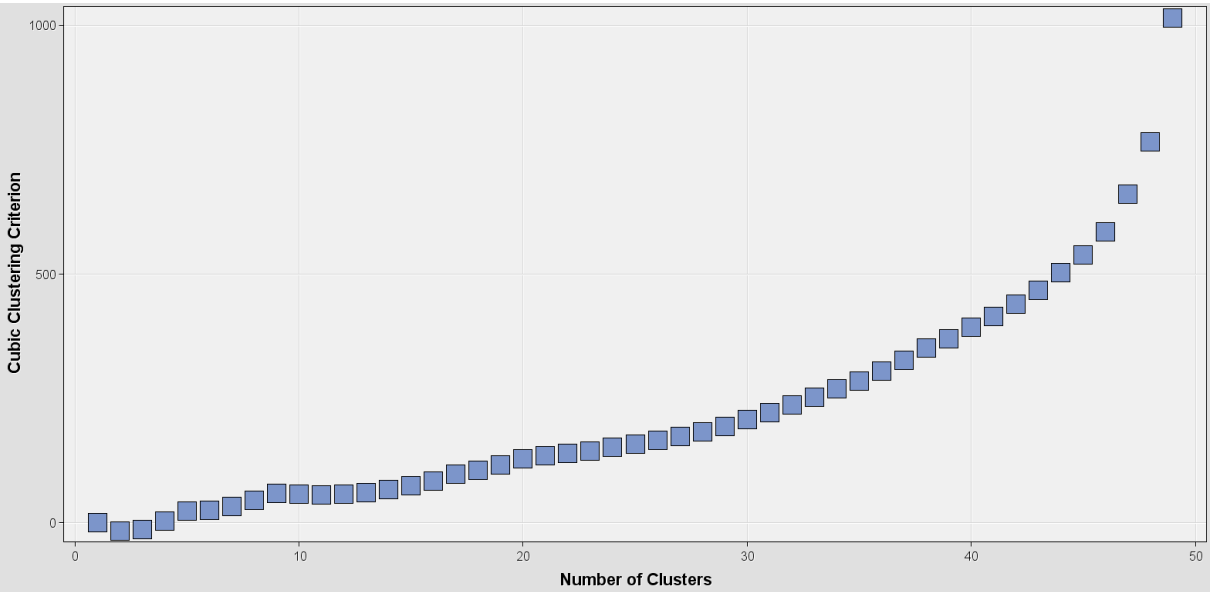


Figura 17 - *There is a local maximum at 9 clusters.*

A segmentação efetuada é sobretudo comportamental, pelo que nos permite agrupar os passageiros em *clusters* de acordo com o seu perfil de consumo. Não nos foi possível efetuar uma segmentação a nível sociodemográfico, pois não tivemos acesso a observações com detalhe sociodemográfico para o corrente trabalho. Conforme referido nas conclusões deste documento, esse será um dos próximos passos em etapas futuras desta pesquisa cujo desafio é compreender cada vez melhor o comportamento da procura nas várias rotas da TAP.

Todas as variáveis do *dataset* foram utilizadas para efetuar a clusterização, excluindo as variáveis *DOW* e *StayDuration_Days*. Estas variáveis, conforme se verá na árvore de Decisão 1 produzida neste trabalho, são variáveis determinantes do *pricing* dos bilhetes vendidos. Assim, optámos por removê-las da fase de clusterização de forma a focar a análise noutras variáveis de interesse como o momento da compra do bilhete (*Days_to_flight*) ou o ponto de venda (*POS_Country*) e a respetiva interação com a variável target do modelo preditivo, o *pricing* dos bilhetes (*RDB_Value).*

3.4 MODELO PREDITIVO: ÁRVORES DE DECISÃO

As árvores de decisão conjugam poder explicativo com simplicidade, tanto a nível conceptual como de potencial interpretativo. Um conjunto de observações denominado *training set* é dividido em *subsets*, de forma a agrupar observações com determinadas características semelhantes nos mesmos nós da árvore. O objetivo é estabelecer regras de decisão que ajudem a prever o valor assumido por uma variável *target*. Assim, a árvore de decisão é um modelo preditivo.

3.4.1 Algoritmos preditivos: o exemplo do algoritmo ID3

Para criar uma árvore de decisão é necessário um *training set* que permita ao algoritmo apreender quais são as características das observações que podem assumir uma função preditiva. Este *training set* é constituído por um conjunto de observações caracterizadas pelas mesmas variáveis de interesse que o grupo de validação. O *training set* é utilizado pelo algoritmo preditivo para definir as regras de decisão que vão vigorar na árvore de decisão. Estas regras serão aplicadas às observações que fazem parte do grupo de validação e permitirão prever o comportamento de observações futuras quanto aos valores assumidos pela variável *target*.

O *SAS Entreprise Miner* utiliza uma variedade de algoritmos, como por exemplo os algoritmos CHAID, ID3 e CRT. A abordagem do SAS para a criação de árvores de decisão incorpora aspetos dos algoritmos mencionados, entre outros. Para o efeito deste trabalho, iremos aflorar o funcionamento do algoritmo ID3, de forma a descortinar a mecânica dos algoritmos preditivos em geral.

O algoritmo ID3 segue a lógica de "dividir para conquistar". O algoritmo procura identificar características que sejam comuns a observações em que a variável target assumiu determinado valor. Depois de identificada a variável de decisão, o algoritmo irá dividir o *dataset* em *subsets*. Esta cisão dá origem aos ramos da árvore. Mais à frente iremos discutir quais são os critérios que permitem identificar a variável de decisão responsável pelo *split*[7].

Depois de criados os ramos, o algoritmo identifica se estes são *pure subsets*. Um *pure subset* é um *subset* em o *outcome* da variável *target* foi semelhante para todas as observações. Se assim for, o algoritmo faz a leitura de que o valor assumido pela variável de decisão no ramo em causa tem um carácter preditivo quanto ao *outcome* na *target variable*. Se não tiver sido atingido um *pure subset*, o algoritmo repetirá este processo até que tal aconteça. Assim, o algoritmo ID3 tem um carácter recursivo, já que vão sendo criados ramos até que os novos *subsets* sejam puros.

[7] Cisão que dá origem aos ramos da árvore, a partir de um nó.

O algoritmo ID3 pode ser sistematizado através das seguintes regras (Quinlan, 1986):

Split (node, (exemples):

- *A ← the best attribute for splitting the (examples)*
- *Decision attribute for this node ← A*
- *For each new child node*
- *Split training (examples) to child nodes*
- *For each child node / subset:*
    - *If subset is pure: STOP*
    - *Else: Split (child_node, (subset))*

Na presença de novas observações, percorrendo os ramos da árvore podemos obter uma previsão do valor esperado da variável *target*. A árvore de decisão configura então um conjunto de regras que devem ser seguidas para fazer previsões quanto a novas observações.

3.4.2 Entropia, *information gain* e variância: identificação de variáveis de decisão

Depois de construída uma árvore de decisão, temos então um conjunto de regras ditadas pelos valores das variáveis preditivas escolhidas pelo algoritmo. No entanto, existe informação valiosa para além dos valores assumidos pelas variáveis de decisão. O número de observações por *outcome* da *target variable* em cada *subset* é também de grande valia para determinar o grau de pureza de cada subset. Esta informação permite não só efetuar uma previsão, como também assignar um determinado nível de confiança a essa previsão, conforme veremos de seguida.

É importante perceber como medir a "pureza" da divisão da árvore por ramos. Um maior grau de "pureza" estará associado a uma maior certeza quanto à fiabilidade da regra de decisão que estamos a definir para o *validation set*. É de notar que necessitamos de uma medida de "pureza" que seja agnóstica quanto aos valores assumidos pela variável *target*. De facto, queremos atingir *subsets* puros, independentemente dos valores em causa.

A entropia (Wang and Suen, 1984) é uma medida de incerteza que respeita o requisito da simetria[8]. A entropia de um subset é dada pela seguinte expressão:

$$H(s) = -p_+ \times log_2(p_+) - p_- \times log_2(p_-) \quad (1)$$

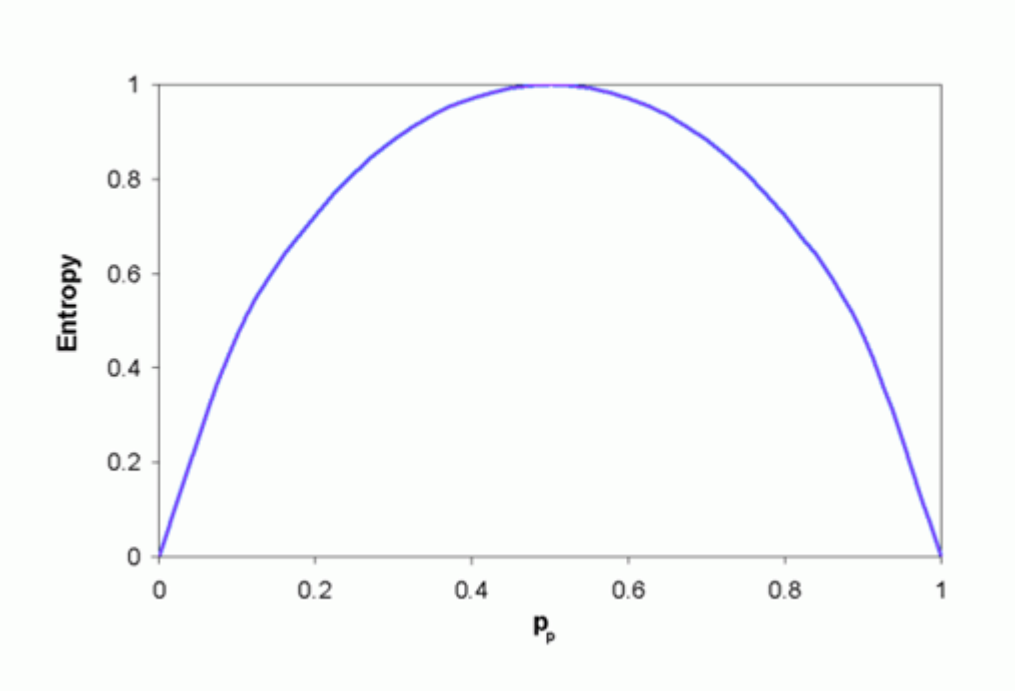


Figura 18 - A entropia é maior quanto maior for a incerteza

A equação 1 adequa-se apenas a um modelo preditivo em que a variável *target* seja binária[9]. A entropia é interpretada como o número de bits necessários para prever o valor assumido pela *target variable*. Assim, o objetivo é escolher variáveis de decisão que criem *subsets* com a menor entropia possível (o mais próximo possível de um *pure subset*). A entropia de um *pure subset* é 0, enquanto a entropia de um *subset* com máximo grau de incerteza (moeda ao ar) será 1.

O ganho de informação dá-nos informação agregada sobre a pureza de vários *subsets*. É calculado efetuando um somatório dos níveis de entropia de cada ramo, ponderado pelo tamanho do *subset* originado:

$$Gain(S, A) = H(S) - \sum_{V \in Valus(A)} \frac{|S_V|}{|S|} \times H(S_V) \quad (2)$$

[8] É necessária uma medida de pureza que valorize da mesma forma um *outcome* positivo e um negativo. O que importa medir é o grau de certeza quanto a essa previsão.

[9] Ou seja, em que a árvore de decisão apenas prevê se a variável *target* assume valor "Sim" ou "Não", por exemplo, em que p_+ é a probabilidade do evento positivo e p_- é a probabilidade do evento negativo. Em casos em que a variável *target* assuma mais do que 2 valores a equação 1 não se adequa.

Em que V é o conjunto de possíveis valores do atributo A, S_V é o tamanho de um dado subset, S é o número total de exemplos e H (S_V) é a entropia do subset. De reparar que H(S) é a entropia antes da divisão em novos ramos, sendo $\sum_{V \in Valus(A)} \frac{|S_V|}{|S|} \times H(S_V)$ a entropia depois de feita a divisão.

Logo, o ganho de informação não é mais do que a diminuição de entropia observada depois de dividir a árvore em novos ramos.

Esta diminuição da entropia é interpretada como um aumento de certeza quanto aos outputs da árvore de decisão (medido em bits). Assim, o algoritmo analisa as variáveis de decisão disponíveis e escolhe como variável de decisão para criar novos ramos a variável que apresenta um maior ganho de informação.

O mecanismo de identificação de incerteza através da entropia e ganho de informação entre os diferentes níveis da árvore permite-nos escolher quais os atributos que mais aumentarão a qualidade do modelo preditivo ao criar novos nós. No entanto, apresenta um problema: tende a favorecer atributos que assumam muitos valores possíveis. Este tipo de variáveis poderá tornar menos provável que as observações do *validation set* sejam convenientemente enquadradas em ramos da árvore de decisão[10].

No caso do presente trabalho, a variável *target* é a *RDB_Value*. Trata-se de uma variável contínua (numérica), conforme visto na secção 3.2. Assim, a medida mais adequada de confiança e precisão nos *outcomes* da árvore de decisão é a variância, medida pelos erros quadrados médios. Esta *fit statistic* é a mais adequada para a previsão de valores numéricos. É obtida através da seguinte expressão:

$$Average\ Squared\ Error = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2 \qquad (2)$$

Em que N é o número de observações, f_i é o valor indicado pelo modelo e y_i o verdadeiro valor dessa observação.

3.4.3 Overfitting e pruning

O algoritmo aflorado nesta secção – o ID3 – é um algoritmo recursivo, que irá dividir os dados do training set em *subsets* continuamente, até que sejam atingidos *subsets* puros. Isto pode significar que haja divisões até que os nós da árvore tenham apenas uma observação, o que não é necessariamente bom. Este fenómeno poderá ser um sintoma de *overfitting*.

$$GainRatio(S, A) = - \sum_{V \in Valus(A)} \frac{|S_V|}{|S|} log \frac{|S_V|}{|S|} \qquad (3)$$

[10] Existe um mecanismo para penalizar o algoritmo de *Information Gain*, que se encontra, no entanto fora do escopo deste trabalho.

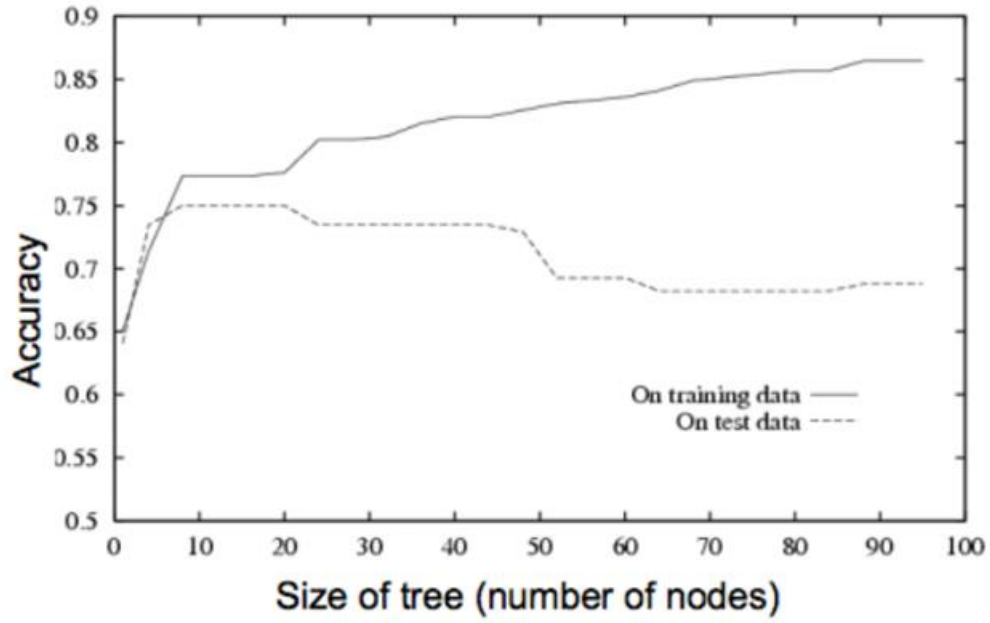


Figura 19 - *Precisão do modelo preditivo pode diferir entre training e validation/test set (fonte: Decision Tree Learning, Duane Lawrence)*

Na figura 19, podemos observar que o nível de precisão da árvore de decisão é incrementado com o aumento do número de nós no *dataset* de treino. No entanto, no *dataset* de validação (teste), o nível de precisão, a partir de dado ponto, cai com o aumento do tamanho da árvore. Isto deve-se ao facto do algoritmo se tornar demasiado específico para o *training set*, sendo incapaz de generalizar.

Existem alguns mecanismos para controlar este fenómeno. Um deles é correr testes de significância de forma a evitar nós originados por um evento contido no training set que tenha ocorrido meramente devido a *randomness*. Outro mecanismo é "podar" a árvore, depois de deixá-la crescer em toda a sua extensão (com ocorrência de *overfitting*). O algoritmo (WF 6.1[11]) simula a remoção de todos os nós, para depois escolher qual o nó que irá ser "podado". De facto, medindo a performance no validation set é possível perceber qual o nó que, quando removido, traz uma maior melhoria na performance da árvore. Este processo é repetido até ao ponto em que a remoção de qualquer um dos nós traz um decréscimo de performance da árvore.

Nas árvores de decisão do trabalho é utilizado o método de minimização dos erros médios quadrados (*Average Squared Errors*), sendo este o método é o mais apropriado para a previsão de valores numéricos (variáveis contínuas).

[11] Algoritmo para pruning.

3.4.4 Random Forest

O Random Forest são uma técnica ensemble, que combina árvores diferentes para obter um modelo mais robusto. O algoritmo desenhado por Leo Breinman em 2001. Uma *random forest* compreende várias árvores de decisão. Uma das características distintivas das árvores de decisão presentes numa random forest é o facto de estas terem uma alta profundidade (*depth* com um valor máximo de 50) e um tamanho pequeno de cada uma das folhas (chegando a 1 observação por folha). O argumento por detrás desta opção metodológica é o facto de se considerar mais robusto utilizar várias árvores de decisão que sofram de *overfitting* do que confiar o modelo preditivo apenas numa árvore de decisão que se acredita afinada ao máximo. É também de notar que os dados utilizados para treinar o algoritmo (*training set*) são uma amostra randomizada do *dataset* completo.

Assim, a principal diferença entre uma *random forest* e uma árvore de decisão normal é o facto de as variáveis de *input* consideradas para a divisão de cada nó serem um *subset* randomizado de todas as variáveis, ao invés da escolha de uma variável apenas para cada *splitting point*. Isto permite reduzir o enviesamento a favor dos fatores com maior influência na variável *target*, permitindo a fatores secundários desempenharem um papel no modelo preditivo.

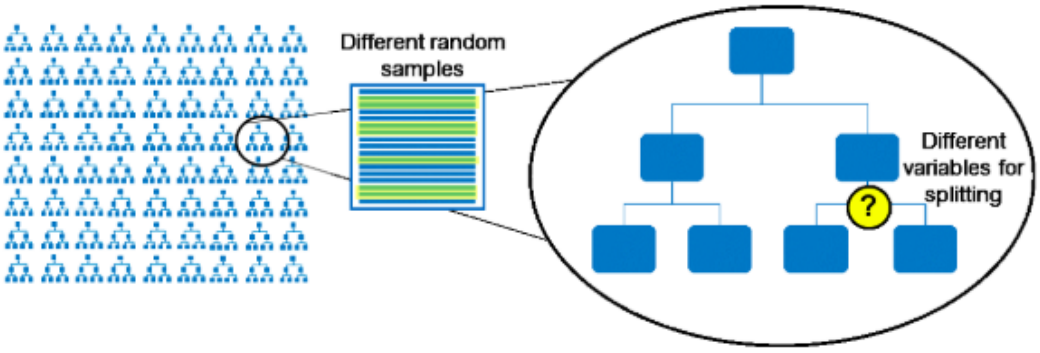


Figura 20 - Uma *random forest* nasce de um conjunto de árvores de decisão (fonte: commnities.sas.com)

Num modelo de *random forest* é efetuada uma média dos *outcomes* dos nós terminais das várias árvores criadas, sendo esta a estimativa do modelo.

4. RESULTADOS

4.1 CLUSTERIZAÇÃO

Foram criados 9 segmentos resultantes da análise de clusterização conduzida. Para a discussão do presente trabalho, considerei relevantes, quer pelo seu peso na amostra, quer pelas suas características comportamentais mais acentuadas, os seguintes segmentos:

- **Segmento 4 – "O Passageiro Organizado"**: este segmento compra o bilhete com antecedência e como tal usufrui de um preço de 381,07€. Trata-se do segmento com maior peso na nossa amostra, correspondendo a um peso de 25,83% das observações recolhidas.
- **Segmento 7 – "O Passageiro *Last-call*"**: este tipo de passageiro faz a decisão de compra perto da partida do voo (em média 60 dias). Como tal, o valor médio do preço do bilhete é de 668,90€, tratando-se do segmento com um valor médio da variável *RDB_Value* mais elevado. Corresponde a um peso de 6,7% na amostra.
- **Segmento 8 – "O Passageiro Oportuno"**: corresponde a um perfil de cliente que consegue adiar a compra do bilhete até uma data próxima da partida do voo (em média 73 dias antes do voo), mas ainda assim consegue um preço médio aproximado da média da amostra (neste caso 391,71€). Corresponde a aproximadamente 25,4% dos passageiros.

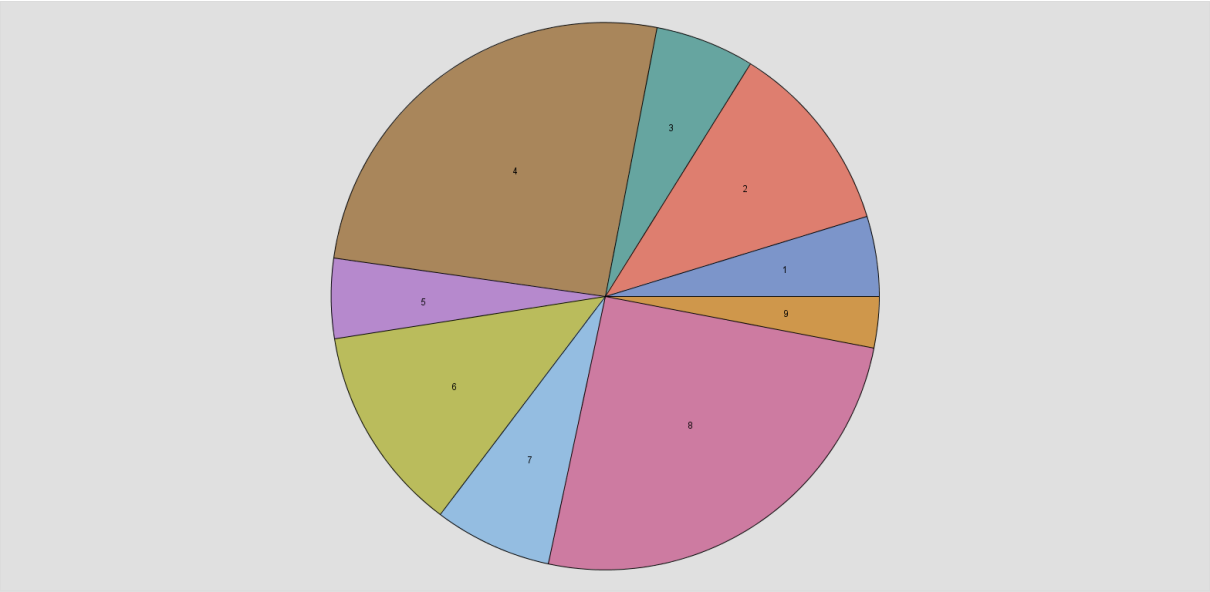


Figura 21 - Os clusters 4 e 8 representam mais de metade dos compradores.

Os restantes *clusters* não são detalhados neste trabalho dado que não oferecem uma interpretação que enriqueça a discussão dos resultados dos modelos preditivos obtidos, ao contrário dos segmentos 4, 7 e 8.

No entanto, parece-nos relevante fazer uma análise global do que cada cluster representa na amostra estudada. Como pode ser observado na figura 21, os *clusters* 1 e 2 contêm sobretudo passageiros que compraram bilhete com um *pricing* médio de 466,73€, sendo dois *clusters* semelhantes na sua distribuição. Os *clusters* 3, 5 e 9 apesar de corresponderem a clientes que pagam um *pricing* irrisório (valores muito próximos de 0€) possuem potencial interpretativo. Isto porque

correspondem a bilhetes *staff* das várias companhias aéreas (*non-revenue*)[12] e também a clientes fidelizados que veem as suas milhas recompensadas com bilhetes em que apenas são pagas as taxas aeroportuárias. Estes *clusters* não são relevantes para a análise conduzida no presente trabalho.

O cluster 6 tem uma distribuição semelhante à dos *clusters* 4 e 8 e é também o terceiro cluster com maior peso na amostra total (12,22%). No entanto, não oferece riqueza interpretativa que justifique incluí-lo na nossa análise inicial.

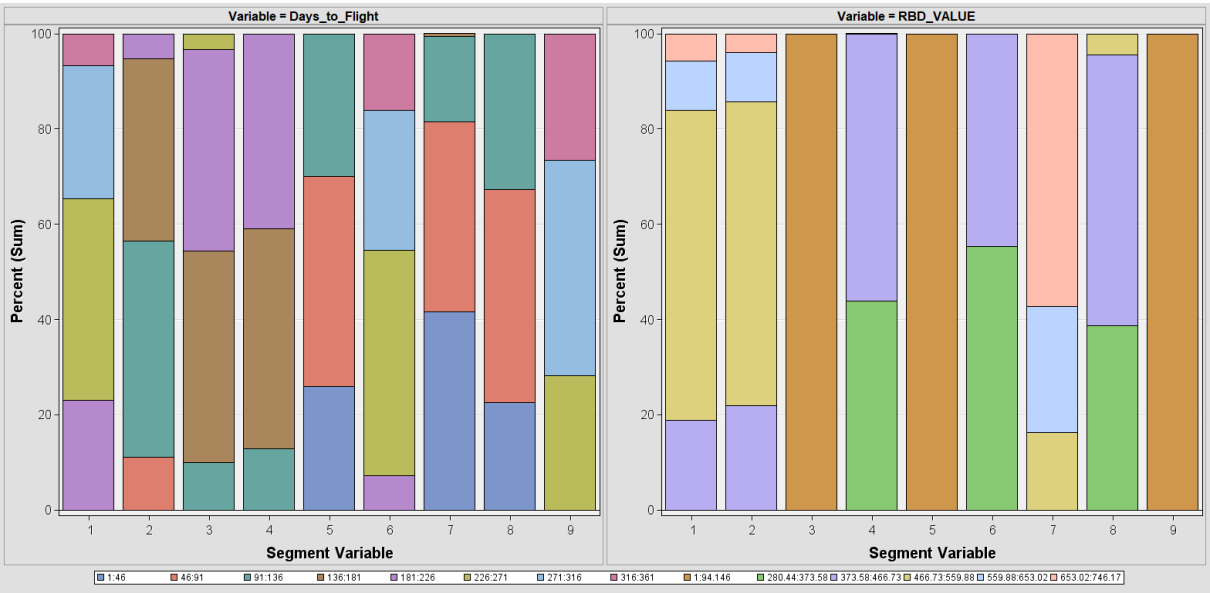


Figura 22 - Os 9 segmentos obtidos, no que diz respeito ao *pricing* e momento de compra.

Os *clusters* 4 e 8 têm um peso conjunto de cerca de 51%. Assim, não é surpresa que apresentem uma distribuição ao nível da variável *RDB_Value* que se aproxime à da amostra total, conforme pode ser verificado na figura 23 abaixo.

Figura 23 - *Os segmentos 4 e (apresentam uma distribuição semelhante à amostra global.*

[12] Esta classe de bilhetes deve-se à existência do ZED (*Zonal Employee Discount*) acordo protocolar entre as várias companhias que permite preços muito baixos para o seu *staff.*

A análise de segmentação efetuada é um preâmbulo para o foco deste trabalho, que é compreender as principais variáveis determinantes do *pricing* para a rota Salvador-Lisboa, bem como aferir a utilidade do *software SAS Enterprise Miner* para o desenvolvimento de um modelo preditivo alternativo ao PROS (*atualmente* utilizado na TAP).

4.2 ESCOLHA DA ÁRVORE DE DECISÃO

Foram construídas 4 árvores de decisão no âmbito deste trabalho de investigação. A árvore escolhida foi a árvore 3, conforme se verá de seguida.

A nossa variável target é a variável RDB_Value. Trata-se de uma variável contínua, medida em euros. Assim, o critério de escolha aplicado é a redução à mínima variância. Assim, o algoritmo C4.5 irá observações, o algoritmo considerará N-1 possíveis *split*ting points. Para cada *split*ting point irá definir um ramo em que as observações assumem valores superiores e outro em que as observações assumem valores inferiores, conforme a figura 24 abaixo.

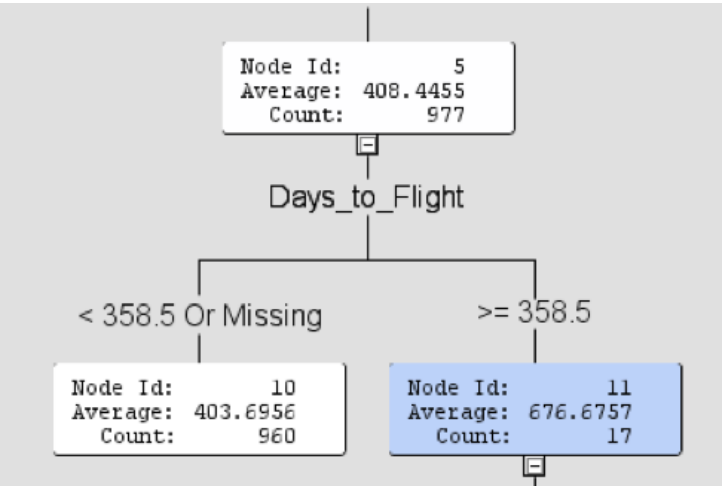


Figura 24 **-** *Splitting point* com árvore de decisão para variável contínua.

Outra questão a considerar é o tamanho mínimo atribuído a cada folha da árvore. Neste trabalho o tamanho mínimo para as folhas da árvore foi definido como 5 para as 4 árvores de decisão produzidas.

4.1.1 Árvore de decisão 1

Para produzir esta árvore o algoritmo C4.5 escolheu as variáveis de decisão mais adequadas de forma automática.

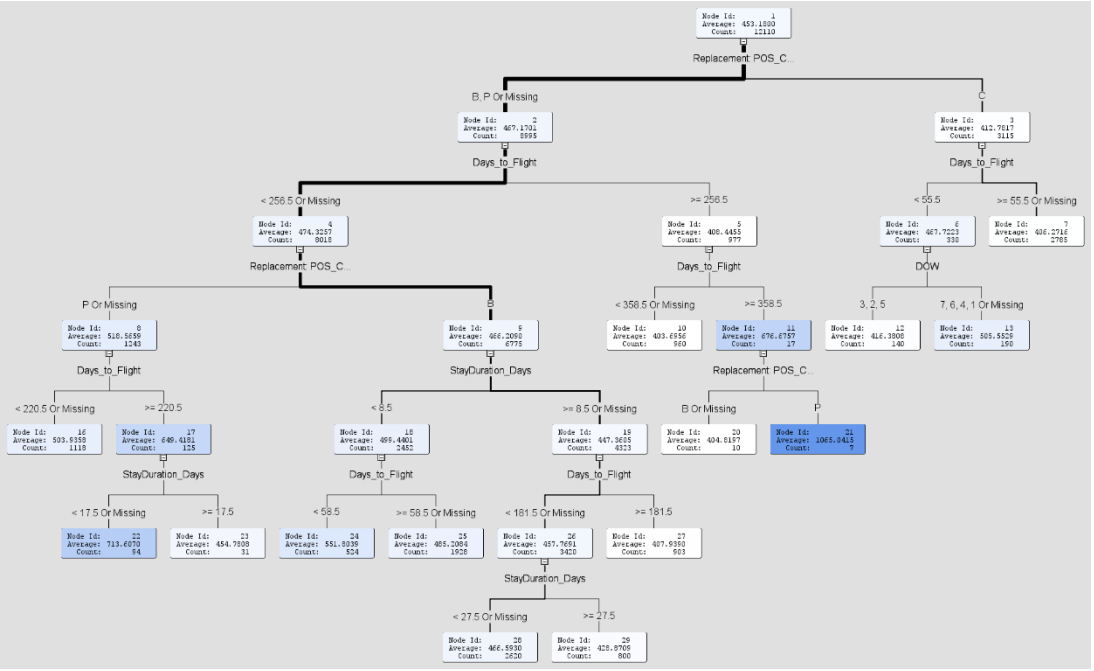


Figura 25 **-** Árvore de decisão 1

Esta árvore apresenta um *Average Squared Error* de 52550,22 no *Validation set.*

4.1.2 Árvore de decisão 2

Para a árvore de decisão 2, as variáveis de decisão foram indicadas utilizando critérios de conhecimento do negócio da TAP e da rota Salvador-Lisboa. Assim, a árvore foi construída com uma indicação prévia quanto às variáveis-chave a considerar para as *splitting rules*. Não foi utilizada a opção *frozen tree*, de forma a criar uma árvore com novos critérios de decisão.

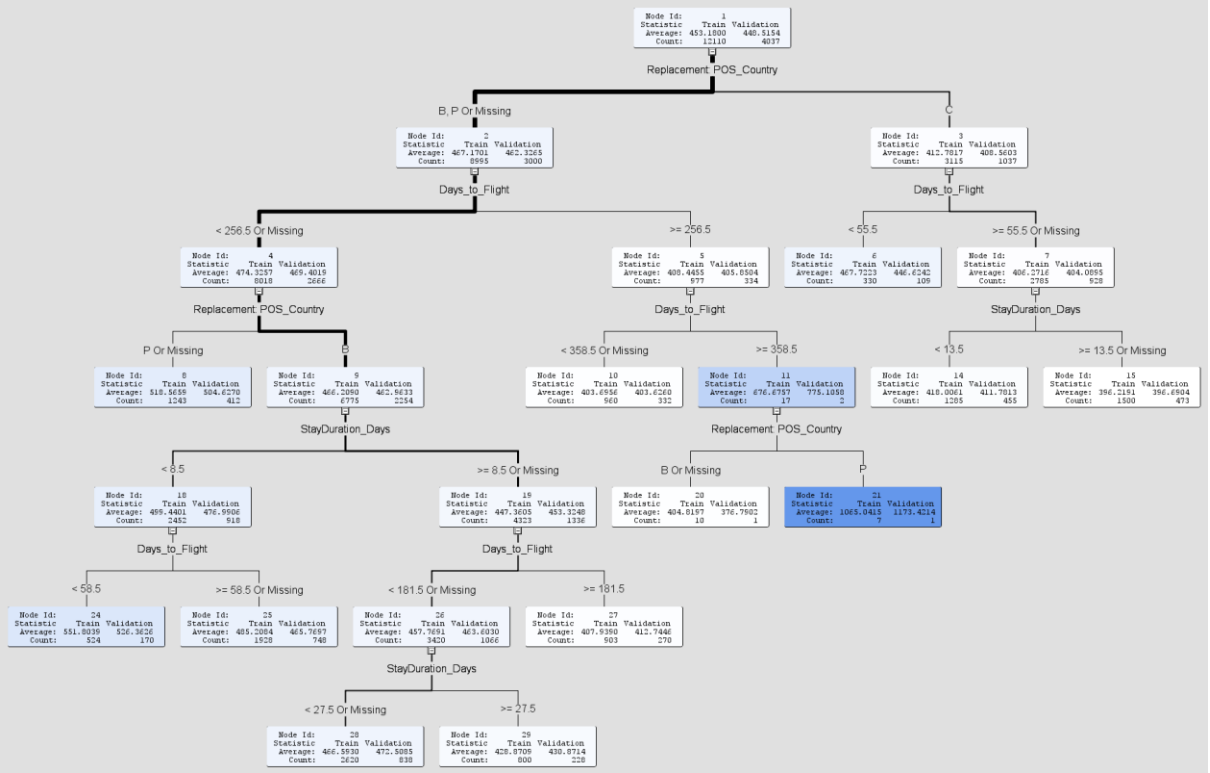


Figura 26 **-** Árvore de decisão 2

Esta árvore apresenta um Average Squared Error de 52067,21 no *Validation set.*

4.1.3 Árvore de decisão 3

Para a árvore de decisão 3, as variáveis de decisão foram indicadas utilizando critérios de conhecimento do negócio da TAP e da rota Salvador-Lisboa, tal como na árvore de decisão 3. Neste caso, foi utilizada a opção *frozen tree,* de forma a importar os critérios de decisão já definidos na árvore de decisão 2.

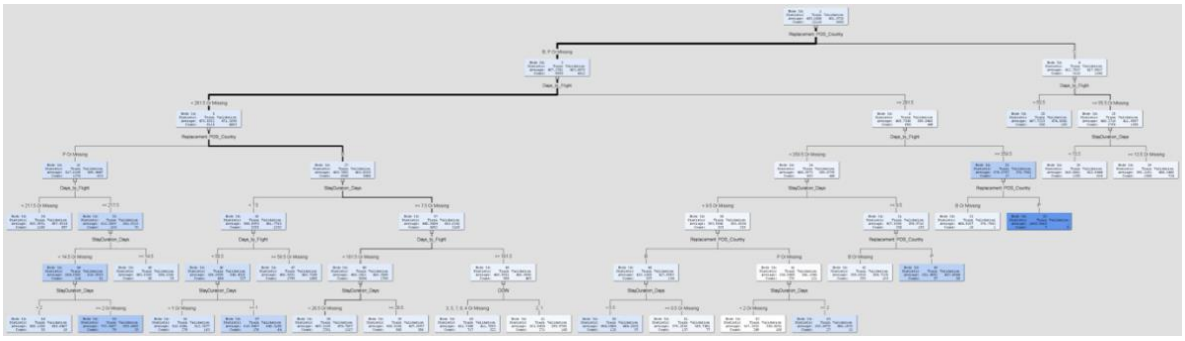


Figura 27 **-** Árvore de decisão 3.

Esta árvore apresenta um *Average Squared Error* de 51.579,9 no *Validation set.*

4.1.4 Árvore de decisão 4

Para a árvore de decisão 4, as variáveis de decisão foram indicadas utilizando critérios de negócio diferentes dos experimentados anteriormente. Assim, a árvore foi construída com o objetivo de testar o efeito das variáveis *POS_Country* (país de venda dos ingressos) e *DOW* (dia da semana em que ocorreu a venda). Foi utilizada a opção *frozen tree*, de forma a importar os critérios de decisão já definidos na árvore de decisão 2.

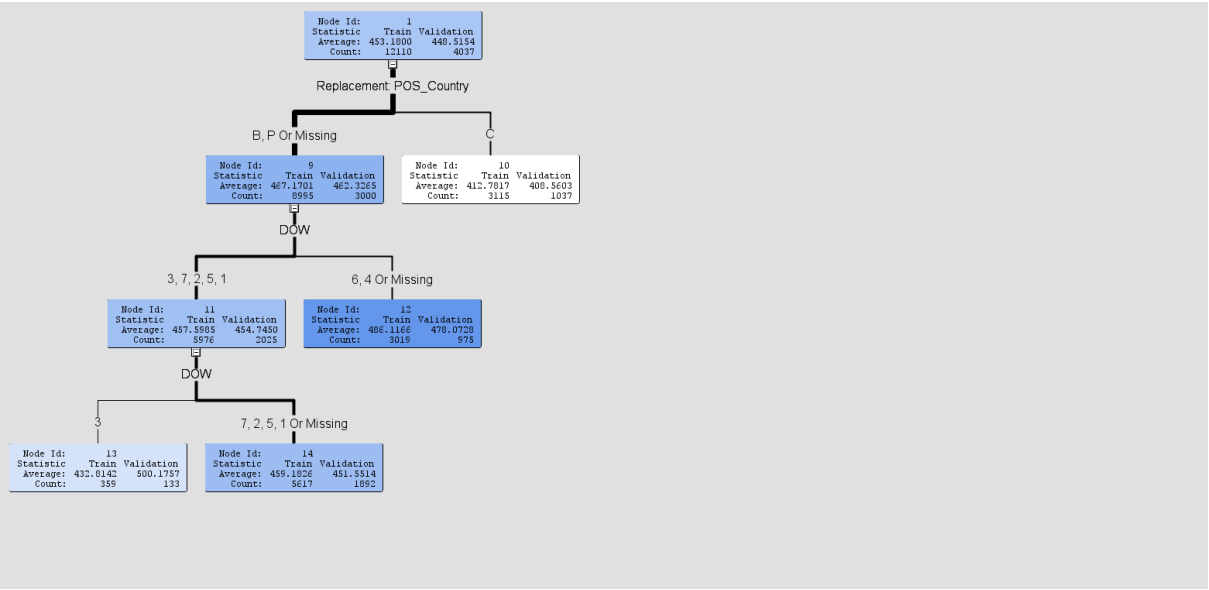


Figura 28 **-** Árvore de decisão 4

Esta árvore apresenta um *Average Squared* Error de 53023,6 no *Validation set.*

A escolha da árvore de decisão mais apropriada foi feita através da análise das *fit statistics* adequadas, após modelar as árvores de decisão. As estatísticas mais frequentemente utilizadas para esta análise são a *misclassification rate* e os erros quadrados médios. A *misclassification rate* é a percentagem de previsões erradas da árvore de decisão. Esta dimensão de análise é particularmente relevante para decisões binárias. Para uma variável target numérica, como é o caso do presente estudo, a estatística de fit mais adequada é a *average squared errors*. Assim, a árvore de decisão escolhida será a árvore com menor valor desta estatística. Das 4 árvores analisadas, a que apresenta um menor *average squared error* é a árvore 3.

4.2 ÁRVORE DE DECISÃO FINAL

A árvore de decisão 3 é a escolhida dado o menor valor da média dos erros ao quadrado que apresenta. Esta árvore tem 22 folhas e 21 nós de decisão, ou seja, 53 nós no total. A figura 34 permite analisar quais as variáveis preditivas encontradas pelo algoritmo em cada um dos nós de decisão, bem como assinalar a presença de nós terminais (*leaves*). Iremos proceder à análise de alguns nós e folhas da árvore de decisão obtida.

A primeira variável de divisão da árvore foi o ponto de venda dos bilhetes *(POS_Country*). Verifica-se que o valor esperado do preço dos bilhetes caso o bilhete tenha sido vendido em Portugal ou no Brasil (países destino e origem da rota, respetivamente) é de 461,09€, sendo o valor esperado de 417,09€. A segunda variável que aporta um maior ganho de informação em ambos os ramos da árvore é o número de dias remanescentes até a data do voo. Naturalmente, verifica-se que uma menor distância temporal influência positivamente o preço. É interessante verificar que para os bilhetes comprados no Brasil ou em Portugal, o *splitting point* escolhido pelo SAS Miner foi 261,5 dias, um valor bastante mais elevado do que o *split*ting point no ramo em que a compra do bilhete foi efetuada num outro ponto de vendas (55,5 dias).

De seguida, verifica-se uma repetição da *splitting variable (POS_Country)*. A variável de decisão seguinte é o número de dias de duração da estadia em Portugal

Na figura 34 (Anexo A) podemos observar uma esquematização das variáveis de decisão relevantes da árvore. Na figura 35 (Anexo A) podemos observar a árvore de decisão final com maior detalhe. De notar que os nós mais brancos são aqueles a que correspondem um maior número de observações (o nó inicial será sempre o mais branco de toda a árvore de decisão). Os nós com tons de azul mais carregado abrangem um menor número de observações.

4.2.1 Um exemplo do poder preditivo da árvore de decisão

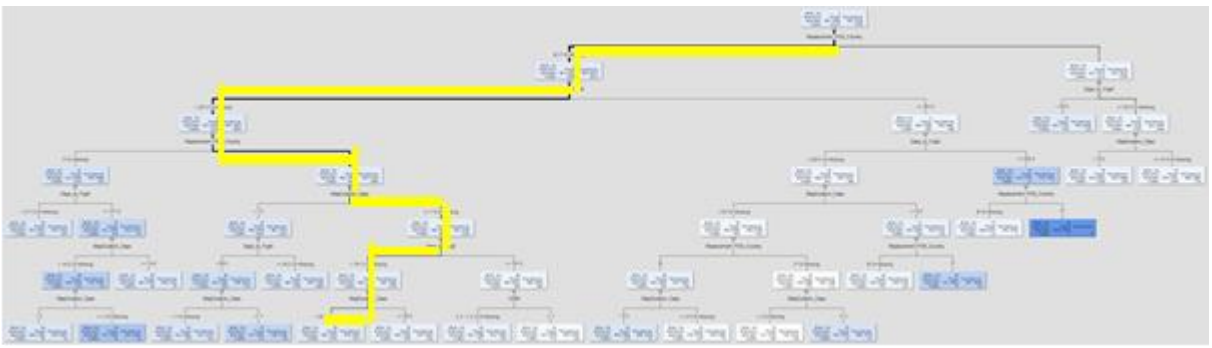


Figura 29 **-** *Exemplo de caminho na árvore de decisão.*

Assim, a título de exemplo a árvore de decisão 3 prevê que uma compra de bilhetes que ocorra no Brasil, a menos de 181,5 dias da partida do voo e cujo tempo de estadia em Portugal seja superior a 7,5 dias terá o valor esperado da variável target (RDB_Value) de 474,79€. Esta previsão corresponde ao “caminho” assinalado a amarelo na figura 29.

4.2.2 *Pruning*: optimização de performance

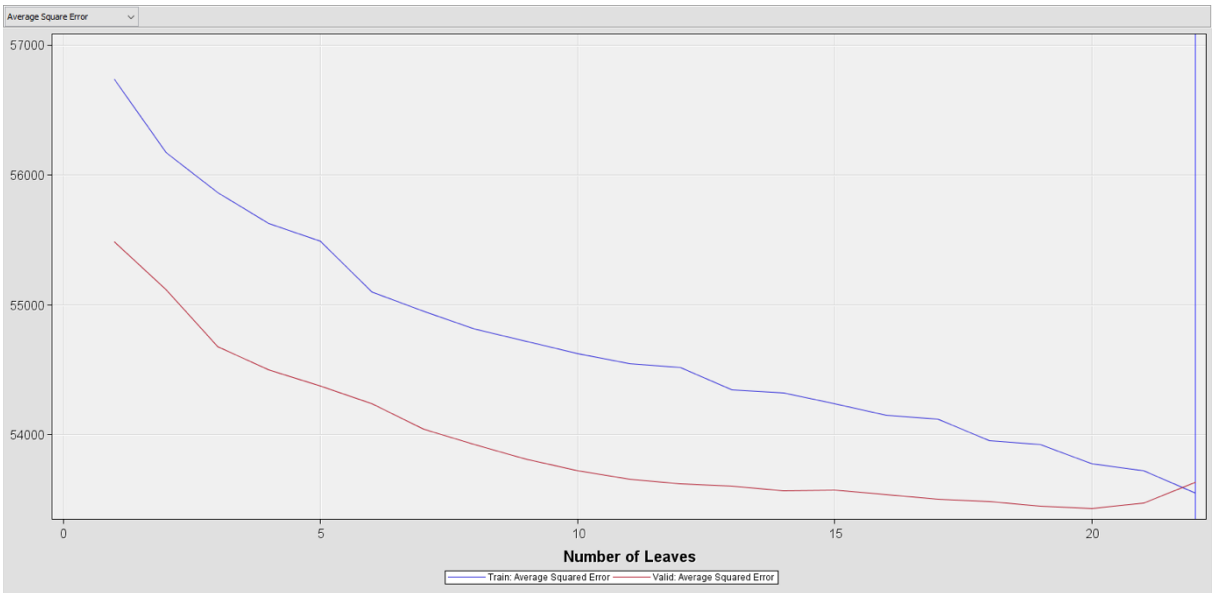


Figura 30 - *Subtree Assessment Plot* – identificação do número de ideal de folhas.

Foi efetuado *pruning* para controlar o problema de *overfitting*. A análise do *subtree assessement plot* permite-nos comparar a performance da árvore de decisão para o set de validação com o set de treino. Mais uma vez, a performance é medida através da variável *Average Square Error.* Na figura 30 é possível observar que 22 folhas (*leaves*) é o número ideal para otimizar a performance da árvore. Para número de folhas superiores a 22 incorremos no problema de *overfitting*, já exposto anteriormente neste trabalho.

4.3. RANDOM FOREST

Uma *random forest* é um conjunto de várias árvores de decisão. O número máximo de árvores geradas pelo processo de *random forest* neste caso foi estabelecido em 100 árvores.

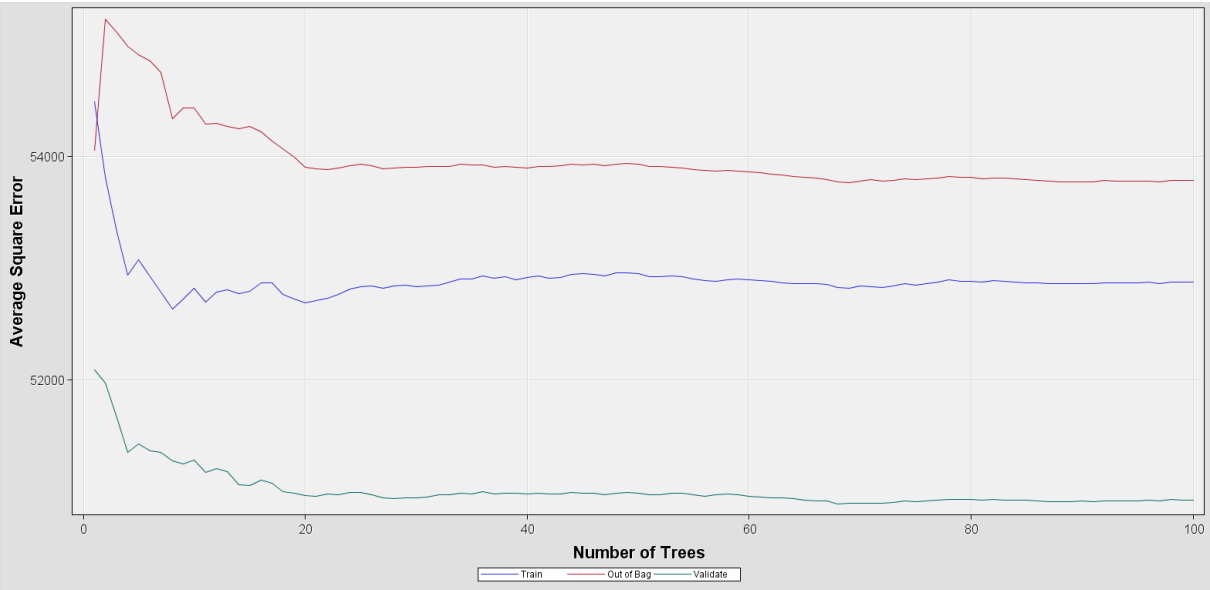


Figura 31 **-** Diferenças entre o set de treino, validação e *out of bag*.

Existem diferenças entre as curvas *Training* set, *Validation* set e *Out of Bag*. A curva *Out of Bag* corresponde aos valores esperados dos erros quadrados médios para um dado subset escolhido de forma independente a partir do *dataset* sete trabalho. Assim, os valores desta curva são considerados as estatísticas de decisão. O facto de a curva do *dataset* de validação ser a que apresenta menores erros quadrados médios sublinha o facto de o método de *random forest* ser orientado uma otimização dos resultados obtidos com o *validation set*.

O número de árvores aumenta a precisão do modelo: na figura 31 é possível observar como os erros quadrados médios diminuem com o aumento do número de árvores. Isto faz com que o modelo produza resultados mais generalizáveis. No entanto, verifica-se que o retorno em termos de aumento de precisão do modelo diminui claramente com o aumento do número de árvores. Em particular, existe um ponto a partir do qual a diminuição do retorno se torna evidente: 20 árvores de decisão.

A escolha das variáveis de decisão utilizadas para fazer o *split* em dois ramos é randomizada no caso de uma *random forest*. Assim, o número de variáveis a considerar para cada nó é uma variável relevante, tal como o número mínimo de observações em cada folha. Neste caso, o número de variáveis são 4, já o número mínimo de observações em cada folha da árvore escolhido foi de 5.

5. DISCUSSÃO DE RESULTADOS

Nesta secção iremos discutir os resultados obtidos, com destaque para a identificação de variáveis-chave para prever o momento de compra de bilhete para a rota Salvador-Lisboa (medido pela variável *RDB_Value*) (1), discussão sobre o contributo das árvores de decisão para as previsões feitas para esta rota (2) e análise à potencial complementaridade entre o *SAS Enterprise Miner* e o PROS (*atualmente* utilizado na TAP para fazer previsões utilizando dados históricos) (3).

5.1 COMPLEMENTARIDADE DO SOFTWARE SAS

A atividade de gestor de rota da TAP exige uma grande rapidez de decisão e capacidade de alavancar na informação disponível para fazer escolhas acertadas tendo em vista a maximização da receita. O facto de o *software* SAS permitir uma leitura fácil da informação, poderá dar um contributo ajustado à necessidade de informação e inteligência rápida, que é característica das funções de um gestor de rota.

Importa assinalar que, no contexto da indústria de aviação, o momento de compra de bilhete está intimamente relacionado com questões de *pricing.* Assim, tendo sido construído um modelo preditivo cujo variável target é o momento de compra de bilhete na rota (medido em número de dias antes da data do voo), os *outcomes* deste terão, naturalmente, valor acrescentado na definição do *timing* de abertura das classes de preços (medidas pela variável *RDB_Value*).

No presente trabalho procurámos investigar a necessidade analítica adicional sobre os fatores determinantes do *pricing* da rota Salvador-Lisboa. *Atualmente*, a previsão da procura é feita com base em dados históricos através de um algoritmo *bayesiano*.

5.2 CLUSTERIZAÇÃO: O MOMENTO DE COMPRA E PERFIL COMPORTAMENTAL DO CLIENTE

É possível constatar que os *clusters* com maior frequência (número de observações), os *clusters* 4 e 8, são aqueles em que o *pricing* médio do cluster está mais próximo dos valores médios praticados na rota Salvador-Lisboa (381,07€ e 391,71€, respetivamente). No entanto, estes correspondem a dois momentos distintos no que concerne ao momento da compra do bilhete. O cluster 4 apresenta um valor médio da variável *Days_to_flight* de aproximadamente 172 dias, enquanto que no caso do cluster 8 os bilhetes são comprados em média 73 dias antes da partida do voo.

Neste momento a TAP não possui inteligência de suporte à decisão de negócio que permite fazer este tipo de constatação. Considero relevante para a gestão de rota da TAP tomar em consideração a existência deste segmento em particular (segmento 8 – "passageiro oportuno"), dado que existe uma aparente assimetria entre o momento da compra (medido pela variável *Days_to_Flight)* e o *pricing* (medido pela variável *RDB_Value*). Este tipo de cliente poderá corresponder a um perfil de cliente com mais orientação para a pesquisa intensiva através de motores pesquisa, o que lhes possibilita aproveitar a abertura de classes RDB´s mais baixas (que correspondem a um *pricing* mais baixo).

É também de destacar que o cluster 7 contém simultaneamente o valor médio mais elevado da variável *RDB_Value* (668,90€) e o valor médio mais baixo da variável *Days_to_flight* (aproximadamente 60 dias). Assim, este grupo corresponde aos consumidores com preferência pela compra dos bilhetes numa data mais próxima do voo, sujeitando-se assim a um *pricing* mais elevado.

5.3 VARIÁVEIS-CHAVE PARA PREVER O *PRICING*

Uma das *researchs questions* fundamentais deste trabalho é perceber quais são as variáveis chave para determinar o *pricing* do bilhete para a rota Salvador-Lisboa, que está – como explanado anteriormente – intimamente ligado ao momento da compra do bilhete. As variáveis escolhidas pelo algoritmo do *software SAS Enterprise Miner* poderão fornecer pistas para perceber melhor o comportamento dos clientes da TAP para esta rota.

As variáveis chave para prever o momento de compra dos ingressos para a rota Salvador-Lisboa são melhor capturadas analisando as variáveis escolhidas pelo algoritmo nos *splits* da árvore de decisão 1, já que nesta árvore o algoritmo escolheu de forma automática todos os ramos.

Verificamos que as variáveis que se assumem com *splitting variables* nos primeiros ramos da árvore são[13]:

- *POS_Country*
- *Days_to_flight*
- *StayDuration_Days*
- *DOW*

Estas variáveis foram também utilizadas nas restantes árvores de decisão, nomeadamente na árvore de decisão 3 – a escolhida como modelo preditivo final. É de assinalar, que a escolha das variáveis preditivas na árvore de decisão 1 (Auto) coincidam com o conhecimento empírico vigente no departamento de gestão de receitas da TAP. Ou seja, o facto de as variáveis *DOW, Days_to_Flight, POS_Country* e *StayDuration_Days* terem sido reconhecidas pelo algoritmo do *software SAS Enterprise Miner* como determinantes para prever a variável *target RDB_Value* valida não só a abordagem utilizada, como os procedimentos de gestão de receita atuais da TAP.
Na figura 32 podemos analisar com maior detalhe alguns *splits* que poderão ter interesse para criar conhecimento sobre a rota Salvador-Lisboa da TAP. A figura 32 deverá ser analisada com suporte da árvore de decisão 3 (figura 27).
O primeiro *split* divide as observações por países de venda do ingresso. Como seria de esperar, a maioria das observações resulta de bilhetes vendidos em Portugal ou no Brasil (países de destino e origem, respetivamente).
Os *splits* seguintes em cada um dos ramos indicam a distribuição das observações de acordo com o dia de compra (medido em dias antes da partida do voo). De entre os bilhetes comprados em Portugal ou no Brasil, a grande maioria é comprada menos de 261,5 dias antes da partida. Nos bilhetes vendidos em pontos de vendas fora dos países de destino e origem o *splitting point* é de 55,5 dias antes do voo e permite-nos perceber que a maioria dos bilhetes são comprados a mais 55,5 dias da data de partida. Investigação subsequente a este trabalho poderá partir deste tipo de análise para segmentar com maior detalhe os perfis de consumo dos clientes da TAP para esta rota.

[13] Ver tabela 1 para explicação detalhada dobre variáveis preditivas.

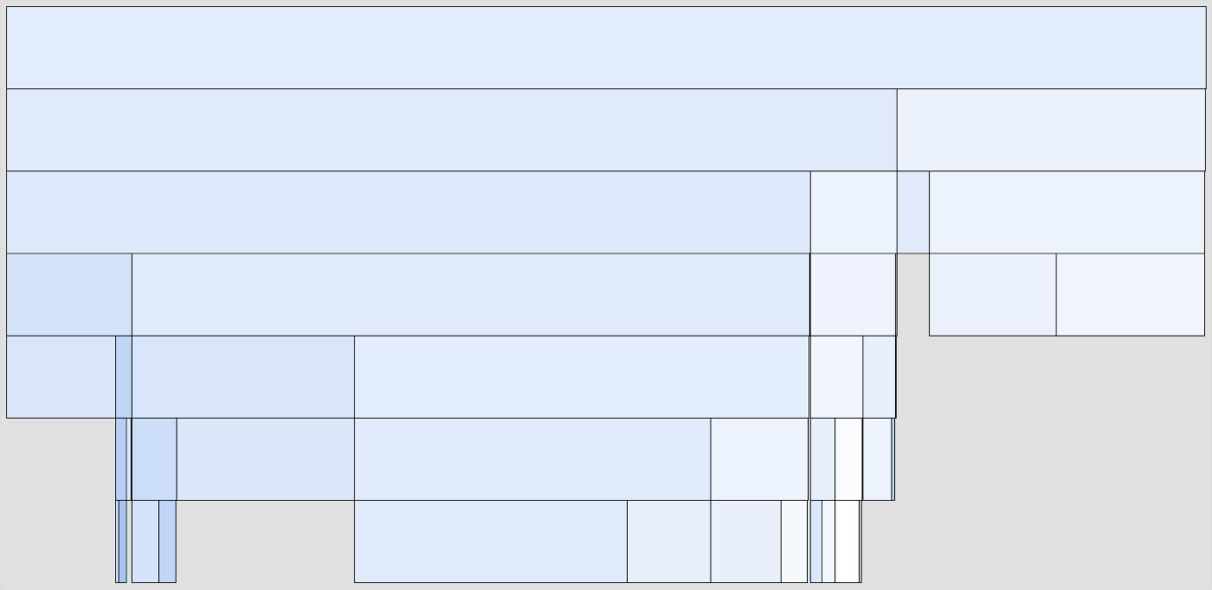


Figura 32 - O *Treemap permite analisar mais facilmente o peso de cada splitting node*

Importa assinalar o peso de um segmento originado por um nó terminal[14]: as observações de bilhetes comprados no Brasil em Portugal ou resultantes de *missing values*, que tenham sido comprados a menos de 181,5 dias[15] da partida e tenham tido um intervalo de pelo menos 7,5 dias[16].

A TAP não tem, neste momento, uma estratégia delineada para fornecer uma oferta focada neste segmento em específico. Esta folha tem na amostra, evidenciado pelos *splits* feito pelo algoritmo da árvore de decisão, A previsão do modelo para a variável *RDB_Value* nesta folha da árvore cifra-se em 474,79€. acreditamos que esta deverá ser uma prioridade para a TAP na gestão da procura pela rota Salvador-Lisboa.

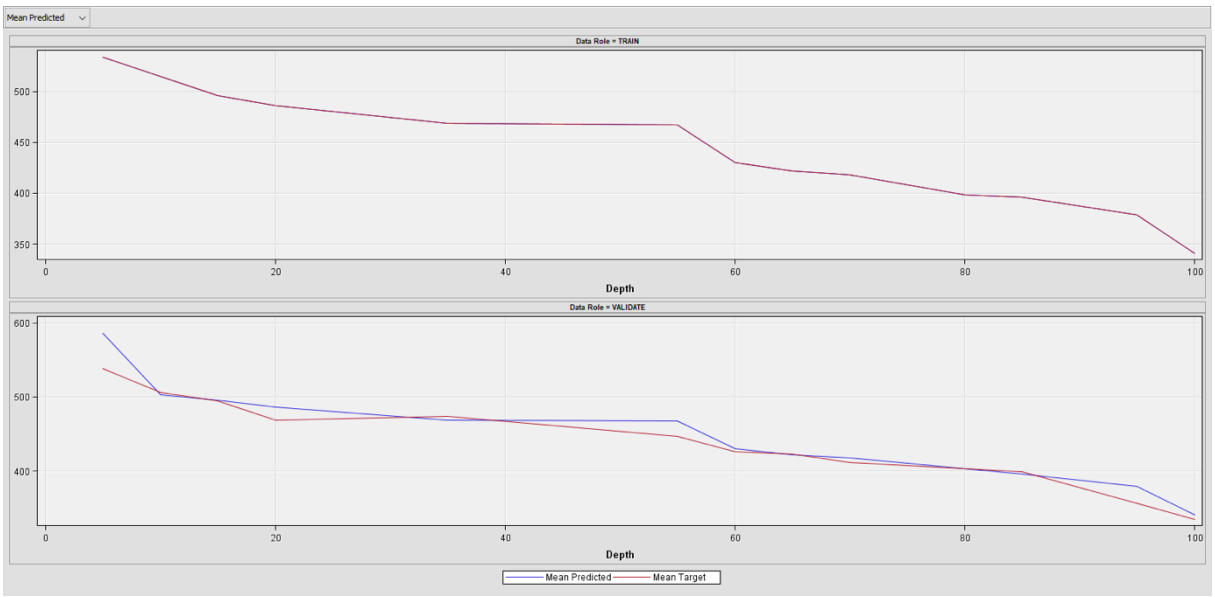


Figura 33 - A *Score Rankings Matrix* permite analisar a distribuição das observações por valor de RDB_Value.

[14] Este *path* já foi assinalado na secção 4.2.

[15] Reparar que este número resulta da intersecção dos conjuntos definidos pelo *split* do nó 3 e nó 37.

[16] Notar que este número resulta da intersecção dos conjuntos definidos pelo *split* do nó 27 e nó 40.

Na figura 33 podemos observar como se distribuem as variáveis por valor assumido pela variável *RDB_Value*, tanto para o *training set* como para o *validation set*. Podemos, por exemplo, verificar que existe uma grande concentração de observações quando a variável *RDB_Value* assume valores entre 475 e 500, bem como entre 400 e 430.

De seguida iremos discutir como as variáveis preditivas identificadas afetam a variável target *RDB_Value*.

Tabela 4 - Importância das variáveis preditivas

Nome da Variável	***Label***	**Número de regras de *splitting***	**Importância *Set* de Treino**	**Importância *Set* de Validação**	**Rácio Importância Validação / Treino**
REP_POS_Country	País de compra do bilhete	5	1,0000	0,7636	0,7636
Days_to_Flight	Dias que faltam para o voo	6	0,8932	1,0000	1,1196
StayDuration_Days	Duração da estadia (dias)	9	0,8722	0,6741	0,7729
DOW	Dia da semana da compra	1	0,2108	0,0000	0,0000

Como se pode verificar a variável *DOW*, apesar de ter sido identificada como uma variável crítica, assume uma importância nula no *set* de validação. Assim, podemos concluir que o dia da semana a que é feita a compra do bilhete não assume carácter preditivo quanto ao momento de compra de bilhete para a rota Salvador-Lisboa.

Por outro lado, a variável *Days_to_Flight* assume um peso de 1,0000 no *set* de validação. Este fenómeno está associado íntima ligação entre o momento de compra (medido em dias antes da partida do voo) e a estratégia de *pricing*, como discutido na Introdução deste trabalho. Isto confirma e valida aquilo que já é uma prática na gestão de procura da TAP para esta rota: o principal determinante para o fecho e abertura de classes é o número de dias até à partida do voo. Esta prática, hoje baseada em heurística e conhecimento histórico da rota, fica aqui validada por esta análise de importância das variáveis. Esta variável tem um impacto positivo no *pricing.*

Os dias de estadia (*StayDuration_Days*) assumem uma importância superior no *set* de treino em relação ao *set* de validação. Verifica-se que os dias de estadia têm um efeito positivo no valor previsto para a variável *RBD_Value* na maioria dos nós de decisão em que é a variável. Isto poderá acontecer devido a um aumento do valor percecionado da viagem com o aumento de dias de estadia. Esta hipótese comportamental poderá ser testada com *datasets* de outras rotas.

O ponto de venda (*REP_POS_Country*) é a variável com maior grau de importância no *set* de treino (1,0000), tendo uma importância de 0,7636 no *set* de validação. O facto de o bilhete ser comprado nos países de origem ou destino (Brasil e Portugal, respetivamente) parece ter um impacto positivo no *pricing* dos bilhetes da rota Salvador-Lisboa. No primeiro *split* o valor esperado da variável

RBD_Value no *Validation Set* é de 417,90€ caso o bilhete seja comprado num país que não Brasil ou Portugal. Caso o bilhete seja comprado nos países de origem ou de destino, o valor esperado sobe para 463,09€. Esta assimetria poderá dever-se a uma maior inevitabilidade das viagens entre Brasil e Portugal nas observações em que o bilhete é comprado nos países de origem ou destino. Esta hipótese deverá ser confirmada em desenvolvimentos futuros deste trabalho através de uma metodologia de regressão linear.

5.4 CONTRIBUTO DAS ÁRVORES DE DECISÃO

Importa também perceber se as árvores de decisão poderão aportar algum valor à estratégia de *pricing* da TAP. Um ponto de partida para esta discussão é perceber quais as vantagens que a utilização de árvores de decisão traz, genericamente.

- Os modelos de decisão baseados em árvores de decisão permitem tratar variáveis nominais sem necessidade de criar várias variáveis *dummy*: os valores que podem ser assumidos pelas variáveis nominais são considerados no momento do *split*, ao contrário do que acontece, por exemplo, num modelo de regressão linear;

- No momento do *split* as árvores de decisão não ignoram observações com *missing values*, alocando-os antes a um dos ramos criados. Este procedimento (alocação a um dos ramos) poderá igualmente acontecer quando, no *validation set* as variáveis de *input* assumem valores distintos dos valores conhecidos no *training set*;

- Os modelos de decisão baseados em árvores de decisão permitem também capturar relações não-lineares entre os dados;

- Por fim, uma árvore de decisão tem uma visualização e interpretação imediatas e intuitivas.

Neste trabalho foi também explorado o modelo de *random forest*. Este modelo permite atribuir importância a variáveis secundárias, mas cujo efeito na variável *target* se quer contabilizado. Apresenta, no entanto, a desvantagem de acarretar dificuldade acrescida ao nível da interpretação.

Como abordado na introdução e *literature review*, as previsões de procura atuais feitas pela TAP socorrem-se sobretudo de dados históricos e conhecimento empírico dos gestores de rota[17]. À luz dos resultados obtidos, é possível perceber que existe uma variável considerada no atual processo de abertura e fecho de classes – o dia da semana em que é comprado o ingresso - que não possui importância no modelo estimado.

Assim, em primeiro lugar a utilização de árvores de decisão pode ser um mecanismo de identificação das variáveis-chave a ter em atenção pelo gestor da rota. As variáveis que o algoritmo seleciona como sendo as variáveis de decisão ideais para gerar o *split* serão as variáveis que melhor descrevem o comportamento do cliente no momento de compra de bilhete.

Adicionalmente, a árvore de decisão permite um mais profundo conhecimento da *willingness to pay* dos clientes, associado ao seu comportamento. Com recurso a árvores de decisão é possível prever qual o *pricing* esperado para clientes com determinadas preferências e necessidades ao nível do

[17] Sendo eu próprio o gestor da rota Salvador-Lisboa, posso testemunhar a importância deste conhecimento empírico e histórico.

momento de compra. Nas árvores de decisão produzidas, analisamos qual é o *pricing* esperado conforme o ponto de venda do bilhete, o dia da compra, os dias de estadia em Portugal e o dia da semana em que a compra é efetuada[18].

No procedimento atual da TAP ao nível de abertura e fecho de classes, é possível ter uma classe aberta apenas para um *point of sales* específico. Esta possibilidade faz com que a análise comportamental dos clientes da rota Salvador-Lisboa permita diferenciar padrões por ponto de venda. Estes padrões deverão ser utilizados pela TAP para adequar a sua oferta às preferências dos passageiros, de forma a maximizar a receita. Vamos utilizar um exemplo concreto para ilustrar esta possibilidade.

Analisando a árvore de decisão 3, verifica-se que os passageiros que compraram o ingresso menos de 261,5 dias antes da partida do voo apresentam um valor esperado de *pricing* de 509,44€ no *validation set* caso tenham comprado o bilhete em Portugal, enquanto os passageiros cujo ponto de vendas é o Brasil têm um valor esperado ao nível do *pricing* que é inferior (463,83€ para o *validation set*). Assim, a TAP ganhará em disponibilizar classes diferenciadas para estes dois pontos de vendas. Aqui se vê o valor acrescentado que uma árvore de decisão pode ter ao nível da gestão de receita de uma companhia aérea, pois permite obter este tipo de *insights* de uma forma simples e visual.

De seguida apresentamos alguns exemplos de análises possibilitadas pela árvore de decisão 3:

- O intervalo de dias de estadia que propicia uma maior *willingness to pay* (695,44€ no *validation set* e 737,54€ no *training set*) é entre 2 e 14,5 dias, para passageiros que comprem bilhete em Portugal e entre 217,5 e 261,5 dias antes da data do voo[19];
- A grande maioria[20] dos passageiros que compra bilhete para a rota Salvador-Lisboa a partir de um ponto de venda que não seja Portugal ou Brasil, fá-lo com antecedência superior a 55,5 dias da partida do voo;
- Uma compra de bilhetes que ocorra no Brasil, a menos de 181,5 dias da partida do voo e cujo tempo de estadia em Portugal seja superior a 7,5 dias terá o valor esperado da variável target (*RDB_Value*) de 474,79€.

[18] A variável *DOW* apresenta, no entanto, um nível de importância baixo no modelo estimado.
[19] Nós 44 e 55 da árvore de decisão 3.
[20] 1308 Observações de 1543 no *validation set* – nós 4 e 20 da árvore de decisão 3.

Acreditamos que as árvores de decisão poderão ser um instrumento útil na definição de uma estratégia de *pricing* em que o conhecimento histórico do comportamento da procura é complementado por uma maior atenção às variáveis que determinam esse comportamento de compra.

No entanto as árvores de decisão não respondem a uma grande parte das necessidades de um gestor de rota no contexto do seu trabalho diário. Algumas necessidades às quais as árvores de decisão não dão resposta são:

- As árvores de decisão não estão orientadas para o *forecast* de um número de passageiros, o que é um elemento fundamental do dia-a-dia de um gestor de rota da TAP;
- Fatores externos relevantes podem não ser considerados nas regras da árvore, pelo que o seu poder explicativo fica limitado por não capturar estes fatores (exemplo: conjuntura político-social).

Podemos aferir que as técnicas de *Machine Learning* e *Data Mining* utilizadas neste projeto podem servir de suporte na obtenção de melhores resultados, numa lógica complementar aos modelos matemáticos existentes, que têm como objetivo a maximização de receita. Estas técnicas permitem descrever com maior riqueza de informação o comportamento esperado da procura. Com a leitura que estas técnicas nos apresentaram para a presente análise da rota Salvador-Lisboa, conseguimos nomear quais das dimensões atualmente utilizadas (com a metodologia "*Bayesian Forecasting"* – Guilhotina) são mais importante e vão dar à companhia maiores benefícios.

Assim, das 11 métricas disponíveis, podemos manipular e dar maior ênfase no nosso trabalho diário, àquelas que, segundo os resultados obtidos pelo SAS, têm maior preponderância na definição da procura. Métricas/dimensões essas que são utilizadas numa base diária pelo analista com a função de *Pricing & Demand*.

As principais ações do analista são: gerir a previsão de procura dos passageiros através do sistema *Origin & Destination* III; incorporar as mudanças dos diversos mercados através da atribuição de influências na procura; criar "*Sponsorships"* através da procura em novos mercados; monotorização dos períodos de férias nos mercados, bem como eventos especiais; rever os alertas que indicam as variações das reservas provenientes das previsões de procura; rever a "performance" da previsão de procura.

Acreditamos que a leitura e análise dos resultados dos modelos utilizados neste trabalho sejam uma mais-valia e suporte fundamental na tomada de decisão das nossas ações diárias, com a finalidade comum de obtenção de receita máxima e adequação da oferta às necessidades dos clientes.

6. CONCLUSÕES E RECOMENDAÇÕES FUTURAS

Conforme discutido na introdução do presente trabalho, o controlo da disponibilidade e preço de lugares num voo é crítico para a otimização de receita. Assim, a previsão do momento de compra de lugares é um processo crítico para o sucesso comercial de uma empresa de aviação.

Neste trabalho procurámos traçar um perfil comportamental de cliente. Este perfil é definido, no caso do corrente trabalho, pelo momento de compra do bilhete, dias da estadia em Portugal, dia da semana em que o bilhete é adquirido e pelo ponto de venda. Uma correta previsão do *pricing* associado a um cliente com determinado perfil comportamental traz valor à TAP, já que lhe permite conhecer com maior profundidade o perfil de cliente associado a cada tarifa.

Complementarmente, procedemos a uma análise de *clusters*, que permitiu traçar segmentos de clientes que correspondem a determinados padrões comportamentais. Assim, identificámos um segmento de clientes que compram bilhete perto da data de voo (Segmento 7), um segmento de clientes que compra com antecedência e usufrui de melhor *pricing* (Segmento 4) e, finalmente, um segmento de clientes que comprando passagem relativamente perto da data de partida do voo, consegue ainda assim usufruir de um *pricing* atrativo. Acreditamos que este poderá corresponder a um tipo de cliente mais orientado para a pesquisa de preços através de motores de busca - *metasearch* relacionados com a venda de bilhetes, com um perfil mais oportunista em relação à compra do mesmo. No entanto, não nos foi possível validar esta hipótese, por limitações dos dados existentes. O aprofundamento destes resultados preliminares obtidos é uma oportunidade para desenvolvimento do corrente trabalho em investigações futuras.

Uma análise com detalhe sociodemográfico dos clientes traria grande valor acrescentado a este trabalho e à TAP, mas não foi possível recolher dados demográficos. Esta limitação deveu-se aos dados disponíveis na área de gestão de receita. Existem dados demográficos noutras áreas da empresa, como CRM, Marketing, Fidelização, entre outras, que – no entanto – não foram disponibilizados para o efeito deste trabalho. Considero que no futuro, caso seja possível realizar um estudo em que todas as secções da empresa estejam *onboard*, o potencial das árvores de decisão poderá sair reforçado. Assim, para além do perfil comportamental no momento da venda, seria possível usar como variáveis preditivas algumas características sociodemográficas que acrescentassem valor à previsão do *pricing* praticado.

Os modelos preditivos em utilização nas diferentes companhias aéreas têm a função de modelar a procura por bilhetes de cada rota aérea, bem como identificar os determinantes dessa procura. O presente trabalho não tem a pretensão de dar origem a um modelo preditivo que mude o paradigma do *forecast* de receitas no sector da aviação. Tem sim a ambição de averiguar quais são as variáveis determinantes para a definição do preço na rota Salvador-Lisboa da TAP, utilizando uma técnica de forecasting pouco utilizada no ramo da aviação: as árvores de decisão.

Consideramos de especial interesse a comparação entre os resultados obtidos através da árvore de decisão escolhida com as previsões de vendas da TAP quanto à rota Salvador-Lisboa. Essa possibilidade deverá ser explorada no futuro. A base de dados que resulta do modelo preditivo *bayesiano* usado pela TAP é dinâmica, sendo que a utilizada neste trabalho é estática. Logo, a base de dados do PROS contemporânea da base de dados utilizada neste trabalho já foi alimentada com

novos *inputs*. Assim, um exercício deste tipo teria de partir de esforço coordenado entre o departamento de gestão de receitas e o departamento de IT.

Este trabalho poderá ser um ponto de partida para uma discussão estratégica para a TAP quanto ao seu modelo de gestão de receitas. Em rotas com maior *track record*, a grande quantidade de dados históricos permite modelar com grande fidelidade o comportamento de clientes que procurem ingressos no futuro para essas mesmas rotas. No entanto, a técnica de árvores de decisão poderá – fazendo uso de um *dataset* relativamente reduzido, usado como *training set* – fornecer *inputs* sobre o comportamento de clientes associado a cada tarifa.

A identificação dos principais determinantes da escolha do momento de compra configura também um passo importante para que a TAP possa conhecer melhor as preferências e comportamentos dos seus clientes. Este conhecimento e inteligência poderá aportar valor tanto numa perspetiva de negócio como numa lógica de melhoria de experiência dos clientes no momento da compra.

A metodologia utilizada carece de futuras validações com outras rotas e o seu interesse poderá ser testado através da comparação com dados históricos. Seria também relevante testar esta metodologia com o sentido oposto da mesma rota (Lisboa-Salvador), algo que foi feito numa fase inicial neste trabalho, sendo depois omisso para efeitos de simplificação e concisão.

Outras abordagens metodológicas serão fundamentais para, no futuro, complementar os insights deste trabalho. Será, nomeadamente, interessante aferir como as variáveis preditivas encontradas influenciam o *pricing* praticado através de uma regressão linear[21].

A análise conduzida através de árvores de decisão é especialmente relevante se tivermos em conta que, com o aparecimento das companhias *low-cost*, se colocam vários desafios às companhias aéreas, que deixaram de competir puramente ao nível do preço. O modelo tradicional foi irrompido e neste momento as companhias aéreas devem focar-se em providenciar a melhor experiência possível ao cliente, bem como garantir que a sua oferta se adequa à procura de mercado. Nesse sentido, considero estratégico para a TAP que haja um maior foco no mapeamento do comportamento dos clientes e sua relação com o *pricing* estabelecido.

[21] Análise dos coeficientes da regressão.

7. BIBLIOGRAFIA

Anderson-Lehman, R., Watson, H.J., Wixom, B.H., & Hoffer, J. A. (2004): *Continental Airlines flies high with real-time business intelligence*. MIS Quarterly Executive, (3)4, pp. 163-176.

Angelis, F. De, Polzonetti, a, & Re, B. (n.d.): *Optimising Performance with Business Intelligence.*

Bação, F. (2016): *Apontamentos da cadeira de Data Mining de Pós-Graduação Gestão do Conhecimento e Business Intelligence.* Nova Information Management School.

Berry, M. Linoff, G. (1997): *Data Mining Techniques, for sales, and customer support, John Wiley and Sons*

Boisot, M. and Canals, A. (2004): *Data, information and knowledge: have we got it right? IN3: UOC*. (Working Paper Series); DP04-002.

Bisson, P. Stephenson. E. and Patrick Vinguerie, S. (2010): The global grid. Mckinsey Quarterly, 1-7.

Breinman, L. (2001): *Random Forests. Machine Learning Journal*, Volume 45, issue 1, pp 5-32.

Davis, G.B. (1974): *Management Information Systems: Conceptuals Foundations*. New York: McGraw-Hill Book Company, 278.

Data Mining Professional Society, Website, http://www.kdnuggets.com/

Dominguez, J. (2009): *The curious case of the chaos report 2009.* Project Smart.

Eman, K., Koru, A.G. (2008): *A replicated survey of IT software Project failures.* IEEE Softw; pp. 84-90.

Evelsen, B., R. Karel, et al. (2010): *Agile BI Out Of The Box, Forrester Research*: pp1.

Fayyad, U. M., G. Piatetsky-Shapiro and P. Smyth (1996): *From Data Mining to knowledge discovery: an overview. Advances in knowledge discovery and Data Mining*. Menlo Park, CA, USA, American Association for Artificial Intelligence: 1-34.

Friedman, J.H. (1998): *Data Mining and statistics: what's the connection*. 29th Symposium on the Interface.

Gartner. *Business Intelligence. (*2013):
http://www.gartner.com/technology/core/products/research/topics/businessIntelligence.jsp (acedido em 15 de janeiro de 2013).

Glass, R.L. (2006). The Standish report: *does it really describe a software crisis? Communications of the ACM*. Volume 49, issue 8, pp 15-16.

Han, J., Kamber, M. (2001): *Data Mining – Concepts and Techniques*, Morgan

Hand D.J., (1998): *Data Mining: statistics and more? The American Statistician*, 52, 112-118

Hsu, C.C., & Ho, C.C. (2012): *The design and implementation of a competency-based intelligent mobile learning system. Expert Systems with Apllications*.

Inmon, Willianm H. (1997): *Como construir o data warehouse*. Rio de Janeiro, Campus.

Khandekar, A., Sharma, A. (2006): *Organizational Learning and Performance: Understanding the Indian Scenario in Present Global Context. Education & Training, 48* (8): 682-692.

Kimball, R. (2002): *Data Warehouse Designer-Two Powerful Ideas, The Foundation for Modern Data Warehousing*, 3 pages.

Kimball, R. e Ross, M. (2002): *The Data Warehouse Toolkit: the complete guide to dimensional modelling*, 2nd ed. John Wiley and Sons, Inc.

Kononenko, I. and Matjaz, K. (2007): *Machine Learning and Data Mining. Elsevier.*

Laney, D. (2001): *3-D Data Management: Controlling Data Volume, Velocity and Variety*. META Group Research Note.

Lavalle, S., Hopkins, M. S., Lesser, E., Shockley, R., & Kruschwitz, N. (2010): *Analytics: The New Path to Value.* MIT Sloan Management Review, 1–24. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Analytics+:+The+New+Path +t o+Value#0

Lawrence, R.D. (2003): *Passenger-based predictive modelling of airline no-show rates. KDD '03 Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and Data Mining*. Pages 397-406.

Leme Filho, T. (2006): *O Business Intelligence como apoio á formulação estratégica*. Centro Universitário Nove de Julho – UNINOVE.

Luhn, Hans. (1958): *A Business Intelligence System. IBM Journal of Research and Development.* Volume 2, Issue 4, pp 314-319.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011): Big Data: The next frontier for innovation, competition, and productivity. Report McKinsey Global Institute.

Marr.B (2013): *Big Data, what is it?* Retirado de: http://pt.slideshare.net/BernardMarr/140228-big-data-slide-share

Mazhelis,Laney, D. (2001): 3D Data Management: Controlling Data Volume, Velocity, and Variety. Meta Group Inc.

McKinsey. (2011): *Big Data: the next frontier for innovation, competition, and productivity.*

McKinsey. (2012): *Delivering large-scale IT projects on time, on budget, and on value.*

Miller, H. J. and J. Han (2001): *Geographic Data Mining and Knowledge Discovery*, CRC Press.

Moss, L.T. and Adelman, S. (2000): *Data Warehouse Project Management. Addison-Wesley Information Technology Series.*

Moss, L.T. and S. Atre (2003): *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications, Boston, MA: Addison-Wesley.*

Nisbet, R., Elder, J., & Miner, G. (2009): *Handbook of Statistical Analysis and Data Mining Applications.* California: Elsevier Inc.

Rockart, John F. (1979): *Chief Executives Define Their Own Data Needs*. Harvard Business Review. Volume 57, issue 2, pp. 81-93.

Takeuchi, H.; Nonaka, E. (1998): *Criação de conhecimento na empresa. Como as empresas japonesas geram a dinâmica da inovação*. Rio de Janeiro: Campus.

TDWI. (2013): *Business Intelligence Journal*. Vol. 18, No. 4.

Pereira, J.C.R. (2001): *Análise de dados qualitativos – estratégias metodológicas para as ciências da saúde, humanas e sociais*. 3ª ed. São Paulo: EDUSP.

Ponchirolli, O., Fialho, Francisco António P. (2005): *Gestão estratégica do conhecimento como parte da estratégia empresarial*. Revista FAE, Curitiba, v. 8, n. 1, pp. 127-138.

Ponjuán Dante, G. (2004): Gestión de información: dimensionaes e implementación para el éxito organizacional. Rosario: Nuevo Paradigma, 218p.

Potts, W.J.E. (1998): *Data Mining Primer: Overview of Applications and Methods*. SAS Institute Inc.

Prieto, I., Revilla, E. (2006): *Assessing the Impact of Learning Capability on Business Performance: Empirical Evidence From Spain*. *Management Learning, 37*(4): 499-522.

Pujari, A. K. (2001): *Data Mining Techniques. Hyderabad, India: Universities Press (India) Private Limited.*

Qi, F. and A. X. Zhu (2003): *Knowledge discovery from soil maps using inductive learning.* International Journal of Geographical Information Science 17(8): 771-795.

Quinlan, J.R. (1986): *Induction of Decision Trees.* Machine Learning, Volume 1, Issue 1.

Riwo-abudho, Marcella (2013): *Strategic Change and Competitiveness: Analysis Of The Airline Industry*. LAP Lambert Academic Publishing.

Robertson, J. (2005): *Dez princípios de Gestão Eficaz da Informação*. (Tradução livre do autor da dissertação, 26 de março de 2012). [. Em linha]. Disponível em http://www.steptwo.com.au/papers/kmc_effectiveim/index.html. (consultado em 26 março de 2012)

Silltow, J. (2006): *Data Mining 101: tools and techniques.* Retirado de: www.theiia.org/intAuditor/itaudit/archives/2006/august/data-mining-101-tools-and-techniques/

Tuomi, I. (1999): *Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory*. Journal of Management Information Systems, Vol. 16, No. 3 (Winter, 1999/2000), pp. 103-117.

Turban, E. Valentim, M. L. P. (2002): *Inteligência competitiva em organizações: dado, informação e conhecimento*. *DataGramaZero*, Rio de Janeiro, v.3., n.4.

Valentim, M. L. P. et al. (2003): *O processo de inteligência competitiva em organizações*. *DataGramaZero*, Rio de Janeiro, v. 4, n. 3, p. 1-23.

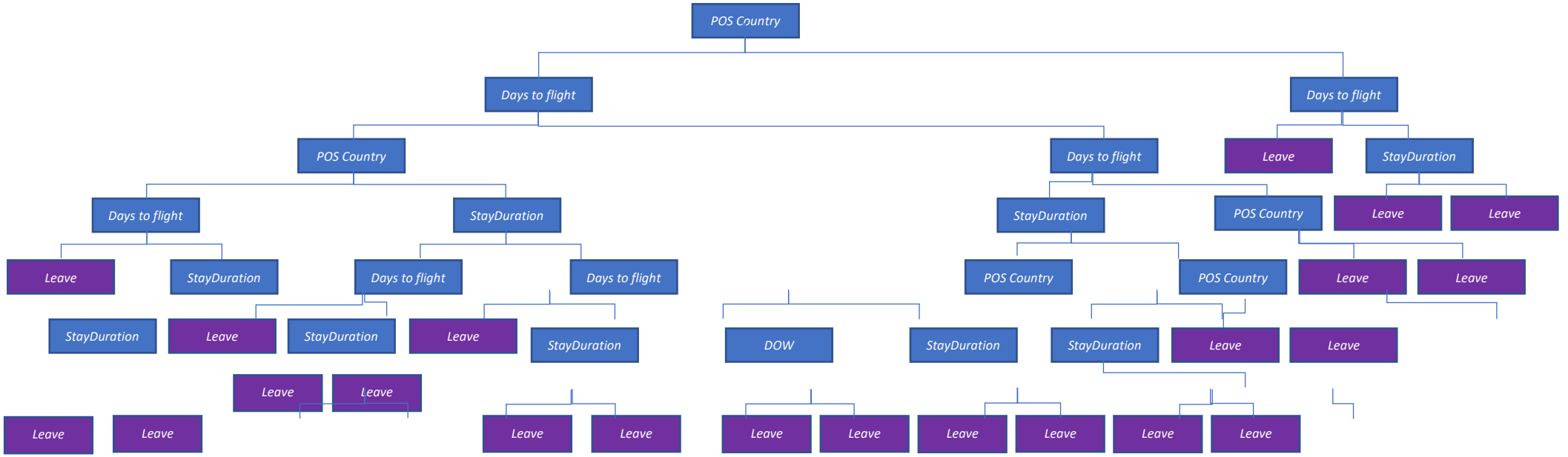
Wang, Q.R. and Suen, C.Y. (1984): *Analysis and Design of a Decision Tree Based on Entropy Reduction and Its Application to Large Character Set Recognition*. Volume PAMI-6 Issue: 4.

Witten, I. H., Frank, E., & Hall, M. a. (2011): *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. doi:10.1002/1521-3773(20010316)40:63.3.CO;2-C

Zhang et al (2011): *Applications of Business Intelligence Technology in the Airports and Airlines Companies.* International Journal of Applied Science and Technology. 5 (1), 74-78.

8. ANEXOS

ANEXO A. Variáveis de decisão na árvore 3



Diagrama| *Variáveis preditivas nos nós de decisão e nós terminais (leaves)*

ANEXO B. Detalhe da árvore 3

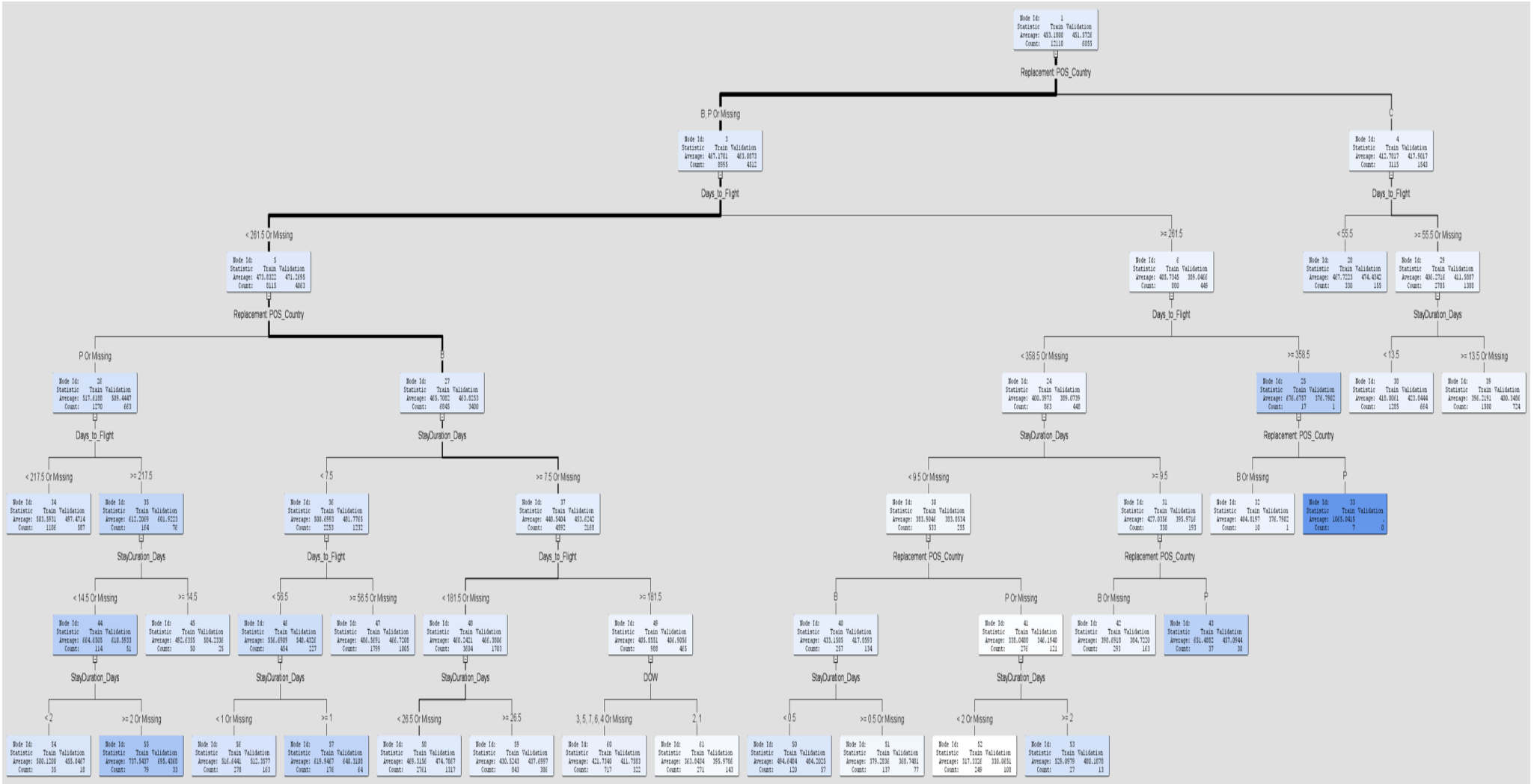


Figura 34 **-** Variáveis preditivas nos nós de decisão e nós terminais (*leaves*), extraídas diretamente do SA

ANEXO C. Aplicações do Software SAS

O *software SAS* é uma das ferramentas de *Business Intelligence,* atualmente disponíveis no mercado, e integra várias componentes num único produto de *software*. Destas componentes destacam-se o ***Enterprise Guide*, o *Enterprise BI Server* e o *Enterprise Miner***.

Figura 35 - Variantes do *software* SAS

Com as suas poderosas capacidades de *Data Mining*, o SAS assume uma posição de liderança na área de *software* de negócios disponível. Agora, habilitado para a Web com novas soluções de "*e-intelligence*", o SAS continua a permanecer na vanguarda da indústria de *software* de negócios.

O reconhecimento da qualidade dos seus produtos continuou a vir de várias fontes em todo o mundo, incluindo *Datamation*, *Data Warehousing World*, *Software Magazine*, *ComputerWorld Brasil* e *PC Week*, juntamente com a prestigiada associação de analistas franceses *Yphise* e a *Australian Corporate Research Foundation*. Além disso, a *Food and Drug Administration* dos EUA reconheceu a integridade do *software* SAS, selecionando a sua tecnologia como padrão para novas aplicações de drogas. Para além do reconhecimento tecnológico, o SAS continua a ser reconhecido como um ótimo lugar para trabalhar, recebendo prémios das revistas *Fortune*, *Working Mother*, *BusinessWeek* e

Mother Jones, juntamente com uma importante cobertura da imprensa nos Estados Unidos, Europa e Austrália.
As componentes do SAS são aplicações gráficas interativas, que funcionam sob uma mesma filosofia. Esta consiste essencialmente na definição de *processos* que são compostos por uma sequência de tarefas a executar sobre os dados. Estas tarefas correspondem a tipos específicos de análises ou relatórios que podem ser aplicados aos dados. Associado a cada tarefa existe um *bloco de código* SAS, que é executado sobre os dados analisados pela tarefa, na sequência definida pelo fluxo do processo.

De uma forma simplista, a interação com as aplicações SAS pode ser vista como uma sequência de quatro etapas:

1. Criação de um projeto;

2. Adição dos dados a analisar;

3. Execução das tarefas de análise;

4. Visualização dos resultados / relatórios criados.

De modo a facilitar a definição e execução dos processos, as aplicações têm ambientes de trabalho semelhantes, compostos por várias janelas, cada uma das quais desempenhando um objetivo específico. Em particular, todas apresentam:

- Uma *Explorer Area* em que se listam as fontes de dados disponíveis, organizadas em bibliotecas, que por sua vez se localizam em servidores virtuais;
- Uma *Task Area* em que se listam as tarefas disponíveis para utilização;
- E uma *Process Area* em que se define e visualiza o processo a aplicar. Esta área tipicamente dá acesso ao fluxo de processos, aos *logs* originados durante a execução do processo e ao código SAS gerado.

As janelas são acompanhadas por um conjunto de menus e *toolbars*, dependentes do contexto, pelo que o seu uso é por vezes difícil, uma vez que dependendo da janela ativa as opções disponíveis são significativamente diferentes.

2.4.2. SAS ENTERPRISE GUIDE

O *SAS®* Enterprise Guide é uma ferramenta *point-and-click*, que possibilita aos utilizadores aceder, transformar, analisar e exportar dados.
O SAS® Enterprise Guide dispõe de um enorme catálogo de funcionalidades que dão aos seus utilizadores a capacidade de realizar quase todas as tarefas de um processo *end-to-end* de preparação de dados, assim como de uma interface simples que permite a qualquer utilizador começar a criar os seus processos de dados.
Apesar da sua interface completamente visual, por detrás de cada tarefa que é arrastada para o processo, é gerado todo o *script* que a suporta e que, posteriormente, permite suportar quer a sua reutilização, quer a automatização da sua execução.

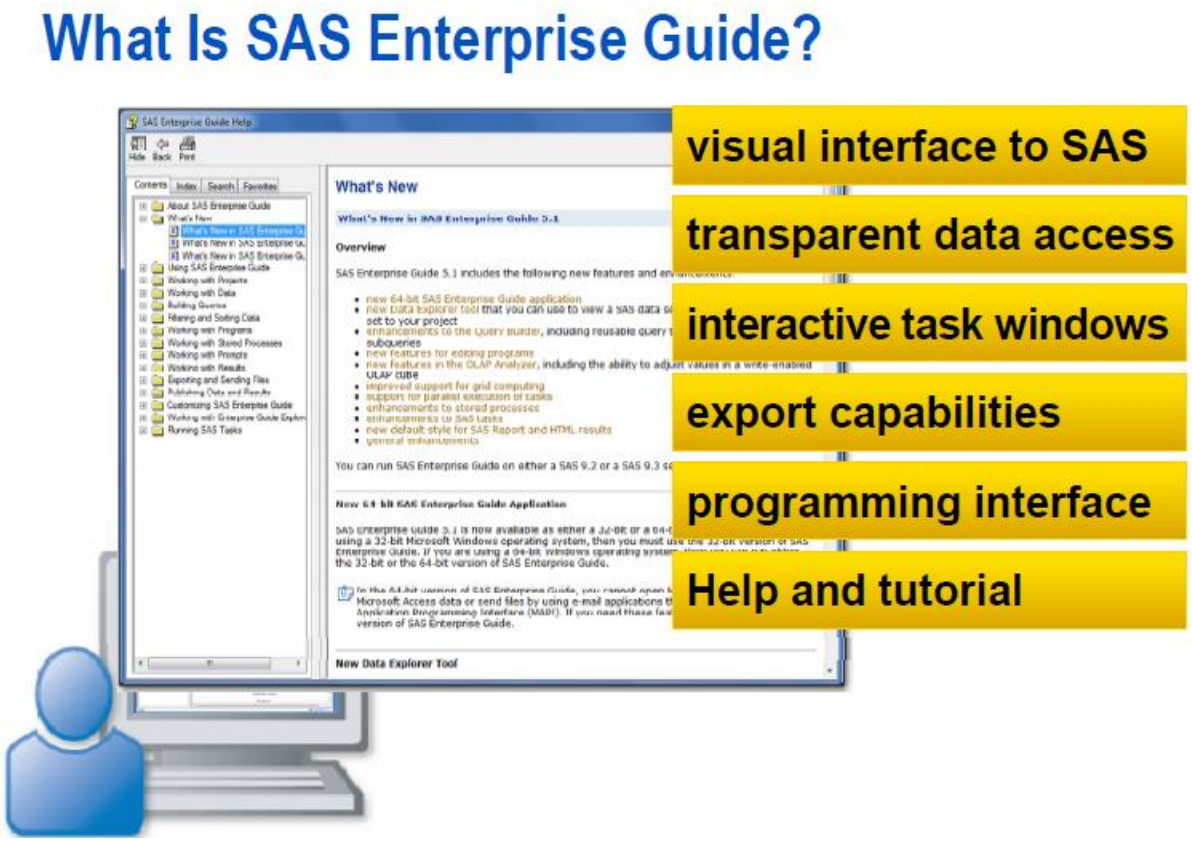


Figura 36 - SAS® Enterprise Guide *Layout*

O SAS® Enterprise Guide:

- Proporciona um ambiente de análise self-service: integra uma ampla gama de análises, numa interface eficiente e *user-friendly*. Os analistas podem produzir análises e distribuir relatórios libertando assim as TI para outros projetos estratégicos;
- Fornece segurança centralizada baseada em funções para gerir o acesso aos dados da organização, garantindo os privilégios adequados a cada utilizador;
- Facilita o acesso às fontes de dados corporativas pelos diferentes utilizadores da organização.

ENTERPRISE BI SERVER

O *SAS Enterprise BI Server* é um pacote de ferramentas que integra a construção, exploração e visualização de dados, permitindo a sua organização segundo modelos mais adequados ao apoio à decisão.

O SAS OLAP Cube Studio permite a definição e criação de cubos, funcionando em ligação com o servidor de metadados e o servidor de OLAP (que devem estar a correr em background).

O SAS Management Console é apenas uma ferramenta que permite a gestão dos vários utilizadores e serviços, nomeadamente a gestão do funcionamento dos servidores referidos.

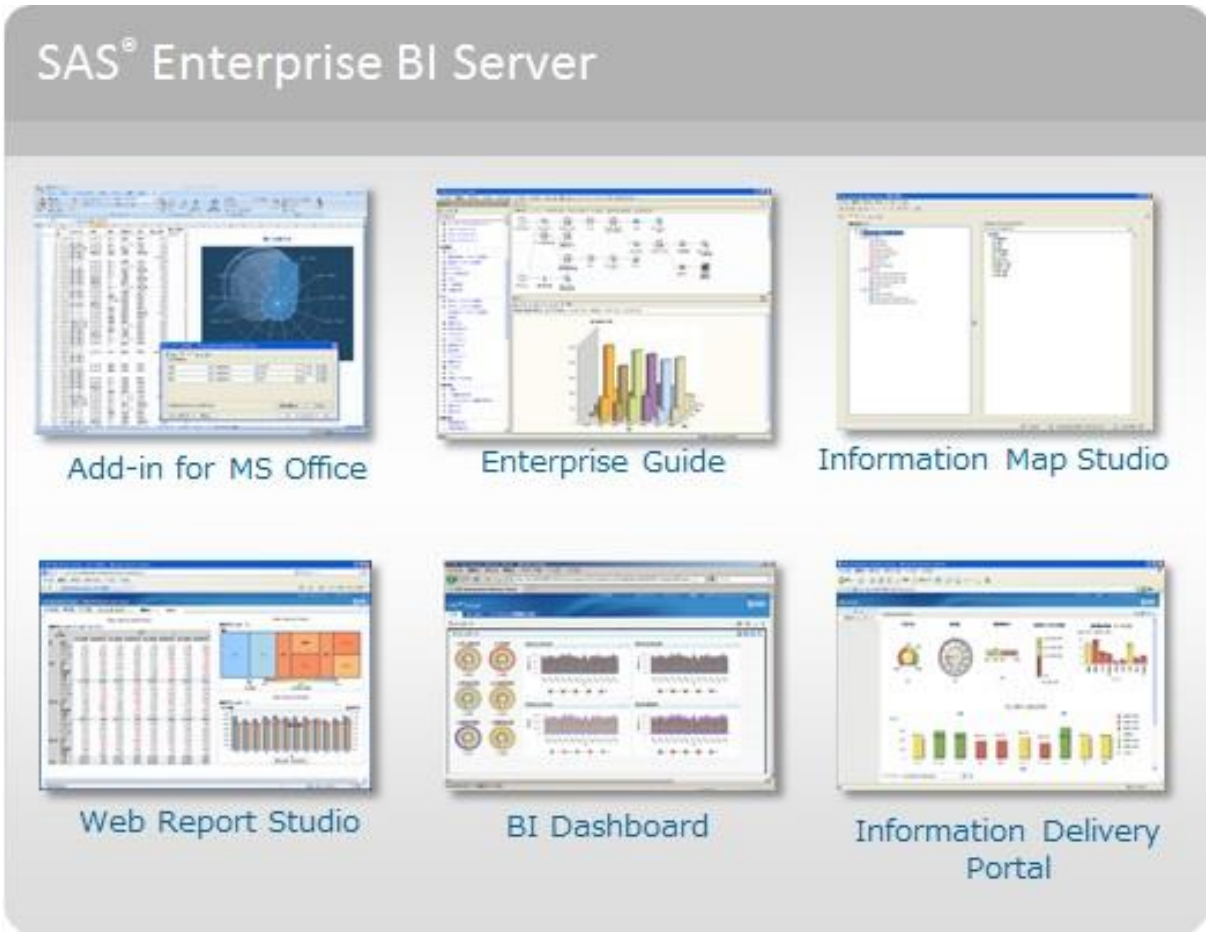


Figura 37 **-** Enterprise BI Server

ENTERPRISE MINER

O *Enterprise Miner* é o pacote do SAS para *Data Mining*, ou seja, que executa processos de extração de informação, desde o acesso aos dados até à visualização da informação descoberta. Sendo um pacote, funciona dentro do *SAS Base*.

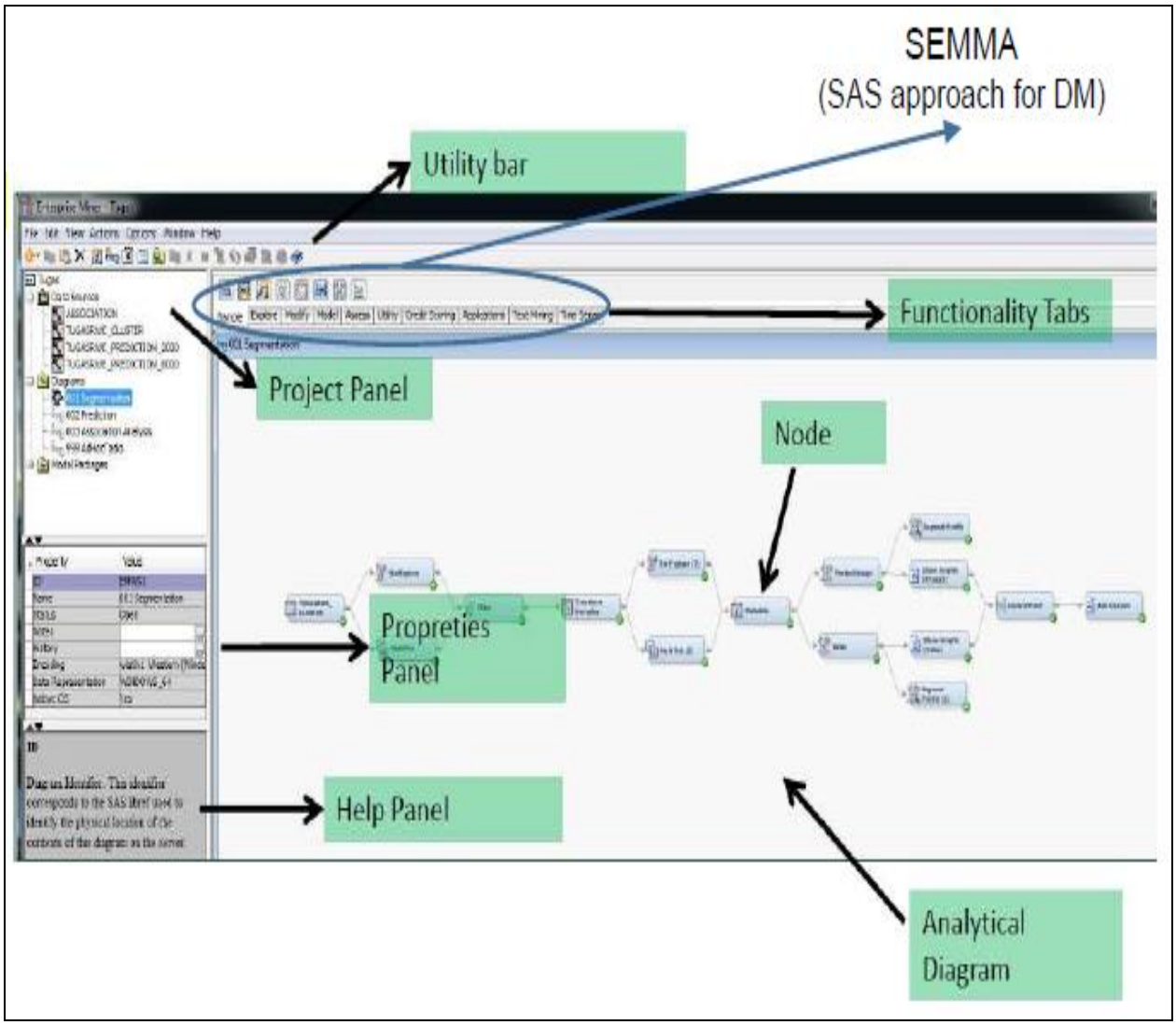


Figura 38 - SEMMA

O Instituto SAS define *Data Mining* como "o processo de Selecionar (*Sample*), Explorar (*Explore*), Modificar (*Modify*), Modelar (*Model*) e Avaliar (*Assess*) – **SEMMA** grandes quantidades de dados, para descobrir padrões previamente desconhecidos":

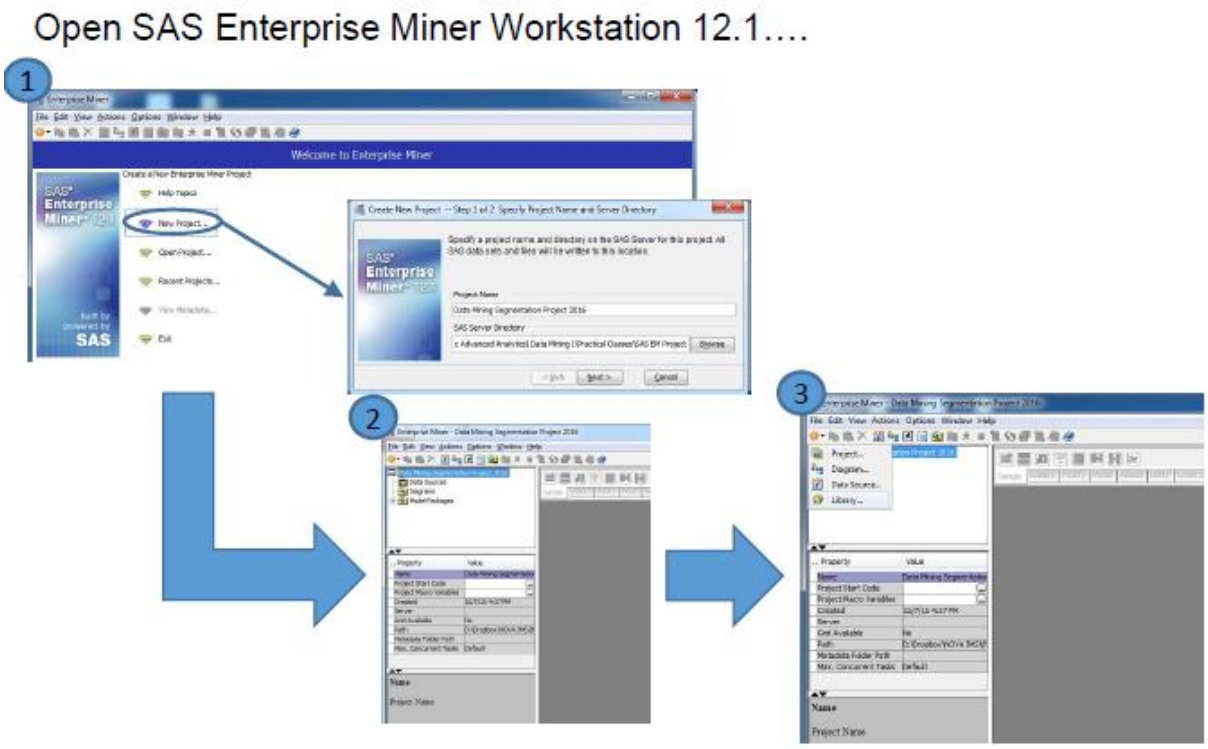
· A **Seleção** dos dados é efetuada com a criação de tabelas. Estas devem conter dados suficientes e significativos, mas ter um tamanho que não inviabilize o processo de descoberta (por o tornar demasiado lento ou inoperante).

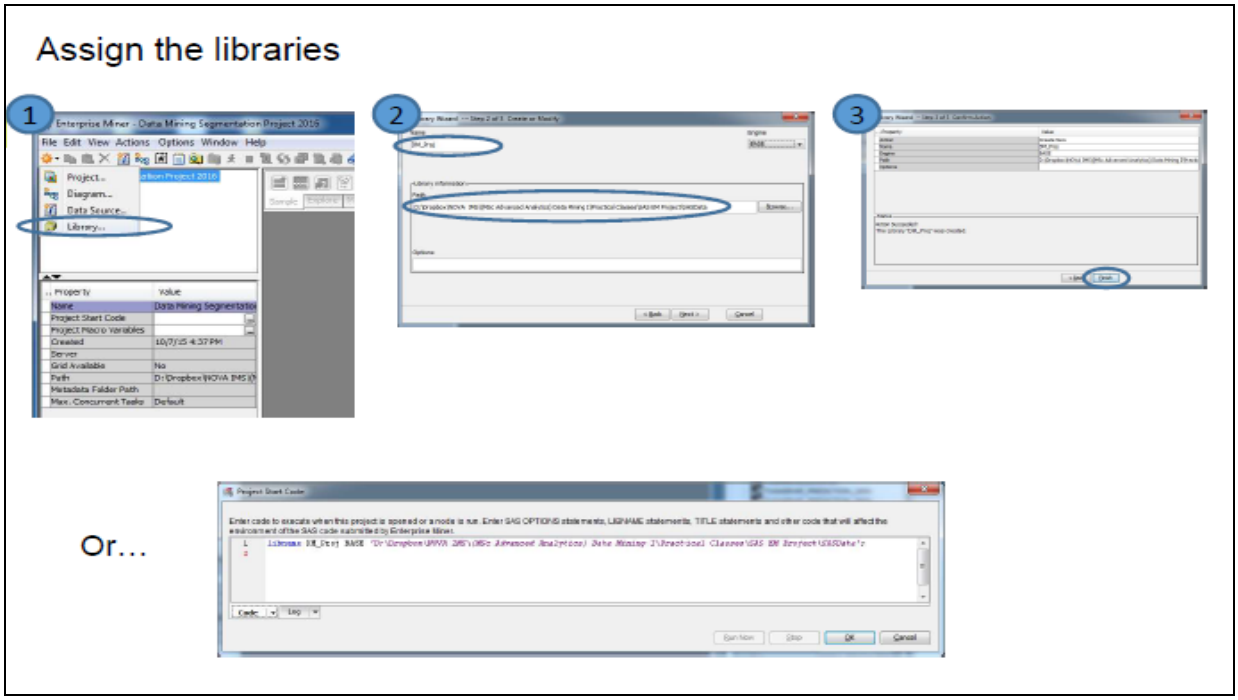
· A **Exploração** dos dados consiste na análise manual dos dados, de forma a adquirir algum conhecimento prévio que ajude na definição dos objetivos do processo.

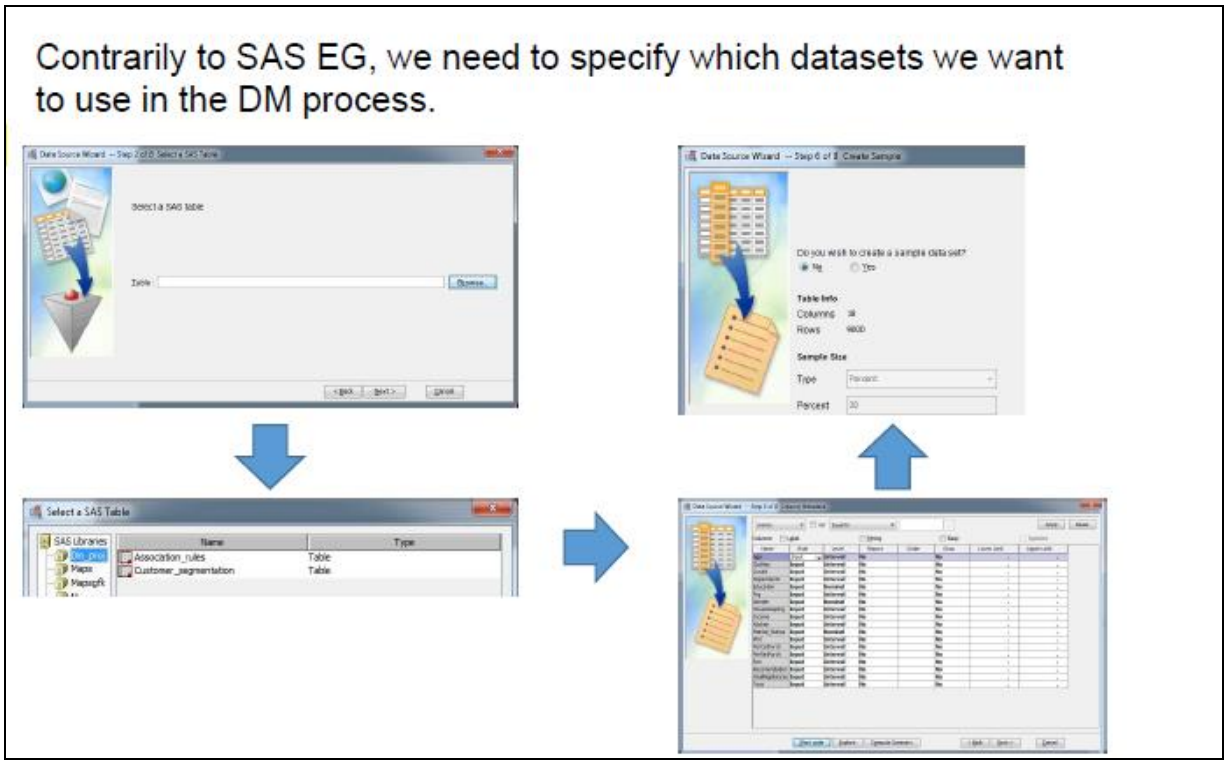
· A **Modificação** dos dados é realizada pela criação, seleção e transformação das variáveis envolvidas no problema, de modo a ajudar a escolher o melhor modelo a usar no processo.

· A **Modelação** dos dados é concretizada pela aplicação das ferramentas de análise disponíveis no pacote, nomeadamente árvores de decisão, redes neuronais, entre outros. É esta a etapa responsável pela descoberta de informação, propriamente dita.

· A **Avaliação** consiste em analisar os resultados obtidos no passo anterior, de modo a determinar a sua utilidade e fiabilidade.







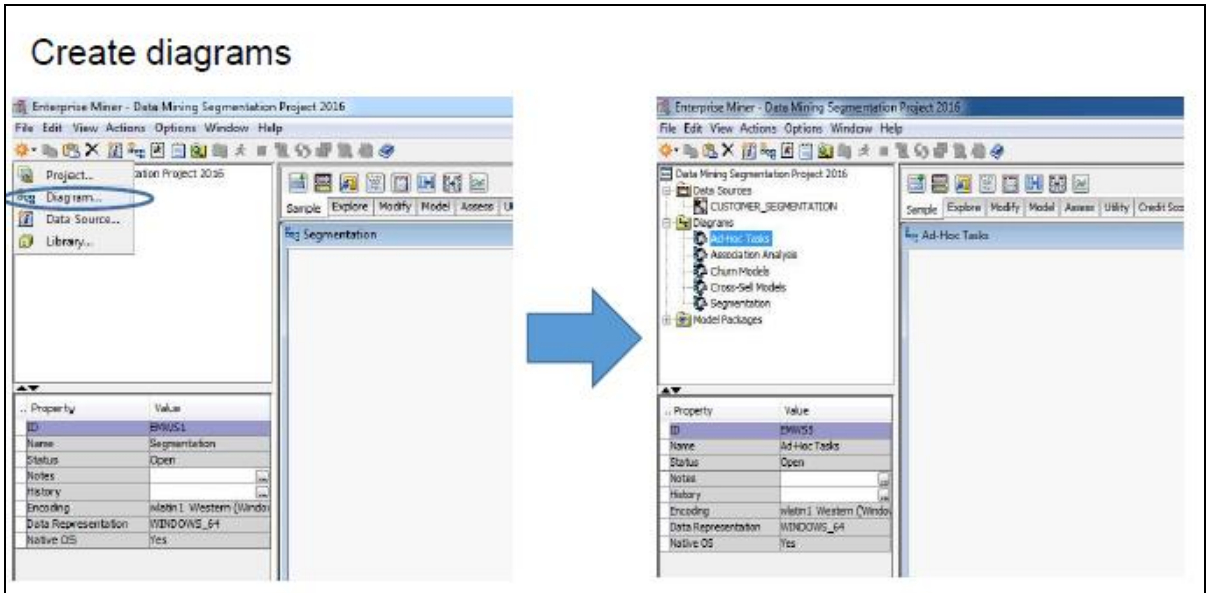


Figura 39 **-** Sequência de procedimentos do Projeto.