NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Ageing Workforce

Determinar e caracterizar os diferentes clusters de colaboradores para uma melhor compreensão da sua diversidade

Ana Margarida Sabino Guerreiro

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

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

2018

Ageing Workforce - Determinar e caracterizar os diferentes clusters de colaboradores para uma melhor compreensão da sua diversidade

Ana Margarida Sabino Guerreiro

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

AGEING WORKFORCE

DETERMINAR E CARACTERIZAR OS DIFERENTES CLUSTERS DE COLABORADORES PARA UMA MELHOR COMPREENSÃO DA SUA DIVERSIDADE

por

Ana Margarida Sabino Guerreiro

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: Professor Roberto Henriques

Coorientador*:* Professor Guilherme Martins Victorino

Novembro 2018

AGRADECIMENTOS

Aos colegas e novos amigos que tive a oportunidade de conhecer e partilhar experiências de trabalho em grupo durante este ciclo, Paula Torrão, Pedro Nina, Patrícia Brito entre outros.

À tutora da empresa, agradeço a orientação e disponibilidade total que teve para me apoiar no desenvolvimento do trabalho. Agradeço ainda todas as aprendizagens que tive a oportunidade de ter com ela dada a sua experiência na área de Recursos Humanos.

Ao professor Roberto Henriques e Guilherme Vitorino por terem aceite orientar este trabalho e pelo acompanhamento no desenvolvimento do mesmo.

Aos meus colegas de trabalho e à minha chefia atual por me terem permitido ter os tempos que considerei necessários para trabalhar na tese durante a fase final sem nunca levantarem obstáculos quanto aos dias que escolhi para me ausentar.

À minha família, sem dúvida o pilar mais importante na minha formação pessoal e a qual considero o pilar de maior importância em todos os aspetos da minha vida. Obrigada a todos os que de uma forma ou de outra contribuíram para a realização deste caminho académico.

RESUMO

O presente trabalho consiste numa análise de clusters com foco na importância que a temática do Ageing Workforce assume atualmente na organização em estudo. Para explorar a relevância do tema são abordados conceitos inerentes ao People Analytics e também algumas técnicas de Data Mining utilizadas em contexto organizacional.

A bibliografia disponível aponta para uma mudança na função de Recursos Humanos que tem vindo a ser registada ao longo dos últimos anos. Esta mudança resulta da necessidade de tornar os dados de Recursos Humanos úteis para uma gestão estratégica da organização. Estes deixam de ser úteis apenas para a função de Recursos Humanos e passam a ser utilizados também em outras áreas de negócio de forma estratégica e assegurando a evolução de desempenho de uma organização.

Comprova-se a importância crescente de ferramentas que tenham por base o conceito de People Analytics na função de Recursos Humanos. É importante que as organizações tenham conhecimento sobre os seus colaboradores, desta forma é crucial garantir o armazenamento dos dados de forma a assegurar a sua qualidade e disponibilidade.

Os objetivos da análise de clusters passam por: analisar a estrutura dos dados, verificar e relacionar os aspetos dos dados entre si e ajudar a caracterizar os colaboradores. O trabalho desenvolvido permite à organização aumentar o conhecimento sobre o perfil dos colaboradores, com vista a uma análise das práticas vigentes de Recursos Humanos e uma eventual adequação das mesmas.

PALAVRAS-CHAVE

People Analytics; Ageing Workforce; Data Mining; Análise de Clusters; k-means.

ABSTRACT

This study shows the results of a cluster analysis focused on the importance that the Ageing Workforce has nowadays, namely in the context of the organization considered for this study. To study the relevancy of the thematic this study addresses concepts related to People Analytics and some Data Mining techniques used in an organizational context.

The worldwide bibliography available points to a change in the Human Resources function which has been registered in the last few years. This change comes from a need of finding Human Resources Data useful to support the strategic management of the organization. Data becomes useful in other business areas, not only in the Human Resources context, which ensures the performance improvement of an organization.

It's possible to prove the growing importance of tools based on the People Analytics concept in a Human Resources function. It's always considered important that organizations have an awareness of their employees, so it is critical to guarantee an efficient data storage in order to provide quality and availability of data.

The objective of the cluster analysis presented in this study is to analyze the data structure verified and relate data with each other in order to improve the employee's characterization. It allows an increase in the knowledge about the employees in the organizational context with the goal to analyze the current Human Resources policies and an eventual adequation of these policies.

KEYWORDS

People Analytics; Ageing Workforce; Data Mining; Cluster analysis; k-means.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ACRÓNIMOS

CCC	Cubic Clustering Criterion
CRISP-DM	Cross-Industry Standard Process for Data Mining
SEMMA	Sample, Explore, Modify, Model, Assess
DCBD	Descoberta de Conhecimento em Base de Dados
RH	Recursos Humanos
RMSD	Root Mean Square Deviation

1. INTRODUÇÃO

1.1. ENQUADRAMENTO E RELEVÂNCIA

O envelhecimento da população mundial tem vindo a revelar-se um desafio para as próximas décadas. Segundo dados do Eurostat (Eurostat, 2017), as baixas taxas de natalidade bem como o aumento da esperança média de vida registados na união europeia estão na base do fenómeno de envelhecimento. Num futuro próximo, a sociedade será fortemente afetada pelas consequências do aumento de longevidade, quer no que respeita ao estado de saúde quer na participação da população na sociedade (Cabral & Ferreira, 2014).

Desta forma, o prolongamento da vida ativa representa um verdadeiro desafio durante as próximas décadas na medida em que é necessário manter as pessoas integradas e sem quebras na produtividade laboral até ao fim das suas carreiras.

A distribuição de idades da população Europeia é um tema de estudo que tem ganho importância na literatura ao longo dos últimos anos. Este facto deve-se sobretudo às alterações demográficas que têm vindo a ser registadas. No gráfico abaixo (Figura 1.1), é possível verificar as alterações registadas na população residente em Portugal no período 2012-2016.

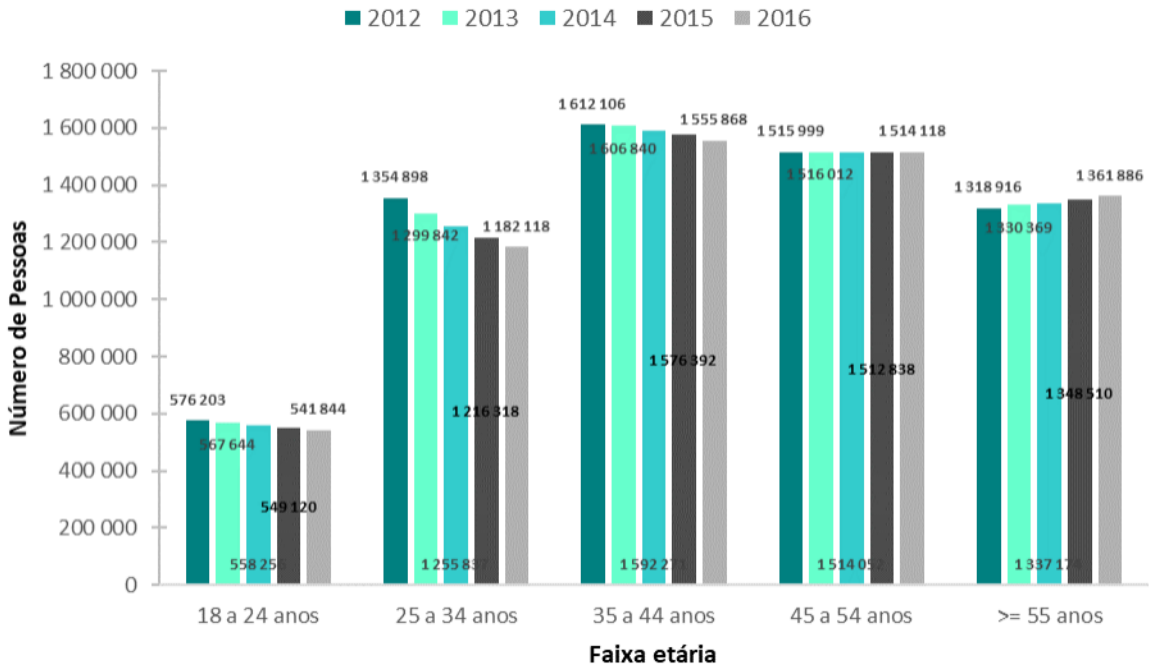


Figura 1.1 – Distribuição da população residente em Portugal por grupo etário (INE, 2017).

Devido às alterações demográficas registadas, é expectável que durante as próximas décadas existam mudanças ao nível da gestão de pessoas, com a tendência para manter os colaboradores no ativo até mais tarde.

Paralelamente, existe também a evolução das Tecnologias de Informação (TI) (e o consequente crescimento das indústrias de informação e conhecimento), as quais têm representado as mais recentes inovações do século, pois permitem aumentar o conhecimento das organizações em

diferentes contextos. Todo esse conhecimento é crucial como forma de apoiar o processo de tomada de decisão, o que resulta em novas indústrias e modelos de negócio que têm vindo a sofrer alterações. De acordo com (Santos & Ramos, 2009) uma gestão eficaz do conhecimento permite aumentar o valor construído a partir do histórico da organização, o que resulta numa aprendizagem organizacional responsável pela mudança organizacional.

A informação que uma organização tem representa uma vantagem competitiva. É um ativo e é, ao mesmo tempo um risco, mas permite às organizações tomar decisões de forma mais eficiente, adequada e relevante. Aliado ao conceito de Data Mining existe a ideia de que os dados do passado contêm informação útil para a tomada de decisão futura. O objetivo do Data Mining relaciona-se sobretudo com a descoberta de padrões no histórico de dados que permitam explicar as necessidades, preferências e propensões (Berry & Linoff, 2004). Existem diversas técnicas de Data Mining que permitem fazer a análise exploratória dos dados, os quais podem surgir de diferentes contextos de informação.

A necessidade de desenvolver esta análise surge devido ao facto de a organização atribuir uma elevada importância ao estudo sobre seus colaboradores. A análise do comportamento adotado pelas pessoas poderá permitir a identificação de oportunidades de melhoria para o negócio da organização no futuro.

1.2. MOTIVAÇÃO

Com recurso a técnicas de análise as organizações podem trabalhar para aprender e inovar criando impacto nas condições de negócio, o que por outro lado anteriormente era feito apenas com base na eficiência e eficácia alcançada pela componente operacional de uma organização. Desta forma, as organizações podem utilizar dados, os quais depois de analisados podem ser importantes e/ou ter impacto em diferentes contextos organizacionais.

O objetivo deste trabalho consiste em desenvolver uma análise descritiva, recorrendo a técnicas de Data Mining, com vista a uma melhor compreensão dos diferentes perfis de colaboradores, tendo em conta a visão do envelhecimento e desgaste profissional dos mesmos.

Este tipo de análise permite sobretudo obter um conhecimento orientado aos dados acrescentando valor ao processo de gestão de pessoas de forma objetiva (Sothmann & Mehta, 2017).

Desta forma, será desenvolvida uma análise para determinar e caracterizar os diferentes clusters de colaboradores, por forma a identificar e caracterizar os diferentes perfis existentes, tendo em conta as variáveis consideradas. Para a realização desta análise serão utilizadas variáveis que caracterizam as diferentes dimensões do colaborador, tais como informação demográfica (como a faixa etária, estado civil, formação académica), informação de contexto profissional na companhia (como o número de anos na companhia, número de horas de formação), informação relativa à loja em que o colaborador se encontra no período estudado e informação relativa à sua chefia (igualmente dados demográficos e de contexto profissional na Companhia).

A análise de clusters é um método exploratório de dados que permite um agrupamento dos indivíduos em grupos/clusters homogéneos. Desta forma, os indivíduos com determinadas características pertencem a um grupo assegurando que clusters diferentes incluem pessoas com perfis diferentes. Será também possível relacionar fatores pessoais e fatores que relacionam a pessoa com a sua função

e identificar padrões de absentismo, acidentes de trabalho, e/ou envolvimento. Associados à análise de clusters existem fatores importantes a considerar como a seleção das variáveis, a definição do critério de semelhança e o contexto dos dados.

Numa primeira vertente, este trabalho pretende contribuir para reforçar a importância que o fenómeno de envelhecimento da população e o consequente aumento do tempo de longevidade têm atualmente no ativo laboral.

Numa segunda vertente o trabalho desenvolvido pretende dar enfâse à importância de uma gestão de conhecimento eficaz, na medida em que os dados do passado possuem informação útil para a tomada de decisão futura. Assim sendo, e aplicado ao contexto organizacional, mais concretamente ao nível da gestão de pessoas do caso de estudo, este trabalho consiste no desenvolvimento de uma análise segmentada da base de dados dos colaboradores.

No capítulo de revisão de literatura - capítulo 2 - são abordados os conceitos utilizados no desenvolvimento do trabalho, tais como Analytics, People Analytics e Ageing Workforce, Informação e Conhecimento, Data Science, Data Mining e análise de clusters, de forma a referir todos os conceitos teóricos importantes para a realização do trabalho.

No capítulo seguinte - capítulo 3 - é abordada a metodologia utilizada no desenvolvimento da análise de clusters. Depois de uma primeira análise descritiva aos dados, é apresentada a análise de clusters seguida da interpretação dos resultados obtidos. Os resultados obtidos são analisados de forma critica e identificam os diferentes perfis de colaboradores existentes na organização, tendo em conta as variáveis consideradas.

No capítulo 4, é feita uma análise das características dos clusters e é abordada a temática da Gestão na Liderança numa organização, para contextualização dos resultados obtidos. Seguidamente são apresentadas as conclusões – capítulo 5 – e identificadas as limitações e possíveis trabalhos futuros – capítulo 6.

2. REVISÃO DA LITERATURA

2.1. ANALYTICS

Analytics é definida como a área de estudo que permite a interseção entre a engenharia, ciências informáticas, tomada de decisão e métodos quantitativos que contribuem para organizar, analisar e dar significado a grandes volumes de dados provenientes de diversos sectores (Mortenson, Doherty, & Robinson, 2015).

O termo Analytics pode ser utilizado pelo menos de três formas relacionadas (Watson, 2013). Primeiramente, por volta de 1970, surgiu através de sistemas de apoio à tomada de decisão. Mais tarde foi popularizado através do termo Business Intelligence (BI) - por volta de 1990 - e mais recentemente surgiu o termo Analytics.

Segundo (Watson, 2014) hoje em dia as organizações recolhem, armazenam e analisam grandes volumes de dados, referidos como Big Data devido ao seu volume, velocidade e variedade. A chave para a criação de valor a partir de Big Data passa pela utilização de técnicas analíticas (Analytics). De acordo com o autor existem diferentes tipos de técnicas utilizadas em Analytics, tais como:

1) Análise Descritiva, que recorre a algoritmos de Data Mining para obter conhecimento sobre os dados.
2) Análise Preditiva, a qual prevê o que ocorre no futuro através de técnicas de regressão, machine learning ou redes neuronais, e que pode ser também realizada com recurso ao software SAS Enterprise Miner.
3) Análise Prescritiva, permite identificar a melhor solução e é muitas vezes utilizada como complemento à análise descritiva e preditiva.

Em suma a análise descritiva é a forma primária de obter uma retrospetiva sobre o que aconteceu criando uma base de fundamento para transformar dados em informação. A previsão é uma técnica de análise mais avançada para obtenção de resultados futuros e a prescrição interpreta os resultados e recomenda uma ação.

Hoje em dia, áreas como a banca, o sector do retalho ou dos seguros recorrem a técnicas de analytics para a gestão de pessoas.

Exemplos da utilização de Analytics por uma organização

Ao longo do seu trabalho (Watson, 2014) apresenta exemplos práticos resultantes da utilização de analytics no dia a dia de uma organização. Dos exemplos apresentados, destacam-se a introdução de um novo produto pela organização Starbucks, um exemplo aplicado ao sector dos seguros e outro relativamente ao sector das telecomunicações. O sector dos seguros automóveis utiliza técnicas de analytics de forma a definir os preços, identificar o risco associado a cada cliente, deteção de fraudes e também na resposta rápida a reclamações dos clientes. O sector das telecomunicações analisa os padrões de serviços e a rentabilidade dos clientes através das redes sociais de forma a minimizar a rotatividade dos seus clientes mantendo-os satisfeitos.

O maior valor da análise preditiva está no facto de permitir a análise de grandes quantidades de dados para identificar padrões de eventos, ou atividades, que preveem as ações das pessoas. Outras

aplicações deste tipo de análise podem ser a renovação de um contrato de telefone, no sector das telecomunicações, ou o encerramento de uma conta corrente, para o sector bancário. Quando uma organização tem a oportunidade de prever este tipo de ação, poderá ter também a oportunidade de interceder com uma oferta ou uma possibilidade de mudança de forma a manter o cliente fidelizado.

Exemplos da utilização de Analytics em Recursos Humanos

Na área de Recursos Humanos (RH) a utilização de dados pode ser útil em diferentes contextos como para a análise de processos de recrutamento, desempenho ou mobilidade de colaboradores. O valor criado a partir deste tipo de análise poderá permitir:

- O desenvolvimento de programas de gestão de talentos para manter equipas bem preparadas, motivadas e dinâmicas;
- O investimento em formação e desenvolvimento dos colaboradores, com ênfase nas competências e segmentos chave da organização, apostando na certificação dos conhecimentos;
- A adoção de políticas de remuneração e desenvolvimento de carreira (compensação e benefícios) e fortalecimento da liderança por parte dos mais dotados;
- O acesso a programas de reforma antecipada.

2.1.1. People Analytics

Segundo (Laurence Collins, David R. Fineman, 2017), o conceito de People Analytics é definido como uma disciplina que permite a análise do envolvimento e retenção aplicado a grupos de pessoas. As organizações utilizam esta disciplina para planeamento organizacional e construção de novas soluções empresariais digitais, as quais permitem conduzir uma análise aprofundada e em tempo real de acordo com as necessidades identificadas pela organização. Este conceito envolve a utilização de ferramentas digitais e dados para medir, estudar/reportar e ganhar conhecimento sobre o comportamento das pessoas. No entanto, People Analytics, tem vindo a sofrer alterações na medida em que deixou de se tratar da identificação de informação relevante para conhecimento dos gestores, e passou a ter uma função orientada ao negócio que se foca na utilização de dados para apoiar o contexto operacional de uma organização.

O conjunto de dados sobre as pessoas que uma organização possui permite atribuir significado ao conceito de People analytics. Segundo (Bersin, 2016) os dados de pessoas são muitas vezes inconsistentes, incorretos, desatualizados e geralmente encontram-se armazenados em bases de dados diferentes, resultando num problema de integridade de dados. Para comprovar o recente elevado impacto que o People analytics tem, (Bersin, 2017) afirma que no ano de 2017 cerca de 70% das organizações focaram-se em consolidar dados de RH provenientes de diversas fontes para a construção de repositórios integrados enquanto em anos anteriores apenas cerca de 10-15% das organizações adotaram essa estratégia.

Estas alterações devem-se sobretudo ao crescimento exponencial da era digital. No entanto, a gestão de dados continua a ser um desafio para as organizações. É imperativo que as organizações se adaptem às necessidades atuais para acelerar, ajustar, facilitar a aprendizagem e obter um percurso empreendedor.

O autor propõe um modelo analítico que explora os pressupostos essenciais a considerar para uma análise bem-sucedida. Este modelo é composto por 4 níveis aos quais é atribuída uma percentagem

de utilização diferente e pode ser utilizado como uma abordagem analítica para o desenvolvimento de uma organização (Figura 2.1).

Figura 2.1 - Modelo Analítico proposto por (Bersin, 2016).

A maioria das organizações, cerca de 66%, encontra-se nos níveis 1 e 2, os quais correspondem a relatórios operacionais e relatórios avançados, respetivamente. Continua a ser um desafio o facto de as organizações possuírem mais do que um repositório de dados, o que compromete os relatórios obtidos e também a integração dos dados (Bersin, 2016).

O People Analytics alia a utilização de tecnologia para processamento de dados na área do *Business Intelligence* e a existência em abundância de dados de pessoas a ferramentas analíticas para melhorar o desempenho da sua função, ou seja, alavancar o seu papel na organização através do People Analytics. Para tal, é necessário apostar numa abordagem sistematizada e integrada.

No entanto, com o avanço das tecnologias e também com o contributo da era digital é expectável que as organizações melhorem a qualidade dos seus repositórios de dados e desta forma se recorra com maior frequência aos dados para desenvolver análises avançadas e modelações preditivas.

Técnicas de Data Mining aplicadas aos Recursos Humanos

A utilização de técnicas de Data Mining permite caracterizar e prever os diferentes comportamentos adotados num grupo, o que contribui para que a função de Recursos Humanos seja mais eficiente. A seleção da técnica de Data Mining mais adequada pressupõe que sejam conhecidas as características identificadas por (Berry & Linoff, 2004) e (D. J. Hand, 1998).

Técnicas de Data Mining	Características
Clustering	• Permite identificar padrões nos dados, contribuindo para o aumento do conhecimento sobre os dados, com recurso a algoritmos de Data Mining.
Redes Neuronais (Berry & Linoff, 2004)	• Adequado para estudos de clustering, classificação e previsão; • Aprende com padrões de exemplo produzindo uma aproximação.
Árvores de Decisão (D. J. Hand, 1998)	• Adequado para classificação e previsão; • Aplica-se a variáveis contínuas e de classe; • Produz um modelo que representa regras fáceis de interpretar a quem tem a responsabilidade de decidir; • Não exige conhecimentos específicos nem a definição de parâmetros iniciais; • Aplicável a grandes volumes de dados; • Os modelos produzidos têm boa precisão.

Existem também organizações que são casos de sucesso, sendo a Google exemplo disso dado que é uma das grandes organizações que têm sido bem-sucedidas no mercado com as suas estratégias focadas na gestão de pessoas. (Sullivan, 2013) explica como a Google mudou a sua estratégia para que esta se focasse na gestão de pessoas e como essa mudança afetou a produtividade da organização. Segundo o autor, "*All people decisions at Google are based on data and analytics*", sendo este um dos fatores chave que tem contribuído para o sucesso do processo de gestão. People Analytics apoia a tomada de decisão no contexto de gestão de pessoas e desta forma as decisões mais importantes podem ser tomadas com base num conhecimento exato que é obtido sobre os dados. O autor identifica ainda alguns pontos que considera importantes na análise de pessoas, de entre os quais se destacam: características e funções dos gestores, modelação preditiva, melhoria no universo de colaboradores e local de trabalho.

People Analytics é uma área de estudo essencial para assegurar o futuro dos colaboradores de forma estratégica contribuindo para a melhoria do desempenho nas organizações (Angrave, Charlwood, Kirkpatrick, Lawrence, & Stuart, 2016). O autor propõe quatro ideias chave para o conceito de HR Analytics:

(1) os gestores devem ter um conhecimento claro sobre a contribuição das pessoas para o sucesso da organização;
(2) deve existir um conhecimento profundo sobre os dados e contexto dos mesmos;
(3) as métricas e ferramentas utilizadas (na medida em que permitem segmentar os grupos de colaboradores);
(4) tomada de decisão orientada aos dados.

Os dados de colaboradores (HR data) podem ser usados para criar, capturar, potencializar e proteger o valor de uma organização. Além disso, podem depois ser utilizados para responder a questões complexas com recurso a modelação multivariada, a qual permite quantificar métricas e medidas importantes para a organização. As análises avançadas podem ser aplicadas na parte operacional, de

gestão e também no recrutamento e seleção permitindo identificar, atrair, desenvolver e manter o talento nas organizações. People Analytics apoia os gestores na medida em que melhora o poder dos dados, aumentando o rigor e coerência na tomada de decisão e no desempenho. No entanto, o desafio relaciona-se com o benefício alcançado através desses dados.

O relatório anual publicado pela (Boston Consulting Group, 2014) também é favorável quanto à importância da função de Recursos Humanos, defendendo que esta está diretamente correlacionada com o desempenho económico de uma organização. A gestão de talentos, liderança, o envolvimento e comportamento das pessoas e a cultura de gestão são identificados como fatores de sucesso para que uma organização tenha bons resultados. No mesmo estudo, a consultora propõe um ranking de 27 subtópicos chave em Recursos Humanos, de entre os quais se destacam: a liderança, a gestão de talentos, o comportamento e cultura, o RH e a estratégia de pessoas, o envolvimento dos colaboradores, o planeamento estratégico da força de trabalho - Strategic Workforce Planning, os modelos de carreira e competências, HR communication, a gestão de performance, a formação e a aprendizagem. Segundo o relatório a conhecida marca PepsiCo é reconhecida por ser um exemplo de investimento na formação para a liderança. A organização deu a oportunidade de os gestores adquirirem conhecimentos importantes através da formação, os quais podem posteriormente ser aplicados para otimizar o sucesso da organização

Mais recentemente, alguns autores, (Carla Arellano, Alexander DiLeonardo, 2017), afirmam que o People Analytics permite o desenvolvimento de análises avançadas em grandes conjuntos de dados de forma a medir a gestão de talentos. Nos dias de hoje, é frequente as organizações recorrerem a esta disciplina em processos como o recrutamento, retenção, descoberta de talentos e perceções não intuitivas sobre o desempenho dos colaboradores. Além destes processos os autores, (Momin & Mishra, 2015) identificam também o acompanhamento de projetos, do absentismo, a monitorização e gestão das tarefas.

A utilização de software para a gestão em RH, aplicações para telemóvel, vídeo e também o conceito de analytics têm permitido a introdução de grandes alterações no contexto organizacional (Laurence Collins, David R. Fineman, 2017). Desta forma, os sistemas de informação estão cada vez mais inteligentes, o que também altera a forma como as organizações geram, lideram e se organizam. Os autores afirmam também que estas transformações alteram as responsabilidades atribuídas aos departamentos de RH e às tecnologias de informação nas organizações. Por esse motivo, as organizações têm vindo a adotar as suas soluções tecnológicas de forma a assegurar que as mesmas são soluções integradas que asseguram uma gestão eficaz. Assim, os sistemas de informação utilizados têm vindo a promover alterações na gestão de pessoas. O conceito de People Analytics tem vindo também a diminuir o seu foco apenas em RH. A Ford é uma das organizações que tem expandido a disciplina People Analytics por outros segmentos do negócio tais como, o sector financeiro, recursos humanos e operações (Laurence Collins, David R. Fineman, 2017).

De acordo com (Michael J. Kavanagh, Mohan Thite, 2011) p.8, muitas organizações não utilizam a tecnologia apenas como forma de suporte para a tomada de decisão em RH, mas também como uma ferramenta que permite que a tomada de decisão seja feita de forma coerente. Devido ao aumento da disponibilidade dos dados, é possível aplicar métricas de RH para avaliar os objetivos traçados em termos de eficiência e eficácia (Dulebohn & Johnson, 2013). Os autores defendem ainda que os sistemas de apoio à tomada de decisão são cada vez mais utilizados por gestores e colaboradores como

parte integrante dos sistemas de informação, os quais integram métricas e ferramentas de análise que contribuem para a resolução de problemas comuns. Os sistemas de apoio à decisão integram dados e modelos que auxiliam os colaboradores e gestores na tomada de decisão.

Aliado ao contexto dos sistemas de apoio à tomada de decisão para RH (Michael J. Kavanagh, Mohan Thite, 2011) defendem que um dos maiores desafios deste tipo de sistema é a captura dos dados que servem de suporte a auditorias, produção de relatórios de gestão e comunicação da eficácia do processo de gestão de RH.

2.1.2. Conceito de Ageing Workforce

Estima-se que em 2050 a população de trabalhadores com mais de 55 anos (55-64 anos) na Europa aumente até aos 60% (Carone and Costello, 2006). Para dar resposta às alterações económicas daí resultantes, as pessoas terão de trabalhar até mais tarde, resultando no aumento da idade de reforma. Estas alterações originam ambientes de trabalho mais diversificados, o que justifica o crescente interesse pelo estudo do conceito de Ageing Workforce dado que permite conhecer os critérios de satisfação, desempenho ou motivação das pessoas em diferentes idades. O aumento desse conhecimento permite adaptar os ambientes de trabalho para que as pessoas sejam bem-sucedidas independentemente da sua idade.

Um estudo mais recente desenvolvido por (Zytkowiak, 2015) prevê que as alterações demográficas que têm vindo a ser registadas tenham impacto nas organizações, alterando as práticas de gestão de pessoas seguidas nos dias de hoje. O autor defende ainda que a percentagem de população no ativo laboral será insuficiente para sustentar os padrões da sociedade atual, o que poderá ter efeitos negativos na economia. Para que uma organização seja sustentável é importante olhar para o futuro com base na informação do passado, o que permite refletir e adaptar os comportamentos futuros.

Segundo o Gabinete de Estatísticas da União Europeia (Eurostat, 2017), em 2030 a união europeia terá mais de 123 milhões de pessoas acima dos 65 anos de idade, enquanto em anos mais recentes (2016-2020) se tem verificado uma taxa de cerca de 87 milhões. É ainda importante referir que em 2080 a união europeia prevê um total de cerca de 290 milhões de pessoas acima dos 65 anos (Anexo III – Gráfico representativo da expectativa de evolução da população entre 2016 e 2080).

Para uma gestão de pessoas efetiva, os fatores relacionados com a idade devem ser tidos em consideração na gestão diária incluindo planos de trabalho e tarefas individuais para que todas as pessoas, independentemente da sua idade, se sintam habilitadas a atingir os seus objetivos individuais e de equipa (Ilmarinen, 2012), p.2.

Para um melhor entendimento do contexto de negócio e dos objetivos do trabalho é importante perceber o conceito de Ageing Workforce. No presente trabalho, este conceito não se foca apenas no grupo etário em que a pessoa se insere, mas também no desgaste que a pessoa apresenta, que é frequentemente associado ao tipo de cargo que tem.

Uma definição geral para o conceito de colaborador mais velho é a de uma pessoa que esteja no grupo etário >= 45 anos (Brooke, 2003).

Desta forma, o conceito de envelhecimento ativo revela-se importante de forma a assegurar que as pessoas nessa faixa etária tenham acesso a condições de trabalho flexíveis, locais de trabalho

saudáveis, formação contínua e planos de reforma (Union, 2012), p.37. De forma a garantir a produtividade pretende-se que todas as pessoas, independentemente da faixa etária e do tipo de atividade que têm, se sintam confortáveis com as suas atividades individuais e de equipa, sendo essencial providenciar boas condições de trabalho. Para isso é importante que os gestores tenham conhecimento sobre o universo de colaboradores que gerem e implementem práticas relacionadas com a gestão de idades. (Beck, 2008), p.10, acredita que o nível de produtividade é reduzido pelo facto de possuírem competências ultrapassadas e não pela idade.

De acordo com (Aitken et al., 2014) existem fatores de grande relevância quando se pretende analisar os colaboradores, tendo por base o conceito de Ageing Workforce:

- São necessárias novas estratégias para manter as pessoas no ativo;
- É necessário assegurar métodos de transferência de conhecimento eficazes;
- Os colaboradores com maior idade têm mais responsabilidades económicas e sociais;
- Os colaboradores com maior idade podem ter a sua eficiência diminuída devido a problemas de saúde, o que representa uma maior taxa de absentismo;
- Existe um conflito entre gerações.

Com base nas alterações e tendências que se têm registado, os autores afirmam ainda ser crucial que os profissionais na área de gestão de pessoas ponham em prática técnicas que assegurem a transferência de conhecimento entre gerações, identifiquem as necessidades específicas das pessoas e explorem opções de reforma faseada, o que poderá ser uma forma de mitigar os problemas eventualmente criados por uma força produtiva envelhecida.

Para uma gestão de pessoas eficaz é crucial que esta seja definida de acordo com as necessidades e objetivos estratégicos do negócio (Čiutienė & Railaitė, 2014).

Os efeitos da idade na produtividade são difíceis de quantificar segundo (Boenzi, Digiesi, Mossa, Mummolo, & Romano, 2015). Por esse motivo os autores propõem um modelo relacionado com a idade dos colaboradores, o qual tem como objetivo construir um sistema para a rotação de funções em ambientes de trabalho caracterizados por tarefas com elevada repetibilidade. O modelo tem a particularidade de incorporar a idade das pessoas num esquema tradicional de rotatividade. Os efeitos produzidos pela idade inevitavelmente afetam o desempenho individual, contudo alterações ao nível de funções cognitivas e físicas podem ocorrer em qualquer altura independentemente da idade. Ainda segundo os mesmos autores, num ambiente de trabalho caracterizado por esforços físicos, manter os colaboradores saudáveis significa não só aumentar as suas capacidades, mas também reduzir o risco de lesões. Além dos efeitos do envelhecimento, existem ainda outros fatores que afetam a produtividade, os quais envolvem características individuais, e que por esse motivo são mais difíceis de quantificar, tais como: fatores cognitivos, motivações socioeconómicas e *learning-forgetting phenomena*.

Fonte	Exemplos concretos da utilização do conceito People Analytics
(Donald M. Truxillo, David M. Cadiz, 2012)	Os autores propõem uma abordagem de planeamento do trabalho orientada à idade do colaborador e dependente das caraterísticas do trabalho para explicar a relação entre a idade e os resultados obtidos. Defendem ainda que o planeamento do trabalho é uma abordagem uniformizada e dependente de vários fatores, como estratégias de compensação, características das tarefas e consequente conhecimento para a sua execução ou características interpessoais no trabalho.
(Thomas H. Davenport & Shapiro, 2009)	Os autores identificam casos concretos onde foi utilizado o conceito do People Analytics em RH, tais como: **(1)** a previsão e acompanhamento do desempenho financeiro e do envolvimento dos colaboradores na JetBlue; **(2)** a identificação das áreas organizacionais que necessitam de melhoria através de sistemas de apoio à tomada de decisão, caso de estudo aplicado à Lockheed Martin; **(3)** a identificação dos fatores que levam os colaboradores a sair ou manter-se numa organização, o qual foi aplicado ao caso de estudo desenvolvido pela Google tendo por base dados sobre os seus colaboradores;

2.2. INFORMAÇÃO E CONHECIMENTO

De acordo com (Santos & Ramos, 2009) p.7, a gestão de informação e conhecimento são duas atividades de gestão importantes para que uma organização possa manter a informação como uma vantagem competitiva e tirar partido das competências que integra. A informação que as tecnologias de informação permitem guardar para disponibilizar aos membros de uma organização representa a base para o conhecimento organizacional. As autoras afirmam que os sistemas de apoio à gestão de conhecimento têm vindo a ser desenvolvidos para apoiar a criação de novo conhecimento, melhoria de processos, partilha de experiências, bem como transformação da informação contida em grandes volumes de informação e também identificação e desenvolvimento de competências associadas à organização.

A experiência quotidiana dos membros de uma organização é um processo contínuo que contribui para o desenvolvimento da inteligência organizacional. Através deste processo contínuo, os membros da organização adquirem conhecimento sobre o negócio, construindo assim uma equipa capaz de tomar decisões, analisar soluções e melhorar processos e políticas em diferentes contextos da organização, contribuindo positivamente para otimizar as condições de negócio. A inteligência organizacional pode ser definida segundo (Santos & Ramos, 2009), p. 73, como "capacidade coletiva, distribuída pelos vários membros da organização, para aplicar o conhecimento e as competências coletivas na produção de novas respostas para problemas que ameaçam a sobrevivência e bem-estar económico, social e ambiental da organização".

O Data Science e Data Mining são ambos conceitos aliados à gestão de conhecimento que envolvem conhecimentos tecnológicos.

2.3. DATA SCIENCE E DATA MINING

Data Science é um conceito multidisciplinar utilizado para descrever a transformação de dados em conhecimento. De acordo com (Jurney, 2013), o objetivo do Data Science é a análise de dados e consequente extração de conhecimento através de conceitos estatísticos, técnicas de Data Mining e algoritmos de Machine Learning.

A análise de dados (Data Analytics) caracteriza-se por permitir a descoberta de valor escondido em grandes volumes de dados (Fitz-enz & John R. Mattox II, 2014).

Data Mining é uma técnica que permite a análise de dados com o objetivo de encontrar padrões e modelos que permitam sumariar os dados de uma forma percetível e útil (D. Hand, Mannila, & Smyth, 2001). A Descoberta de Conhecimento em Base de Dados (DCBD) é um processo que aplica técnicas de Data Mining com o objetivo de identificar relacionamentos, padrões, tendências ou modelos nos dados armazenados (Santos & Ramos, 2009) p. 127. A DCBD é definida por (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996) como sendo "o processo não trivial de identificação de padrões válidos e potencialmente úteis, percetíveis a partir dos dados". Este conjuga fundamentos provenientes de diferentes áreas, tais como inteligência artificial, aprendizagem automática, estatística, reconhecimento de padrões, bases de dados, ciências da informação, entre outras. A Figura 2.2 representa o processo de descoberta do conhecimento em base de dados.

Figura 2.2 - Processo de DCBD (Santos & Ramos, 2009) p. 105.

Existem diferentes métodos de aprendizagem para a obtenção de conhecimento a partir dos dados, como é o caso da aprendizagem supervisionada (previsão) e não-supervisionada (descrição).

2.3.1. Aprendizagem não supervisionada – Modelação Descritiva

A aprendizagem não supervisionada, também designada como modelação descritiva, é feita com base em factos observados que permitam obter um conhecimento sumariado. Neste tipo de aprendizagem, não são definidas classes pelo que o algoritmo de Data Mining utilizado identifica os padrões existentes permitindo aumentar o conhecimento à cerca dos dados. Este tipo de aprendizagem pode ser de diferentes tipos, tais como:

- **Análise de clusters**: técnica que tem como objetivo agrupar de forma homogénea os objetos e/ou variáveis.
- **Regras de associação** ou Market Basket Analysis: técnica que permite identificar itens que ocorrem em conjunto num determinado evento ou registo;
- **Visualização**: técnica que permite a representação gráfica dos dados.

2.3.1.1. A importância da análise de clusters na classificação

A análise de clusters é uma técnica aplicada no campo da estatística descritiva (não inferencial) que permite a construção de grupos de entidades semelhantes entre si. Contrariamente aos testes

estatísticos (t-teste, ANOVA, …) que têm como objetivo confirmar hipóteses, a análise de clusters é usada para perceber os padrões existentes nos dados, através do agrupamento de entidades.

Na análise de clusters não existem exemplos pré-classificados. Os algoritmos agrupam as entidades de acordo com um critério de semelhança.

A análise de clusters surge frequentemente associada ao processo de Data Mining dado que constitui um dos primeiros passos do processo de extração do conhecimento de grandes quantidades de dados (Han & Kamber, 2006). Os autores defendem ainda que, nos dias de hoje, as técnicas de análise de clusters dividem-se em:

- Métodos de partição ou otimização**,** que englobam o método k-means, k-medoids;
- Métodos hierárquicos**,** os quais criam uma divisão hierárquica do conjunto de objetos, tendo como principal desvantagem o facto de não se poder voltar atrás;
- Métodos de densidade;
- Métodos baseados em grelhas;
- Métodos baseados em modelos;

A análise de clusters implica a definição do conjunto de variáveis a utilizar, definição de um critério de semelhança/dissemelhança entre entidades, aplicação do algoritmo de clustering e análise e validação da solução final (Bação, n.d.).

Para o desenvolvimento deste trabalho serão utilizados métodos não hierárquicos devido às vantagens que apresentam, como a fácil aplicação em grandes conjuntos de dados o que por sua vez não é possível a partir dos métodos hierárquicos. Existem vários métodos não hierárquicos os quais englobam o algoritmo de partição k-means, SOM, etc. Para a realização deste trabalho optou-se por escolher o método k-means, apresentado na subsecção seguinte.

2.3.1.2. Algoritmo de Partição: k-means

O algoritmo de partição k-means é uma das técnicas mais utilizadas na análise de clusters. A partir de um conjunto de dados, o algoritmo constrói uma partição, isto é, um conjunto de objetos cuja totalidade constitui o conjunto inicial. O algoritmo inicia-se com um número de clusters (k grupos) pré-definidos. Com recurso a centroides pré-estabelecidos (seeds), o algoritmo agrupa os elementos por k clusters e recursivamente recalcula os centroides. A melhor partição deverá satisfazer os critérios de homogeneidade, coesão interna, isolamento dos grupos e heterogeneidade entre grupos. De uma forma geral é um processo iterativo, que se sintetiza da seguinte forma:

1- Seleciona uma partição de n objetos em k clusters, definidos à priori;
2- Calcula os centroides para cada k cluster e posteriormente calcula a distância do centroide a cada ponto;
3- Agrupa os objetos cujos centroides se encontram mais próximos, formando k clusters, e depois disso é calculada a nova média de cada cluster;
4- O processo continua para o ponto 2 até a função objetivo convergir, construindo k clusters o mais compacto e separado possível (isto é, até que não ocorra uma variação significativa na distância mínima de cada indivíduo da base de dados a cada um dos centroides).

A eficiência computacional e a fácil aplicação em grandes conjuntos de dados são as principais vantagens do algoritmo, sendo ainda possível implementar o algoritmo com diferentes centroides iniciais, o que produz soluções diferentes para o mesmo número de grupos.

Este método só pode ser aplicado quando é possível calcular a média de um cluster, ou seja, é aplicado em variáveis contínuas. O cálculo da distância é baseado na distância euclidiana, o que faz com que o algoritmo tenha tendência a encontrar clusters esféricos, de dimensão e densidade semelhante (Bação, n.d.). É ainda de salientar que a variância à volta do centroide do cluster deverá ser minimizada.

A distância Euclidiana (1) entre dois elementos (i, j) é obtida através da raiz quadrada do somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis (v=1, 2,...., p):

$$d_{ij} = \sqrt{\sum_{v=1}^{p} \left(X_{iv} - X_{jv}\right)^2} \qquad (1)$$

Devido ao facto de utilizar a distância euclidiana como medida de distância, o algoritmo funciona melhor em variáveis com distribuição normal.

2.3.2. Aprendizagem Supervisionada – Modelação Preditiva

A aprendizagem supervisionada também conhecida como modelação preditiva, permite aprender um critério de decisão para a classificação de exemplos novos e desconhecidos, isto é, são utilizados modelos capazes de prever o valor de uma variável com base na informação de outra variável existente nos registos. Pode ser considerada a classificação – árvore de decisão - quando pertence a determinada classe, o que acontece no caso de variáveis categóricas, ou na regressão quando aplicada a variáveis continuas.

As árvores de decisão são frequentemente utilizadas em métodos de frequência indutiva, e permitem obter resultados com base em exemplos pré-classificados – conjunto de treino. Um conjunto de dados é dividido em subconjuntos através da aplicação de regras que promovem a homogeneidade dos conjuntos de acordo com a variável dependente, target, (Berry & Linoff, 2004).

As árvores são compostas pelos nós, ramos e nós terminais (também conhecidos por folhas). De acordo com (D. J. Hand, 1998), os nós representam os testes ou atributos, o ramo corresponde à resposta ao teste e os nós terminais correspondem a um conjunto de dados homogéneo. As árvores de decisão seguem uma abordagem descendente, *top-down,* para seleção dos atributos que constituem as regras de um modelo. O conjunto de dados de treino é dividido sucessivamente até formar conjuntos homogéneos (D. J. Hand, 1998).

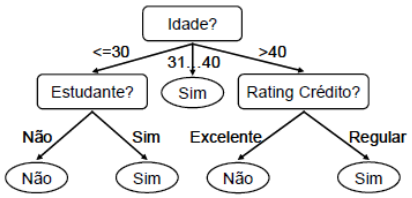


Figura 2.3 – Exemplo de classificação de uma árvore de decisão com base no algoritmo C4.5. (Bação, n.d.)

Atualmente existem alguns algoritmos utilizados na construção das árvores de decisão, de entre os mais populares destacam-se o algoritmo CART (Classification and Regression Trees) e o algoritmo C4.5, o qual surge como uma versão melhorada do algoritmo ID3.

A utilização de uma abordagem descendente implica a utilização de medidas adequadas para a seleção dos atributos, isto é, deve ser assegurada a máxima capacidade discriminante. Ainda de acordo com o mesmo autor, (D. J. Hand, 1998), o critério de divisão de cada nó da árvore em novos ramos reconhece o melhor atributo a partir do qual deve ser feita a divisão em cada nó e também os ramos do nó que devem crescer de acordo com o objetivo do modelo. A divisão sucessiva em vários subconjuntos termina quando os nós terminais da árvore apresentam grupos de classes homogéneas. A existência de uma única classe garante um grau de pureza ideal no conjunto de dados.

2.3.2.1. Árvore de decisão aplicada a Recursos Humanos

As árvores de decisão são um método de classificação popular devido à fácil interpretação dos resultados obtidos que resultam em regras interpretáveis e lógicas. As regras extraídas a partir de uma árvore de decisão podem ser usadas para previsões futuras.

No estudo apresentado por (Jantan, Razak Hamdan, & Ali Othman, 2010), os autores avaliam a aplicabilidade das árvores de decisão em modelos de previsão, aplicados a Recursos Humanos, mais especificamente na retenção de talentos numa organização. No caso de estudo mencionado acima, os dados são provenientes de várias instituições e são relativos aos professores universitários em categorias distintas.

Na construção de uma árvore de decisão o processo inicia-se através da identificação da variável dependente (target), que foi definida pelos autores como “Recomenda a promoção?” e a qual assume os valores “Sim” ou “Não”, que significa a recomendação ou não, respetivamente, de um professor. Relativamente às variáveis independentes foram consideradas variáveis de caracterização do colaborador como o género, a categoria, a qualificação e a avaliação. Os autores reconhecem a Gestão de Talento como um processo essencial numa organização, dado que este permite identificar as áreas e colaboradores essenciais ao sucesso da organização o que conduz ao desenvolvimento da organização de forma a reter e aumentar o envolvimento.

O estudo foi realizado com base em vários subconjuntos de dados, divididos em dados de treino e teste, os quais foram obtidos de forma aleatória. A escolha do algoritmo recaiu sobre o algoritmo C4.5, desenvolvido por Ross Quinlan, o qual constrói árvores de decisão a partir de um conjunto de dados de treino e o qual tem por base o conceito de entropia. Para a classificação e previsão de talentos académicos, o processo geral de obtenção de conhecimento através de ferramentas de Data Mining teve por base as ferramentas WEKA e ROSETTA. O estudo dividiu-se em três fases: (1) Recolha dos dados, incluindo o tratamento e pré-processamento dos dados; (2) Obtenção das regras de classificação para o conjunto de treino, com base no algoritmo C 4.5, o que incluiu todos os atributos possíveis identificados pelos autores; (3) Avaliação e interpretação das regras de classificação com o objetivo de determinar a precisão da classificação.

Segundo os autores, o modelo de classificação desenvolvido obteve bons resultados para o conjunto de dados utilizado, no entanto, consideram que a redução de atributos deve ser feita de forma a reduzir o tempo de processamento na melhoria da precisão do modelo. Além disso, o classificador

proposto deve ser testado com diferentes conjuntos de dados para comprovar se o seu potencial se mantém com a elevada precisão que o modelo permite obter.

Para o desenvolvimento de árvores de decisão com boa precisão, considera-se importante a utilização de conjuntos de dados com elevado número de observações, além disso o pré-processamento dos dados, identificação de outliers e valores em falta, também contribui para a melhoria da precisão do modelo.

3. METODOLOGIA

Existem diferentes metodologias aplicadas aos estudos de Data Mining. As autoras (Santos & Ramos, 2009) apresentam as principais diferenças existentes entre as metodologias de descoberta de conhecimento em bases de dados (DCBD), a metodologia CRISP-DM e o processo SEMMA.

Além de iterativo o processo DCBD é também interativo, uma vez que necessita que a tomada de decisão seja feita pelo utilizador. (Fayyad et al., 1996) p. 6, definem DCBD como "o processo não trivial de identificação de padrões válidos potencialmente úteis, percetíveis a partir dos dados".

A metodologia CRISP-DM é a mais completa de todas e apresenta-se como um modelo de referência que define as fases a seguir, as tarefas a executar e os resultados esperados com base no ciclo de vida dos projetos de Data Mining. É também um processo iterativo, à semelhança da DCBD, e apresenta as seguintes fases: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e desenvolvimento.

O processo SEMMA corresponde à metodologia desenvolvida pelo SAS Institute e está orientado para auxiliar a execução das tarefas de Data Mining através da ferramenta SAS Enterprise Miner. As etapas do projeto são Amostragem (Sample), Exploração (Explore), Modificação (Modify), Modelação (Model) e Avaliação (Assess).

Na Tabela 3.1 é apresentado um resumo de correspondências entre as metodologias apresentadas anteriormente.

<table>
<tr><th>DCBD</th><th>SEMMA</th><th>CRISP-DM</th></tr>
<tr><td>Pré DCBD</td><td>----------</td><td>Conhecimento do negócio</td></tr>
<tr><td>Seleção</td><td>Amostragem</td><td rowspan="2">Conhecimento sobre os dados</td></tr>
<tr><td>Pré processamento</td><td>Exploração</td></tr>
<tr><td>Transformação</td><td>Modificação</td><td>Preparação dos dados</td></tr>
<tr><td>Data Mining</td><td>Modelação</td><td>Modelação</td></tr>
<tr><td>Interpretação/Avaliação</td><td>Análise/Verificação</td><td>Avaliação</td></tr>
<tr><td>Pós DCBD</td><td>----------</td><td>Implementação</td></tr>
</table>

Tabela 3.1 – Resumo de correspondências entre metodologias (Azevedo & Santos, 2008).

3.1. O PROCESSO SEMMA

O processo SEMMA permite um fácil entendimento do processo e também uma adequada organização, desenvolvimento e manutenção de projetos de Data Mining (Azevedo & Santos, 2008). A metodologia adotada para o desenvolvimento deste trabalho corresponde ao processo SEMMA. Nas subsecções seguintes é apresentada a metodologia utilizada no desenvolvimento da tese de acordo com as fases de desenvolvimento definidas por este processo.

3.1.1. Dados utilizados - Amostragem

A fase de amostragem consiste na seleção dos dados a analisar a partir do conjunto de dados disponíveis. Geralmente a amostra é dividida em: treino (conjunto de dados utilizados na identificação

do modelo), validação (conjunto de dados utilizados na avaliação do modelo) e teste (conjunto de dados utilizados para analisar a capacidade de generalização do modelo.

3.1.2. Identificação da base de dados

Para o desenvolvimento deste estudo foi utilizada uma base de dados disponibilizada pela organização com os identificadores de colaboradores codificados não permitindo a identificação do colaborador. Estes dados contém um total de 1800 registos e 76 variáveis associadas ao colaborador, no período janeiro a junho de 2017. O principal objetivo da análise de clusters é agrupar os colaboradores com base nas características comuns, isto é, nos valores que possuem nas diferentes variáveis e desta forma conhecer melhor os diferentes perfis de colaboradores.

3.1.3. Identificação das variáveis

Por forma a manter o anonimato sobre os dados dos colaboradores, a informação disponibilizada foi previamente codificada pela organização sendo desta forma assegurada que não é transmitida qualquer informação que possa identificar o seu titular.

As variáveis disponibilizadas contêm dados que caracterizam as três dimensões consideradas: colaborador, gestor de loja e loja. Para a dimensão colaborador, são utilizadas variáveis para caracterizar a informação demográfica (faixa etária, género, estado civil, número de filhos, formação académica) e informação de contexto profissional (como o tipo de vínculo, número de anos na organização, número de horas de formação, absentismo). Relativamente à dimensão gestor de loja e loja, as variáveis contêm informação relativa à loja em que o colaborador se encontra e informação relativa à sua chefia (igualmente dados demográficos e de contexto profissional).

Para uma melhor compreensão de cada variável individualmente existe um glossário de variáveis, no entanto, como se considera ser informação confidencial não pode ser apresentado neste documento.

Antes de prosseguir com a análise exploratória dos dados através do software SAS, foi necessário definir roles e levels das variáveis.

3.1.4. Análise Exploratória dos dados - Exploração

A fase de exploração dos dados surge logo após a fase de amostragem e contribui para o aumento do conhecimento sobre os dados em estudo.

Para o conjunto de dados utilizado, os colaboradores são maioritariamente do sexo masculino e a maioria possui o ensino básico. A maioria dos colaboradores encontram-se divididos pelas faixas etárias acima dos 35 anos de idade.

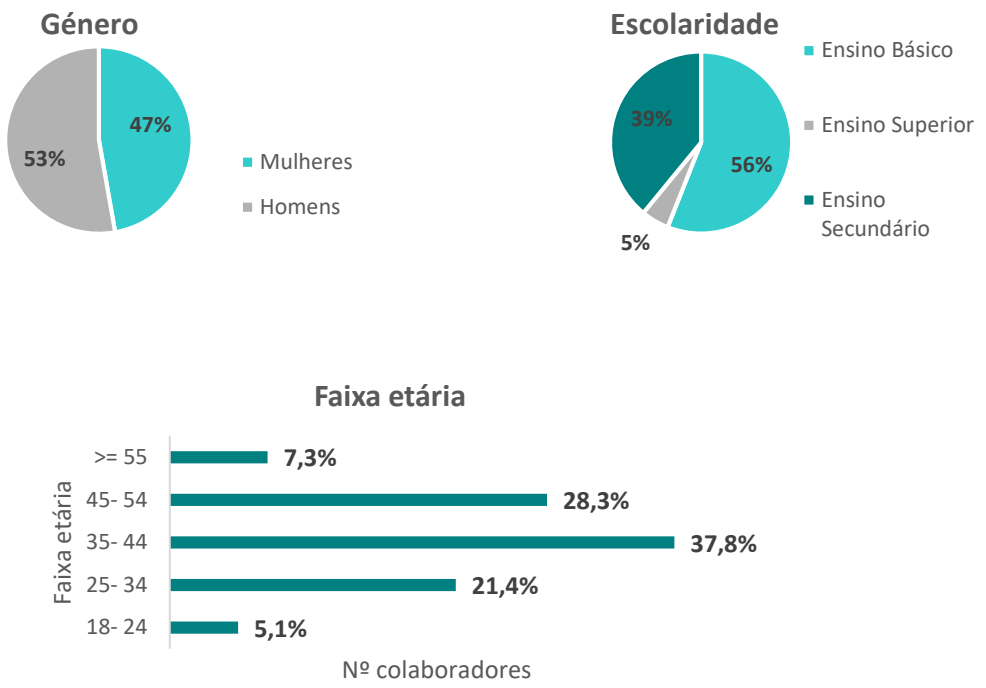


Figura 3.1 – Contextualização do universo de colaboradores em estudo.

O SAS Enterprise Miner permite de uma forma rápida obter as estatísticas descritivas mais importantes para cada variável, das quais se podem considerar:

- Estatísticas descritivas das variáveis intervalares (número de valores existentes no dataset (non-missing values), número de valores em falta (missing values), total, valor mínimo, médio, mediana, máximo e desvio padrão para cada variável;
- Tabela de frequência das variáveis categóricas, a qual permite obter o número de valores em falta, a moda e a frequência da moda.
- Histograma de frequência para as variáveis intervalares e categóricas;

Desta forma, foi possível fazer uma análise para cada variável e obter as estatísticas descritivas, bem como os histogramas de frequência. É ainda de referir que durante esta fase foi possível identificar e retirar registos incorretos e perceber a distribuição associada a cada variável através da análise visual dos histogramas, os quais permitem identificar outliers. Esta fase permite também conhecer com algum detalhe os dados e por isso serve de auxílio no processo de definição das variáveis relevantes para a construção dos clusters. Foi ainda possível obter o número total de registos na base de dados, o número de variáveis em cada categoria (nominal, binária, intervalar e ordinal) e o número de valores em falta (missing values).

As variáveis disponibilizadas, embora tivessem como referência o período de janeiro a junho, continham informação dividida em variáveis trimestrais. Desta forma, para assegurar a qualidade dos dados, optou-se por fazer uma análise para variáveis referentes ao semestre janeiro a junho.

Optou-se por não considerar os registos com valores em falta para as variáveis de avaliação sobre o envolvimento (engagement) e os desafios de cada função (job challenge) por loja, dado que as lojas

que apresentavam valores em falta não participaram nos questionários de avaliação. Considerar essas variáveis poderia ser um risco na medida em que a amostra poderia ficar enviesada.

Foram também eliminadas 2 lojas do conjunto de dados, devido ao facto de serem consideradas outliers, 12% e 0,8% da amostra, respetivamente.

Para as estatísticas descritivas a análise foi feita essencialmente com base na tendência central (média e mediana), forma de dispersão (desvio padrão e coeficiente de variação) e forma de distribuição (máximo, mínimo e skewness).

Foi possível aferir que a maioria das variáveis apresentam uma distribuição não normal. Após experimentar algumas transformações optou-se por fazer a transformação Maximum Normal, que aplica diferentes transformações de forma a maximizar a normalidade dos dados em cada variável. No entanto, para a construção de clusters optou-se por utilizar variáveis não transformadas devido ao risco de enviesar a solução obtida.

Além das estatísticas descritivas foi possível aferir, pela análise da matriz de correlações (Anexo II – Matriz de correlações), que existem variáveis altamente correlacionadas e por isso optou-se por eliminar uma das variáveis do par, sendo que o nível de corte foi definido como uma correlação >= 0,8. Na fase seguinte é identificado o nó que permitiu eliminar as variáveis altamente correlacionadas.

Considerou-se também importante durante a análise exploratória a criação de algumas variáveis, tais como:

- Divisão da faixa etária em 2 grupos – colaboradores/gestores de loja com menos de 45 anos e colaboradores/gestores de loja com 45 anos ou mais (<45 e >=45 anos);
- Assimetria de idade entre os colaboradores por loja (faixa etária <45 anos e >=45 anos);

A decisão pela criação destas variáveis/segmentos resultou do aumento do conhecimento sobre o conjunto de dados disponibilizados e pelas diferentes tentativas de implementação de um modelo com resultados de acordo com o conhecimento adquirido.

Através do SAS é possível explorar as variáveis de uma forma visual com recurso a gráficos produzidos através do nó Graph Explore. A título de exemplo, foi possível analisar os dias perdidos por acidentes de trabalho em função da idade dos colaboradores. É possível aferir que os colaboradores com maior idade apresentam mais dias perdidos por acidentes de trabalho, no entanto, são também os que apresentam menos horas de formação.

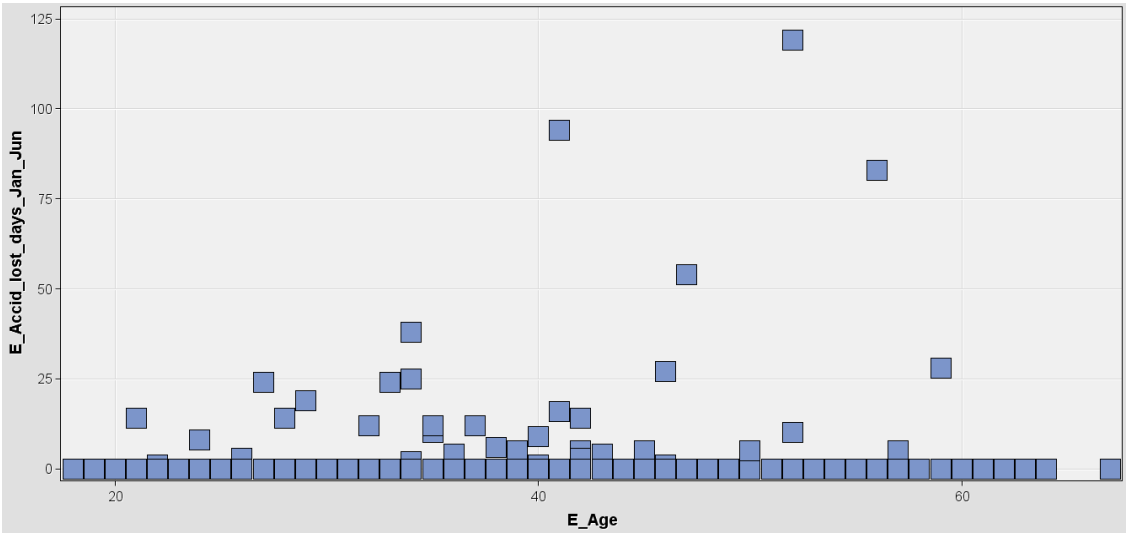


Figura 3.2 – Gráfico representativo dos dias perdidos por acidente de trabalho em função da idade do colaborador.

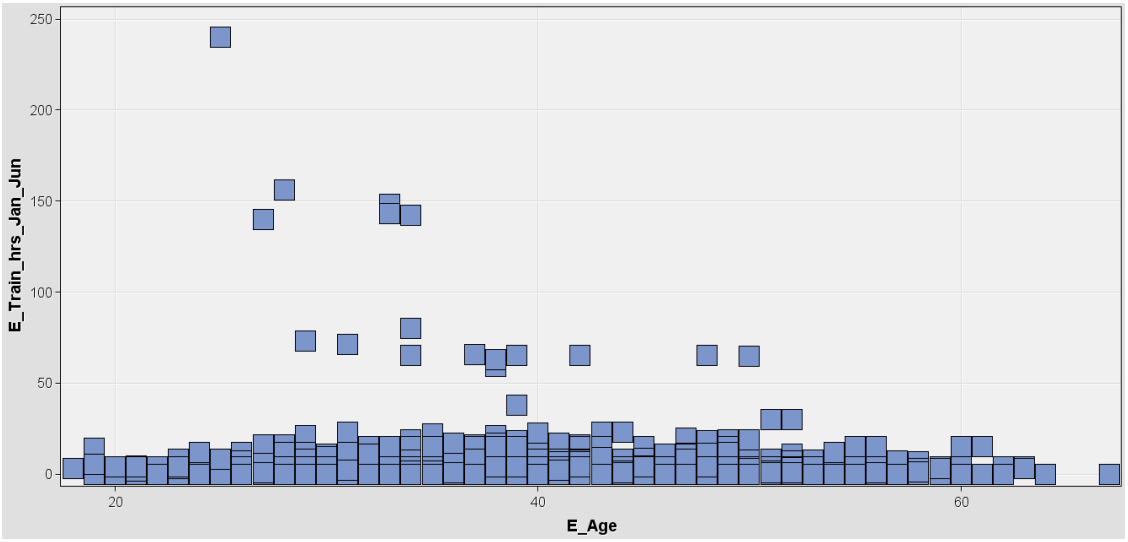


Figura 3.3 - Gráfico representativo das horas de formação em função da idade do colaborador.

Além da utilização do SAS Enterprise Miner, durante a fase exploratória foi também feita uma análise em Excel com recurso a tabelas pivot e gráficos de forma a ganhar conhecimento sobre os dados. Através dessa análise, foi possível ganhar conhecimento sobre os dados contidos na base de dados, nomeadamente acerca dos seguintes pontos chave:

(1) Verifica-se que existe uma variação entre os dias de ausência por acidente de trabalho nas diferentes faixas etárias, bem como de ausência em geral.

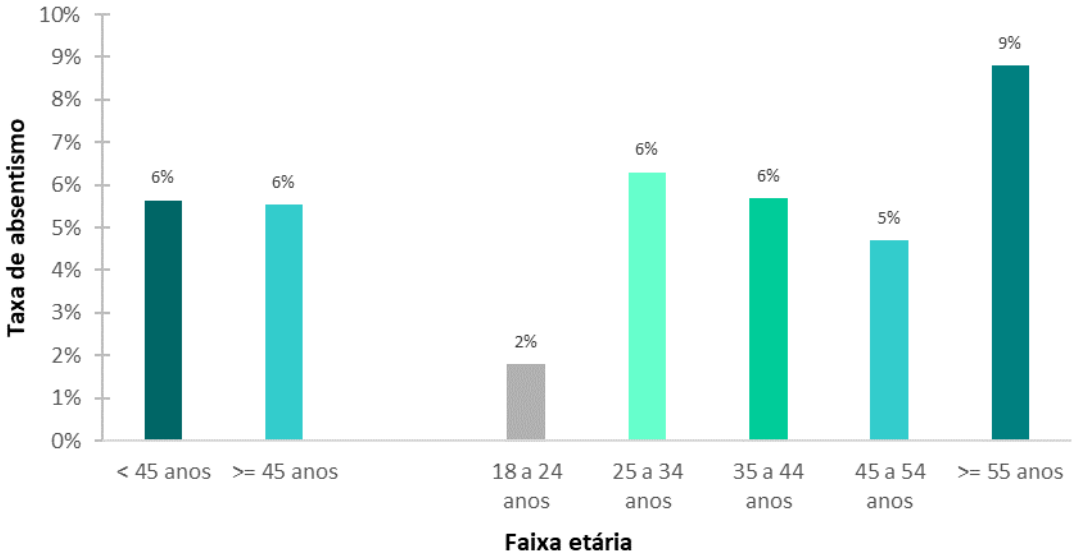


Figura 3.4 – Gráfico representativo da taxa de absentismo por faixa etária.

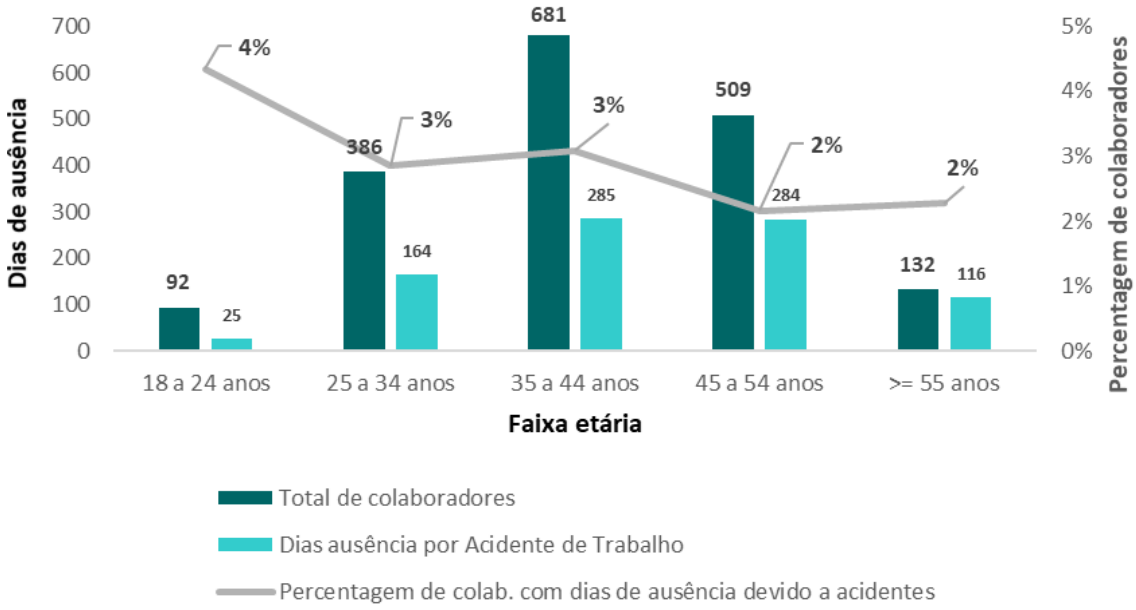


Figura 3.5 – Gráfico representativo dos dias perdidos por acidentes de trabalho e da percentagem de colaboradores com dias perdidos por acidentes por faixa etária.

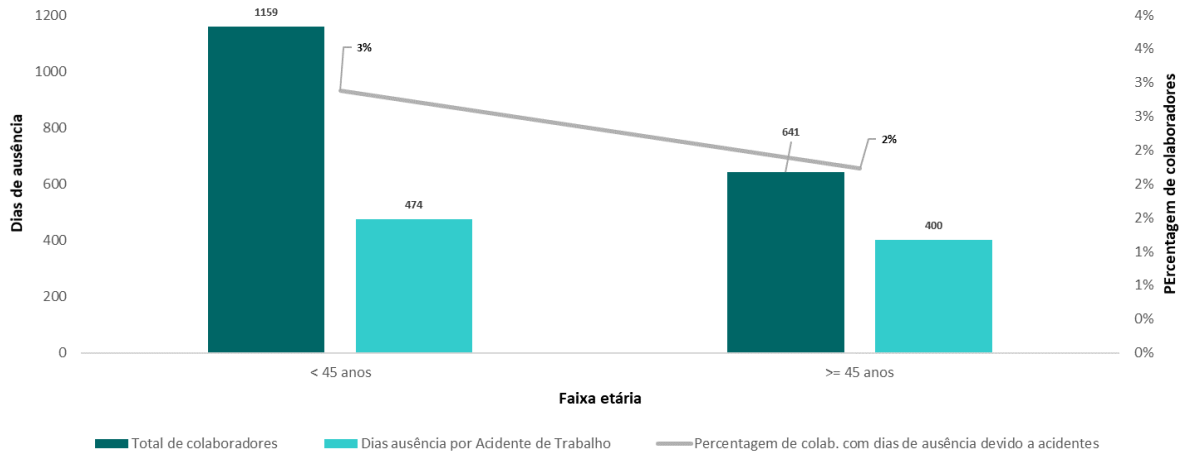


Figura 3.6 – Gráfico representativo dos dias perdidos por acidentes de trabalho e da percentagem de colaboradores com dias perdidos por acidentes por faixa etária (<45 e >= 45 anos).

(2) Verifica-se que os colaboradores com maior idade estão frequentemente associados a lojas geridas por gestores de loja com uma senioridade mais elevada na organização, e consequentemente também com maior idade.

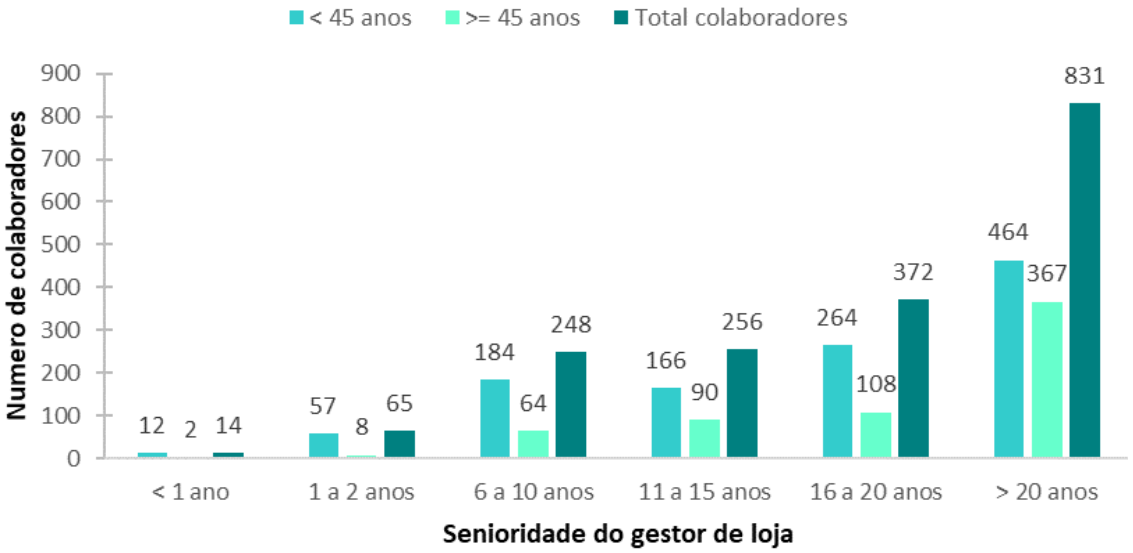


Figura 3.7 – Gráfico representativo da distribuição do número de colaboradores por faixa etária e por senioridade do gestor de loja.

No gráfico apresentado na Figura 3.7, o número total de colaboradores apresentado não inclui um conjunto de 14 colaboradores para os quais a senioridade do gestor de loja não foi disponibilizada.

(3) Verifica-se que o perfil dos gestores de loja corresponde frequentemente a pessoas na faixa etária >= 45 anos e com mais de 20 anos na organização.

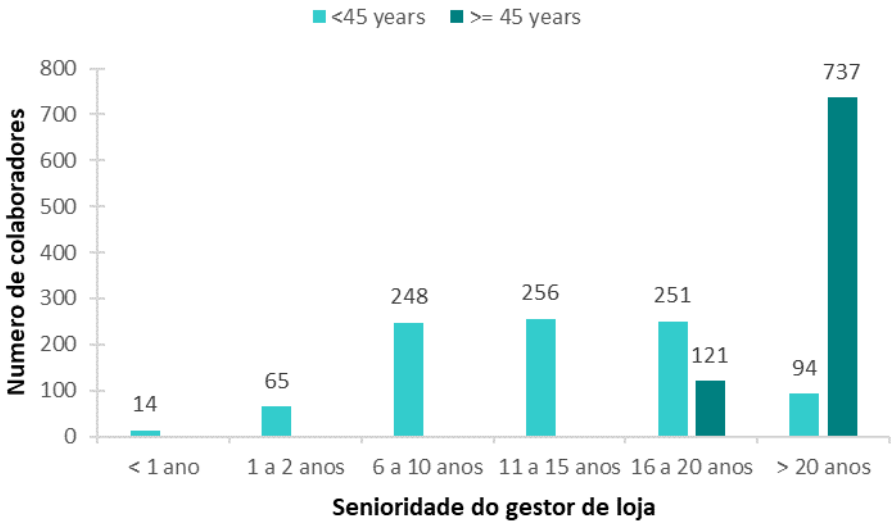


Figura 3.8 – Gráfico representativo da distribuição do número de colaboradores por faixa etária do gestor de loja e por senioridade do gestor de loja.

3.1.5. Modificação

A fase de modificação permite a criação de novos atributos, novas variáveis, e também a transformação das variáveis existentes.

Através do nó Drop foi possível eliminar variáveis consideradas não relevantes para a análise. O conhecimento sobre os dados e também o facto de possuírem valores em falta foram os critérios seguidos. Através deste nó, foram também eliminadas as variáveis que se considerou serem altamente correlacionadas, considerando um nível de correlação >=0,8 - Anexo II – Matriz de correlações entre variáveis obtida através do SAS Enterprise Miner.

3.1.6. Construção dos clusters

A construção de clusters durante a fase de modelação tem como objetivo agrupar os colaboradores que possuem perfis semelhantes. Desta forma é possível caracterizar os diferentes tipos de colaboradores existentes na amostra o que contribui para uma análise descritiva consistente.

De entre os métodos de clustering existentes na bibliografia, optou-se por utilizar o método de clustering não hierárquico, k-means, devido à facilidade com que o algoritmo pode ser implementado computacionalmente em grandes conjuntos de dados. Esta implementação foi dividida em 2 fases:

1ª fase: Escolha das variáveis a incluir no cluster;
2ª fase: Escolha do modelo de cluster com significado de interpretação.

A primeira fase caracteriza-se por ser um processo iterativo para definição das variáveis de input. A decisão sobre quais as variáveis a incluir é da responsabilidade do utilizador e é feita com base no conhecimento sobre os dados e nos grupos obtidos. Durante esta fase foram testadas diferentes combinações de variáveis. Logo após a decisão sobre quais as variáveis a incluir, a escolha do número de clusters é feita aplicando a regra do cotovelo, a qual permite analisar a distância ao centroide de cada cluster.

O SAS disponibiliza um método automático para definição do número de clusters e um método alternativo em que o número de clusters é definido à priori pelo utilizador.

Inicialmente optou-se por seguir o método automático, para definição do número de clusters, disponibilizado pelo SAS, o que deu origem a 4 clusters. Para a aplicação deste método, existem 3 formas distintas para o cálculo da distância entre clusters: **Average** (distância média entre 2 pares de observações), **Centroid** (distância euclidiana entre dois centroides) e **Ward** (método utilizado por defeito, no qual a distância entre 2 clusters é dada pela soma dos quadrados entre dois clusters mais a soma global das variáveis).

Relativamente ao processo de inicialização das sementes (seeds), este pode ser de 3 formas distintas: **MacQueen**, **First** e **Princomp**. O primeiro método é utilizado por defeito e tem por base o algoritmo k-means para definir a semente inicial do cluster; o método First define os primeiros casos completos como as sementes iniciais e o método Princomp tem por base a análise das componentes principais para definir a inicialização das sementes. Os restantes métodos disponibilizados pelo SAS, Full Replacement e Partial Replacement não se aplicam a esta análise devido ao facto de serem indicados para identificação de outliers.

Os resultados obtidos tendo por base o método automático encontram-se registados na Tabela 3.2, através da qual se verifica que uma possível solução poderá ser entre 4 a 20 clusters.

Método de clustering	Inicialização da seed	Nº Clusters	CCC
Average	MacQueen	16	0,636
Average	Princomp	20	0,568
Average	First	4	0,765
Centroid	MacQueen	17	0,618
Centroid	Princomp	4	0,744
Centroid	First	7	0,701
Ward	MacQueen	20	0,607
Ward	Princomp	20	0,568
Ward	First	20	0,57

Tabela 3.2 – Resultados obtidos a partir do método automático de construção de clusters.

De entre estes 3 métodos para inicialização das sementes apenas um pode ser escolhido. Essa escolha deve ser feita tendo por base a premissa de que o melhor modelo é aquele que tem a menor distância máxima à semente inicial, isto é, maximiza as diferenças entre clusters e as semelhanças dentro de cada cluster.

Optou-se por utilizar a distância euclidiana - **Centroid** - para o cálculo da distância entre clusters. Relativamente ao cálculo das seeds iniciais optou-se por seguir o método das componentes principais – **Princomp** - que para o mesmo número de clusters apresenta o segundo valor maior no parâmetro Cubic Clustering Criterion.

Embora a solução Average/First apresente um valor maior para o parâmetro CCC, a solução obtida não apresenta uma frequência por cluster homogénea e por isso não foi considerada para a construção de

clusters. Relativamente ao método ward, a solução que se obtém não tem significado válido devido ao elevado número de clusters, 20.

O método das componentes principais tem como vantagem o facto de permitir que as seeds iniciais não fiquem muito juntas. Verificou-se também que o método MacQueen obtém clusters com frequências muito diferentes. Enquanto que o método das componentes principais permite obter clusters com frequências mais equilibradas sugerindo também clusters mais homogéneos.

Para a construção dos clusters, foi utilizado o nó Cluster, o qual permite a utilização de variáveis binárias, nominais, ordinais e intervalares, dado que os três primeiros tipos de variáveis referidos são codificados em variáveis dummy numéricas para a construção dos clusters.

Seguidamente foi testado o método em que o número de clusters é definido pelo utilizador. Foram testadas várias soluções, com diferentes números de clusters definidos à priori (k=8,7,6,5,4), com base na análise da variância explicada. A análise da variância é feita através do método elbow "cotovelo" (Figura 3.9), que se caracteriza por ser um método visual. Este método baseia-se no facto de que o aumento do número de clusters pode ajudar a reduzir a soma das variâncias dentro do cluster, devido ao facto de que a existência de um número maior de grupos permite a captura dos grupos que apresentam maior semelhança entre si. No entanto, a soma das variâncias (dentro do cluster) pode baixar se muitos grupos forem formados, porque a divisão de um conjunto coeso em dois origina uma redução. Desta forma, a escolha do número ideal de clusters pode ser feita tendo por base o ponto de viragem na curva da soma de variâncias (dentro do cluster), em relação ao número de grupos.

Através do gráfico abaixo é possível verificar que, para um certo valor de k a curva representada no gráfico diminui, ou seja, o ganho em termos de coesão dos clusters deixa de justificar a criação de um cluster adicional (Bação, n.d.).

Nº clusters	RMSDv	|Decréscimo|
8	4,29	
7	2,91	1,38
6	3,36	0,45
5	2,74	0,62
4	3,37	0,63
3	3,58	0,21
2	4,83	1,24
1	4,39	0,43

Tabela 3.3 – Valores utilizados para a construção do gráfico cotovelo.

Figura 3.9 – Gráfico cotovelo.

Em ambos os métodos (Automatic e User Specify) foram utilizadas as variáveis que se consideraram mais importantes. Essa decisão foi tomada com base nos segmentos obtidos/perfil de colaboradores obtidos, e também no conhecimento adquirido durante a fase de exploração dos dados.

Os resultados obtidos através do método automático coincidiram com a análise feita através do gráfico cotovelo, pelo que as variáveis abaixo identificadas (Figura 3.10) representam a melhor solução correspondente a k=4.

Name	Use
ST_years	Yes
SM_numb_functions	Yes
ST_Assimetria_Age_range_NEW	Yes
ST_Per_Empl_acc_lostdays	Yes
ST_Avg_acc_lostdays	Yes
ST_terminations	Yes
ST_empl_avg_legal_sen	Yes
E_Marital_status	Yes
E_N_Children	Yes
E_Gender	Yes
E_Sen_Legal_yrs	Yes
E_Numb_fuctions	Yes
E_Age	Yes
E_Age_Range	Yes
E_Accid_lost_days_Jan_Jun	Yes
E_Educ_level	Yes
E_Avg_yrs_function	Yes
SM_Gender	Yes
SM_Marital_status	Yes
SM_Educational_Level	Yes
SM_avg_years_function	Yes
SM_N_Filhos	Yes
Engagement___fav_	Yes
Job_Challenge___fav_	Yes
E_Train_hrs_Jan_Jun	Yes
SM_Age_range	Yes
SM_Age	Yes

Figura 3.10 – Variáveis incluídas no cluster colaboradores.

No subcapítulo seguinte são apresentados os diferentes segmentos obtidos para uma solução com 4 clusters.

3.1.7. Avaliação dos clusters

A avaliação dos clusters obtidos assenta na minimização das diferenças dentro do cluster e maximização das diferenças entre clusters. Para avaliar o perfil dos clusters obtidos através do algoritmo k-means, foi utilizado o nó *Segment Profile*.

Este nó permite examinar os segmentos de dados gerados, os clusters, e identificar os pontos que diferenciam os segmentos do conjunto de dados. A análise é feita com base nos diferentes critérios que o nó disponibiliza para exploração dos resultados. Para esta análise os critérios considerados mais relevantes foram:
- as **tabelas de frequência**, onde é possível verificar a distribuição de registos por cluster;
- os **gráficos de perfil**, os quais permitem visualizar a distribuição das variáveis por cluster comparativamente com a amostra de dados;
- o **perfil de importância das variáveis para a árvore de decisão**, no qual as variáveis com mais importância, isto é, maior valor atribuído no critério worth são identificadas como sendo as que têm maior poder discriminatório para o cluster obtido;
- os **histogramas** com o valor calculado do critério worth para cada cluster e a importância das variáveis por cluster, baseado no valor de worth que a variável tem;

Relativamente ao funcionamento do nó a importância das variáveis num determinado cluster é decidida através da criação de uma pseudo variável target, a qual se baseia numa medida designada adesão ao segmento. O nó disponibiliza dois métodos para determinar a diferenciação entre variáveis. Para esta análise optou-se por utilizar o método default disponível, o qual atribui uma ordem de importância às variáveis intervalares e às variáveis de classe dependentes, baseada no parâmetro *logworth value.* Por sua vez, o valor de worth é baseado na pseudo target variable e as variáveis intervalares são escolhidas de forma a obter o valor máximo para o critério logworth. Assim, as variáveis com maior poder discriminatório (maior valor worth) para a árvore de decisão aparecem em primeiro lugar nas estatísticas de perfil de importância das variáveis.

Assim, através da técnica de segmentação é possível dividir a amostra em segmentos e identificar as variáveis com maior contributo para cada segmento/cluster. No contexto do estudo, um segmento representa colaboradores com as mesmas características, isto é, faixa etária semelhante, número de funções semelhante, etc. Uma segmentação bem-sucedida deverá permitir obter grupos de colaboradores com o mesmo perfil em diferentes amostras.

Através das **tabelas de frequência** é possível verificar como se distribui a amostra pelos clusters, o que pode também ser representada através da Figura 3.11, na qual é possível observar uma distribuição uniforme, a qual varia de 21,2% a 27,9%.

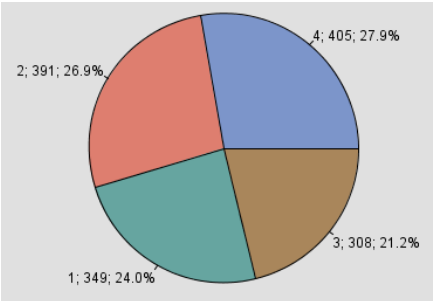


Figura 3.11 – Distribuição da amostra por cluster.

Na análise através dos **gráficos de perfil**, as variáveis de classe são representadas num gráfico circular representado por 2 anéis concêntricos. O anel interno representa a distribuição da amostra total, enquanto o anel exterior representa a distribuição de um dado segmento. Relativamente às variáveis intervalares, as mesmas são representadas por um gráfico de barras - histograma. As barras representadas a azul correspondem à distribuição da amostra num dado segmento e o contorno a vermelho representa a distribuição da amostra. O contributo das variáveis em cada segmento diminui no sentido da direita para a esquerda, sendo a variável com maior contributo para o segmento apresentada à esquerda.

Relativamente aos 4 segmentos obtidos, os mesmos foram divididos em:
Segmento 1: Gestores de loja mais novos que estão em lojas mais recentes;
Segmento 2: Gestores de loja mais novos (do que o segmento 1) e solteiros;
Segmento 3: Gestores de loja mais experientes, colaboradores mais novos e boa avaliação para o engagement e job challenge;
Segmento 4: Gestores de loja mais experientes, colaboradores com maior idade e boa avaliação para o engagement e job challenge;

Segmento 1:
- Representa 24% da amostra;
- Lojas com uma senioridade média dos colaboradores mais baixa;
- Lojas com uma assimetria entre as faixas etárias elevada;
- A avaliação para o engagement e job challenge por loja é baixa, assumindo valores acima da média da amostra nas avaliações medianas;
- Gestores de loja mais novos e consequentemente com menos anos na mesma função;
- O numero de saídas nestas lojas é elevado;
- Representa lojas mais novas e com alguns dias perdidos por acidentes de trabalho (não tantos como no segmento 2), por colaborador;
- Colaboradores com uma senioridade baixa;

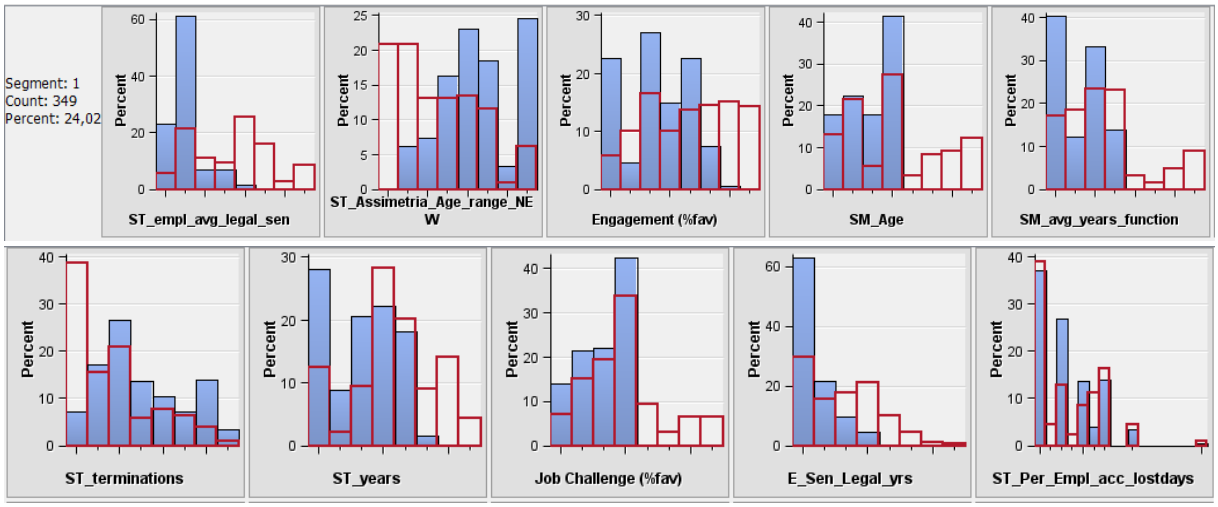


Figura 3.12 – Gráficos de perfil obtidos através do SAS Enterprise Miner para o segmento 1.

Segmento 2:
- Representa 27% da amostra;
- Gestores de loja mais novos (do que no segmento 1);
- Lojas com uma assimetria de idades entre as faixas etárias mais elevada para valores médios;
- Lojas com uma senioridade média dos colaboradores elevada para valores médios;
- Os gestores de loja têm um nível de escolaridade mais elevado – ensino superior;
- Representa as lojas com mais dias perdidos por acidentes de trabalho;
- Os gestores de loja são maioritariamente solteiros;
- A média de dias perdidos por acidentes de trabalho por colaborador é elevada (mais elevada do que no segmento 1);
- O engagement tem uma avaliação média (melhor do que no segmento 1);
- Job challenge tem uma avaliação semelhante ao segmento 1;

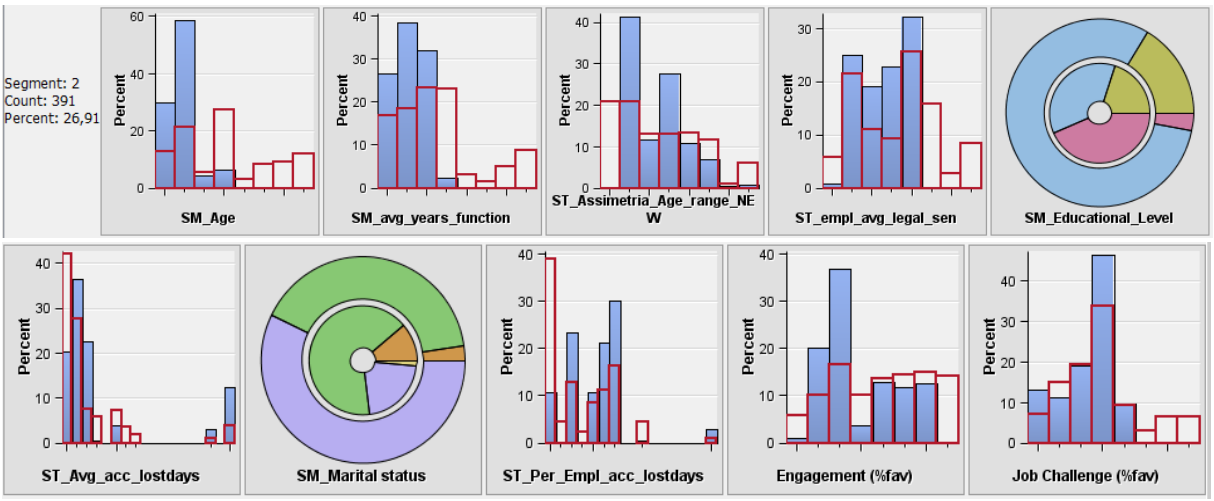


Figura 3.13 - Gráficos de perfil obtidos através do SAS Enterprise Miner para o segmento 2.

Segmento 3:
- Representa 21% da amostra;
- Os gestores de loja estão há cerca de 2-3 anos na mesma função;
- Representado por gestores de loja com idade >= 45 anos e com representatividade significativa na faixa etária >= 55 anos (são com maior idade do que no segmento 4);
- Lojas com uma senioridade média dos colaboradores elevada (não é tão elevada como no segmento 4);
- Representa as lojas com uma avaliação job challenge semelhante à da amostra total;
- Colaboradores são mais novos;
- Representa as lojas com uma avaliação elevada para o engagement;
- Representa lojas com poucas saídas;
- Representado por lojas mais antigas e com uma assimetria elevada de idades dos colaboradores;
- Lojas com uma assimetria entre as faixas etárias elevada para valores médios (não é tão elevada como o segmento 1);

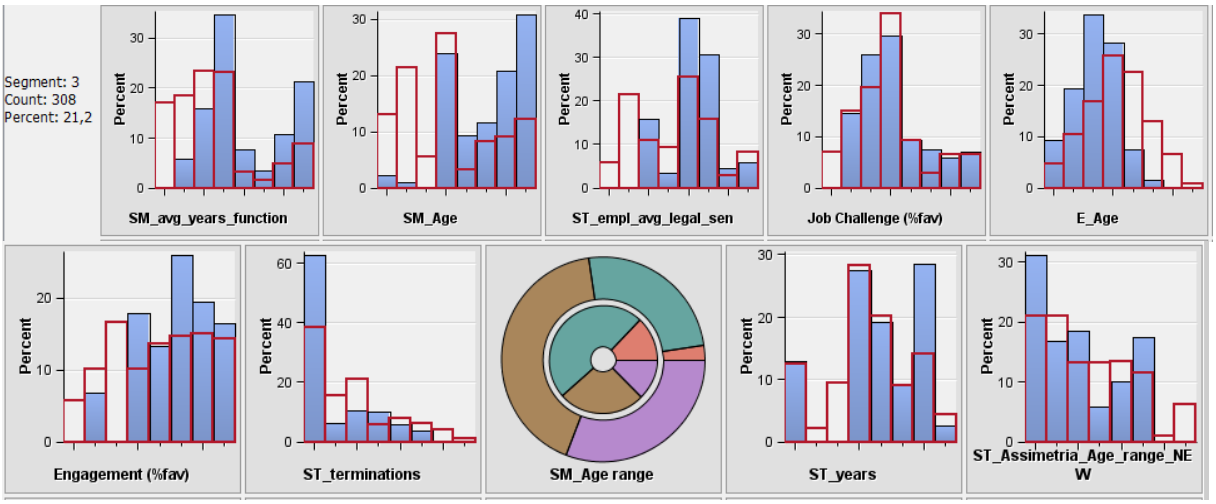


Figura 3.14 - Gráficos de perfil obtidos através do SAS Enterprise Miner para o segmento 3.

Segmento 4:
- Representa 21% da amostra;
- Lojas com uma senioridade média dos colaboradores elevada;
- Representado por gestores de loja com idade entre 45-54 anos -51%- (mais novos do que no segmento 3);
- Em geral os gestores de loja que estão há 4 anos ou mais na mesma função;
- A avaliação job challenge é elevada;
- A assimetria de idades dos colaboradores é mais baixa do que nos restantes segmentos;
- Colaboradores com maior idade, predomina a faixa etária 45-54 anos;
- Representa lojas com poucas saídas;

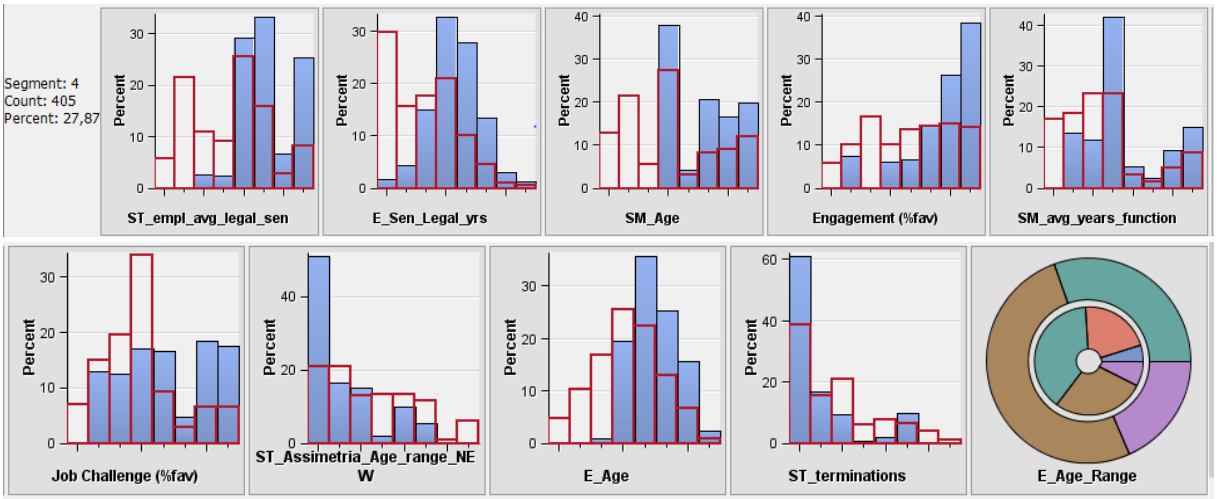


Figura 3.15 - Gráficos de perfil obtidos através do SAS Enterprise Miner para o segmento 4.

Relativamente ao perfil de importância das variáveis para a árvore de decisão, este critério permite obter informação sobre as variáveis mais importantes para os clusters obtidos. Através da análise da importância das variáveis por cluster, foi possível aferir sobre a importância das variáveis para o conjunto de dados da amostra. Na Tabela 3.4 são apresentadas variáveis que permitem caracterizar os clusters e também é identificada a dimensão a que pertencem, gestor de loja, loja e colaborador.

As variáveis são apresentadas por ordem de importância com base no valor do critério worth. O critério worth apresentado é o resultado mais alto obtido para cada variável considerando os diferentes clusters obtidos. Para cada uma das variáveis são também identificados os valores que que assumem nos clusters.

Variável	Worth	Dimensão	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
SM_Age	0,220	Gestor de loja	[31-45]	[31-45]	[34-59]	[42-59]
ST_empl_avg_legal_sen	0,216	Loja	4,947-14,8	5-15,08	9,882-21,878	2,075-6,745
SM_avg_years_function	0,192	Gestor de loja	0,605-3,518	0,605-3,518	1,634-6,745	9,882-21,878
ST_Assimetria_Age_range_NEW	0,161	Loja	0,159-0,947	0,159-0,843	0-0,647	0-45
E_Sen_Legal_yrs	0,159	Colaborador	0-28	-	-	0-0,647
Engagement___fav_	0,148	Loja	61-89	61-89	70-97	70-97
Job_Challenge___fav_	0,136	Loja	47-64	47-66	54-81	54-81
SM_Educational_Level	0,132	Gestor de loja	-	-	-	-
ST_terminations	0,119	Loja	0-0,214	-	0-0,134	0-0,134
ST_years	0,115	Loja	7,0-24	-	9,0-33	-
ST_Avg_acc_lostdays	0,112	Loja	-	0-83	-	-
SM_Marital_status	0,112	Gestor de loja	-	-	-	-
E_Age	0,110	Colaborador	-	-	19-51	
ST_Per_Empl_acc_lostdays	0,106	Loja	0-0,142	0-0,142	-	-
E_Age_Range	0,099	Colaborador	-	-	-	-
SM_Age_range	0,052	Gestor de loja	-	-	-	-

Tabela 3.4 – Importância das variáveis e valores que assumem por cluster.

Variável	Ordem de importância	Descrição
SM_Age	1	Idade do Gestor de Loja
ST_empl_avg_legal_sen	2	Senioridade média dos colaboradores por loja
SM_avg_years_function	3	Média de anos na função do gestor de loja
ST_Assimetria_Age_range_NEW	4	Assimetria entre as faixas etárias <45 e >=45 anos por loja
E_Sen_Legal_yrs	5	Senioridade do colaborador na organização
Engagement___fav_	6	Engagement
Job_Challenge___fav_	7	Job challenge
SM_Educational_Level	8	Nível de escolaridade do gestor de loja
ST_terminations	9	Rescisões por loja
ST_years	10	Anos por loja
ST_Avg_acc_lostdays	11	Média de dias perdidos por acidentes de trabalho por loja
SM_Marital_status	12	Estado civil do gestor de loja
E_Age	13	Idade do colaborador
ST_Per_Empl_acc_lostdays	14	% de colaboradores com dias perdidos por acidente por loja
E_Age_Range	15	Intervalo de idades do colaborador
SM_Age_range	16	Intervalo de idades do gestor de loja

Tabela 3.5 – Ordem de importância e descrição do conteúdo das variáveis.

Legenda para a dimensão de cada variável:
Gestor de loja
Loja
Colaborador

Com base na Tabela 3.4 é possível aferir que a idade do Gestor de loja é a variável com maior valor no critério worth, e por esse motivo tem o maior contributo na definição da ordem de importância das variáveis. Na Tabela 3.5, é apresentada a ordem de importância e também a descrição de cada uma das variáveis identificadas.

3.1.7.1. Extração de Regras para interpretação dos resultados obtidos

A construção de uma árvore de decisão tem como vantagem o facto de representar regras que podem facilmente ser entendidas por pessoas. Numa árvore de decisão as regras podem ser obtidas nas folhas, as quais representam os nós finais.

Quando uma árvore de decisão é utilizada na previsão de uma classificação, é uma vantagem possuir um elevado numero de folhas porque permite obter resultados com maior precisão, no entanto, quando o objectivo passa simplesmente por gerar regras é recomendável que estas sejam em menor número, dado que quanto menos regras existirem mais facil será compreender o problema (Berry & Linoff, 2004).

Com base nos valores que cada variável assumiu por cluster e no critério de worth foi possível identificar os critérios que as variáveis assumem de forma a melhor compreender as regras subjacentes à construção dos clusters obtidos. Desta forma pretende-se que as regras identificadas contribuam para a interpretação dos resultados obtidos para os 4 clusters (Anexo IV – Critérios para a extração de regras).

4. RESULTADOS E DISCUSSÃO

4.1. ANÁLISE DAS CARACTERÍSTICAS DOS CLUSTERS

A área HR Analytics é hoje em dia muito importante na gestão de Recursos Humanos e por isso contribui para o sucesso de uma organização na medida em que permite um retorno do investimento da organização.

A Gestão na Liderança assume um papel fundamental dado que é através dos gestores, também designados de líderes, que as práticas de Gestão de Recursos Humanos são passadas aos colaboradores. Por esse motivo existe uma relação entre os comportamentos adotados pelos gestores/estilos de liderança e o clima organizacional.

Através dos resultados obtidos neste estudo em particular foi possível perceber que o Gestor de loja assume uma elevada importância na caracterização do perfil dos colaboradores em grupos, dado que as variáveis que caracterizam o gestor de loja têm um poder discriminatório mais elevado. Por esse motivo, a temática da gestão na liderança, muito em voga nos dias de hoje, deve ser também considerada um fator importante a considerar pela organização. Por ser considerado um fator importante para a organização em estudo, no subcapítulo 4.1.1 é abordada a temática do clima organizacional e a gestão na liderança tendo por base a literatura considerada relevante.

4.1.1. O Clima Organizacional e a Gestão na Liderança

Segundo (Culture and engagement, 2015) o envolvimento dos colaboradores é considerado um fator de elevada importância nas organizações. As organizações que têm por base uma cultura marcada pelo trabalho, envolvimento dos colaboradores, trabalho apto à organização e fortes competências de liderança conseguem superar as organizações concorrentes e têm maior capacidade para atrair melhores talentos. O envolvimento dos colaboradores é, hoje em dia, considerado um fator de grande relevância; (Crabtree, 2013) defende que o envolvimento dos colaboradores é baixo, e demonstra no seu estudo que apenas cerca de 13% dos colaboradores estão altamente envolvidos.

Devido a estes resultados é cada vez mais importante perceber o que está a acontecer nas organizações, como se caracterizam as equipas, analisar os resultados obtidos e também o comportamento dos líderes com o objetivo de melhorar o clima organizacional e os resultados atingidos por uma organização.

(Bersin, Geller, Wakefield, & Walsh, 2015) no artigo “Culture and engagement: The naked organization” identifica quatro pontos chave para explicar o baixo envolvimento dos colaboradores numa organização:

- O mercado de trabalho é hoje em dia muito dinâmico, em parte devido à contribuição das redes sociais Linkedin, Facebook, entre outras, pelo que a probabilidade de rotatividade nas organizações aumenta quando existem colaboradores insatisfeitos com o ambiente na organização;

- A falta de conhecimento dos líderes para entenderem que a cultura da organização começa neles próprios, isto é, de cima para baixo em termos de responsabilidade.

- As alterações que têm sido registadas ao longo dos últimos anos mudaram também a forma de garantir o envolvimento dos colaboradores. A flexibilidade, responsabilização, desenvolvimento e mobilidade total definem a cultura organizacional.

- As motivações dos colaboradores têm registado mudanças e os colaboradores têm hoje em dia novos objetivos para a sua carreira profissional (Deloitte, 2015). Assiste-se atualmente ao aumento da importância na paixão pelo trabalho e à diminuição da ambição, segundo (John Hagel, 2014) 12% e 5% respetivamente, o que reforça a necessidade de os líderes assumirem um papel de maior relevância de forma a garantirem um ambiente de trabalho envolvente para a sua equipa

A cultura de uma organização e o envolvimento dos colaboradores tem vindo a aumentar a sua importância ao longo dos últimos anos. O caso de sucesso a que se tem vindo a assistir na Google, caracterizada como uma das melhores organizações para trabalhar, é um dos exemplos que comprovam a importância da cultura numa organização. Acredita-se que os bons resultados são consequência da favorável cultura organizacional que se vive.

O foco para uma gestão com resultados tem por base o envolvimento, capacidade para atingir resultados e ambiente organizacional. O facto de hoje em dia os sistemas de informação permitirem às organizações o acesso a informação em tempo real facilita o processo de gestão por parte dos líderes, na medida em que quando existem problemas, salvo exceções, estes são identificados de forma mais rápida do que seriam no passado. (Bersin, 2014) propõe um modelo que explica as mudanças que se têm registado na gestão de pessoas e carreiras profissionais desde os anos 90 até aos dias de hoje.

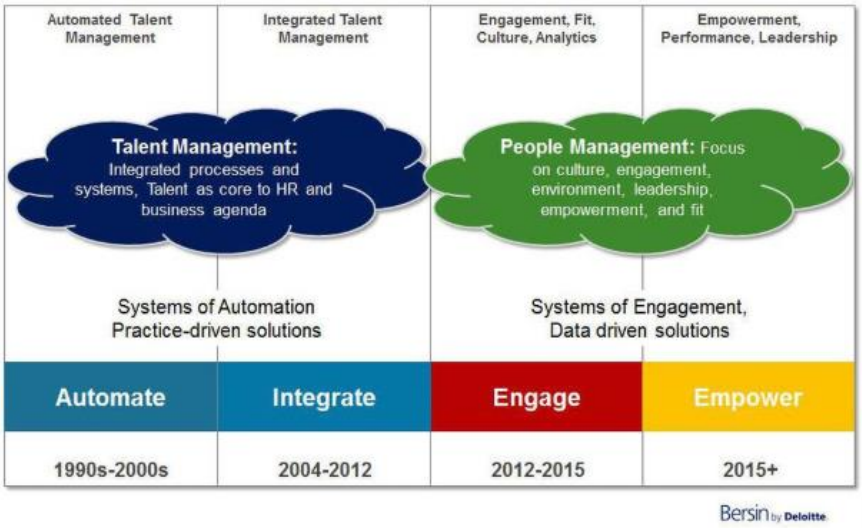


Figura 4.1 – Mudanças na gestão de pessoas e carreiras profissionais (Bersin, 2014).

Na organização em estudo, a caracterização dos gestores de loja, em conjunto com outras variáveis do colaborador, permite compreender os diferentes perfis de colaboradores (tendo em conta as variáveis consideradas para o propósito referido anteriormente).

As características pessoais centram um conjunto variado de competências individuais e relacionais sobre o individuo, e definem a forma como a pessoa se relaciona com o meio envolvente. De acordo

com (Hamidah Jantan, Abdul Razak Hamdan, 2009) as características pessoais incluem a liderança, habilidade para organizar, disciplina, proatividade e inovação. Além destas, também *skills* de gestão e o conhecimento e expertise contribuem para a definição dos fatores de competências individuais.

Na organização em estudo, sugere-se a realização de análises mais aprofundadas para perceber as razões subjacentes a estes comportamentos, seja através da observação em loja ou através de estudos qualitativos/quantitativos. Verifica-se que o perfil do gestor de loja bem como o perfil do colaborador são características de elevada importância no padrão de comportamento dos clusters. Os gestores de loja dividem-se em gestores de loja jovens e gestores de loja mais experientes.

Segundo o estudo publicado pela (Portugal, 2018), os gestores reconhecem que a interligação entre as competências humanas e a tecnologia digital avançada permitirá melhorar os resultados de uma organização. A organização referida no estudo acima, identifica as principais tendências na gestão de talento em Portugal para o ano 2018:

- **Velocidade na mudança**. Inclui formação, aposta em estruturas de gestão mais horizontais, descentralização de processos de tomada de decisão, eliminação de funções de menor impacto e criação de equipas de projeto. Cerca de 94% dos gestores em Portugal refere a inovação como o ponto chave para 2018;

- **Trabalhar com um propósito**. De entre as razões apontadas para o sucesso de um colaborador destacam-se: uma remuneração justa e competitiva, oportunidades de desenvolvimento de carreira, líderes que definem uma direção clara para a sua equipa e organização, possibilidade de trabalhar com pessoas de topo, oportunidades de formação e trabalho em projetos com um propósito bem definido. Uma evidência do estudo é que colaboradores com incentivos de carreira demostram quatro vezes mais compromisso para com a organização.

- **Flexibilidade permanente.** Foi destacada a necessidade de desenvolver líderes para o futuro. Um dos fatores apontados como sendo um dos que tem mais impacto nas empresas em 2018 é a valorização e reforço da experiência do colaborador.

- **Digital de fora para dentro.** Tem-se verificado o aumento da importância dada às ferramentas digitais na atividade profissional. Pretende-se que seja feito investimento na gestão de conhecimento, na melhoria da eficácia de vendas e também na eficiência das equipas.

Relativamente ao desafio que a temática do Ageing Workforce impõe atualmente nas organizações, é importante que estas adotem estratégias de gestão de recursos humanos com ênfase na questão da idade. De acordo com (Čiutienė & Railaitė, 2014), a temática do Ageing Workforce apresenta-se como um desafio para as organizações e por isso estas devem organizar-se para assegurar uma gestão de recursos humanos alinhada com as necessidades e objetivos estratégicos definidos. No mesmo estudo, os autores identificam os fatores que consideram importantes para uma gestão eficaz:

- **Condições de trabalho e respetiva melhoria das mesmas.** É identificado como um fator que encoraja os colaboradores a sentirem-se bem no seu local de trabalho durante mais tempo;

- **Competências dos colaboradores.** As organizações devem dar importância às competências dos colaboradores sem se focarem apenas no aspeto da idade.

- **Transferência de conhecimento.** Deve ser assegurada a transferência de conhecimento entre gerações, dado que é frequente os colaboradores mais novos apresentarem lacunas relativamente às competências práticas e experiência adquirida.

Num estudo mais recente, apresentado por (Hirsch, 2017) são identificados como fatores chave para superar os problemas resultantes da temática do Ageing Workforce:

- **Encorajar os colaboradores mais experientes a ficar na organização.** A saída dos colaboradores deve ser planeada de forma a evitar saídas repentinas e em grande volume;

- **Desenvolver uma cultura de transferência de conhecimento.** Os colaboradores mais experientes devem passar o seu conhecimento às gerações mais novas antes de saírem da organização, o que deve ser planeado através de programas de gestão desenhados para o efeito;

- **Investir no desenvolvimento da carreira dos colaboradores.** As organizações que investem no desenvolvimento dos seus colaboradores têm mais hipóteses de atrair e manter os colaboradores considerados "top talent". A rotatividade é vista pelos autores como resultado de uma falha ao nível do desenvolvimento das oportunidades de carreira;

- **Dar valor aos colaboradores mais novos (Geração Millenials).** Com o aumento da representatividade desta geração é importante que as organizações tenham conhecimento das prioridades que estes valorizam como, equilíbrio entre a vida profissional e pessoal, oportunidades de carreira, horários flexíveis, objetivos bem definidos e/ou programas de formação.

De acordo com (Weber, 2010) é imperativo manter as organizações competitivas e alinhadas com os seus objetivos estratégicos. Considera-se importante melhorar e/ou manter a satisfação dos colaboradores, reter talentos e evitar uma elevada taxa de rotatividade, o que pode ser assegurado através da adoção de práticas de compensação e benefícios adicionais. É também importante assegurar o planeamento e desenvolvimento de carreiras para evitar que os colaboradores se sintam pouco envolvidos e procurem outras soluções.

5. CONCLUSÕES

O trabalho desenvolvido descreve o enquadramento do Ageing Workforce nas organizações. De forma a perceber a importância do conceito na função de Recursos Humanos foi necessário abordar alguns temas inerentes ao Ageing Workforce, tais como o People Analytics e a análise de dados.

De uma forma geral o tema apresentado neste trabalho de projeto é importante na Gestão de Recursos Humanos, no entanto, assume uma importância ainda maior quando se trata de organizações com funções que exigem um desgaste físico dos colaboradores.

A identificação atempada dos motivos que levam ao desgaste, redução de produtividade ou acidentes de trabalho, diminuição do envolvimento e satisfação dos colaboradores é um objetivo da organização a longo prazo, pelo que se torna imperativo aumentar o conhecimento sobre os seus colaboradores. Para isso, no trabalho apresentado recorreu-se à análise de clusters para fazer a exploração dos dados e uma análise descritiva dos colaboradores.

A obtenção de um modelo preditivo e de uma análise descritiva dos colaboradores mais detalhada foi condicionada pela disponibilidade dos dados, o que afetou a diversidade da caracterização dos grupos de colaboradores.

Em termos práticos, os resultados obtidos foram contextualizados com bibliografia que reforça a importância que a Gestão na Liderança assume atualmente. Considera-se que este é um fator chave a considerar pelas organizações no futuro de forma a manter os seus colaboradores no ativo satisfeitos com as suas funções e sem quebras de produtividade. Também a temática do Ageing Workforce foi contextualizada com a bibliografia que se considerou relevante, tendo sido identificados os fatores chave a considerar por uma organização, de uma forma geral, para ultrapassar os problemas daí resultantes.

6. LIMITAÇÕES E POSSÍVEIS TRABALHOS FUTUROS

A utilização de uma base de dados com poucas observações e referente a um período amostral curto revelou-se uma limitação na medida em que limitou as técnicas de Data Mining a utilizar.

Desta forma, a eficácia em alcançar os objetivos propostos foi condicionada pela qualidade dos dados disponíveis para análise, tanto no número de variáveis como no número de amostras disponíveis. Seria desejável ter disponível um conjunto de dados maior, com mais variáveis que caracterizassem os colaboradores e referente a um período amostral maior do que seis meses. O período amostral de seis meses demonstrou ser também uma forte limitação. O número de observações para os colaboradores que registam acidentes de trabalho revelou-se pequeno para caracterizar o desgaste dos colaboradores. Além disso também as variáveis que identificam as competências dos colaboradores revelaram-se uma limitação na medida em que possuem poucas observações.

Uma variável relevante para quantificar a efetividade do negócio é o valor das vendas para cada uma das lojas incluídas no estudo, no entanto, não foi possível obter atempadamente os valores de forma a incluí-los na análise.

De forma geral, considera-se importante no futuro melhorar o conjunto de dados dos colaboradores para realizar uma análise com maior profundidade.

A utilização do People Analytics como uma ferramenta auxiliar para a Gestão de Recursos Humanos poderá permitir de uma forma eficaz definir estratégias que melhorem as práticas de gestão, aumentem o envolvimento dos colaboradores e reforcem um planeamento estratégico para a alocação do trabalho.

Por exemplo, poderá ser desenvolvida uma modelação preditiva para identificar grupos de colaboradores mais suscetíveis a ter faltas devido a acidentes de trabalho, com o objetivo de definir estratégias e praticas preventivas.

7. BIBLIOGRAFIA

Aitken, M., Hedge, J., Ball, K., Cabrera, A., Hinkle-Bowles, P., McFarland, B., … Sweet, S. (2014). *Exutive Roundtable on the Aging Workforce*. Retrieved from www.iwh.on.ca

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: why HR is set to fail the big data challenge. *Human Resource Management Journal*, *26*(1), 1–11. https://doi.org/10.1111/1748-8583.12090

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January), 182–185. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Bação, F. L. (n.d.). Unidade de Aprendizagem - Modulo 4 e 6 - Análise de Clusters/Árvores de Decisão. In *Ciência da Informação*.

Beck, V. (2008). *Older Workers – Older Learners: The Perspectives of Employers in the East Midlands*. Retrieved from https://lra.le.ac.uk/handle/2381/36621

Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques for Marketing, Sales, and Customer Relationship Management*. *Wiley Publishing, Inc.* (2nd Editio). Wiley Publishing, Inc.

Bersin, J. (2014). Why The Talent Management Software Market Will Radically Change. *Forbes*. Retrieved from https://www.forbes.com/sites/joshbersin/2014/12/29/how-and-why-the-talent-management-market-is-changing/#3b7b897e4d20

Bersin, J. (2016). People Analytics Market Growth: Ten Things You Need to Know. Retrieved from http://joshbersin.com/2016/07/people-analytics-market-growth-ten-things-you-need-to-know/

Bersin, J. (2017). People Analytics: Here With A Vengeance. Retrieved from https://joshbersin.com/2017/12/people-analytics-here-with-a-vengeance/

Bersin, J., Geller, J., Wakefield, N., & Walsh, B. (2015). Global human capital trends 2015. *Deloitte University Press*, 112. https://doi.org/http://www2.deloitte.com/us/en/pages/human-capital/articles/employee-engagement-culture-human-capital-trends-2015.html

Boenzi, F., Digiesi, S., Mossa, G., Mummolo, G., & Romano, V. A. (2015). Modelling workforce aging in job rotation problems. *IFAC-PapersOnLine*, *28*(3), 604–609. https://doi.org/10.1016/j.ifacol.2015.06.148

Boston Consulting Group. (2014). *Creating People Advantage 2014-2015*. https://doi.org/Acedemico/material didatico/bibliografia 2105

Brooke, L. (2003). Human resource costs and benefits of maintaining a mature-age workforce. *International Journal of Manpower*, *24*(3), 260–283. https://doi.org/https://doi.org/10.1108/01437720310479732

Cabral, M. V., & Ferreira, P. M. (2014). *O ENVELHECIMENTO ACTIVO EM PORTUGAL - Trabalho, Reforma, Lazer e Redes Sociais*. Fundação Francisco Manuel dos Santos.

Carla Arellano, Alexander DiLeonardo, and I. F. (2017). Using people analytics to drive business performance: A case study. *McKinsey Quarterly*. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/using-people-analytics-to-drive-business-performance-a-case-study

Čiutienė, R., & Railaitė, R. (2014). Challenges of Managing an Ageing Workforce. *Procedia Social and Behaviour Sciences Journal, Presented on 19th International Scientific Conference «Economics and Management – 2014» (ICEM-2014)*, *156*(April), 69–73. https://doi.org/10.1016/j.sbspro.2014.11.121

Crabtree, S. (2013). *Worldwide, 13% of employees are engaged at work*. Retrieved from https://news.gallup.com/poll/165269/worldwide-employees-engaged-work.aspx

Deloitte. (2015). *Business needs to reset its purpose to attract Millennials*. Retrieved from https://www2.deloitte.com/be/en/pages/about-deloitte/articles/fourth-annual-millennial-survey.html

Donald M. Truxillo, David M. Cadiz, J. R. R. (2012). Designing jobs for an Aging Workforce: An Opportunity for Occupational Health. In *Contemporary Occupational Health Psychology* (pp. 109–123). John Wiley & Sons, Inc. Retrieved from https://books.google.pt/books?hl=pt-PT&lr=&id=1soqAIeiQBIC&oi=fnd&pg=PA109&dq=people+analytics+applied+to+ageing+workforce&ots=6KpqTLrss7&sig=_ggHmRh-MWEyqK7oI5CgPedKBlU&redir_esc=y#v=onepage&q=people analytics applied to ageing workforce&f=false

Dulebohn, J. H., & Johnson, R. D. (2013). Human resource metrics and decision support: A classification framework. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2012.06.005

Eurostat. (2017). Population structure and ageing. Retrieved February 2, 2018, from http://ec.europa.eu/eurostat/statistics-explained/index.php/Population_structure_and_ageing

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. The MIT Press.

Fitz-enz, J., & John R. Mattox II. (2014). *Predictive Analytics for Human Resources*. John Wiley & Sons, Inc.

Hamidah Jantan, Abdul Razak Hamdan, Z. A. O. (2009). Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application. Retrieved from https://waset.org/publication/Knowledge-Discovery-Techniques-for-Talent-Forecasting-in-Human-Resource-Application/11782

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Elsevier* (2nd Editio, Vol. 12). Elsevier Inc. Retrieved from http://link.springer.com/10.1007/978-3-642-19721-5

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician*, *52*(2), 112–118. https://doi.org/10.1080/00031305.1998.10480549

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. *Building*. The MIT Press. Retrieved from https://doc.lagout.org/Others/Data Mining/Principles of Data Mining %5BHand%2C Mannila %26 Smyth 2001-08-01%5D.pdf

Hirsch, A. S. (2017). 4 Ways for HR to Overcome Aging Workforce Issues. Retrieved from https://www.shrm.org/resourcesandtools/hr-topics/behavioral-competencies/global-and-cultural-effectiveness/pages/4-ways-for-hr-to-overcome-aging-workforce-issues.aspx

Ilmarinen, J. (2012). *Promoting active ageing in the workplace*. Retrieved from http://www.ipbscordoba.es/uploads/Documentos/promoting-active-ageing-in-the-workplace.pdf

INE, P. (2017). População residente: total e por grupo etário. Retrieved from https://www.pordata.pt/Portugal/População+residente+total+e+por+grupo+etário-10

Jantan, H., Razak Hamdan, A., & Ali Othman, Z. (2010). Human Talent Prediction in HRM using C4.5 Classification Algorithm. *International Journal on Computer Science and Engineering*, *02*(08), 2526–2534. https://doi.org/10.4018/jtd.2010100103

John Hagel. (2014). Passion versus ambition - Did Steve Jobs have worker passion? *Deloitte University Press*. Retrieved from https://www2.deloitte.com/insights/us/en/topics/employee-engagement/employee-passion-ambition.html

Jurney, R. (2013). *Agile Data Science*. O'Reilly Media. Retrieved from http://shop.oreilly.com/product/0636920025054.do

Laurence Collins, David R. Fineman, A. T. (2017). People analytics: Recalculating the route. *Deloitte*. Retrieved from https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2017/people-analytics-in-hr.html

Michael J. Kavanagh, Mohan Thite, R. D. J. (2011). *Human Resource Information Systems: Basics, Applications, and Future Directions* (2nd Editio). SAGE Publications, Inc.

Momin, W. Y. M., & Mishra, K. (2015). HR Analytics as a Strategic Workforce Planning. *International Journal of Applied Research*, *1*(4), 258–260.

Mortenson, M. J. ., Doherty, N. F., & Robinson, S. (2015). Operational research from Taylorism to Terabytes: A research agenda for the analytic sage. *European Journal of Operational Research*, *241*(3), 583–595. https://doi.org/doi:10.1016/j.ejor.2014.08.029

Portugal, M. (2018). *Global Talent Trends 2018 study - Unlocking Growth in the Human Age*. Retrieved from https://www.mercer.pt/our-thinking/career/global-talent-trends-portugal-2018.html

Santos, M. Y., & Ramos, I. (2009). *Business Intelligence - Tecnologias da Informação na Gestão do Conhecimento* (2ª Edição). FCA.

Sothmann, A., & Mehta, S. (2017). Workforce Analytics: The Gap between Rhetoric and Experience. Mercer.

Sullivan, J. (2013). How Google Is Using People Analytics to Completely Reinvent HR. *HR Management, HR News & Trends*. Retrieved from http://docshare01.docshare.tips/files/28758/287584559.pdf

Thomas H. Davenport, J. H., & Shapiro, J. (2009). *Competing on Talent Analytics*. Retrieved from https://hbr.org/2010/10/competing-on-talent-analytics

Union, E. (2012). *Active ageing and solidarity between generations* (2012 Editi). https://doi.org/10.2785/17758

Watson, H. J. (2013). All about Analytics. *International Journal of Business Intelligence Research*, pp.13-28.

Watson, H. J. (2014). Tutorial : Big Data Analytics : Concepts , Technologies , and Applications. *Communications of the Association for Information Systems*, *34*(June), 1246–1269. Retrieved from http://aisel.aisnet.org/cais/vol34/iss1/65

Weber, A. V. M. L. (2010). Práticas de Remuneração como Estratégia para Retenção de Talentos: Um Estudo de caso em uma Empresa de Serviços. *1º Simpósio Brasileiro de Ciência e Serviços*.

8. ANEXOS

Anexo I - Tabela com as estatísticas descritivas das variáveis intervalares.

Anexo II - Matriz de correlações entre variáveis obtida através do SAS Enterprise Miner.

Anexo III - Gráfico representativo da expectativa de evolução da população entre 2016 e 2080.

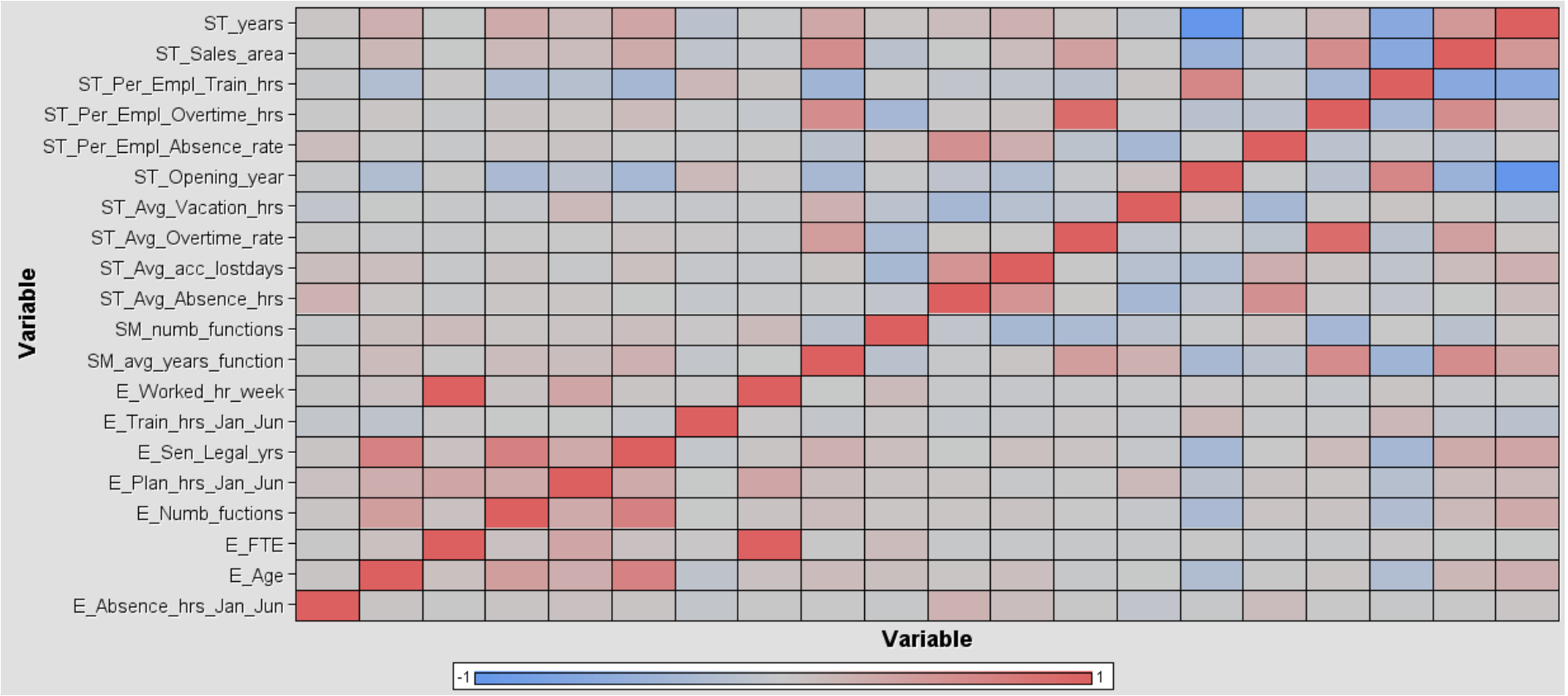
Anexo IV - Critérios para a extração de regras.

Anexo V - Diagrama criado na aplicação SAS Enterprise Miner 14.2 para a modelação descritiva.

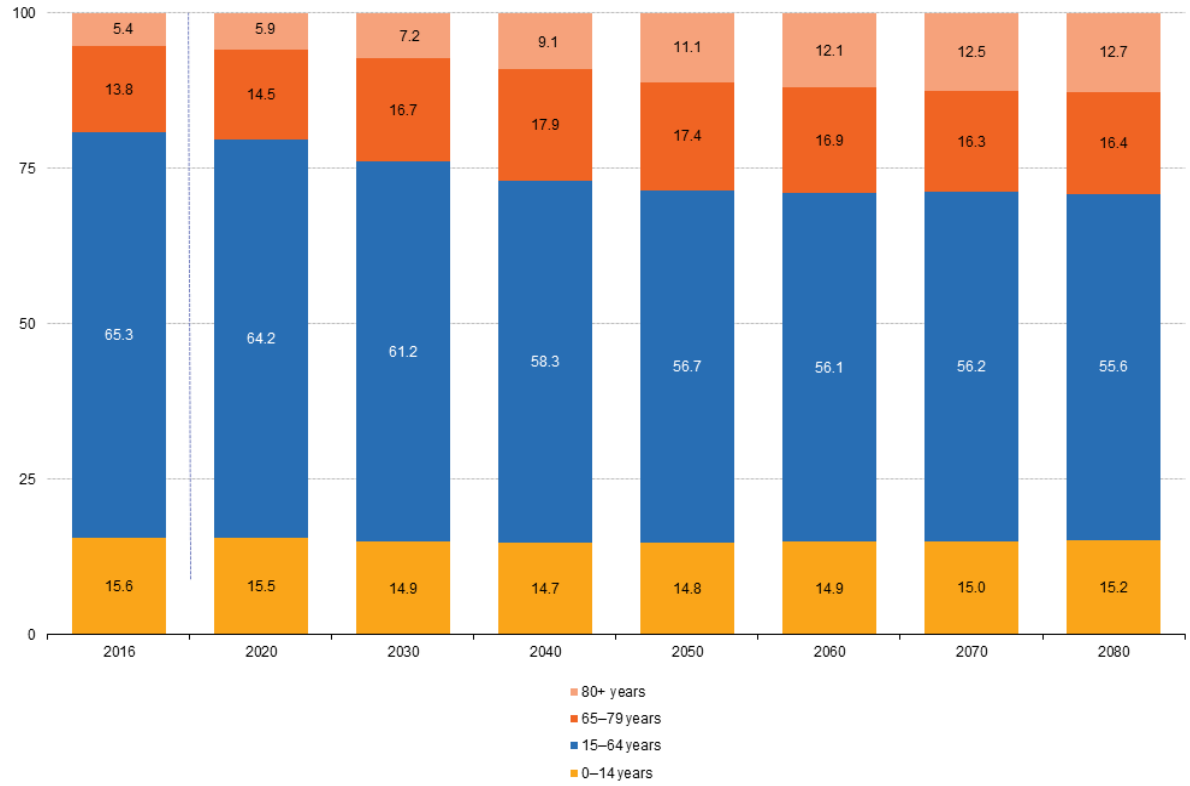
Anexo I – Tabela com as estatísticas descritivas das variáveis intervalares

Nome da variável	Mediana	Nº observ. em falta	Nº observ.	Valor Min.	Valor Max.	Média	StdDev	Skewness	Kurtosis	CV
E_Absence_hrs_Jan_Jun	0	0	1800	0	1032	54,04	169,61	4,27	18,87	3,14
E_Absence_rate_Jan_Jun	0	2	1798	0	1	0,05	0,17	4,25	18,63	3,13
E_Accid_lost_days_Jan_Jun	0	0	1800	0	119	0,49	4,88	16,50	326,21	10,06
E_Age	41	0	1800	18	68	40,72	9,40	-0,05	-0,48	0,23
E_Avg_yrs_function	3,03	0	1800	0,01	17,12	3,18	2,07	1,11	2,84	0,65
E_Comp_1	2,63	319	1481	1,13	4,38	2,67	0,45	0,44	0,21	0,17
E_Comp_3	2,67	320	1480	1	4	2,60	0,50	0,25	-0,43	0,19
E_FTE	1	0	1800	0,2	1	1,00	0,04	-12,11	154,33	0,04
E_N_Children	1	0	1800	0	6	1,17	0,96	0,48	0,22	0,82
E_Numb_fuctions	4	0	1800	1	13	3,96	2,28	0,50	-0,32	0,57
E_Overtime_hrs_Jan_Jun	0	0	1800	0	78,24	0,93	3,92	11,13	180,48	4,21
E_Overtime_rate_Jan_Jun	0	2	1798	0	0,08	0,00	0,00	10,19	147,84	4,20
E_Plan_hrs_Jan_Jun	1009	0	1800	0	1050	964,45	169,26	-4,05	16,12	0,18
E_Sen_Legal_yrs	13	0	1800	0	52	12,95	9,82	0,35	-0,63	0,76
E_Train_hrs_Jan_Jun	0	0	1800	0	241,5	3,49	13,85	11,78	168,52	3,97
E_Vacation_hrs_Jan_Jun	80	0	1800	0	304	79,14	48,75	0,39	0,21	0,62
E_Worked_hr_week	40	0	1800	8	40	39,82	1,76	-11,83	149,33	0,04
SM_Age	44	14	1786	31	59	45,00	7,83	0,08	-1,28	0,17
SM_avg_years_function	2,82	14	1786	0,61	6,75	3,23	1,79	0,65	-0,87	0,55
SM_N_Filhos	2,00	14	1786	0	3	1,53	0,96	-0,25	-0,91	0,62
SM_numb_functions	6	14	1786	2	10	6,37	1,84	0,09	-0,18	0,29
SM_Sen_Legal_yrs	19	14	1786	0	39	20,27	9,47	-0,19	-1,00	0,47
ST_Avg_Absence_hrs	53,90	0	1800	1,69	183,32	54,04	36,41	1,66	3,48	0,67
ST_Avg_Absence_rate	0,04	0	1800	0,00	0,19	0,06	0,04	1,70	2,94	0,72
ST_Avg_acc_lostdays	7,5	0	1800	0	83	11,64	16,48	2,76	8,93	1,42
ST_Avg_Overtime_hrs	0,05	0	1800	0	8,825833	0,93	1,41	2,63	10,69	1,52
ST_Avg_Overtime_rate	0,00	0	1800	0	0,017471	0,00	0,00	2,57	10,13	1,52
ST_Avg_Train_hrs	2,90	0	1800	0,68	18,16	3,49	2,88	2,88	11,17	0,82
ST_Avg_Vacation_hrs	79,15	0	1800	30,17	116,51	79,14	15,36	0,04	1,52	0,19
ST_empl_avg_legal_sen	13,84	0	1800	2,36	21,88	12,99	4,68	-0,21	-0,83	0,36
ST_Per_Empl_Absence_hrs	0,33	0	1800	0,10	0,79	0,32	0,12	0,41	0,24	0,38
ST_Per_Empl_Absence_rate	0,22	0	1800	0,06	0,71	0,22	0,11	0,97	2,44	0,48
ST_Per_Empl_acc_lostdays	0,02	0	1800	0	0,142857	0,03	0,03	1,12	1,37	1,03
ST_Per_Empl_Overtime_hrs	0,03	0	1800	0	0,75	0,16	0,22	1,23	-0,09	1,44
ST_Per_Empl_Overtime_rate	0,03	0	1800	0	0,75	0,16	0,22	1,23	-0,09	1,44
ST_Per_Empl_Train_hrs	0,37	0	1800	0,16	1	0,42	0,19	1,01	1,17	0,45
ST_Per_Empl_Vacation_hrs	0,91	0	1800	0,49	1	0,89	0,09	-2,13	6,95	0,10
ST_terminations	0,03	0	1800	0	0,21	0,05	0,05	1,15	0,73	0,90

Anexo II – Matriz de correlações entre variáveis obtida através do SAS Enterprise Miner



Anexo III – Gráfico representativo da expectativa de evolução da população entre 2016 e 2080



Note: 2016: estimate, provosional. 2020–80: projections (EUROPOP2015).
Source: Eurostat (online data codes: demo_pjangroup and proj_15ndbims)

(Eurostat, 2017)

Anexo IV – Critérios para a extração de regras

Nome do Cluster	Idade do Gestor de Loja (anos)	Senioridade média dos colaboradores por loja (anos)	Média de anos na função do gestor de loja	Assimetria entre as faixas etárias < 45 e >=45 anos por loja	Senioridade do colaborador na organização (anos)	Engagement	Job_Challenge	Nível de escolaridade do gestor de loja
Gestores de loja mais novos que estão em lojas mais recentes	[31-45]	[4,947-14,8]	[0,605-3,518]	[0,159-0,947]	[0-28]	[61-89]	[47-64]	-
Gestores de loja mais novos (do que o segmento 1) e solteiros	[31-45]	[5-15,08]	[0,605-3,518]	[0,159-0,843]	-	[61-89]	[47-66]	[Degree] - 80%
Gestores de loja mais experientes, colaboradores mais novos e boa avaliação para o engagement e job challenge	[34-59]	[9,882-21,878]	[1,634-6,745]	[0-0,647]	-	[70-97]	[54-81]	-
Gestores de loja mais experientes, colaboradores com maior idade e boa avaliação para o engagement e job challenge	[42-59]	[2,075-6,745]	[9,882-21,878]	[0-45]	[0-0,647]	[70-97]	[54-81]	-

Nome do Cluster	Rescisões por loja	Anos por loja	Média de dias perdidos por acidentes de trabalho por loja	Estado civil do gestor de loja	Idade do colaborador	% de colaboradores com dias perdidos por acidente por loja	Intervalo de idades do colaborador	Intervalo de idades do gestor de loja
Gestores de loja mais novos que estão em lojas mais recentes	[0-0,214]	[7,0-24]	-	-	-	[0-0,142]	-	-
Gestores de loja mais novos (do que o segmento 1) e solteiros	-	-	[0-83]	[Single] - 57%	-	[0-0,142]	-	-
Gestores de loja mais experientes, colaboradores mais novos e boa avaliação para o engagement e job challenge	[0-0,134]	[9,0-33]	-	-	[19-51]	-	-	[45-54 anos] - 42%
Gestores de loja mais experientes, colaboradores com maior idade e boa avaliação para o engagement e job challenge	[0-0,134]	-	-	-	-	-	[45-54 anos] - 51%	-

Legenda para a dimensão de cada variável:

Gestor de loja
Loja
Colaborador

Anexo V – Diagrama criado na aplicação SAS Enterprise Miner 14.2 para a modelação descritiva