



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Previsão das saídas voluntárias de colaboradores: Aplicação a um Grupo Económico Português

Sofia Luís Rolim Poço

Mestrado em Business Analytics

Orientador:

Professor Doutor Raul Manuel da Silva Laureano, Professor Associado, Iscte – Instituto Universitário de Lisboa

outubro, 2023



BUSINESS
SCHOOL

Departamento de Métodos Quantitativos para Gestão e
Economia

**Previsão das saídas voluntárias de colaboradores: Aplicação
a um Grupo Económico português**

Sofia Luís Rolim Poço

Mestrado em Business Analytics

Orientador:

Professor Doutor Raul Manuel da Silva Laureano, Professor
Associado, Iscte – Instituto Universitário de Lisboa

outubro, 2023

Agradecimentos

A presente investigação representa a finalização de mais um ciclo do meu percurso académico e esta não se teria realizado sem o apoio de diversas pessoas. Desta forma, quero agradecer a todos os que de forma direta ou indireta contribuíram para a sua elaboração, nomeadamente:

Ao professor Doutor Raul Laureano, que me acompanha desde o início do Mestrado em Business Analytics, por toda a ajuda, orientação, compreensão, partilha de conhecimentos e disponibilidade.

Aos profissionais da área de Recursos Humanos do Grupo pela ajuda, disponibilidade e incentivo, crucial para o desenvolvimento e término da investigação.

À minha família, especialmente aos meus pais e irmãs, por toda a motivação e apoio incondicional ao longo do meu percurso académico.

Ao Rúben por toda a paciência, compreensão, apoio nos momentos mais difíceis e pela ajuda necessária para os ultrapassar.

Aos meus amigos mais chegados pelo companheirismo, pela ajuda e motivação, essencial para concluir esta etapa.

A todos, o meu sincero obrigada.

Resumo

A existência de elevadas taxas de *turnover* nas empresas é um problema capaz de pôr em risco a continuidade destas no mercado, dada a sua associação a mais despesas, perda de conhecimento e de produtividade. No entanto, é sobre o *turnover* voluntário que as empresas devem prestar maior atenção, visto que acarreta elevados custos e a decisão está do lado dos colaboradores.

Assim, importa perceber como é que as empresas podem conseguir antecipar a saída voluntária dos seus colaboradores e, por essa razão, este estudo avalia a relação entre os dados armazenados pelos empregadores, sobre os seus colaboradores, e a decisão, dos colaboradores, em abandonarem os seus empregadores.

A presente investigação é sustentada por uma revisão sistemática da literatura, que inclui 24 artigos, e, seguindo a metodologia CRISP-DM, pela aplicação de técnicas de *machine learning*, nomeadamente árvores de decisão, comparando diferentes algoritmos. Os modelos desenvolvidos foram aplicados a um conjunto de dados extraídos de um *software* de gestão de cadastro de colaboradores, que inclui características pessoais e profissionais destes. No total foram utilizadas mais de 60 variáveis e uma amostra superior a 31000 colaboradores.

Os resultados permitem concluir que os dados armazenados pelos departamentos de recursos humanos são capazes de prever o *turnover* voluntário com uma precisão de 80,2%. Desta forma, o principal contributo deste estudo é ajudar o tecido empresarial, nomeadamente os profissionais de Recursos Humanos, a identificar e prevenir o *turnover* voluntário dos seus colaboradores, de forma mais eficaz.

Palavras-Chave: *Turnover* voluntário, Recursos Humanos, CRISP-DM, *Analytics*, Árvores de decisão

Classificação JEL: C67 e M54

Abstract

The existence of high turnover rates in companies is a problem that can jeopardize their continuity in the market, given its association with more expenses, loss of knowledge and productivity. However, it is voluntary turnover that companies need to pay more attention to, since it entails high costs and the decision lies with the employees.

It is therefore important to understand how companies can anticipate the voluntary departure of their employees and, for this reason, this study evaluates the relationship between the data stored by employers about their employees and the employees' decision to leave their employers.

This research is supported by a systematic literature review, which includes 24 articles, and, following the CRISP-DM methodology, by the application of machine learning techniques, namely decision trees, comparing different algorithms. The models developed were applied to a set of data extracted from employee registration management software, which includes their personal and professional characteristics. In total, more than 60 variables and a sample of more than 31000 employees were used.

The results allow us to conclude that the data stored by human resources departments can predict voluntary turnover with an accuracy of 80.2%. Thus, the main contribution of this study is to assist the business community, especially HR professionals, in identifying and preventing voluntary turnover among their employees more effectively.

Keywords: Voluntary turnover, Human Resources, CRISP-DM, Analytics, Decision Trees

JEL Classification: C67 and M54

Índice

Agradecimentos	i
Resumo	iii
Abstract	v
Índice	vii
Índice de figuras	ix
Índices de tabelas	xi
Abreviaturas e siglas.....	xiii
Capítulo 1 - Introdução.....	1
1.1. Contextualização do problema.....	1
1.2. Questão de investigação e objetivos.....	2
1.3. Abordagem metodológica	3
1.4. Estrutura da dissertação.....	3
Capítulo 2 – Revisão da literatura	5
2.1. Protocolo da revisão sistemática da literatura	5
2.1.1. Objetivos e questões de pesquisa	5
2.1.2. Processo de seleção da literatura	6
2.1.3. Processo de revisão de literatura	9
2.1.4. Avaliação da qualidade da literatura da RSL	9
2.2. Síntese do conteúdo da literatura.....	10
2.2.1. Contexto dos estudos.....	10
2.2.2. Metodologias aplicadas.....	13
2.2.3. Avaliação dos resultados	18
2.2.4. Contributos e implicações	21
2.3. Avaliação dos artigos.....	22
Capítulo 3 – Metodologia	25
3.1. Compreensão do negócio.....	26
3.2. Compreensão dos dados	28
3.3. Preparação dos dados	31
3.4. Modelação e avaliação	35
3.5. Implementação	39
Capítulo 4 – Resultados.....	41
4.1. Caracterização do <i>turnover</i> voluntário.....	41
4.2. Modelos preditivos.....	43

4.2.1. Avaliação dos modelos e seleção do melhor	44
4.3. Discussão e implicações	49
Capítulo 5 – Conclusão	53
5.1. Contributos.....	54
5.2. Limitações e pistas futuras de investigação	55
Referências bibliográficas.....	57
Anexos	62

Índice de figuras

Figura 2.1. Processo de seleção da literatura.....	7
Figura 2.2. Nuvem de palavras do <i>abstract</i> dos artigos incluídos na RSL	8
Figura 3.1. Fases do CRISP-DM.....	25
Figura 4.1. Gráfico de ganhos dos melhores modelos por algoritmo	45
Figura 4.2. Curva ROC dos melhores modelos por algoritmo	45
Figura 4.3. Gráfico de importância do modelo J	47
Figura 4.4. Distribuição do <i>turnover</i> voluntario pelo tipo de contrato e anos de efetividade	47
Figura 4.5. Distribuição do <i>turnover</i> voluntário pela antiguidade e avaliação média	48

Índices de tabelas

Tabela 2.1. Critérios de exclusão de literatura.....	6
Tabela 2.2. Critérios de inclusão de literatura	7
Tabela 2.3. Artigos incluídos na RSL, anos 2023 e 2022.....	7
Tabela 2.4. Artigos incluídos na RSL, entre 2021 e 2018	8
Tabela 2.5. Critérios de avaliação da literatura da RSL	9
Tabela 2.6. Âmbito e objetivos dos artigos incluídos na RSL	13
Tabela 2.7. Caracterização da amostra e base de dados dos artigos incluídos na RSL.....	15
Tabela 2.8. Metodologia dos artigos incluídos na RSL	18
Tabela 2.9. Resultados dos artigos incluídos na RSL	20
Tabela 2.10. Contributos e limitações dos artigos incluídos na RSL	21
Tabela 2.11. Avaliação dos artigos incluídos na RSL	22
Tabela 2.12. Pontuação por critério de avaliação	23
Tabela 3.1. Evolução dos colaboradores ativos e saídas, de 2005 a 2023	27
Tabela 3.2. Dicionário de dados das variáveis sociodemográficas extraídas.....	29
Tabela 3.3. Dicionário de dados das variáveis profissionais extraídas.....	30
Tabela 3.4. Dicionário de dados das variáveis salariais extraídas	30
Tabela 3.5. Dicionário de dados das variáveis sociodemográficas e profissionais para modelação	32
Tabela 3.6. Dicionário de dados das variáveis salariais para modelação.....	33
Tabela 3.7. Caracterização da variável <i>target</i>	34
Tabela 3.8. Parametrizações dos melhores modelos de cada algoritmo.....	37
Tabela 3.9. Matriz de confusão	38
Tabela 3.10. Métricas de avaliação usuais para modelos de classificação	38
Tabela 4.1. Distribuição das características sociodemográficas por <i>turnover</i> voluntário	42
Tabela 4.2. Distribuição das características salariais por <i>turnover</i> voluntário.....	43
Tabela 4.3. Resultados dos melhores modelos por algoritmo	44
Tabela 4.4. Top cinco variáveis mais importantes por modelo.....	46
Tabela 4.5. Recomendações aos principais <i>stakeholders</i>	51

Abreviaturas e siglas

AUC – Area Under de ROC Curve (Área abaixo da curva ROC)

CART – Classification And Regression Tree (Árvore de classificação e regressão)

CHAID – Chi-square automatic interaction detection

CRISP-DM – Cross Industry Standard Process for Data Mining

I&D – Investigação e Desenvolvimento

PCCC – Percentagem de Casos Corretamente Classificados

PRISMA – Preferred Reporting Items for Systematic reviews and Meta-Analyses (Metodologia para revisões sistemáticas e meta-análises)

QUEST – Quick, Unbiased, Efficient Statistical Tree

RH – Recursos Humanos

ROC – Receiver Operating Characteristic

RSL – Revisão Sistemática da Literatura

TI – Tecnologias da Informação

Capítulo 1 - Introdução

1.1. Contextualização do problema

O *turnover* de colaboradores, definido como a saída de um colaborador da organização pela rescisão do seu contrato (Shaw *et al.*, 2005), quando observado em demasia, ou seja, existindo uma taxa de *turnover* elevada, representa um problema que pode ser fatal para as organizações (Berengueres, Duran & Castro, 2017). Como consequência das elevadas taxas de *turnover* encontram-se o aumento de despesas, a perda de experiência e conhecimentos importantes, a diminuição da produtividade e a perda de crença na cultura da organização (O'Connell & Kung, 2007). Contudo, é o *turnover* voluntário a questão mais problemática para as empresas, pois a decisão está fora do seu controlo (Sexton *et al.*, 2005).

O *turnover* voluntário de colaboradores é dispendioso monetariamente para as organizações, onde estimativas vão desde um ano de salário por cada trabalhador que deixa a empresa (Boroş & Curşeu, 2013) até milhões de dólares gastos em formação, recrutamento e perda de produtividade (Sagie *et al.*, 2002). Além disto, este tipo de rotatividade afeta os restantes colaboradores que permanecem na empresa, pois é expectável que estes assumam o trabalho extra até ser selecionado e admitido um substituto e, com isso, experienciem a realidade que levou à saída do colega (Nandialath *et al.*, 2018). Em consideração deve estar, também, o facto de este acontecimento apresentar efeitos de contágio entre os colaboradores (Felps *et al.*, 2009). Deste modo, conseguir identificar quais os colaboradores com maior propensão para o despedimento voluntário permite uma melhor gestão de carreiras e uma consequente diminuição de custos (Wasmuth & Davis, 1983).

Conhecer e entender quais as razões que levam ao *turnover* voluntário dos colaboradores é não só um tema relevante a nível organizacional e de recursos humanos, mas também económico e psicológico. Uma abordagem possível para o seu estudo é a aplicação de algoritmos preditivos que permitam gerar conhecimento útil para o tema (Kuzior, Kettler & Rab, 2022).

Os estudos, quer sobre o *turnover* voluntário, quer sobre a intenção de *turnover*, definido pela tendência comportamental dos colaboradores em tentar deixar a organização, o que pode levar ao *turnover* efetivo (Chen *et al.*, 2014), baseiam-se, principalmente, em perceções, obtidas através da realização de questionários (Ma *et al.*, 2020) ou na combinação da aplicação dos questionários e da utilização de dados pessoais e profissionais dos colaboradores (Frederiksen, 2017). No entanto, a realização de questionário é demorada, está associada a baixas taxas de respostas, o que torna as conclusões pouco generalizáveis (Rombaut & Guerry, 2018), e enquadra-se no estudo da intenção de *turnover* e não do *turnover* real. Por outro lado, como as organizações têm vindo a recolher cada vez mais dados e, por consequência, os próprios departamentos de recursos humanos têm acesso a uma

grande variedade de dados (Harris *et al.*, 2011), já existem estudos que se dedicam à investigação da previsão de *turnover* com base nos dados armazenados em bases de dados de recursos humanos, sem associação a inquéritos (Rombaut & Guerry, 2018). Contudo as variáveis utilizadas são limitadas, tendo em conta o percurso histórico pessoal e profissional dos colaboradores.

Assim, tendo em consideração o problema dos custos elevados do *turnover* para o tecido empresarial, a presente investigação centra-se na previsão do *turnover* voluntário para que este possa ser minimizado.

1.2. Questão de investigação e objetivos

Considerando a problemática enunciada e por forma a orientar a presente dissertação, é definida a questão de investigação: Como se pode prever a saída voluntária de um colaborador?

Formulada a questão de investigação, a dissertação apresenta como principal objetivo de negócio gerar conhecimento útil para a prevenir e, consequentemente, diminuir o *turnover* voluntário, melhorando a taxa de retenção de colaboradores e diminuindo os custos associados ao *turnover*.

A investigação é guiada por dois tipos de objetivos diferentes, mas necessariamente relacionados: o objetivo de negócio, já definido, e objetivos analíticos. Os objetivos analíticos estão alinhados com o objetivo de negócio, para que se verifique o seu cumprimento, e são identificados como:

- 1) Caracterizar demograficamente e profissionalmente o *turnover* voluntário.
- 2) Construir modelo preditivo de classificação para o *turnover* voluntário.
- 3) Identificar perfis de colaboradores com elevada propensão ao *turnover* voluntário.
- 4) Identificar os fatores mais associados ao *turnover* voluntário.

O alcance dos objetivos e, por consequência, a obtenção da resposta para a questão apresentada contribuem, de forma prática, para prevenir o *turnover* voluntário, identificando quem são os colaboradores que tomam a decisão de se despedir da empresa empregadora, qual o seu perfil e quais os principais fatores associados a essa decisão.

Adicionalmente, a identificação dos perfis e dos fatores mais associados ao *turnover* voluntário permitem gerar conhecimento relevante para a definição de melhores estratégias de retenção de colaboradores, com novas ações que possam prevenir os despedimentos voluntários, e agilizar os processos que um despedimento deste tipo desencadeia, como o recrutamento de um novo trabalhador.

Por fim, a aplicação de uma nova abordagem analítica na área de Recursos Humanos, pode ser capaz de gerar novo conhecimento generalizável e útil de ser aplicado a outras realidades e, academicamente, apresentar novas perspetivas para futuras investigações.

1.3. Abordagem metodológica

Tendo em consideração os objetivos traçados, a investigação apresenta um carácter quantitativo. A dissertação tem por base a utilização de uma extensa base de dados, que contempla informações pessoais e profissionais de colaboradores e ex-colaboradores de um Grupo Económico Português, sem que seja exposta a identidade dos mesmos.

Consequentemente, trata-se de um estudo de caso em que o Grupo Económico Português, detentor de diversas empresas que atuam em diferentes setores de negócio, com presença em diversos países e com um *headcount* próximo dos 10.000 colaboradores, é alvo de estudo.

Os dados recolhidos são analisados, numa primeira fase, com recurso a técnicas de estatística descritiva para avaliar a relevância das relações entre as diferentes variáveis e a tomada de decisão associada ao despedimento voluntário. Segue-se a aplicação de técnicas de *data mining*, concretamente a árvores de decisão, para a criação de modelos preditivos de *turnover* voluntário.

De referir que as análises adotadas são suportadas por uma revisão sistemática da literatura (RSL), que segue um protocolo rigoroso, de forma a concretizar os objetivos definidos e, por consequência, responder à questão de investigação.

Por fim, as diversas etapas metodológicas enquadram-se na metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM) de Chapman *et al.* (2000) e que se divide em seis fases. Esta é a metodologia utilizada, pois, além do seu foco no negócio, a sua aplicação é considerada como bem-sucedida, uma vez que se baseia na experiência prática e real da condução de projetos em *data mining* (Chapman *et al.*, 2000).

1.4. Estrutura da dissertação

A estrutura da dissertação está dividida em cinco capítulos que correspondem às diferentes etapas da dissertação, incluindo esta introdução que enquadra a problemática em estudo, contextualizando-a, definindo a questão que guia a investigação e os objetivos da mesma.

Segue-se o segundo capítulo onde é apresentada a RSL, expondo-se o enquadramento teórico que suporta a investigação. A revisão da literatura foca-se no estudo do *turnover*, identificando as causas que lhe estão mais associadas e que métodos são aplicados para a sua previsão. Neste capítulo, é descrito o protocolo da RSL, é efetuada a sumarização do conteúdo dos artigos científicos analisados por forma a responder às questões formuladas à literatura, é efetuada a avaliação de cada artigo e, por fim, são apresentadas as implicações da revisão para a parte empírica desta dissertação.

No terceiro capítulo é apresentada e descrita a metodologia utilizada, o CRISP-DM, evidenciando todas as ações desenvolvidas e que se enquadram em cada uma das seis fases da metodologia, que vão desde a compreensão do negócio até à implementação (Schröer, Kruse & Gómez, 2021).

No capítulo seguinte, Capítulo 4, são apresentados e discutidos os resultados, pela ordem dos objetivos desta investigação. Por fim, no último capítulo, Conclusão, apresenta-se uma síntese de toda a investigação e responde-se à questão de investigação formulada. Adicionalmente, apresentam-se os contributos da investigação, as limitações e as pistas para futuras investigações.

Capítulo 2 – Revisão da literatura

2.1. Protocolo da revisão sistemática da literatura

Para a revisão de literatura sobre o *turnover* voluntário é escolhido o método da RSL que se baseia na identificação, avaliação e interpretação de toda a literatura disponível sobre a área em estudo, sendo apontado com uma das principais razões para a sua utilização a intenção de resumir as conclusões já existentes sobre uma determinada metodologia, com identificação dos contributos e limitações dos métodos utilizados (Kitchenham, 2004).

Deste modo, as estratégias de pesquisa, recolha e revisão desta RSL são baseadas nas diretrizes propostas por Barbara Kitchenham (2004) e no protocolo Preferred Reporting Items for Systematic Review (PRISMA). Este protocolo, ou lista, que se define como um plano detalhado para uma revisão sistemática (PRISMA, 2020), foi desenvolvido para facilitar a elaboração de relatórios completos deste método de revisão (Page *et al.*, 2021).

Tendo em consideração o descrito, a revisão sistemática inicia com a definição dos objetivos e quais as questões a que a literatura deve ser capaz de responder. Para que sejam selecionados os artigos apropriados é definida uma *query* de pesquisa e designados quais os critérios para a inclusão e exclusão dos mesmos. Terminado o processo de seleção, é analisado e sintetizado o conteúdo de cada artigo de forma a ser possível responder às questões colocadas e, por fim, os artigos são avaliados segundo diferentes critérios.

2.1.1. Objetivos e questões de pesquisa

A RSL é realizada com o principal objetivo de perceber e sumarizar o conhecimento já existente sobre a previsão do *turnover*, com maior incidência no despedimento voluntário de colaboradores, para que seja compreendido qual o contexto, de forma ampla e atualizada, da área em estudo. Identificada como crucial, esta RSL contribui para sustentar, guiar e melhorar a investigação. Tendo em consideração o objetivo definido, pretende-se que a RSL consiga responder à questão “No âmbito empresarial, como é que as empresas preveem e, consequentemente, identificam antecipadamente as saídas voluntárias dos seus colaboradores?”.

Para que seja possível organizar e esquematizar a resposta, segmentou-se a questão anterior em quatro questões de pesquisa mais específicas:

- I) Em que contexto são realizados os estudos?
- II) Qual a metodologia aplicada nos estudos realizados?
- III) Quais são os resultados obtidos?
- IV) Quais os contributos e implicações dos estudos?

2.1.2. Processo de seleção da literatura

Após a definição dos objetivos e das questões às quais a RSL deve ser capaz de responder, dá-se início ao processo de seleção dos artigos, que é dividido em três fases distintas: identificação, triagem e inclusão da literatura. Para se dar início à primeira fase é necessário recolher a literatura que sustenta a RSL, recorrendo a bases de dados científicas. A pesquisa de literatura nestas bases de dados deve ser realizada através da agregação de termos chave capazes de filtrar entre a literatura existente, identificando os estudos relevantes para o âmbito da investigação em causa.

Para a fase de identificação são eleitas duas bases de dados científicas, *Web of Science* e *Scopus*, por serem as bases de dados que abrangem mais áreas de estudo, sendo frequentemente utilizadas para a pesquisa de literatura (Guz & Rushchitsky, 2009), existindo em ambas uma grande oferta nas diferentes áreas de investigação (Vera-Baceta *et al.*, 2019). A pesquisa de literatura aplicada nas duas bases de dados selecionadas permite a identificação de diferentes estudos, reforçando a pertinência da utilização de ambas neste processo.

A literatura selecionada, proveniente de ambas as bases de dados, resulta da aplicação da seguinte *query* de pesquisa, validada por dois especialistas da área de recursos humanos: *(hr or people or work* or employ* or "human resources" or "human capital") and (model* or algorithm* or predict* or software) and (voluntary or intention) and (turnover or resignation or rotating or dismissal or quits) and (analytics or "deep learning" or "data science" or "computational intelligence" or "machine learning" or "data mining" or "artificial intelligence")*. A partir desta foram identificados 25 estudos do Scopus e 30 do Web of Science. A *query* encontra-se dividida em cinco grupos de termos chaves, considerados necessários para a identificação dos estudos que vão constituir a revisão da literatura. Quer ao nível dos grupos de termos, quer ao nível de cada termo em si, a *query* foi testada diversas vezes para que fosse ao encontro de literatura relevante, tendo sido adicionados termos observados em artigos considerados relevantes e eliminados termos que nutriam de um carácter redundante para a pesquisa. A partir desta *query*, aplicada ao título, *abstract* e palavras chaves dos artigos são identificados 25 estudos no Scopus e 30 no Web of Science.

Terminada a fase de identificação, segue-se a triagem e inclusão da literatura, pela aplicação dos critérios de exclusão e inclusão, identificados nas Tabelas 2.1 e 2.2 e usualmente aplicados em estudos (Mu *et al.*, 2023; Tóth *et al.*, 2023; Salehi *et al.*, 2023), selecionando-se, assim, os estudos a analisar.

Tabela 2.1. Critérios de exclusão de literatura

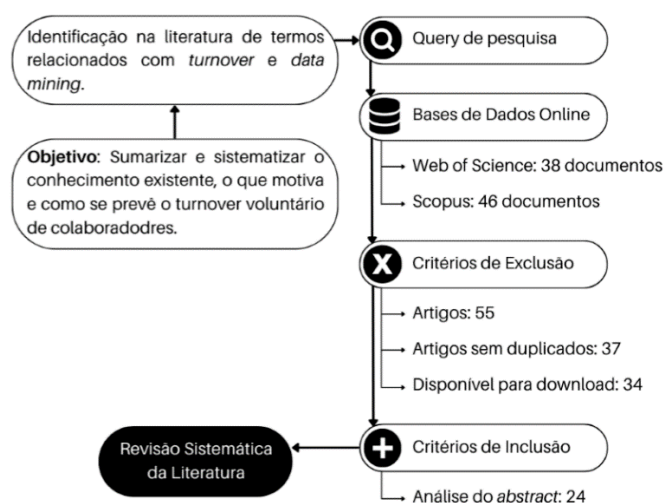
Critérios de Exclusão
• Documentos que não sejam artigos científicos • Documentos que não têm como idioma o inglês • Artigos duplicados • Artigos sem acesso

Tabela 2.2. Critérios de inclusão de literatura

Critérios de inclusão
- Abordem o tema da previsão de <i>turnover</i> e/ou <i>turnover</i> voluntário - Estudem as motivações do <i>turnover</i> e/ou <i>turnover</i> voluntário - Estudem e prevejam <i>turnover</i> e/ou <i>turnover</i> voluntário

Tendo sido selecionados 37 artigos aplicáveis à RSL, após a triagem, é necessário verificar se estes vão de encontro ao âmbito de estudo pretendido e, aplicando os critérios de inclusão, consideram-se 24 artigos para integrarem a RSL, identificados nas Tabelas 2.3 e 2.4. Todo o processo está sintetizado na Figura 2.1.

Figura 2.1. Processo de seleção da literatura



Fonte: adaptado do diagrama de seleção do PRISMA

Tabela 2.3. Artigos incluídos na RSL, anos 2023 e 2022

ID	Ano	Título	Autores	Jornal
1	2023	Diagnostic model for preschool workers' unwillingness to continue working: Developed using machine-learning techniques	Matsuo, M; Matsumoto, K; Higashijima, M; Shirabe, S; Tanaka, G; Yoshida, Y; Higashi, T; Miyabara, H; Komatsu, Y; Iwanaga, R	
2	2023	A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks	Park, J; Kwon, S; Jeong, SP	Journal of Big Data
3	2023	Research Topic Trends on Turnover Intention among Korean Registered Nurses: An Analysis Using Topic Modeling	Lee, JL; Kim, Y	Healthcare
4	2022	Early Prediction of Employee Turnover Using Machine Learning Algorithms	Atef, M; Elzanfaly, DS; Ouf, S	International Journal of Electrical na Computer Engineering Systems
5	2022	Predictions of attrition among US Marine Corps: Comparison of four predictive methods	Vanegas, JMA; Wine, W; Drasgow, F	Military Psychology
6	2022	Turnover intention among Indian police: Do organizational and community stressors matter?	Anand, V; Verma, L; Santhanam, N; Grover, A	Journal of Criminal Justice
7	2022	Predicting job-hopping motive of candidates using answers to open-ended interview questions	Jayarathne, M; Jayatilake, B	Journal of Computational Social Science
8	2022	Information system use antecedents of nursing employee turnover in a hospital setting	Thompson, SC; Holmgren, AJ; Ford, EW	Health Care Management Review
9	2022	Green talent management and turnover intention: the roles of leader STARA competence and digital task interdependence	Ogbeibu, S; Jabbar, CJC; Burgess, J; Gaskin, J; Renwick, DWS	Journal of Intellectual Capital

Tabela 2.4. Artigos incluídos na RSL, entre 2021 e 2018

ID	Ano	Título	Autores	Jornal
10	2021	Predictors of Turnover Intention in US Federal Government Workforce: Machine Learning Evidence That Perceived Comprehensive HR Practices Predict Turnover Intention	Kang, IG; Croft, B; Bichelmeyer, BA	Public Personnel Management
11	2021	Building prediction models with grouped data: A case study on the prediction of turnover intention	Yuan, S; Kroon, B; Kramer, A	Human Resource Management Journal
12	2021	Know Your Stars Before They Fall Apart: A Social Network Analysis of Telecom Industry to Foster Employee Retention Using Data Mining Technique	Younis, S; Ahsan, A	IEEE Access
13	2021	Retaining, resigning and firing: bibliometrics as a people analytics tool for examining research performance outcomes and faculty turnover	Ryan, JC	Personnel Review
14	2021	Determinants of voluntary turnover: A data-driven analysis for blue and white collar workers	Rombaut, E; Guerry, MA	Work - A Journal of Prevention Assessment & Rehabilitation
15	2020	Predicting Employee Attrition Using Machine Learning Techniques	Fallucchi, F; Coladangelo, M; Giuliano, R; De Luca, EW	Computers
16	2020	The effectiveness of employee retention through an uplift modeling approach	Rombaut, E; Guerry, MA	International Journal of Manpower
17	2019	Using Machine Learning to Translate Applicant Work History Into Predictors of Performance and Turnover	Sajjadiani, S; Sojourner, AJ; Kammeyer-Mueller, JD; Mykerezzi, E	Journal of Applied Psychology
18	2019	Predict employee attrition by using predictive analytics	Raman, R; Bhattacharya, S; Pramod, D	Benchmarking - An International Journal
19	2019	Developing HRIS for predictive attrition and retention management of indian it engineers- using ANN, ANOVA and smart PLS	Nijjer, S., Singh, J., Raj, S.	International Journal of Innovative Technology and Exploring Engineering
20	2019	An Improved Random Forest Algorithm for Predicting Employee Turnover	Gao, X., Wen, J., Zhang, C.	Mathematical Problems in Engineering
21	2019	Human capital talentum analytics-a focus study on schools of business (Sobs) in Telangana & Karnataka	Bhanu Prakash, K., Adishesha Reddy, A., Siva Reddy, P.	International Journal of Recent Technology na Engineering
22	2019	Prediction of employee retention using cassandra and ensemble learning	Karande, S., Shelake, A., Sivagami, M., Sophia, S.	International Journal of Recent Technology na Engineering
23	2018	Predicting voluntary turnover through human resources database analysis	Rombaut, E; Guerry, MA	Management Research Review
24	2018	Modeling the determinants of turnover intentions: a Bayesian approach	Nandialath, A.M., David, E., Das, D., Mohan, P.	Evidence-based HRM

De forma a demonstrar a pertinência dos artigos selecionados para a revisão de literatura do estudo, sobre os motivos do *turnover* e como é que este pode ser previsto aplicando-se técnicas de *data mining*, apresenta-se na Figura 2.2 as palavras mais frequentes (frequência > 4) no *abstract* dos artigos.

Figura 2.2. Nuvem de palavras do *abstract* dos artigos incluídos na RSL



Fonte: elaboração própria

2.1.3. Processo de revisão de literatura

Selecionados os artigos a analisar, para que seja possível ir de encontro ao objetivo da revisão e, em simultâneo, obter a resposta à questão de pesquisa já identificada, é necessário proceder à leitura de cada um dos estudos. O conteúdo de cada um é sintetizado e resumido tendo em consideração as quatro perguntas específicas de pesquisa. Assim, para todas as questões são apresentadas as respostas dadas, em formato de tabela e por tópicos, por cada um dos artigos.

Ao longo da leitura e interpretação dos diversos estudos é expectável o surgimento de conceitos ambíguos ou de difícil interpretação pelo que são envolvidos especialistas nas áreas de recursos humanos e *business analytics* capazes, pelo seu nível de conhecimento e experiência, guiar da melhor forma a revisão da literatura. Depois de recolhidas as respostas, de cada um dos estudos, e estruturadas em tabelas, interpretam-se estas tabelas, retirando-se conclusões e identificando-se tendências com recurso à análise descritiva.

2.1.4. Avaliação da qualidade da literatura da RSL

No âmbito da RSL existe, ainda, a fase de avaliação da literatura selecionada, tendo sido definido um conjunto de critérios para cada uma das questões de pesquisa, com o propósito de avaliar quais os artigos que melhor respondem a estas. Os 12 critérios a aplicar a cada artigo estão apresentados na Tabela 2.5 e avaliam os artigos tendo em consideração o objeto de estudo, os fatores relacionados com as saídas voluntárias, a metodologia, as variáveis e técnicas preditivas utilizadas. A cada critério está associada uma pontuação que varia entre 0 e 1. Assim, um artigo é avaliado com 1 se responder de forma completa, é avaliado com 0,5 se responder de forma parcial e com 0 se não responder à questão.

Tabela 2.5. Critérios de avaliação da literatura da RSL

Questão	Critério
I	1. Apresenta, de forma clara, o âmbito do estudo?
	2. Define, claramente, qual(ais) o(s) objetivo(s) do estudo?
II	3. Esclarece qual a fonte dos dados incluídos no estudo?
	4. Caracteriza, detalhadamente, a amostra?
	5. Detalha e justifica a(s) técnica(s) preditiva(s) utilizada(s)?
	6. Enumera, justificando, quais as variáveis a considerar no estudo?
	7. Define explicitamente que critérios são utilizados para avaliar os resultados dos modelos preditivos?
III	8. Descreve, de forma clara, o desempenho dos modelos face aos critérios definidos, utilizando diferentes métricas de avaliação?
	9. Compara claramente os resultados de diferentes algoritmos?
	10. Apresenta e justifica quais as principais variáveis explicativas do estudo?
IV	11. Enumera, de forma clara, quais os contributos do estudo?
	12. Descreve, claramente, as limitações do estudo?

2.2. Síntese do conteúdo da literatura

De forma a ser possível perceber como é estudado e previsto o *turnover* na literatura existente, quais os principais âmbitos de aplicação e as abordagens adotadas dos estudos é analisado em detalhe, neste subcapítulo, o conteúdo dos 24 artigos selecionados.

2.2.1. Contexto dos estudos

Na literatura existente, o maior foco de análise é sobre a intenção de *turnover* de colaboradores (Park, Kwon & Jeong, 2023; Lee & Kim, 2023; Anand *et al.*, 2022), que pode ser denominada pela intenção do colaborador em sair da empresa ou despedir-se (Nandialath *et al.*, 2018), ao invés do *turnover* efetivo (Gao, Wen & Zhang, 2019; Karande *et al.*, 2019). Ogbeibu *et al.* (2022) consideram a intenção de *turnover* como a vontade deliberada e consciente dos colaboradores e intenção comportamental dos mesmos em deixarem a organização para a qual trabalham¹.

Dos 24 artigos em análise, aproximadamente, 42% (10 artigos) apresentam como objeto de estudo a intenção de *turnover* de colaboradores (Tabela 2.6). Kang, Croft e Bichelmeyer (2021) justificam a escolha do estudo da intenção de *turnover* ao indicar que este é aceite pelos investigadores como uma aproximação ao *turnover* efetivo, referindo que de um ponto de vista prático é mais económico medir a intenção utilizando questionários do que estudar o *turnover* efetivo através de uma abordagem longitudinal, que adicionalmente pode representar um problema ético, ao indicar informação pessoal sobre a saída das organizações. De um ponto de vista teórico, os autores evidenciam que o comportamento de uma pessoa é determinado pela intenção de o efetivar, o que torna a intenção de *turnover* no preditor mais importante do *turnover* efetivo. Outros autores, como Nandialath *et al.* (2018), também apoiam esta perspetiva e identificam a intenção *turnover* como o principal preditor do *turnover* efetivo.

Por outro lado, na literatura, existem investigadores que dedicam os seus estudos ao *turnover* efetivo, que, para Gao, Wen e Zhang (2019), representa a entrada e saída de colaboradores de uma empresa², e que se pode dividir em *turnover* voluntário, onde é o colaborador que decide sair da empresa ou se reforma, e *turnover* involuntário, representado pela decisão do empregador em despedir o colaborador (Karande *et al.*, 2019). Rombaut e Guerry (2018) defendem que utilizar uma abordagem de investigação baseada em dados longitudinais, em vez da utilização de questionários, é mais vantajoso, pois permite estudar o *turnover* efetivo e as suas causas, sem a necessidade de recolher dados suplementares. Acrescentam ainda que permite a observação de um grupo maior de colaboradores, visto que os questionários estão associados a baixas taxas de resposta e limitam a

¹ Definição baseada no autor Ngo-Henha (2017).

² Definição baseada nos autores Denvir e McMahon (1992).

generalização dos resultados. De entre os artigos cujo objeto de estudo é o *turnover* efetivo, que totaliza oito artigos, como o de Ryan (2021), dois focam a sua investigação no *turnover* voluntário por considerarem que a sua previsão é crucial para as organizações, pela sua associação a custos elevados e à diminuição de produtividade (Rombaut & Guerry, 2018, 2021).

Além do estudo do *turnover* e intenção de *turnover*, existem autores que investigam acontecimentos semelhantes ou relacionados, como sejam:

- a *attrition*, que em contexto militar é apresentada por Vanegas, Wine e Drasgow (2022) como uma forma de *turnover* associada à não adaptação ao serviço militar e consequente abandono³, enquanto Fallucchi *et al.* (2020) assumem *attrition* como a demissão por parte do colaborador ou a sua reforma;
- o *job-hopping*, identificado por Jayaratne e Jayatilleke (2022) como um tipo de *turnover* voluntário que inclui os indivíduos com maior predominância para a mudança frequente de emprego⁴;
- O *turnover hazards*, descrito por Sajjadian *et al.* (2019) como a velocidade esperada do *turnover*, indicando se e quando ocorreu o *turnover* dos colaboradores⁵.

Adicionalmente, dos artigos incluídos na RSL, destacam-se ainda dois artigos cujos objetos de estudo não se enquadram nos já enunciados, apesar de se relacionarem com estes, onde um estuda a eficácia de estratégias de retenção de colaboradores (Rombaut & Guerry, 2020) e o outro analisa o impacto de uma ferramenta de recursos humanos na gestão de talentos (Prakash *et al.*, 2019).

Na literatura em análise verifica-se que 83% dos estudos têm como âmbito a previsão. Em concordância, no que toca aos objetivos, cerca um terço identifica como objetivo principal aplicar um ou vários modelos preditivos de forma a estudar o objeto definido (Atef, Elzanfaly & Ouf, 2023; Sajjadiani *et al.*, 2019; Gao, Wen & Zhang, 2019), enquanto 30% define como objetivo primordial a identificação de padrões que possam explicar o objeto em estudo (Ogbeibu *et al.*, 2022; Younis & Ahsan, 2021; Bhanu, Adishesha & Siva, 2019). Ambos os objetivos foram identificados para aproximadamente 37% da literatura (Anand *et al.*, 2022; Jayaratne & Jayatilleke, 2022; Thompson, Holmgren & Ford, 2022). Ainda que os objetivos se enquadrem nesta categorização, ao serem aplicados a diferentes contextos todos eles apresentam as suas especificidades, como, por exemplo, avaliar a utilização de métodos preditivos para a intenção de *turnover* durante o processo de recrutamento (Atef, Elzanfaly & Ouf, 2022), avaliar se os registos eletrónicos profissionais de colaboradores podem prever o *turnover* voluntário (Thompson, Holmgren & Ford, 2022), avaliar a utilização de indicadores bibliométricos de investigações para analisar decisões de recursos humanos

³ A definição utilizada por este autor é de Stark *et al.* (2011).

⁴ A definição usada por este autor é de Lake, Highhouse e Shrift (2018).

⁵ A definição utilizada por este autor é da autoria de Dickter *et al.* (1996).

como *turnover* (Fallucchi *et al.*, 2020) ou avaliar a utilização dos padrões de comunicação de colaboradores via email para prever o *turnover* dos mesmos (Raman, Bhattacharya & Pramod, 2019).

Em seguimento e no que diz respeito ao contexto em que as investigações foram conduzidas, realça-se que a maioria utilizou como alvo de estudo os colaboradores de uma organização, estudando assim o *turnover* de uma forma mais generalizada, não limitando as conclusões à função específica do colaborador, mas sim ao empregador (Kang, Croft & Bichelmeyer, 2021; Yuan, Kroon & Kramer, 2021; Fallucchi *et al.*, 2020). Em contrapartida, outros autores identificaram como contributo para a literatura investigar o *turnover*, ou respetivos acontecimentos semelhantes, em alvos mais específicos (50%), como foi o caso de Anand *et al.* (2022), ao investigarem a intenção de *turnover* de agentes da polícia, analisando a influência dos fatores organizacionais e da comunidade, ou Thompson, Holmgren e Ford, (2022), que estudaram o *turnover* de enfermeiros através dos seus registos de trabalho eletrónicos. Atualmente, verifica-se uma maior tendência na literatura para o estudo em alvos específicos e constata-se a existência de um maior interesse na investigação do *turnover* em professores e colaboradores que desempenham funções nas áreas de tecnologia e informação⁶.

Noutra perspetiva, os artigos selecionados basearam as suas investigações, principalmente, em dados recolhidos na década de 2010-2020, onde os mais recentes apresentam dados de 2019 e 2021 (Yuan, Kroon & Kramer, 2021; Lee & Kim, 2023) e os mais antigos são da década anterior, de 2006 e 2007 (Rombaut & Guerry, 2018; Sajjadi *et al.*, 2019). Destacam-se ainda os artigos de Lee e Kim (2023) e Rombaut e Guerry (2018) por utilizarem registos de um período igual ou superior a 10 anos. De entre os artigos que contextualizaram a investigação sobre o país ou região, cerca de 26% foram conduzidos na Índia (Anand *et al.*, 2022; Raman, Bhattacharya & Pramod, 2019), 21% nos Estados Unidos da América (Vanegas, Wine & Drasgow, 2022; Thompson, Holmgren & Ford, 2022) e 16% na Bélgica (Rombaut & Guerry, 2018, 2020, 2021). Dos restantes 37%, a maioria enquadra-se no continente asiático (Matsuo *et al.*, 2023; Park, Kwon & Jeong, 2023; Lee & Kim, 2023).

Com base na análise realizada é possível responder à primeira questão específica da RSL: o âmbito das investigações tende a ser maioritariamente a previsão da intenção de *turnover*, seguindo-se o *turnover* efetivo, com o principal objetivo de o estudar através da aplicação de modelos preditivos a dados de colaboradores, na sua generalidade, ou a aplicação a um grupo específico que desempenha funções semelhantes, durante a segunda década do século XXI.

⁶ Raman, Bhattacharya e Pramod (2019) referem que o *turnover* do corpo docente é um problema recorrente nas escolas e instituições de ensino principalmente nos primeiros anos de carreira, sendo a sua retenção vital para assegurar o papel significativo destas instituições no desenvolvimento social, cultural e económico das comunidades (Ryan, 2021). Já autores como Nijjer, Singh e Raj (2019) focam o estudo em colaboradores do setor da tecnologia e informação por referirem que neste setor o *turnover* tem sido um problema habitual, em todas as funções e que aumenta com o nível de competências.

Tabela 2.6. Âmbito e objetivos dos artigos incluídos na RSL

ID Artigo	Âmbito	Objeto em estudo	Objetivo(s)	Alvo	Período	País ou Região
1	Prever	Intenção de <i>turnover</i>	Aplicar modelo(s) preditivo(s)	Professores do ensino pré-escolar	2018	Japão
2	Prever	Intenção de <i>turnover</i>	Aplicar modelo(s) preditivo(s)	Recém-formados	2019	Coreia do Sul
3	Analisar	Intenção de <i>turnover</i>	Identificar padrões	Artigos sobre enfermagem	2010 a 2021	Coreia do Sul
4	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s)	Candidatos em processo de recrutamento	N.E.	N.E.
5	Prever	<i>Attrition</i>	Aplicar modelo(s) preditivo(s)	Recrutados para o Corpo de Fuzileiros Navais	N.E.	EUA
6	Prever	Intenção de <i>turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Agentes da Polícia	N.E.	Índia
7	Prever	<i>Job-Hopping</i> - motivação para a mudança de emprego	Aplicar modelo(s) preditivo(s) e Identificar padrões	Candidatos a empregos	N.E.	N.E.
8	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Enfermeiros	Março de 2017 e Março de 2018	EUA
9	Prever	Intenção de <i>turnover</i>	Identificar padrões	Colaboradores dos departamentos de RH, I&D e TI	N.E.	Nigéria
10	Prever	Intenção de <i>turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Colaboradores	2018	EUA
11	Prever	Intenção de <i>turnover</i>	Aplicar modelo(s) preditivo(s)	Colaboradores	2017 a 2019	N.E.
12	Prever	Intenção de <i>turnover</i>	Identificar padrões	Colaboradores de departamentos de engenharia	2019	Paquistão
13	Analisar	<i>Turnover</i>	Identificar padrões	Corpo Docente	2010 a 2015	Região Árabe
14	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Colaboradores	2015	Bélgica
15	Prever	<i>Attrition</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Colaboradores	N.E.	N.E.
16	Analisar	<i>Personalised Treatment Effect</i>	Identificar padrões	Colaboradores	2016	Bélgica
17	Prever	<i>Turnover hazards</i>	Aplicar modelo(s) preditivo(s)	Candidatos a cargos de professores	2007 a 2013	EUA
18	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Corpo Docente	2010 a 2017	Índia
19	Prever	Intenção de <i>turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Engenheiros de <i>software</i>	N.E.	Índia
20	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s)	Colaboradores	N.E.	China
21	Analisar	<i>Human Capital Talentum Analytics (HCTA)</i>	Identificar padrões	Escolas de Gestão	N.E.	Índia
22	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s) e Identificar padrões	Colaboradores	N.E.	N.E.
23	Prever	<i>Turnover</i>	Aplicar modelo(s) preditivo(s)	Colaboradores	2006 a 2016	Bélgica
24	Prever	Intenção de <i>turnover</i>	Identificar padrões	Colaboradores	N.E.	Índia

Nota: N.E – Não específica

2.2.2. Metodologias aplicadas

De entre a literatura selecionada a escolha predominante para o estudo do *turnover*, ou conceitos similares, incidiu na utilização de questionários a colaboradores (Tabela 2.7), tendo sido utilizados em cerca de 46% dos estudos (Matsuo *et al.*, 2023; Anand *et al.*, 2022; Nijjer, Singh & Raj, 2019). Em concordância com aquilo que foi mencionado no subcapítulo anterior, a opção metodológica de realizar questionários encontra-se associada ao estudo da intenção de *turnover*: mais de 80% das investigações que recolheram dados através de questionários definiram como objeto de estudo a intenção de *turnover* (Park, Kwon & Jeong, 2023; Ogbeibu *et al.*, 2022; Yuan, Kroon & Kramer, 2021).

Entre os estudos onde a base de dados utilizada é proveniente da realização de questionários verifica-se, também, que as investigações cuja fonte destes é de elaboração própria (55%) estão associadas a amostras mais pequenas comparativamente a outros estudos (Ogbeibu *et al.*, 2022;

Younis & Ahsan, 2021; Nandialath *et al.*, 2018). Desta forma, observa-se que as amostras de menor dimensão (menores que 1000 indivíduos), encontram-se, na sua maioria, em investigações que utilizam questionários realizados e distribuídos pelos próprios autores, confirmando o que é defendido por Rombaut e Guerry (2018): a realização de questionários limita os estudos por estarem associados a baixas taxas de reposta. Contudo, um cenário diferente é identificado aquando da utilização de questionários realizados e distribuídos por outras entidades. No estudo de Kang, Croft e Bichelmeyer (2021) são utilizadas as respostas do Federal Employee Viewpoint Survey (2018), um questionário realizado e aplicado pelo U.S. Office of Personnel Management aos colaboradores federais dos Estados Unidos da América, que apesar de ter registado uma taxa de resposta inferior a 50%, apresenta uma amostra de 567838 indivíduos. Efetivamente, dos artigos selecionados e que apresentam uma maior amostra (maior ou igual que 10000 indivíduos), 80% utilizou as respostas de questionários conduzidos por terceiros (Park, Kwon & Jeong, 2023; Jayaratne & Jayatilleke, 2022).

Em adição à utilização de questionários como base de dados, os autores de um terço da literatura analisada optaram pela utilização de dados relativos ao histórico pessoal e profissional recolhidos pelas empresas empregadoras, sendo o seu principal objeto de estudo o *turnover* efetivo (Atef, Elzanfaly & Ouf, 2022; Gao, Wen & Zhang, 2019). Esta opção metodológica permitiu que os estudos obtivessem, maioritariamente, amostras entre 1000 e 5000 indivíduos (Fallucchi *et al.*, 2020; Rombaut & Guerry, 2018), destacando-se o estudo de Rombaut & Guerry (2021) que apresenta uma amostra de 385680 indivíduos. Verifica-se, ainda, que o estudo do *turnover*, a partir de dados pessoais e/ou profissionais de colaboradores recolhidos pelo empregador, é realizado, maioritariamente, em empresas do setor privado (Rombaut & Guerry, 2020; Gao, Wen & Zhang, 2019).

Para além destes dois tipos de bases de dados, na literatura existem autores que investigam os objetos de estudo definidos através da utilização de artigos publicados sobre a intenção de *turnover* em enfermeiros (Lee & Kim, 2023), *metadata* de um sistema de registo eletrónico de trabalho utilizado por enfermeiros num centro médico (Thompson, Holmgren & Ford, 2022), *emails* e *metadata* associada do corpo docente de uma *Business School* (Raman, Bhattacharya & Pramod, 2019) e entrevistas conduzidas em escolas de gestão (Bhanu, Adishesha & Siva, 2019). Estes últimos estudos encontram-se associados, de uma forma geral, a amostras reduzidas e ao setor público.

Tabela 1.7. Caracterização da amostra e base de dados dos artigos incluídos na RSL

ID Artigo	Dimensão da Amostra	Organização	Base de Dados	Fonte
1	551	Jardins-de-infância e creches e instituições de acolhimento	Questionário	Elaboração própria
2	12.202	N.E.	Questionário GOMS	Korea Employment Information Service's
3	390	N.A.	Artigos publicados	Repositório de artigos nacionais e internacionais*
4	1.470	N.E.	Base de dados de <i>turnover</i> de colaboradores	IMB
5	3.9043	USMC	Questionário	AFQT, TAPAS
6	492	Delhi Police	Questionário	Elaboração própria
7	45.899	PredictiveHire	Questionário	Plataforma PredictiveHire
8	1.836	Centro Médico Académico	Metadata do sistema de registo médico eletrónico	Data warehouse da instituição
9	372	Empresas industriais	Questionário	Elaboração própria
10	567.838	Governo Federal dos Estados Unidos	Questionário: Federal Employee Viewpoint Survey	U.S. Office of Personnel Management
11	1.454	199 PME's	Questionário	Elaboração própria
12	169	Ufone e Zong - operadoras de telecomunicações	Questionário	Elaboração própria
13	684	Universidade	Dados bibliométricos das investigações e dados sobre o corpo docente.	Scopus, Site e HRIS da universidade.
14	385.680	Empresas Belgas	Base de dados de recursos humanos	Empresa de processamento de salários
15	1.470	N.E.	Base de dados de <i>turnover</i> de colaboradores	IMB
16	1.606	Empresa	Base de dados de recursos humanos	Departamento de Recursos Humanos
17	16.071	MPS	TEES; registos das candidaturas (formulários); dados relativos aos candidatos que ficaram no agrupamento	Agrupamento de escolas
18	126	Business School	Emails	Repositório de emails
19	581	Empresas de TI	Questionário	Elaboração própria
20	2.000	Empresa de comunicações	Base de dados interna da empresa sobre os colaboradores	Empresa alvo do estudo
21	60	Escolas de Gestão	Entrevista	Elaboração própria
22	N.E.	N.E.	Base de dados de <i>turnover</i> de colaboradores	Kaggle
23	4.041	Sucrusal de empresa privada especializada em RH	Base de dados de recursos humanos	Departamento de Recursos Humanos
24	123	Call-center	Questionário	Elaboração própria

Notas: N.A. - Não aplicável; N.E. - Não específica; USMC - United States Marine Corps; PME's - Pequenas e Médias Empresas; MPS - Minneapolis Public School District; TI - Tecnologias de Informação; GOMS - Graduates Occupations Mobility Survey; AFQT - The Armed Forces Qualification Test; TAPAS - The Tailored Adaptive Personality Assessment System; HRIS - human resource information system; TEES - Teacher-effectiveness evaluation system

* - Nurimedia, National Science and Technology Information Center, Research Information Sharing Service, Korean Studies Information Service System, PubMed, CINAHL

Aprofundando a metodologia escolhida pelos diferentes autores, no que toca às técnicas de análise de dados utilizadas, para estudar e prever o *turnover* e acontecimentos similares, constata-se a predominância da utilização de técnicas de *machine learning* (Park, Kwon & Jeong, 2023; Karande *et al.*, 2019). A Tabela 2.8 evidencia a presença deste tipo de análise em 75% dos estudos, o que vai de encontro aos objetivos identificados anteriormente, demonstrando, também, a relevância do processo de pesquisa e seleção dos mesmos. Na literatura, prevalece a utilização de técnicas supervisionadas (Kang, Croft & Bichelmeyer, 2021; Gao, Wen & Zhang, 2019), à exceção do artigo de Lee e Kim (2023) que utiliza *topic modelling*, uma técnica não supervisionada, para identificar *clusters* de termos, em artigos, relacionados com a intenção de *turnover* de enfermeiros.

Dos estudos que aplicam técnicas de *machine learning*, cerca de 78% apresentam um problema de classificação associado à utilização de uma variável *target* dicotómica, como saiu ou não saiu da

empresa, ou qualitativa não dicotómica enquanto *target* (e.g., três diferentes *status* de colaboradores) (Vanegas, Wine & Drasgow, 2022; Ryan, 2021). Neste âmbito, verifica-se uma maior predominância de variáveis *target* dicotómicas (67%), que se encontram mais associadas às investigações que estudam o *turnover* efetivo de colaboradores (Atef, Elzanfaly & Ouf, 2022; Thompson, Holmgren & Ford, 2022). De entre estes artigos (18), metade dos autores procedeu à utilização de diferentes algoritmos de forma a poder comparar a performance de diferentes modelos preditivos (Matsuo *et al.*, 2023; Karande *et al.*, 2019), enquanto os restantes aplicaram e avaliaram apenas um modelo preditivo (Jayaratne & Jayatilleke, 2022; Nijjer, Singh & Raj, 2019). Destacam-se os artigos de Fallucchi *et al.* (2020), Sajjadiani (2019) e Gao, Wen e Zhang (2019) que optaram pelo desenvolvimento e comparação de pelo menos cinco modelos (Tabela 2.9).

No que toca aos algoritmos aplicados, na literatura selecionada, sobressaem a regressão logística, aplicada em 44% das investigações que utilizam técnicas de *machine learning*, e *random forest*, presente em 33% dessas investigações. Rombaut e Guerry (2018) justificam a utilização da regressão logística pelo facto de esta ser flexível e de fácil utilização, os seus parâmetros permitirem estimar resultados relevantes e ser uma técnica conhecida e bastante utilizada. Já Gao, Wen e Zhang (2019) justificam a utilização do algoritmo *random forest* afirmando que este é considerado, por muitos estudos empíricos, um algoritmo de elevada precisão e com boa tolerância ao possível ruído das variáveis utilizadas. Destaca-se, ainda, a aplicação de outros algoritmos baseados em árvores de decisão, que vão desde CART (Vanegas, Wine & Drasgow, 2022), C4.5 (Gao, Wen & Zhang, 2019) ou XGBoost Tree (Kang, Croft & Bichelmeyer, 2021) e que, todos juntos, apresentam uma representatividade de 39%.

Para avaliar a performance dos diferentes modelos de classificação, os autores decidiram, na sua maioria, calcular as métricas associadas à matriz de confusão: percentagem de casos corretamente classificados (PCCC) ou acurácia, precisão, sensibilidade ou *recall*, especificidade e F-score (Park, Kwon & Jeong, 2023; Fallucchi *et al.*, 2020). Autores como Matsuo *et al.* (2023) e Atef, Elzanfaly e Ouf (2022) optaram pela inclusão de outras métricas como a AUC (Area under the ROC curve). O estudo de Gao, Wen e Zhang (2021) refere que as métricas mais comuns de performance de modelos preditivos de classificação, são a PCCC, sensibilidade, precisão e a AUC.

Associado à aplicação destas técnicas verifica-se, ainda, a utilização da análise estatística: descritiva, inferencial ou ambas. Mais de metade dos artigos que aplicam técnicas de *machine learning*, aplicam também análise estatística. Os autores Atef, Elzanfaly e Ouf (2022) e Fallucchi *et al.* (2020) utilizam estatística descritiva com o intuito de analisar as variáveis e identificar quais afetam de forma mais significativa o *target*. Por outro lado, estudos como o de Anand *et al.* (2022) utilizam estatística inferencial para identificar as correlações existentes entre as variáveis ou se as mesmas são estatisticamente significativas (Rombaut & Guerry, 2020).

Já as investigações que optaram por analisar e prever os respetivos objetos de estudo apenas através da aplicação da análise estatística, utilizaram maioritariamente a análise descritiva e inferencial (Ogbeibu *et al.*, 2022; Ryan, 2021), ainda que autores como Nandialath *et al.* (2018) tenham optado, por exemplo, pela execução adicional da regressão linear. Estes estudos estão associados principalmente à investigação da intenção de *turnover*, utilizando variáveis qualitativas como *target* (Ogbeibu *et al.*, 2022; Nandialath *et al.*, 2018).

Por fim, no que toca ao tipo de variáveis utilizadas pelas investigações, existe uma maior predominância da inclusão de variáveis do foro profissional dos colaboradores, onde, dependendo da base de dados utilizada, se encontram diferentes grupos de características. Dos estudos que optaram pela utilização de uma base de dados proveniente das respostas a questionários identifica-se as seguintes variáveis, relacionadas com características a nível profissional, como as mais comuns: satisfação, remuneração e apoio organizacional (Park, Kwon & Jeong, 2023; Nandialath *et al.*, 2018).

No que diz respeito às variáveis mais utilizadas, por estes estudos, categorizadas como pessoais destacam-se o género, a idade e a escolaridade dos colaboradores (Yuan, Kroon & Kramer, 2021). Analisando os estudos que, por outro lado, utilizam bases de dados provenientes dos empregadores, as variáveis que se destacam ao nível profissional são a remuneração, a antiguidade na empresa e a distância entre a casa e o trabalho (Rombaut & Guerry, 2020, 2021). Já as variáveis pessoais mais utilizadas dizem respeito à idade, ao género e ao estado civil (Atef, Elzanfaly & Ouf, 2022). Os artigos que se baseiam na investigação através de outro tipo de bases de dados que não as anteriores, utilizam consequentemente na análise outro tipo de variáveis como sejam os títulos, os resumos e palavras-chave de artigos (Lee & Kim, 2023), ou os dados bibliométricos, como o número de publicações, de artigos publicados por professores (Ryan, 2021).

Na literatura verifica-se que o número médio de variáveis utilizadas para as investigações é de 12, destacando-se os artigos de Anand *et al.* (2022), Ryan (2021) e Karande *et al.* (2019) que utilizam apenas quatro e cinco características, respetivamente, e o artigo de Fallucchi *et al.* (2020) que utiliza o maior número de variáveis (34).

Em resposta à segunda questão específica da RSL, os estudos que se focam no *turnover* efetivo de colaboradores utilizam, na sua maioria, os dados recolhidos e armazenados pelos respetivos empregadores e focam a análise na generalidade dos colaboradores. Já o estudo da intenção de *turnover*, ou de outros acontecimentos, é baseado nas respostas a questionários e tendem a focar-se num grupo de colaboradores específico. Para a previsão e análise do objeto de estudo são utilizadas estatística descritiva, inferencial e/ou técnicas de *machine learning* para classificação, que originam modelos avaliados, maioritariamente, através da matriz de confusão.

Tabela 2.8. Metodologia dos artigos incluídos na RSL

ID Artigo	Técnicas	Tipo de problema	Variáveis (nº)			Target	Classificação do target	Critérios de avaliação
			Pessoais	Profissionais	Outras			
1	Estatística inferencial, Regressão Logística e ML - supervisionado	Classificação	8	6	-	Intenção de sair do emprego.	Dictômico	AUROC, Sensibilidade e Especificidade
2	ML - supervisionado	Classificação	-	13	-	Intenção de sair do emprego.	Dictômico	PCCC, Precisão, Sensibilidade e F1-score
3	TM e Topic modeling	Clusters	-	-	3 tipos de elementos de artigos de enfermagem	N.A.	N.A.	N.E.
4	Estatística descritiva e ML - supervisionado	Classificação	4	1	-	Saída do emprego.	Dictômico	PCCC, Matriz de Confusão, AUC, ROC, Precisão, Sensibilidade, F1 score, F-score, Especificidade e FPR
5	Estatística descritiva e inferencial e ML - supervisionado	Classificação	2	16	-	Abandono/dispensa do programa.	Dictômico	Phi coefficient, Sensibilidade, Especificidade, PPV, NPV, Clinical efficiency, AUC
6	Estatística descritiva e inferencial e ML - supervisionado	Classificação	-	3	2 - contexto social	Burnout e intenção de sair do emprego.	Qualitativo	Precisão, Recall, F1 score e Hamming Loss
7	Estatística descritiva e inferencial, TM e ML - supervisionado	Regressão	N.E.	N.E.	N.E.	Scores Job-Hopping	Quantitativo	Accuracy, medido através do coeficiente de correlação
8	Estatística descritiva e ML - supervisionado	Classificação	-	5	-	Saída do emprego.	Dictômico	Sensibilidade e Especificidade
9	Estatística descritiva e inferencial	N.A.	-	8	-	Intenção de sair do emprego.	Qualitativo	N.A.
10	ML - supervisionado	Classificação	4	superior a 20	-	Intenção de sair do emprego.	Dictômico	AUC, Precisão, Sensibilidade, F-1 score, PCCC
11	Estatística descritiva, Regressão linear e ML - supervisionado	Regressão	3	15	-	Intenção de sair do emprego.	Quantitativo	R ²
12	Regressão Múltipla	Regressão	8	5	-	Intenção de sair do emprego.	Quantitativo	R ²
13	Estatística descritiva e inferencial	N.A.	-	-	4 - dados bibliométricos de investigações	Status do colaborador: retido, apresentou demissão ou despedido pelo empregador.	Qualitativo	N.A.
14	Estatística descritiva e inferencial e ML - supervisionado	Classificação	6	11	-	Saída voluntária do emprego.	Dictômico	AUC
15	Estatística descritiva e ML - supervisionado	Classificação	5	29	-	Saída do emprego.	Dictômico	PCCC, Precisão, Sensibilidade, Especificidade, F1-score, Matriz de confusão
16	Estatística inferencial e ML - supervisionado	Classificação	6	15	-	Saída voluntária do emprego.	Dictômico	Qini curve
17	Estatística descritiva, TM, ML - supervisionado	Classificação	-	7	3 - variáveis de controlo	Nível de performance e Turnover Hazards	Quantitativo	PCCC
18	Estatística inferencial e TM	N.A.	-	-	15 - emails e metadata associada	Saída do emprego.	Dictômico	N.E.
19	Estatística inferencial e ML - supervisionado	Regressão	13	4	-	Intenção de sair do emprego, satisfação com o trabalho e fit entre pessoa-organização	Quantitativo	N.E.
20	ML - supervisionado	Classificação	2	12	-	Saída do emprego.	Dictômico	Sensibilidade, F-measure, área ROC, PCCC, Matrix de Confusão
21	Estatística (n.e.)	N.A.	-	-	17 - práticas e desempenho da faculdade	N.A.	N.A.	N.A.
22	ML - supervisionado	Classificação	-	4	-	Saída do emprego.	Dictômico	Matriz de confusão: PCCC, precisão, sensibilidade, especificidade
23	ML - supervisionado	Classificação	7	7	-	Saída voluntária do emprego.	Dictômico	AUC
24	Estatística descritiva e inferencial e Regressão Linear	N.A.	6	10	-	Intenção de sair do emprego.	Qualitativo	N.A.

Notas: N.E. - Não específica; N.A. - Não aplicável; ML - Machine Learning; TM - Text Mining; AUC - Area under the ROC curve; PPV - Positive predictive value; NPV - negative predictive value; FPR - False-positive rate; AFQT - The Armed Forces Qualification Test; TAPAS - The Tailored Adaptive Personality Assessment System

2.2.3. Avaliação dos resultados

Considerando os estudos que procederam à aplicação de modelos preditivos e à respetiva avaliação, verifica-se que: os modelos baseados na aplicação do algoritmo *XGBoost* são considerados os modelos com melhor desempenho, quando comparados com outros modelos de algoritmos diferentes (Matsuo *et al.*, 2023; Park, Kwon & Jeong, 2023). Já o modelo com melhor desempenho, desenvolvido por Vanegas, Wine e Drasgow (2022) e Anand *et al.* (2022), foi baseado no algoritmo *random forest*, no entanto este algoritmo foi comparado em outros quatro estudos e não originou o modelo com a

melhor performance (Sajjadiani *et al.*, 2019; Gao, Wen & Zhang, 2019). No estudo de Fallucchi *et al.* (2020), que comparou o maior número de modelos, baseados em diferentes algoritmos, considerou que o modelo com melhor desempenho foi criado a partir do *Gaussian Naive Bayes*. É, ainda, de realçar que entre todos os resultados das diferentes métricas de avaliação dos modelos identificados na literatura, destaca-se o modelo de Sajjadiani *et al.* (2019) e o modelo de Gao, Wen e Zhang (2019) que obtiveram um PCCC de 0,988 e 0,928, utilizando os algoritmos *Naive Bayes* e *Weighted Quadratic random forest*, respetivamente.

No que toca às variáveis com maior capacidade explicativa para a previsão ou análise dos diferentes acontecimentos estudados, como o *turnover* ou a intenção de *turnover*, percecione-se pela Tabela 2.9, que são as variáveis relacionadas com o trabalho as que influenciam, em maior número, os objetos estudados (Rombaut & Guerry, 2018; Nandialath *et al.*, 2018). Na literatura destacam-se as variáveis idade, rendimento, distância entre a casa e o trabalho e a antiguidade na empresa como as mais relevantes para a predição e explicação do *turnover* ou acontecimentos similares estudados. De entre todas as variáveis consideradas como significativas pelos diferentes estudos, estas são as mais mencionadas na literatura em análise (Rombaut & Guerry, 2018, 2020).

A variável idade demonstrou ter a mesma influência sobre o *turnover* nos diversos artigos, que concluíram que o aumento da idade dos colaboradores diminui a probabilidade de estes abandonarem a organização (Atef, Elzanfaly & Ouf, 2022; Fallucchi *et al.*, 2020). O mesmo é verificado para a antiguidade dos colaboradores, que também influencia negativamente o *turnover*, ou seja, à medida que aumenta a antiguidade diminui a probabilidade de *turnover* (Rombaut & Guerry, 2021; Gao, Wen & Zhang, 2019). No entanto, Rombaut e Guerry (2018) concluíram, ainda, que no início a probabilidade de *turnover* aumenta com o aumento da antiguidade até o colaborador atingir os cinco anos de serviço, altura em que começa a diminuir. Por outro lado, o rendimento tem a influência oposta sobre o *turnover*. Segundo a literatura, à medida que o salário aumenta a probabilidade de o colaborador deixar a organização diminui (Atef, Elzanfaly & Ouf, 2022; Fallucchi *et al.*, 2020; Gao, Wen & Zhang, 2019).

Ainda assim, a investigação de Rombaut e Guerry (2018) demonstra que o efeito do rendimento pode ser o oposto, isto é, em colaboradores com salário mais alto, aqueles que têm carro da empresa, apresentam maior probabilidade de *turnover*, já em colaboradores com salários mais baixos, os que têm carro da empresa têm menor probabilidade de *turnover*. Por fim, a maioria dos autores, que considerou a distância entre o trabalho e a residência do colaborador como variável explicativa, concluem que à medida que a distância aumenta a probabilidade de *turnover* também aumenta (Atef, Elzanfaly & Ouf, 2022; Rombaut & Guerry, 2021). Contrariamente, o estudo de Fallucchi *et al.* (2020) verifica que os colaboradores que vivem mais perto das instalações da organização são, em número absoluto, quem mais abandona o empregador.

Estas variáveis foram maioritariamente identificadas em estudos sobre o *turnover* (Rombaut & Guerry, 2018, 2021), pois os estudos que se focaram na intenção de *turnover*, ou noutros acontecimentos similares, ao utilizarem respostas a questionários e focarem a análise num alvo específico, identificaram como variáveis explicativas características específicas ao contexto da investigação e, por isso, menos generalizáveis (Vanegas, Wine & Drasgow, 2022; Ogbeibu *et al.*, 2022).

Tabela 2.9. Resultados dos artigos incluídos na RSL

ID	Algoritmos Preditivos										Modelo com melhor desempenho	Variáveis significativas
Artigo	KNN	XGBoost	RF	LASSO	CART	NB	RL	DT	SVM	Outro		
1		X	X								XGBoost: AUROC = 0,769; S = 0,793 e E = 0,679	Não viver com crianças; Problemas relacionais com a chefia; Risco de <i>mental distress</i> e anos de trabalho
2	X	X					X				XGBoost: PCCC = 0,823; Precisão = 0,838; S = 0,951 e F1-score = 0,891	N.E.
3											N.E.	Tópicos mais recorrentes: "job", "burnout", "workplace bullying", "job stress" e "emotional labor"
4	X		X								KNN: PCCC = 0,84; AUC = 0,79; Precisão = 0,47; S = 0,12; F-score = 0,187; FPR = 0,025, E = 0,974	Salário, Idade e Distância
5			X	X	X		X				RF: PCCC = 0,783; S = 1; E = 0,566; PPV = 0,25; NVP = 1; ϕ 0.3763	Pontuação AFQT, anos de escolaridade (e/ou possivelmente a idade), o Compromisso de servir, Coragem, Sociabilidade, Responsabilidade e Condição Física
6										Multi-label classification com RL e RF	RF: Precisão entre 0,809 e 1; S entre 0,828 e 1; F1-score entre 0,85 e 0,98; Haming Loss = 6,1	Apoio organizacional, Supervisão abusiva e Intervenção política.
7			X								Correlação de 0,35	Linguagem utilizada nas respostas sobre o comportamento passado e julgamento situacional. Os traços de personalidade "Aberto a experiências" e "Amabilidade".
8						X					S = 0,734 e Es = 0,841	Potencial de poupança de tempo, Duração média do turno, Total de turnos efetuados
9											N.A.	Gestão verde rígida de talentos, Gestão verde e flexível de talentos, Competência STARA do líder, Ambiente de trabalho e Cultura organizacional
10					X					XGBoost tree	CART: AUC = 0,716; Precisão = 0,79; S = 0,86; F1-score = 0,82; PCCC = 0,7496	Satisfação com o trabalho, com a organização, com a informação da administração, com a oportunidade de promoção e com o salário, Lealdade, Realização pessoal, Envolvimento nas decisões, Significado do trabalho, Oportunidade de desenvolvimento de competências e de liderança, Posse na organização, Promoção com base no mérito, Utilização de talentos, Desenvolvimento dos trabalhadores
11			X							Regression Trees	R ² = 0,32	Perceção de justiça, Oportunidades de carreira, Satisfação com o salário, Idade e Partilha entre líderes e colaboradores
12											R ² = 0,523	Out-degree centrality dos colaboradores-chave
13											N.A.	Individual annualized h-index. (h1a)
14							X				N.E.	Antiguidade, Salário, Distância trabalho-casa, Carro da empresa, Trabalho noturno e Doença.
15	X		X			X	X	X	X	GNB e LSVM	GNB: PCCC = 0,825; Precisão = 0,386; S = 0,541; E = 0,845; F1-score = 0,446	Rendimento mensal, Idade, Horas extraordinárias, Distância do domicílio.
16										CCIF	Valores entre -4,14% e 21,29 %	Estratégias mais relevantes para a retenção de trabalhadores: reconhecimento e a compensação. Variáveis significativas para essas estratégias: salário fixo, idade, comparativo, doença, nível de educação, distância até trabalho e salário condicional
17	X		X			X	X	X			NB: PCCC = 0,988	Experiência profissional relevante e histórico de saídas para encontrar melhores empregos
18											N.A.	N.E.
19										ANN	N.E.	As variáveis relacionadas com as atitudes no trabalho são mais significativas na previsão do turnover, enquanto que as características pessoais são mais significativas para a previsão da satisfação com o trabalho e o fit organização-colaborador
20			X				X			WQRF, C4.5 e BPNN	WQRF: PCCC = 0,928; Área ROC = 0,881; F-score = 0,711; S = 0,653	Rendimento mensal, Horas extraordinárias, Idade, Distância de casa, Anos na empresa e Percentagem de aumento salarial
21											N.A.	N.A.
22							X		X	MLP e Ensemble Model	Ensemble Model: PCCC = 0,84; Precisão = 0,682; S = 0,517; E = 0,932	N.E.
23							X				AUC = 0,7432	Género, Idade, Antiguidade, Nacionalidade, Estado civil, Salário, Carga de trabalho, Carro e telemóvel da empresa
24											N.A.	Satisfação com o trabalho e Suporte organizacional percebido

Notas: LASSO - least absolute shrinkage and selection operator regression; XGBoost - eXtreme gradient boosting; KNN - K-nearest neighbor; RF - Random Forests; RL - Regressão Logística; CART - Classification and Regression Tree; SVM - Support Vector Machines; LSVM - Linear Support Vector Machines; CCIF - Causal Conditional Inference Forests; ANN - Artificial neural networks; WQRF - Weighted quadratic random forest; MLP - Multilayer Perceptron; NB - Naive Bayes; GNB - Gaussian Naive Bayes; DT - Decision tree; BPNN - back propagation neural network; N.E. - Não específica; N.A. - Não aplicável; FPR - False-positive rate; S - Sensibilidade; E - Especificidade; PPV - Positive predictive value; NPV - negative predictive value; PCCC – percentagem de casos corretamente classificados

Respondendo à terceira questão específica da RSL: a literatura, a partir das investigações realizadas, identifica as variáveis relacionadas com o percurso profissional dos colaboradores como as mais explicativas para o *turnover* e restantes objetos de estudo, identificando em maior número as variáveis: idade, rendimento, distância entre a casa e o trabalho e a antiguidade na empresa.

2.2.4. Contributos e implicações

Na tabela 2.10 encontra-se sumariado os principais contributos e limitações identificados pelos artigos em análise, que responde à quarta e última questão específica da RSL.

Os contributos dos estudos podem ser, por exemplo, a investigação de um novo contexto de aplicação, que representa um terço dos contributos identificados e que diz respeito à aplicação de metodologias existentes a novas realidades. O estudo de Lee e Kim (2023) é uma das investigações que se enquadra neste âmbito pela utilização de *topic modelling* para prever o *turnover* de enfermeiros coreanos. Este contributo verifica-se em maior número nos estudos mais recentes. Outro contributo, também identificado em um terço da literatura, é a utilização de novas dimensões de variáveis, por exemplo, através da utilização de registos médicos para a criação de um modelo preditivo do *turnover* voluntário (Thompson, Holmgren & Ford, 2022).

Tabela 2.10. Contributos e limitações dos artigos incluídos na RSL

ID Artigo	Contributos	Limitações
1	Novo contexto de aplicação	Enviesamento pela amostra, Possibilidade de sobreajustamento, Generalização de resultados, Não inclusão de outras variáveis
2	Nova abordagem	N.E.
3	Novo contexto de aplicação	N.E.
4	Novo contexto de aplicação	N.E.
5	Novo contexto de aplicação	Interpretabilidade dos modelos, amostra não balanceada, enviesamento da amostra, generalização dos resultados, motivos para <i>attrition</i>
6	Nova dimensão de variáveis	Enviesamento da amostra, Não inclusão de outras variáveis
7	Novo contexto de aplicação	N.E.
8	Nova dimensão de variáveis	Generalização dos resultados, Seleção da amostra, Não comparação entre diferentes algoritmos
9	Nova dimensão de variáveis	Generalização dos resultados, Não inclusão de outras variáveis
10	Nova abordagem	Enviesamento pela amostra, Generalização dos resultados, Não inclusão de outras variáveis. O modelo não prevê o <i>turnover</i> real.
11	Nova abordagem	Não inclusão de outras variáveis, generalização de resultados, utilização de outros algoritmos
12	Novo contexto de aplicação	Generalização dos resultados, enviesamento pela amostra, avaliação de outras variáveis como mediadoras
13	Nova dimensão de variáveis	Generalização dos resultados, Enviesamento pela amostra, Assume a influência da performance das investigações nas decisões de contratação, retenção ou despedimento.
14	Nova dimensão de variáveis	Amostra limitada, generalização dos resultados, não identificação das razões para as diferenças encontradas.
15	Nova dimensão de variáveis	Aumentar e atualizar a amostra
16	Nova abordagem	Generalização dos resultados, amostra disponível, não balanceamento da amostra
17	Nova dimensão de variáveis	Imputação de valores em falta, generalização dos resultados, estudo focado apenas nas relações entre preditores e <i>target</i> .
18	Novo contexto de aplicação	Amostra limitada, generalização dos resultados, não identificação das razões para as diferenças encontradas.
19	Nova dimensão de variáveis	N.E.
20	Novo algoritmo preditivo	Generalização dos resultados, tempo de modelação
21	Novo contexto de aplicação	N.E.
22	Nova abordagem	N.E.
23	Nova abordagem	Generalização dos resultados, amostra disponível, não inclusão de outras variáveis
24	Nova abordagem	Limitação da amostra, não comparação de métodos, generalização dos resultados

Notas: N.E. - Não específica

Já as limitações dizem respeito a, por exemplo, a questões relacionadas com a amostra, que pode ser limitada ao nível do número de registos, enviesada ou não balanceada (Rombaut & Guerry, 2021; Matsuo *et al.*, 2023; Venegas, Wine & Drasgow, 2022). Apesar de cerca de 30% dos estudos não identificar limitações, a mais mencionada é a não generalização dos resultados obtidos.

2.3. Avaliação dos artigos

As pontuações de cada critério de avaliação, tal como definido na secção 2.1.4., são de 0, 0,5 ou 1 e encontram-se discriminadas na Tabela 2.11.

Através da avaliação dos artigos alvo da RSL é possível perceber que os artigos mais relevantes para a presente dissertação são o 1, 5, 6, e 20 (Matsuo *et al.*, 2023; Vanegas, Wine & Drasgow, 2022; Anand *et al.*, 2022; Gao, Wen & Zhang, 2019). O artigo melhor classificado nesta revisão foi o de Gao, Wen & Zhang (2019), que prevê o *turnover* através da aplicação de vários modelos preditivos baseados em diferentes algoritmos, utilizando a base de dados de uma empresa, com o objetivo de encontrar o modelo com melhor performance. Já os artigos 1, 5 e 6 dedicam as investigações à previsão do *turnover*, utilizando questionários como fonte de dados. No entanto, ao nível metodológico, também utilizam técnicas de *machine learning*, comparando modelos de diferentes algoritmos e identificando qual obteve o melhor desempenho. Todos eles evidenciam a avaliação dos diferentes modelos e quais as variáveis mais explicativas.

Tabela 2.11. Avaliação dos artigos incluídos na RSL

ID Artigo	Critérios												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
1	1	0,5	1	1	0,5	1	1	1	1	1	1	1	11
2	1	1	1	0,5	1	1	1	1	1	0	1	0	9,5
3	1	1	1	1	0	1	0	0	0	0,5	1	0	6,5
4	0,5	1	1	1	1	1	1	1	1	1	1	0	10,5
5	0,5	1	1	0,5	1	1	1	1	1	1	1	1	11
6	0,5	1	1	1	0,5	1	1	1	1	1	1	1	11
7	0,5	1	1	1	1	0,5	0,5	0,5	1	1	1	0	9
8	1	1	1	1	1	1	0,5	1	0	1	1	1	10,5
9	0,5	1	1	1	0	1	0	0	0	1	1	1	7,5
10	1	1	1	0,5	1	0,5	1	0,5	0,5	1	1	1	10
11	0,5	1	1	1	1	1	1	0	1	1	1	1	10,5
12	1	1	1	1	0	1	0	0,5	0	0,5	1	1	8
13	1	1	1	1	0	1	0	0	0	1	1	1	8
14	1	1	1	1	1	1	1	0	0	0,5	1	1	9,5
15	0,5	1	1	1	1	1	1	1	1	0,5	0,5	0,5	10
16	1	1	1	1	1	1	1	0,5	0	1	1	1	10,5
17	1	1	1	1	0,5	1	0,5	0,5	1	0,5	1	1	10
18	1	1	1	1	0,5	1	0	0	0	0,5	1	1	8
19	0,5	1	1	0	1	0,5	0	0	0	1	1	0	6
20	0,5	1	1	1	1	1	1	1	1	1	1	1	11,5
21	0,5	1	1	0,5	0	0,5	0	0	0	0	0,5	0	4
22	0,5	1	1	0	1	1	1	1	1	0	0	0	7,5
23	1	1	1	0,5	1	1	1	0,5	0	1	1	1	10
24	0,5	1	1	1	0	1	0	0	0	1	1	1	7,5

A partir da Tabela 2.12 é ainda possível identificar os critérios (operacionalizados em questões) com maior pontuação, que dizem respeito à identificação dos objetivos (critério 2) e da fonte dos dados utilizados (critério 3), à enumeração e justificação das variáveis utilizadas (critério 6) e indicação dos contributos (critério 11). Por outro lado, questões os critérios com menor pontuação são o 8 e o 9, que correspondem à identificação do desempenho dos modelos face aos critérios de avaliação e a comparação da performance de diferentes modelos.

Por fim, relativamente às questões específicas da RSL, a melhor avaliada é a primeira, que corresponde ao contexto de aplicação dos estudos, em que os artigos, na sua maioria, identificam de forma clara o seu âmbito e objetivos. Em oposição, a questão pior avaliada diz respeito aos resultados dos artigos, já que os critérios com menor pontuação dizem respeito a esta dimensão.

Tabela 2.12. Pontuação por critério de avaliação

Questões	Q1		Q2					Q3			Q4	
Critérios	1	2	3	4	5	6	7	8	9	10	11	12
Pontuação total	18	23,5	24	19,5	16	22	14,5	12	11,5	18	22	16,5
Média	20,75		19,2					13,83			19,25	

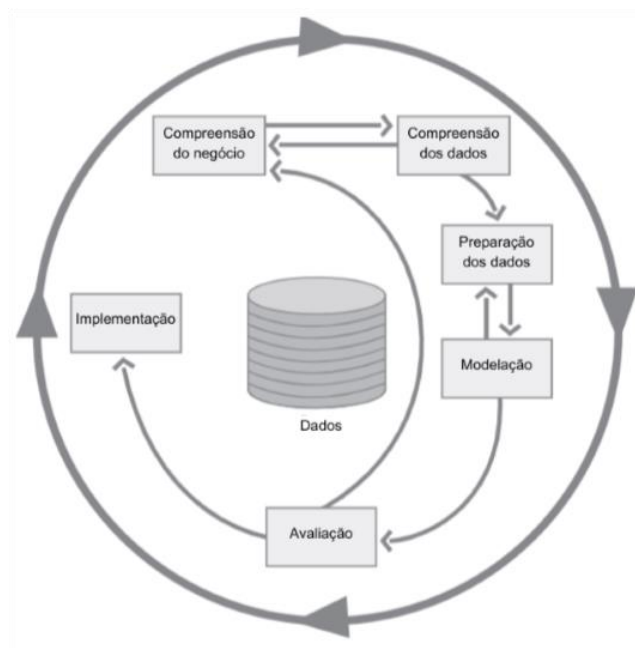
Capítulo 3 – Metodologia

A metodologia adotada é o CRISP-DM por seguir uma abordagem orientada por objetivos e, devido à sua maturidade, é uma metodologia amplamente aceita em projetos de *data mining* que recorrem à aplicação de algoritmos de *machine learning* (Ayele, 2020). Adicionalmente, esta metodologia é considerada a ideal para a descoberta de novos conhecimentos através da utilização de bases de dados (Martinez-Plumed *et al.*, 2019), enquadrando-se na presente investigação.

O CRISP-DM permite a obtenção de conhecimento ao nível do negócio e ao nível dos dados, guiando a preparação dos dados, modelação e validação dos modelos (Solano *et al.*, 2021), através de seis fases interativas que não seguem uma sequência rígida. Avançar e recuar entre as diferentes fases é considerado um requisito pois o resultado de cada uma determina qual a próxima fase a ser desenvolvida (Chapman *et al.*, 2000). As diferentes fases e a forma como estas interagem entre si encontram-se ilustradas na Figura 2.

Neste capítulo são descritas, de forma sistematizada, as tarefas realizadas em cada uma das fases, com maior ênfase nas 5 primeiras, que vão até à avaliação dos modelos. Todas as fases apresentam um grande foco no negócio, não só pela natureza da metodologia adotada, mas também por ser considerado fundamental para o sucesso da investigação a interação entre o negócio o processo de *data mining*.

Figura 3.1. Fases do CRISP-DM



Fonte: adaptado Chapman *et al.* (2000, p.12)

3.1. Compreensão do negócio

Tal como referido no Capítulo 1, esta investigação utiliza dados pessoais e profissionais de colaboradores e ex-colaboradores de diversas empresas, durante o período compreendido entre janeiro de 2005 e agosto de 2023, disponibilizados por um Grupo Económico Português. Atualmente, o Grupo apresenta um *headcount* próximo dos 10.000 colaboradores, mas apenas são considerados para este estudo os colaboradores que, desde 2005, estão ou estiveram vinculados às empresas portuguesas do Grupo.

O Grupo caracteriza-se pela detenção de várias empresas cujas atividades se enquadram em diferentes setores de negócio e com operações em diferentes países do mundo. Ao longo da sua história, foram criadas, adquiridas e vendidas diversas empresas sendo que, durante o período em análise, pertenciam ao Grupo vinte empresas portuguesas, cuja atuação se enquadrava em, pelo menos, uma das seguintes áreas: automóvel, concessões e serviços, construção, corporativo, distribuição, hotelaria e imobiliária. No que concerne aos países de atuação os colaboradores estão, ou estiveram, dispersos por vinte e quatro países, entre a Europa, África, América e o Médio Oriente. A nível nacional, as operações do grupo cobrem todos os distritos portugueses.

Na Tabela 3.1 encontra-se registado o número de colaboradores ativos e as saídas de colaboradores, quer voluntárias, quer involuntárias, ao longo dos 19 anos. Associado a estes registos apresentam-se, por ano, a taxa de *turnover*, a taxa de *turnover* voluntário e o peso das saídas voluntárias nas saídas totais do grupo, calculados a partir das fórmulas praticadas pelos profissionais do Grupo. As saídas involuntárias englobam motivos de cessação de contratos, como sejam revogação de contrato por mútuo acordo, caducidade do contrato a termo, extinção do posto de trabalho ou a denúncia do período experimental por parte do empregador. Já nas saídas voluntárias os motivos considerados são demissão por parte do trabalhador, denúncia do período experimental pelo trabalhador e abandono do trabalho.

No que concerne à evolução das saídas voluntárias, evidencia-se que estas registaram o pico de crescimento, quer a nível absoluto, quer a nível do seu peso no total das saídas, nos anos de 2009 e 2010. É importante referir que este período se enquadra nos anos de crescimento do número de colaboradores ativos, que à medida que foi decrescendo, a partir de 2014, também o número de saídas, voluntárias e involuntárias, diminuiu. Contudo, verifica-se que nos últimos anos as saídas voluntárias têm ganho cada vez mais expressão. Num período em que o Grupo estabilizou o despedimento de colaboradores e promoveu a efetivação da maioria da sua mão de obra, começou a registar um maior fluxo de saídas voluntárias. Em 2022 e nos oito primeiros meses de 2023 houve mais colaboradores a saírem de forma voluntária do que involuntária, o que significa que mais de metade das saídas desse período não foram planeadas pelas empresas do Grupo. Segundo os profissionais de

recursos humanos destas empresas, esta situação pode vir a comprometer seriamente o desempenho do negócio, pois existe, por um lado, a dificuldade de reter os colaboradores e, por outro, a dificuldade em contratar em tempo útil quem consiga assumir as funções e as tarefas de quem saiu. Desta forma, é importante para o Grupo entender o que leva os seus colaboradores a apresentarem a demissão.

Tabela 3.1. Evolução dos colaboradores ativos e saídas, de 2005 a 2023

Anos	Total Ativos	Saídas			Turnover (%)	Turnover voluntário (%)	Tx. Saídas voluntárias
		Total	Involuntárias	Voluntárias			
2005	2754	222	145	77	7,0%	2,4%	34,7%
2006	2931	611	344	267	19,3%	8,4%	43,7%
2007	3083	776	398	378	24,5%	11,9%	48,7%
2008	3271	732	315	417	23,1%	13,1%	57,0%
2009	3401	802	408	394	25,3%	12,4%	49,1%
2010	3237	862	597	265	27,2%	8,4%	30,7%
2011	3239	790	586	204	24,9%	6,4%	25,8%
2012	3505	604	462	142	19,0%	4,5%	23,5%
2013	3363	835	655	180	26,3%	5,7%	21,6%
2014	3600	761	451	310	24,0%	9,8%	40,7%
2015	3165	1129	871	258	35,6%	8,1%	22,9%
2016	2997	851	684	167	26,8%	5,3%	19,6%
2017	2798	745	489	256	23,5%	8,1%	34,4%
2018	2863	535	312	223	16,9%	7,0%	41,7%
2019	2391	835	616	219	26,3%	6,9%	26,2%
2020	2108	637	457	180	20,1%	5,7%	28,3%
2021	2233	336	191	145	10,6%	4,6%	43,2%
2022	2387	403	171	232	12,7%	7,3%	57,6%
2023	2469	294	143	151	9,3%	4,8%	51,4%

Notas: *Turnover* = total de saídas/efetivo médio; *Turnover voluntário* = saídas voluntárias/efetivo médio; Tx. Saídas voluntárias = saídas voluntárias/total de saídas

Para o desenvolvimento desta investigação, o Grupo disponibilizou diversos relatórios extraídos diretamente da sua plataforma de cadastro de colaboradores, que incluem informações relativas à sua vida pessoal e profissional. A plataforma apenas conta com registos a partir do ano 2005, limitando o período passível de ser estudado. Além disso, os relatórios que permitem extrair dados não contêm todas as informações associadas aos colaboradores e que podem influenciar a sua decisão de sair ou não das empresas, como sejam posse de telemóvel da empresa, acesso a carro da empresa, absentismo, dias de férias gozados e por gozar e horas efetivas de trabalho.

A compreensão e caracterização do problema em causa, bem como a definição dos objetivos que guiam esta investigação e que se enquadram nesta fase da metodologia foram já apresentados na introdução e são, agora, recordados. O principal objetivo de negócio diz respeito à obtenção de conhecimento útil capaz de prevenir e diminuir o *turnover* voluntário, sendo que o objetivo é cumprido com sucesso se, no mínimo, exista uma diminuição de 15% nos próximos dois anos face aos valores do ano de 2023. Já os objetivos analíticos correspondem à (1) caracterização demográfica e profissional

do *turnover* voluntário, isto é, os colaboradores que ficam na empresa e os que saem voluntariamente, (2) construção de modelo preditivo de classificação para o *turnover* voluntário, (3) identificação de perfis de colaboradores e (4) identificação das características dos colaboradores mais explicativas do *turnover* voluntário.

O primeiro objetivo analítico é cumprido aquando da identificação das características dos colaboradores que têm impacto significativo no *turnover* voluntário, para um nível mínimo de significância 0,05. O segundo objetivo é atingido com o recurso a algoritmos baseados em árvores de decisão, uma vez que o negócio considera essencial entender as razões que levam um colaborador a ficar na empresa ou a sair voluntariamente. Adicionalmente, as métricas de avaliação de modelos devem ser superiores a 75%, com destaque para a sensibilidade, precisão, especificidade e PCCC, em alinhamento com os estudos prévios, Tabela 2.9. Já o terceiro objetivo está cumprido com a interpretação das regras proposicionais de decisão extraídas do modelo que, de acordo com o negócio, têm de apresentar um suporte mínimo de 50 e uma confiança mínima de 90%. Por fim, o último objetivo é cumprido com a identificação dos atributos com maior importância relativa, atribuída pelo modelo, para a sua explicabilidade.

3.2. Compreensão dos dados

A segunda etapa da metodologia refere-se à compreensão dos dados, que inclui a sua recolha, descrição, exploração e verificação de qualidade. No seguimento do já enunciado, o conjunto de dados foi extraído do SAP, *software* utilizado pelas empresas do Grupo, que armazena as informações dos colaboradores desde a sua contratação até à cessação do seu contrato, em que os profissionais de recursos humanos registam informações, como sejam agregado familiar, situação fiscal, atualizações salariais, avaliação de desempenho ou até documentação profissional e pessoal relevante (contrato de trabalho, carta de condução e fichas de aptidão médica). O *software* vai armazenando todas as alterações realizadas ao cadastro dos colaboradores e permite exportar, para Excel, relatórios personalizados sobre qualquer informação relativa aos colaboradores.

No período em análise a informação encontra-se armazenada em dois servidores de SAP distintos. Um dos servidores guarda dados dos colaboradores ativos nos anos compreendidos entre 2005 e 2021 e o segundo servidor guarda dados dos ativos desde o ano 2021, o que leva necessariamente à produção de pelo menos dois relatórios para ser possível analisar todos os anos. Adicionalmente, os relatórios produzem tantas linhas quantas alterações tenha guardadas por colaborador. Por exemplo, se um colaborador desde que foi admitido alterou a sua categoria profissional três vezes, então no relatório existem três linhas para este colaborador, cada uma com uma categoria profissional distinta. Por consequência, adicionarem-se várias informações distintas no mesmo relatório, ao longo de

diversos anos, o grau de complexidade e de processamento do *software* aumenta consideravelmente. Por esta razão opta-se pela extração dos dados dividida em diversos relatórios.

No total extraem-se 28 ficheiros Excel que contêm informações de um total de 16084 colaboradores entre os anos de 2005 e 2023. Os ficheiros dividem-se pelos seguintes tipos de informações: abonos (3.265.866 linhas), atribuição organizacional (107.704 linhas), categoria profissional e habilitações (48.058 linhas), contratos (34.311 linhas), dados pessoais (31.426 linhas), datas (75.202 linhas), deficiência (29.557 linhas), local de residência (30.594 linhas), média de avaliação (18.113 linhas), medidas (44.951 linhas), remuneração (85.753 linhas) e tempo teórico de trabalho (176.754 linhas). Cada tipo conta com dois ficheiros diferentes, que correspondem a um relatório por servidor SAP e que, posteriormente, são agregados numa única tabela. A exceção são os abonos que pela quantidade de dados extraídos têm de ser divididos em três relatórios por servidor. Todas as tabelas dos ficheiros apresentam o número de colaborador e as datas de registo (data de início e data fim) associados a cada informação. Nas Tabelas 3.2, 3.3 e 3.4 são identificadas e descritas cada variável extraída, associada à sua tabela de origem e divididas em três dimensões: sociodemográfica, profissional e salarial.

Adicionalmente cada variável está caracterizada quanto ao seu tipo, papel no desenvolvimento da investigação e em relação às não respostas. No geral, a grande maioria das variáveis são qualitativas nominais e não apresentam não respostas. Quanto ao seu papel, existem variáveis consideradas diretamente para a fase de modelação e outras, que pelo tipo de informação que apresentam, são incluídas apenas para a fase de preparação. Nesta fase de análise, verifica-se também que parte das variáveis devem ser excluídas.

Tabela 3.2. Dicionário de dados das variáveis sociodemográficas extraídas

Dimensão Sociodemográfica					
Tabela Origem	Designação	Descrição	Tipo	Papel	Não Respostas (%)
Dados Pessoais	N_colab	Número de identificação	QN	Campo Chave	0,00%
Dados Pessoais	Género	Género	QN	Modelação	0,00%
Dados Pessoais	País_Nascimento	País de nascimento	QN	Preparação	0,00%
Dados Pessoais	Nacionalidade	Nacionalidade	QN	Preparação	0,00%
Dados Pessoais	N_Filhos	Número de filhos	QD	Modelação	0,00%
Dados Pessoais	Estado_Civil	Estado civil	QN	Modelação	0,10%
Dados Pessoais	Data_Nascimento	Data de nascimento	QC	Preparação	0,00%
Deficiência	%_Deficiência	Percentagem de deficiência	QC	Modelação	0,00%
Local residência	Localidade	Localidade de residência	QN	Excluída	0,60%
Local residência	País	País de residência	QN	Excluída	0,03%
Local residência	Cód Postal	Código postal da residência	QN	Excluída	0,60%
Local residência	num_cod_postal	Número do código postal	QN	Preparação	0,60%
Local residência	ext_cod_postal	Extensão do código postal	QN	Excluída	0,60%
Categ.Profiss. e habilitações	Habilitação_Literá	Habilitação literária	QN	Modelação	0,00%

Notas: QC - Quantitativa Contínua; QD - Quantitativa Discreta; QN - Qualitativa Nominal

Tabela 3.3. Dicionário de dados das variáveis profissionais extraídas

Dimensão Profissional					
Tabela Origem	Designação	Descrição	Tipo	Papel	Não Respostas (%)
Categ.Profiss. e habilitações	Categoria profissional	Categoria profissional	QN	Preparação	0,00%
Atribuição Organizacional	Empresa	Empresa	QN	Excluída	0,03%
Atribuição Organizacional	Área_RH	Identificação do local de trabalho (para RH)	QN	Modelação	0,03%
Atribuição Organizacional	Local de trabalho	Localidade do local de trabalho	QN	Excluída	1,20%
Atribuição Organizacional	Relação_Emprego	Tipo de relação de emprego	QN	Modelação	0,03%
Atribuição Organizacional	Setor	Setor da empresa empregadora	QN	Modelação	0,03%
Atribuição Organizacional	Atividade	Atividade desempenhada	QN	Excluída	47,00%
Atribuição Organizacional	País_Trab	País do local de trabalho	QN	Preparação	0,20%
Atribuição Organizacional	Concelho_Trab	Concelho do local de trabalho	QN	Modelação	0,03%
Atribuição Organizacional	Distrito_Trab	Distrito do local de trabalho	QN	Modelação	0,03%
Tipo de contrato	Tipo_Contrato	Tipo de contrato	QN	Modelação	0,00%
Medidas	Tipo de medida	Tipo de alterações no contrato	QN	Preparação	0,00%
Medidas	Motivo da medida	Motivo das alterações no contrato	QN	Preparação	0,00%
Especificações de datas	Tipo de data	Tipo de datas associadas ao contrato	QN	Preparação	0,00%
Especificações de datas	Data	Data associada ao tipo de data	QC	Preparação	0,00%
Tempo teórico	Hrs.mês	Total de horas indicativas de trabalho, por mês	QC	Excluída	0,00%
Tempo teórico	HrsSm	Total de horas indicativas de trabalho, por semana	QC	Excluída	0,00%
Tempo teórico	HorDiár.	Total de horas indicativas de trabalho, por dia	QC	Modelação	0,00%
Tempo teórico	Dias p semana	Total de dias indicativos de trabalho, por semana	QC	Modelação	0,00%
Tempo teórico	HrTrabAno	Total de horas indicativas de trabalho, por ano	QC	Excluída	0,00%
Média Avaliação	Avaliação_média	Avaliação média	QC	Modelação	0,00%

Notas: QC - Quantitativa Contínua; QD - Quantitativa Discreta; QN - Qualitativa Nominal

Tabela 1.4. Dicionário de dados das variáveis salariais extraídas

Dimensão Salarial					
Tabela Origem	Designação	Descrição	Tipo	Papel	Não Respostas (%)
Remuneração	Rúbrica salarial	Rúbrica salarial da remuneração base	QN	Preparação	1,80%
Remuneração	Montante	Montante pago associado à rúbrica salarial	QC	Preparação	1,80%
Remuneração	Moeda	Moeda associada ao montante pago	QN	Excluída	1,80%
Abonos	Rúbrica salarial	Código da rúbrica salarial relativa a abonos pagos	QN	Excluída	0,00%
Abonos	Descr.RubricaSalarial	Descrição da rúbrica salarial	QN	Preparação	0,00%
Abonos	Montante	Montante pago associado à rúbrica salarial	QC	Preparação	1,40%

Notas: QC - Quantitativa Contínua; QD - Quantitativa Discreta; QN - Qualitativa Nominal

Por se considerar relevante estabelecer uma ligação entre a residência dos colaboradores e os seus locais de trabalho, são apenas considerados os concelhos e os distritos das localidades identificadas, pois as localidades detalham até ao nível das aldeias, dificultando a comparação. Assim, no caso da tabela do local de residência apenas é considerada a parte do código postal que permite identificar a que concelho e distrito pertence a localidade de residência e é, também, excluído o local de trabalho.

Adicionalmente excluem-se as variáveis: atividade por apresentar uma taxa elevada de não respostas (47%), moeda, pois apresenta apenas uma categoria de resposta, o euro, Hrs.mês, HrsSm salarial, HrTrabAno e Rúbrica Salarial dos abonos, pelo seu carácter redundante face a outras variáveis consideradas para a modelação e preparação e, por fim, empresa por não ser relevante para a previsão das saídas, uma vez que a análise é feita para o Grupo como um todo e este pode vir a adquirir ou vender mais empresas.

Ainda nesta fase é necessário analisar a qualidade dos dados de forma a garantir que a informação é o mais consistente, o mais precisa e com o menor ruído possível. Neste sentido os registos são analisados de forma a identificar duplicados, erros na inserção de dados no *software*, informações semelhantes escritas de forma diferente, inconsistência das informações e os motivos para a existência de valores omissos.

3.3. Preparação dos dados

Na preparação dos dados é realizada a seleção, limpeza, transformação dos dados e a criação de novas variáveis consideradas relevantes para a fase de modelação. Por fim, toda a informação é integrada e fundida numa única tabela, a ser utilizada na próxima fase, a da modelação.

No seguimento da análise aos dados realizada na fase anterior, são eliminados registos duplicados e corrigidos erros de registo, principalmente ao nível das datas, onde se encontram registos com o ano de 2222, em vez de 2022, ou datas de nascimento iguais às datas de admissão. Alguns valores são alterados de forma a existir consistência nos dados, garantindo-se que não existe informação similar escrita de forma diferente, como por exemplo, os mesmos países escritos em idiomas distintos. Para evitar ambiguidade, também são alterados registos que, por exemplo, para um colaborador num determinado ano identificam como estando a contrato a termo ao mesmo tempo que apresentam uma data de passagem a efetivo (contrato sem termo) anterior ao ano em análise. Por fim, depois de analisados os motivos da existência de valores omissos, parte dos registos são preenchidos, nas variáveis passíveis de tal alteração, como a variável país, com o valor Portugal se as localidades registadas são nacionais.

Atendendo às variáveis descritas e selecionadas na compreensão dos dados, verifica-se que parte é considerada apenas para esta fase de preparação, ou seja, a sua informação é utilizada para criar novas variáveis. A Tabela 3.5 apresenta todas as variáveis selecionadas para a modelação relativamente à dimensão sociodemográfica e profissional, onde as novas são acompanhadas da fórmula utilizada para a sua obtenção. É importante referir que a tabela identifica, também, a designação e descrição de cada variável, o tipo, o papel e o número de categorias distintas das variáveis e a estatística descritiva associada a estas. Já as variáveis selecionadas relativas à dimensão salarial encontram-se discriminadas na Tabela 3.6.

No que concerne à tarefa de criação de variáveis, verifica-se que atributos associados a datas são pouco relevantes para a modelação, uma vez que apresentam um número de valores únicos muito elevado. No entanto, a sua informação pode ser relevante para a previsão. Assim, a partir da data de nascimento é criada a variável idade em anos, a data de passagem a efetivo (contrato sem termo) dá origem aos anos de efetividade e a data de admissão permite a criação da variável antiguidade. Outras

variáveis são criadas com o objetivo de permitir a análise da evolução ao longo de vários anos, como é o caso da evolução da percentagem de deficiência e da variação da remuneração entre diferentes anos. Para permitir uma comparação entre a residência dos colaboradores e o seu local de trabalho, existem novas variáveis que identificam se o colaborador vive e trabalha no mesmo concelho e no mesmo distrito. Já as medidas e as rubricas dos abonos apresentam, nos dados originais, diferentes informações numa única coluna pelo que existe a necessidade de dividir a informação por várias variáveis distintas, ou seja, é criada uma variável por cada tipo de medida e por cada tipo de rubrica de abono.

Tabela 3.5. Dicionário de dados das variáveis sociodemográficas e profissionais para modelação

Designação	Descrição	Fórmula Excel	Tipo	Papel	Nº categorias distintas	Distribuição
Género	Género	NA	QN	Input	2	HF: Masculino (87%); LF: Feminino (13%)
País nascimento	País de nascimento, agrupado	NA	QN	Input	6	HF: Portugal (90%); LF: América do Norte (0,2%)
Nacionalidade	Nacionalidade, agrupado	NA	QN	Input	5	HF: Portugêses (95%); LF: Asiático (0,1%)
Idade	Idade (anos)	=DATADIF(Data_Nascimento;Data_Análise;"y")	QC	Input	85	M: 42,17; DP: 11,14; Min:17 ; Máx: 105
Nº filhos	Número de filhos (unidade pessoa)	NA	QD	Input	12	M: 1,02; DP: 1,09; Min: 0; Máx: 11
Tem filhos	Ter filhos	=SE(N_Filhos > 0;1;0)	QN	Input	2	HF: Sim (59%); LF: Não (41%)
Estado civil	Estado civil	NA	QN	Input	6	HF: Casado (56%); LF: Separação de facto (0,3)
Grau deficiência	Percentagem de deficiência (%)	NA	QC	Input	19	M: 0,82%; DP: 7,29%; Min: 0%; Máx: 92%
Evolução defeciência	Evolução da percentagem de deficiência	=SE(%_Deficiência N > %_Deficiência anos antes; "Aumentou"; SE (%_Deficiência N < %_Deficiência anos antes; "Diminuiu"; "Igual"))	QN	Input	3	HF: Igual (98%); LF: Diminuição (0,07%)
Habilitação literár.	Habilitação literária	NA	QN	Input	13	HF: 3º Ciclo (25%); LF: Não sabe ler/escrever (0,17%)
Dimensão Profissional						
Nº mudanças categ. Profissional	Número de anos em que mudou de categoria, nos últimos 3 anos	Ver Anexo A	QD	Input	3	M:0,13; DP:0,37; Min:0; Máx:3
Sector	Sector da empresa empregadora	NA	QN	Input	12	HF:Construção (47%); LF:Contr. - Reabilitação (0,14%)
País trabalho	País do local de trabalho, agrupado	NA	QN	Input	5	HF:Portugal (84%); LF:Médio Oriente (0,4%)
Concelho igual	Concelho de residência e de trabalho igual	=SE(PROCV(num_cod_postal;lista concelhos;2;0)=Concelho_Trab;1;0)	QN	Input	2	HF:Não (88%); LF:Sim (12%)
Distrito igual	Distrito de residência e de trabalho igual	=SE(PROCV(num_cod_postal;lista distritos;2;0)=Distrito_Trab;1;0)	QN	Input	2	HF:Não (52%); LF:Sim (48%)
Área de RH	Identificação do local de trabalho (para RH)	NA	QN	Input	4	HF:Continente (82%); LF:Açores (0,3%)
Relação_Emprego	Tipo de relação de emprego	NA	QN	Input	4	HF: Padrão (84%); LF: Lic s/ vencimento (0,04%)
Tipo de contrato	Tipo de contrato	NA	QN	Input	3	HF:Sem termo (60%); LF:Termo Incerto (14%)
Antiguidade	Antiguidade (anos)	=DATADIF(FILTRAR(Data; Tipo de data = Data de Admissão; "NA");Data_Análise;"y")	QC	Input	68	M:8,52; DP:10,19; Min:0; Máx:78
Motivo alteração catg. Prof (ano N-1) anterior	Motivo alteração categoria, ano anterior	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-1))*(Tipo de medida=Alteração da categoria;"NA");;-1)	QN	Input	4	HF:Não aplicável (87%); LF:Promoção legal (2%)
Motivo alteração catg. Prof (ano N-2) antes	Motivo alteração categoria, 2 anos antes	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-2))*(Tipo de medida=Alteração da categoria;"NA");;-1)	QN	Input	4	HF:Não aplicável (95%); LF:Promoção legal (1%)
Motivo alteração catg. Prof (ano N-3) antes	Motivo alteração categoria, 3 anos antes	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-3))*(Tipo de medida=Alteração da categoria;"NA");;-1)	QN	Input	4	HF:Não aplicável (96%); LF:Promoção legal (1%)
Motivo transf. no Grupo (ano N-1) anterior	Motivo transferência no grupo, ano anterior	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-1))*(Tipo de medida=Transferência no Grupo;"NA");;-1)	QN	Input	4	HF:Não aplicável (93%); LF:Fim destacamento (2%)
Motivo transf. no Grupo (ano N-2) antes	Motivo transferência no grupo, 2 anos antes	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-2))*(Tipo de medida=Transferência no Grupo;"NA");;-1)	QN	Input	4	HF:Não aplicável (98%); LF:Fim destacamento (1%)
Motivo transf. no Grupo (ano N-3) antes	Motivo transferência no grupo, 3 anos antes	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-3))*(Tipo de medida=Transferência no Grupo;"NA");;-1)	QN	Input	4	HF:Não aplicável (97%); LF:Fim destacamento (1%)
Motivo manut. Contrato (ano N-1) anterior	Motivo manutenção contrato, ano anterior	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-1))*(Tipo de medida=Manutenção contrato;"NA");;-1)	QN	Input	3	HF:Não aplicável (77%); LF:Passagem a efetivo (4%)
Motivo manut. Contrato (ano N-2) antes	Motivo manutenção contrato, 2 anos antes	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-2))*(Tipo de medida=Manutenção contrato;"NA");;-1)	QN	Input	3	HF:Não aplicável (90%); LF:Passagem a efetivo (2%)
Motivo manut. Contrato (ano N-3) antes	Motivo manutenção contrato, 3 anos antes	=INCLUIR(FILTRAR(Motivo da medida;(N_colab=Nº colab)*(data <= data_análise(n-3))*(Tipo de medida=Manutenção contrato;"NA");;-1)	QN	Input	3	HF:Não aplicável (93%); LF:Passagem a efetivo (2%)
Anos efetivo	Número de anos em contrato efetivo	=SE(FILTRAR(Data; Tipo de data = Passagem a efetivo; "NA")="NA";0;ANO(Data_Análise)-ANO(FILTRAR(Data; Tipo de data = Passagem a efetivo; "NA")))	QC	Input	69	M:7,16; DP:10,59; Min 0; Máx:78
Regresso baixa prolongada	Regresso de baixa prolongada nos últimos 3 anos	=FILTRAR(Data; Tipo de data = Regresso aus.prolong; "NA")	QN	Input	2	HF:Não (99,7%); LF:Sim (0,3%)
Horas de trabalho por dia	Total de horas indicativas de trabalho, por dia	NA	QC	Input	13	M:8,01; DP:0,58; Min:4; Máx:9,7
Dias de trabalho por semana	Total de dias indicativos de trabalho, por semana	NA	QC	Input	7	M:5,01; DP 0,36; Min:2; Máx:7
Avaliação média	Avaliação média	NA	QC	Input	1798	M:3,17; DP:1,77; Min:0; Máx:5

Notas: NA – Não aplicável; QC - Quantitativa Contínua; QD - Quantitativa Discreta; QN - Qualitativa Nominal; M - Média; DP - Desvio Padrão; Min - Mínimo; Máx - Máximo; HF - *Highest Frequency*; LH - *Lower Frequency*; Os valores calculados referem-se à base de dados final.

Tabela 3.6. Dicionário de dados das variáveis salariais para modelação

Dimensão Salarial							
Designação	Descrição	Fórmula Excel	Tipo	Papel	Nº categorias	Distribuição	
Remuneração (ano N-1)	Valor remuneração regular, ano anterior	=SOMA(INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-1))*(Rubrica salarial=Vencimento;"NA");-1);INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-1))*(Rubrica salarial=Insenção de Horário;"NA");-1))	QC	Input	1973	M:1474,34; DP:1254,18; Min:361,85; Máx:11250	
Remuneração (ano N-2)	Valor remuneração regular, 2 anos antes	=SOMA(INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-2))*(Rubrica salarial=Vencimento;"NA");-1);INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-2))*(Rubrica salarial=Insenção de Horário;"NA");-1))	QC	Input	1802	M:1602,75; DP:1334,85; Min:365; Máx:11250	
Remuneração (ano N-3)	Valor remuneração regular, 3 anos antes	=SOMA(INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-3))*(Rubrica salarial=Vencimento;"NA");-1);INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-3))*(Rubrica salarial=Insenção de Horário;"NA");-1))	QC	Input	1836	M:1622,5; DP:1343,54; Min:11250; Máx:375	
Isenção de horário (ano N-1)	Recebimento IHT, ano anterior	=SE(INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-1))*(Rubrica salarial=Insenção de Horário;0);-1)>0;1;0)	QN	Input	2	HF:Sim (51%); LF:Não (49%)	
Isenção de horário (ano N-2)	Recebimento IHT, 2 anos antes	=SE(INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-2))*(Rubrica salarial=Insenção de Horário;0);-1)>0;1;0)	QN	Input	2	HF:Não (68%); LF:Sim (32%)	
Isenção de horário (ano N-3)	Recebimento IHT 3 anos antes	=SE(INCLUIR(FILTAR(Montante;(N_colab=Nº colab)*(data <= data_análise(n-3))*(Rubrica salarial=Insenção de Horário;"NA");-1)>0;1;0)	QN	Input	2	HF:Não (73%); LF:Sim (27%)	
Varição remuneração anos N-1 e N-2	Varição da remuneração entre o ano -1 e -2 (%)	=(RB_N1-RB_N2)/RB_N2	QC	Input	3340	M:6%; DP:10%; Min:-40%; Máx:200%	
Varição remuneração anos N-2 e N-3	Varição da remuneração entre o ano -2 e -3 (%)	=(RB_N2-RB_N3)/RB_N3	QC	Input	3007	M:4%; DP:4%; Min:-69%; Máx:100%	
Anos com horas extra	Total de anos com realização de horas extras, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:0 (48%); LF:2 (5%)	
Último ano com horas extra	Último ano de realização de horas extra, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Horas extra;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	2	HF:1 (52%); LF:0 (48%)	
Anos com ajudas de custo	Total de anos com recebimento de ajudas de custo, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Ajuda custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Ajuda custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Ajuda custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:3 (43%); LF:2 (6%)	
Último ano com ajudas de custo	Último ano com recebimento de ajudas de custo, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Ajudas custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Ajudas custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Ajudas custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Ajudas custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Ajudas custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Ajudas custo;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	4	HF:0 (45%); LF:2 (7%)	
Anos com distribuição de resultados	Total de anos com recebimento de distribuição resultados, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	3	HF:1 (55%); LF:0 (0,01%)	
Último ano com distr. resultados	Último ano com recebimento de distribuição resultados, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Distr.resultados;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	4	HF:3 (47%); LF:2 (8%)	
Anos com prémio	Total de anos com recebimento de prémios, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:1 (55%); LF:0 (0,01%)	
Último ano com prémio	Último ano com recebimento de prémios, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Prémios;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	4	HF:1 (75%); LF:2 e 3 (0,01%)	
Anos com excedente subs. almoço	Total de anos com recebimento de excedente sub. almoço, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Excedente sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Excedente sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Excedente sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:3 (38%); LF:2 (7%)	
Último ano com exced. subs. almoço	Último ano com recebimento de excedente sub. almoço, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Excedente Sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Excedente Sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Excedente Sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Excedente Sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Excedente Sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Excedente Sub.alim.;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	3	HF:1 (76%); LF:2 (0,02%)	
Anos com subsídio de almoço	Total de anos com recebimento de subsídio almoço, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:3 (47%); LF:2 (7%)	
Último ano com subsídio de almoço	Último ano com recebimento de subsídio almoço, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Sub.almoço;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	3	HF:1 (88%); LF:2 (0,01%)	
Anos com subsídio transporte	Total de anos com recebimento de subsídio transporte, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:0 (50%); LF:2 (4%)	
Último ano com subsídio transp.	Último ano com recebimento de subsídio transporte, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Sub.transporte;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	4	HF:0 (50%); LF:2 e 3 (0,003%)	
Anos com trabalho por turnos	Total de anos com realização de turnos, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:0 (84%); LF:2 (2%)	
Último ano com trabalho por turnos	Último ano com realização de turnos, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Turnos;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	2	HF:0 (84%); LF:1 (16%)	
Anos com trabalho noturno	Total de anos com realização de trabalho noturno, nos últimos 3 anos	=SOMA(INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1))	QC	Input	4	HF:0 (88%); LF:2 (1%)	
Último ano com trabalho noturno	Último ano com realização de trabalho noturno, nos últimos 3 anos	=SE(INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);1;1;SE(E(INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);1);2;SE(E(INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-1)));0;-1);0;INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-2)));0;-1);0;INCLUIR(FILTAR(Trab.noturno;(N_colab=Nº colab)*(Ano = ANO(data_análise(n-3)));0;-1);1);3;=)))	QC	Input	2	HF:0 (88%); LF:1 (12%)	

Notas: QC - Quantitativa Contínua; QN - Qualitativa Nominal; NA - Não aplicável; M - Média; DP - Desvio Padrão; Min - Mínimo; Máx - Máximo; HF - Highest Frequency; LH - Lower Frequency; Os valores calculados referem-se à base de dados final.

Ainda no processo de criação de variáveis, é definida e criada a variável *target* que, com base na contextualização da investigação e no conjunto de dados disponível tem por base os motivos da medida de cessão dos contratos. O *target* é uma variável dicotómica que assume o valor 1, nos colaboradores associados à cessação de contrato por motivos voluntários (demissão por parte do trabalhador, denúncia do período experimental pelo trabalhador e abandono do trabalho), e 0, nos colaboradores ativos nos diferentes anos em análise.

Tabela 1.7. Caracterização da variável *target*

<i>Target</i>				
Designação	Tipo	Papel	Nº categorias distintas	Distribuição
<i>Turnover</i> voluntário	QN	<i>Target</i>	2	HF: Ativo (91%); LF: Saída voluntária (9%)

Notas: QN - Qualitativa Nominal; HF - *Highest Frequency*; LH - *Lower Frequency*; Os valores calculados referem-se à base de dados final.

Quanto à transformação dos dados, esta passou pela alteração de valores associados a algumas variáveis, de forma a uniformizar as suas informações. A título de exemplo, a habilitação literária passou da identificação de diferentes licenciaturas e mestrados para identificar apenas o nível de habilitação do colaborador: terceiro ciclo do ensino básico, ensino secundário, licenciatura, mestrado e restantes níveis. Já as variáveis associadas aos países de nascimento, de trabalho e nacionalidades também passaram a identificar apenas o continente ou região a que estes pertencem.

Por fim, a base de dados final para a próxima fase, modelação, é construída por forma a equilibrar o número de colaboradores ativos e colaboradores que saíram voluntariamente. Com esse objetivo contabilizam-se quantos colaboradores saíram voluntariamente em cada um dos anos em análise e criam-se 10 grupos de amostras de colaboradores ativos por ano, selecionados aleatoriamente e com dimensão igual ao número de saídas voluntárias. Por exemplo, em 2022 saíram voluntariamente 232 colaboradores e, por isso, criam-se aleatoriamente 10 grupos de colaboradores ativos no ano de 2022. Esta abordagem permite não só comparar a realidade dos colaboradores por anos, bem como, na modelação, permite que se selecione e se teste a performance dos modelos. A utilização de um menor ou maior número de grupos de amostra permite avaliar o impacto do rácio de colaboradores ativos e colaboradores que saíram voluntariamente, sendo que caso o rácio for de 1, então o *target* está balanceado, ou seja, o número de colaboradores ativos é igual ao número de saídas voluntárias.

Para as variáveis que apresentam informações de histórico, são apenas considerados os últimos três anos face ao ano em análise. Desta forma, a base de dados contém os colaboradores ativos e que saíram voluntariamente desde 2008 a 2023 e os dados de 2005 a 2007 aparecem apenas enquanto histórico.

Como os colaboradores podem estar ativos em mais do que um ano, o mesmo colaborador pode aparecer várias vezes na base de dados. No entanto, as suas informações refletem a sua situação

no ano a que pertence a amostra da qual faz parte, ou seja, um colaborador pode aparecer numa amostra de 2023 e apresentar dois filhos enquanto na amostra de 2020 apresenta apenas um filho. Adicionalmente, por forma a adequar o mais possível o estudo à realidade atual do Grupo, apenas se consideram os colaboradores e ex-colaboradores das empresas que atualmente se encontram ativas e integradas no Grupo. É, no entanto, importante salvaguardar que, como é incluído no estudo o histórico dos colaboradores, existe referência a empresas que, num determinado momento do período em análise, pertenciam ao Grupo e eram as empregadoras de alguns dos colaboradores.

Incluído nesta fase está, também, o tratamento de dados considerados como não respostas e valores extremos. Após análise, verifica-se que a percentagem destes não é significativa para base de dados e considerando que o algoritmo aplicado, baseado em árvores de decisão, lida bem com a existência deste tipo de valores (Constantino, 2020), opta-se por não substituir ou eliminar estes registos.

3.4. Modelação e avaliação

A fase de modelação pressupõe a seleção das técnicas de modelação a utilizar, sempre considerando os objetivos de negócio e analíticos.

Assim, para o objetivo analítico (1) de caracterizar o *turnover* voluntário segundo as características demográficas e profissionais dos colaboradores recorre-se a técnicas de estatística descritiva bivariada (frequências, medidas de síntese, medidas de associação) e, para avaliar a magnitude das diferenças entre quem sai voluntariamente e quem fica na empresa, recorre-se a testes estatísticos, nomeadamente ao teste de independência do qui-quadrado e ao teste t para duas amostras independentes, e considera-se um nível de significância de 0,05. Para classificar o nível de significância, foi considerada uma escala onde, se o nível de significância estiver entre: 0 e 0,15, é considerado muito fraco; 0,15 e 0,3, é considerado fraco; 0,3 e 0,7, é considerado moderado; 0,7 e 0,9, é considerado forte; 0,9 e 1, é considerado muito forte.

Para os restantes três objetivos analíticos (criar modelo preditivo do *turnover*, identificar perfis de colaboradores propensos à saída e, também, a ficarem na empresa, e identificar as características demográficas e profissionais dos colaboradores que mais contribuem para a explicação do *turnover* voluntário) recorre-se a técnicas de análise preditiva com aprendizagem supervisionada, nomeadamente, a árvores de decisão para problemas de classificação, já que é utilizada uma base de dados onde os dados se encontram devidamente identificados para treinar os algoritmos, e em que se pretende classificar os colaboradores como ativos (*target* igual a zero) ou como inativos por saírem voluntariamente (*target* igual a um). Neste caso, os algoritmos, em fase de treino, identificam as combinações de variáveis mais associadas ao *target* e, na fase de teste, classificam, com base na

aprendizagem, novos registos e comparam os resultados com as classificações reais (Larose & Larose, 2015).

De entre os diferentes tipos de algoritmos utilizados para problemas de classificação supervisionada, opta-se pela utilização daqueles que se baseiam em árvores de decisão, pois criam modelos que apresentam uma maior interpretabilidade, uma vez que permitem entender como foram contruídas as regras de decisão (Larose & Larose, 2015). Desta forma é possível perceber o porquê da classificação realizada pelos modelos, característica crucial na ótica do negócio, que pretende entender quais razões que levam os colaboradores a ficar ou a sair voluntariamente do Grupo.

Para a modelação são utilizados vários algoritmos de árvores de decisão para ser possível seleccionar aquele que obtém a melhor performance. Os algoritmos considerados são o CART, CHAID, QUEST, Tree-AS e C5.0, disponíveis para modelação na plataforma analítica SPSS Modeler (versão, 18.4) e que são caracterizados pela International Business Machines Corporation (IBM, 2021) como:

- CART, desenvolvido por Breiman *et al.* (1984), é um algoritmo que gera uma árvore de decisão para classificar futuras observações utilizando repartição recursiva para dividir os registos de treino em segmentos, minimizando impurezas. Um nó da árvore é puro se a totalidade dos casos desse nó se enquadrarem numa categoria específica de resposta. Suporta variáveis qualitativas e quantitativas e todas as suas divisões são binárias (dois subgrupos). São modelos robustos na presença de dados omissos e grande número de variáveis, não necessitam de muito tempo para treinar e tendem a ser mais fáceis de interpretar do que outros modelos;
- CHAID, desenvolvido por Kass (1980), gera árvores de decisão com recurso à estatística do teste de independência do qui-quadrado para identificar as melhores divisões e é capaz de gerar árvores não binárias (as divisões têm mais do que duas ramificações). Tal como o CART, funciona bem com qualquer tipo de variável de entrada;
- QUEST, desenvolvido por Loh e Shih (1997) utiliza um método de classificação binário, com vista à redução do tempo de processamento de grandes análises de classificação e favorecer entradas com mais divisões. As divisões são todas binárias e os campos de entrada suportam qualquer tipo de variável. Utiliza testes estatísticos para decidir se as variáveis são ou não utilizadas e aplica critérios diferentes para a seleção de entrada e da divisão;
- C5.0, desenvolvido por Quinlan (1993), cria uma árvore de decisão através da divisão da amostra com base na variável que oferece mais informações em cada nível e as divisões múltiplas são permitidas. Cria modelos robustos na presença de dados omissos e grande número de variáveis, sem necessidade de utilizar muito tempo na fase de treino;
- Tree-AS é semelhante ao CHAID, contudo é utilizado para processar *big data* de forma a criar uma única árvore.

Para cada um dos algoritmos são utilizadas diferentes parametrizações, reportando-se apenas as melhores parametrizações de cada algoritmo (descritas na Tabela 3.8)⁷. Os modelos podem ser construídos utilizando apenas uma única árvore ou recorrendo a *ensemble learning*, métodos que agrupam um conjunto de aprendizagens e onde se enquadram as técnicas *boosting* e *bagging* (IBM, 2021). Enquanto o *boosting* utiliza um método que visa melhorar a previsão do modelo, o *bagging* procura aumentar a estabilidade do modelo, tornando-o mais generalizável (Opitz & Maclin, 1999). Ainda nas parametrizações, é possível definir a profundidade máxima da árvore, onde uma árvore com menos níveis é menos complexa e mais generalizável, e o número mínimo de casos por Nó Pai e por Nó Filho. Este processo de poda das árvores tem em vista melhorar a previsão, aumentar a generalização e evitar o sobreajustamento dos modelos aos dados de treino, procurando-se fixar valores que permitam não só melhorar a performance dos modelos, mas também que se adequem ao negócio. Adicionalmente, são utilizados, para parte dos modelos, custos de má classificação, aumentando a importância da categoria do *target* menos representativa, por forma a equilibrar a capacidade preditiva dos modelos para os colaboradores que permanecem ativos (especificidade) e para os colaboradores que saem voluntariamente (sensibilidade). Importa referir que considerando o carácter iterativo da metodologia, a possibilidade de refazer tarefas ou até regressar a fases anteriores é uma possibilidade que se mantém em aberto.

Tabela 3.8. Parametrizações dos melhores modelos de cada algoritmo

	Modelos				
	E	J	O	T	Y
Algoritmo	CART	CHAID	QUEST	Tree-AS	C5-0
Ensemble	Boosting	Bagging	Boosting	Não	Não
Profundidade máxima	5	6	7	6	5
Nº casos Nó Pai	2%	50	2%	2%	2%
Nº casos Nó Filho	1%	25	1%	1%	1%
Custos de má classificação	Não	Sim	Sim	Não	Sim
Amostra em análise	0 e 2	0 e 2	0 e 2	0 e 2	0 e 2

Por forma a verificar o comportamento dos modelos face aos diferentes grupos de amostras criados na fase de preparação, são seleccionadas diferentes amostras em cada um dos modelos. A amostra zero é sempre seleccionada, pois contém todos os registos de colaboradores com *turnover* voluntário (*target* igual a 1), e ao ser seleccionada juntamente com mais uma amostra, por exemplo a amostra um que é composta por colaboradores ativos (*target* igual a 0), torna o conjunto de dados seleccionado naturalmente balanceado (o número de casos com *target* igual a 0 e com *target* igual a 1 é igual).

⁷ Ver no Anexo B algumas das parametrizações testadas para cada algoritmo e os respetivos resultados, que conduziram à seleção do melhor modelo de cada algoritmo.

Para além da segmentação dos dados pelos diferentes grupos de amostras, cada conjunto de dados selecionado para a aplicação dos modelos é dividido em amostra de treino (70%) e amostra de teste (30%), prática usual quando se têm grandes amostras (Vrigazova, 2021). Trata-se de uma separação necessária para a avaliação da performance desses mesmos modelos em novos dados de teste.

No que toca à avaliação, esta é a fase seguinte da metodologia adotada, onde os modelos são avaliados através de diferentes métricas baseadas na matriz de confusão. A partir desses resultados os modelos são, ou não, aprovados caso tenham atingidos os objetivos analíticos e de negócio. A matriz de confusão está representada na Tabela 3.9 e a Tabela 3.10 resume as fórmulas utilizadas para calcular cada métrica (Berthold *et al.*, 2020).

Tabela 3.9. Matriz de confusão

		Turnover voluntário previsto	
		0 – ativo (negativo)	1 - saída (positivo)
Turnover voluntário observado	0 – ativo (negativo)	Verdadeiros negativos (VN)	Falsos positivos (FP)
	1 - saída (positivo)	Falsos negativos (FN)	Verdadeiros positivos (VP)

Tabela 3.10. Métricas de avaliação usuais para modelos de classificação

Métrica	Fórmula	Significado
PCCC	$PCCC = \frac{VP + VN}{VP + VN + FP + FN}$	Performance geral do modelo (taxa geral de acerto).
Sensibilidade	$S = \frac{VP}{VP + FN}$	Capacidade de o modelo identificar corretamente as saídas voluntárias.
Especificidade	$E = \frac{VN}{VN + FP}$	Capacidade de o modelo identificar corretamente os colaboradores que não saem.
Precisão	$P = \frac{VP}{VP + FP}$	Proporção entre os casos corretamente classificados como saída e o total de casos classificados como saída.

Outra métrica tida em consideração é a AUC (Area Under the ROC Curve), isto é, a área por baixo da curva ROC, curva que reflete graficamente o desempenho de um modelo de classificação, ou seja, representa a relação entre a sensibilidade e especificidade. Assim, quanto maior for o valor da AUC, melhor é a precisão global do modelo. Todas as métricas são aplicadas aos conjuntos de treino e de teste para que seja possível entender como podem ser melhorados os modelos, nomeadamente face a um possível sobreajustamento, em que os resultados são bons no treino e maus no teste, ou a um possível subajustamento, em que os resultados mostram que o treino pode ser melhorado. Ambos podem ser corrigidos através do processo de poda já descrito ou do ajustamento de outros parâmetros.

Assim, o modelo selecionado deve apresentar valores acima dos 75% nas métricas de avaliação (sensibilidade, precisão, especificidade e PCCC), respeitando o critério de sucesso definido na fase de

compreensão do negócio, e, permitir identificar claramente as características dos colaboradores que mais contribuem para explicar as previsões de *turnover* e, bem como, perfis de colaboradores propensos à saída que correspondam a regras com, pelo menos, 90% de confiança e um suporte de, pelo menos, 50 colaboradores. Em suma, os resultados devem gerar insights que permitam ao negócio tomar medidas que levem à redução da taxa de *turnover* voluntário.

3.5. Implementação

A última fase da metodologia CRISP-DM é denominada por implementação e diz respeito à aplicação da investigação desenvolvida, enunciando conclusões, contributos, limitações e recomendações. Todo este conteúdo encontra-se desenvolvido nos próximos capítulos, Discussão de Resultados e Conclusões, e, em particular, a implementação é resultado da elaboração escrita do documento de dissertação e a apresentação dos resultados da investigação, quer para o Grupo, quer para a comunidade científica e académica.

Adicionalmente, está disponível para o Grupo o modelo final caso este pretenda implementá-lo. Para apoiar a tomada de decisão e identificar que colaboradores, atualmente ativos, estão mais propensos a ficar ou a sair voluntariamente bastava aplicar o modelo aos colaboradores atuais, depois de calcular todos os preditores do *turnover*, para gerar essa identificação.

Capítulo 4 – Resultados

No seguimento da metodologia, o presente capítulo procura apresentar e interpretar os resultados provenientes das ações ilustradas no capítulo anterior. Adicionalmente, os resultados são discutidos tendo em conta os objetivos definidos.

4.1. Caracterização do *turnover* voluntário

Para a caracterização tanto a nível sociodemográfico, como profissional e salarial do *turnover* voluntário, são identificadas as características dos colaboradores, por cada uma das dimensões, que apresentam um impacto significativo no *turnover* voluntário. Desta forma e para se ir ao encontro do primeiro objetivo analítico, são aplicadas técnicas de estatística descritiva e testes estatísticos, identificados na metodologia.

Quanto à distribuição das características demográficas por *turnover* voluntário evidencia-se que todas as variáveis demográficas se relacionam significativamente com o *turnover* voluntário ($p < 0,001$). No entanto, apenas a idade se relaciona de forma fraca com o *turnover* (medida de associação entre 0,15 a 0,30), sendo que em todas as outras a relação é muito fraca (medida de associação menor que 0,15). De facto, a idade, em termos médios, é significativamente superior nos colaboradores que ficam ($M=42,8$ anos) quando comparado com os que saem voluntariamente ($M=36$ anos). Destaca-se, ainda, que no estado civil a maioria dos ativos (57,4%) é casado, seguindo-se os solteiros (32,2%), enquanto nos colaboradores que saem voluntariamente a maioria é solteiro (51,3%), seguindo-se os casados (38,7%). Em todas as outras características as diferenças entre ativos e inativos são mais ténues (Tabela 4.1).

No que concerne à distribuição das características profissionais pelo *turnover* voluntário, verifica-se que entre as 23 variáveis profissionais apenas três não se relacionam significativamente com o *turnover* voluntário ($p > 0,05$), nomeadamente o concelho de residência e de trabalho igual, o número de mudanças de categoria profissional, nos últimos 3 anos, e o regresso de baixa prolongada. Quanto às restantes, destacam-se as variáveis tipo de contrato, antiguidade e avaliação média por apresentarem uma relação fraca ou moderada. No tipo de contrato, a maioria dos ativos tem um contrato sem termo (63,7%), sendo que para os colaboradores que saem voluntariamente este tipo de contrato representa menos de 20%. Nestes colaboradores que saem voluntariamente, o tipo de contrato mais recorrente é o contrato a termo certo (51,9%). Na antiguidade, destaca-se uma grande diferença entre os colaboradores ativos e que saem voluntariamente, onde os ativos apresentam uma antiguidade de 9,2 anos, mais 7,6 anos do que os que saem voluntariamente, que tendem a sair logo ao fim do primeiro ano.

A avaliação média, que apresenta uma relação moderada (medida de associação entre 0,30 a 0,70), também apresenta uma diferença significativa entre os colaboradores ativos, cuja avaliação média é de 3,33, e os colaboradores que saem voluntariamente, onde a sua avaliação média é de 1,61. As distribuições e diferenças das restantes variáveis podem ser consultadas no Anexo C.

Tabela 4.1. Distribuição das características sociodemográficas por *turnover* voluntário

		Ativos (n=28940)		Saídas Voluntárias (n=2897)		Total (n=31837)		Teste estatístico
		Colaboradores	%	Colaboradores	%	Colaboradores	%	Medida de associação
Género	Feminino	3778	13,05	474	16,36	4252	13,36	$\chi^2(1)=24,89$; $p<0,001$ Cramer V=0,028
	Masculino	25162	86,95	2423	83,64	27585	86,64	
País Nascimento	África	1468	5,07	159	5,49	1627	5,11	$\chi^2(5)=130,337$; $p<0,001$ Cramer V=0,064
	América do Norte	48	0,17	3	0,1	51	0,16	
	América do Sul	394	1,36	107	3,69	501	1,57	
	Ásia	56	0,19	8	0,28	64	0,2	
	Europa	666	2,3	116	4	782	2,46	
	Portugal	26308	90,91	2504	86,43	28812	90,5	
Nacionalidade	Africano	722	2,49	85	2,93	807	2,53	$\chi^2(4)=79,817$; $p<0,001$ Cramer V=0,050
	Americano	282	0,97	72	2,49	354	1,11	
	Asiático	39	0,13	4	0,14	43	0,14	
	Europeu	432	1,49	76	2,62	508	1,6	
	Português	27465	94,9	2660	91,82	30125	94,62	
Idade		M=42,79	DP=11,06	M=35,95	DP=9,91	M=42,17	DP=11,14	t(3658,43)=35,016; $p<0,001$ eta=0,191
Habilitação Literária	Não sabe ler nem escrever	55	0,19	0	0	55	0,17	$\chi^2(12)=163,818$; $p<0,001$ Cramer V=0,072
	Sabe ler e escrever	223	0,77	8	0,28	231	0,73	
	1º Ciclo - Ensino Básico (4º ano)	3416	11,8	207	7,15	3623	11,38	
	2º Ciclo - Ensino básico	3122	10,79	329	11,36	3451	10,84	
	3º Ciclo - Ensino Básico (9º ano)	7050	24,36	780	26,92	7830	24,59	
	Ensino Secundário (12º ano)	5656	19,54	641	22,13	6297	19,78	
	Ensino Secundário (Profissional)	49	0,17	15	0,52	64	0,2	
	Escola Profissional (nível II ou III)	347	1,2	64	2,21	411	1,29	
	Ensino Técnico	562	1,94	47	1,62	609	1,91	
	Bacharelato	739	2,55	41	1,42	780	2,45	
	Licenciatura	6242	21,57	556	19,19	6798	21,35	
	Especialização	61	0,21	12	0,41	73	0,23	
	Mestrado	1418	4,9	197	6,8	1615	5,07	
Estado civil*	Casado	16579	57,37	1120	38,73	17699	55,67	$\chi^2(5)=453,107$; $p<0,001$ Cramer V=0,119
	Divorciado	1733	6	160	5,53	1893	5,95	
	Soltei	9307	32,2	1483	51,28	10790	33,94	
	Sp Fac	94	0,33	7	0,24	101	0,32	
	U.Fact	1045	3,62	116	4,01	1161	3,65	
	Viúvo	142	0,49	6	0,21	148	0,47	
Número filhos		M=1,05	DP=1,09	M=0,7	DP=0,95	M=1,02	DP=1,09	t(31835)=18,576; $p<0,001$ eta=0,104
Tem filhos	Não	11346	0,39	1619	0,56	12965	0,41	$\chi^2(1)=303,521$; $p<0,001$ Cramer V=0,098
	Sim	17594	0,61	1278	0,44	18872	0,59	
Grau deficiência		M=0,89	DP=7,61	M=0,06	DP=2,01	M=0,82	DP=7,29	t(14269,715)=14,223; $p<0,001$ eta=0,034
Evolução defic.	Aumento	391	1,35	2	0,07	393	1,23	$\chi^2(2)=35,504$; $p<0,001$ Cramer V=0,033
	Diminuição	19	0,07	2	0,07	21	0,07	
	Igual	28530	98,58	2893	99,86	31423	98,7	

Nota: * - para um número total de casos de 37792.

Por fim, segue-se a distribuição das características salariais por *turnover* voluntário onde apenas uma das características não apresenta uma relação significativa com o *turnover* ($p>0,05$) e que é a variação, em percentagem, da remuneração regular entre os anos N-2 e N-3. Entre as restantes

características, eu se relacionam significativamente com o *turnover* voluntário ($p<0,05$), destacam-se as que dizem respeito à remuneração regular auferida em cada um dos últimos 3 anos e a variação do rendimento regular entre os anos N-1 e N-2, pois apresentam uma relação moderada.

De facto, verifica-se que a média dos rendimentos regulares para os colaboradores ativos é superior em 453€ no ano N-1, 440€ no ano N-2 e 347€ no ano N-3, comparativamente à média de rendimentos dos colaboradores que saem para cada um desses anos. Também na variação entre os rendimentos dos anos N-1 e N-2, existe uma clara diferença entre os dois grupos de colaboradores. Enquanto a variação média dos rendimentos nos colaboradores ativos é de 6,25%, a variação média nos colaboradores que saem voluntariamente é de 4,96%. Os resultados de todas as características encontram-se na Tabela 4.2.

Tabela 4.2. Distribuição das características salariais por *turnover* voluntário

		Ativos (n=28940)		Saídas Voluntárias (n=2897)		Total (n=31837)		Teste estatístico
		Colaboradores	%	Colaboradores	%	Colaboradores	%	Medida de associação
Remuneração (ano N-1)		M=1516,29	DP=1289,66	M=1063,27	DP=708,89	M=1474,34	DP=1254,18	t(5121,512)=29,734;p<0,001 eta=0,320
Remuneração (ano N-2)		M=1627,06	DP=1352,92	M=1187,16	DP=877,09	M=1602,75	DP=1334,85	t(1534,778)=16,176;p<0,001 eta=0,363
Remuneração (ano N-3)		M=1636,28	DP=1358,02	M=1289,47	DP=863,13	M=1622,5	DP=1343,54	t(903,848)=10,445;p<0,001 eta=0,342
Isenção de horário (ano N-1)	Não	13687	47,29	1793	61,89	15480	48,62	$\chi^2(1)=224,618;p<0,001$ Cramer V=0,084
	Sim	15253	52,71	1104	38,11	16357	51,38	
	Total	28940	100	2897	100	31837	100	
Isenção de horário (ano N-2)	Não	19087	65,95	2435	84,05	21522	67,6	$\chi^2(1)=393,845;p<0,001$ Cramer V=0,111
	Sim	9853	34,05	462	15,95	10315	32,4	
	Total	28940	100	2897	100	31837	100	
Isenção de horário (ano N-3)	Não	20535	70,96	2559	88,33	23094	72,54	$\chi^2(1)=399,116;p<0,001$ Cramer V=0,112
	Sim	8405	29,04	338	11,67	8743	27,46	
	Total	28940	100	2897	100	31837	100	
Variação remuneração anos N-1 e N-2 (%)		M=6,25	DP=9,8	M=4,96	DP=8,67	M=6,18	DP=9,74	t(1366,877)=4,926;p<0,001 eta=0,428
Variação remuneração anos N-2 e N-3 (%)		M=4,15	DP=7,01	M=4,72	DP=7,97	M=4,17	DP=7,06	t(729,403)= -1,949;p=0,052 eta=0,450
Anos com horas extra		M=1,2	DP=1,32	M=0,77	DP=1,02	M=1,16	DP=1,31	t(3946,546)=21,004;p<0,001 eta=0,136
Último ano com horas extra		M=0,52	DP=0,5	M=0,46	DP=0,5	M=0,52	DP=0,5	t(3503,727)=6,044;p<0,001 eta=0,034
Anos com ajudas de custo		M=1,75	DP=1,24	M=0,93	DP=1	M=1,67	DP=1,24	t(3846,895)=40,856;p<0,001 eta=0,191
Último ano com ajudas de custo		M=0,78	DP=0,41	M=0,6	DP=0,49	M=0,76	DP=0,43	t(3324,270)=19,167;p<0,001 eta=0,122
Anos com distribuição resultados		M=1,34	DP=1,33	M=0,59	DP=1,03	M=1,27	DP=1,32	t(3938,765)=36,222;p<0,001 eta=0,165
Último ano com distr. resultados		M=0,57	DP=0,49	M=0,3	DP=0,46	M=0,55	DP=0,5	t(3602,135)=29,966;p<0,001 eta=0,156
Anos com prémio		M=1,85	DP=1,26	M=0,94	DP=1,07	M=1,76	DP=1,27	t(3743,154)=42,768;p<0,001 eta=0,208
Último ano com prémio		M=0,77	DP=0,42	M=0,55	DP=0,5	M=0,75	DP=0,43	t(3323,222)=22,899;p<0,001 eta=0,146
Anos com excedente subs. almoço		M=1,65	DP=1,22	M=1,1	DP=0,99	M=1,6	DP=1,21	t(3837,172)=27,841;p<0,001 eta=0,153
Último ano com exced. subs. almoço		M=0,77	DP=0,42	M=0,71	DP=0,45	M=0,76	DP=0,42	t(3414,596)=6,860;p<0,001 eta=0,041
Anos com subsídio de almoço		M=1,96	DP=1,13	M=1,3	DP=0,98	M=1,9	DP=1,13	t(3707,092)=34,305;p<0,001 eta=0,181
Último ano com subsídio de almoço		M=0,89	DP=0,32	M=0,82	DP=0,38	M=0,88	DP=0,33	t(3310,36)=8,650;p<0,001 eta=0,056
Anos com subsídio de transporte		M=1,05	DP=1,23	M=0,62	DP=0,94	M=1,01	DP=1,22	t(3974,279)=22,751;p<0,001 eta=0,106
Último ano com subsídio de transp.		M=0,51	DP=0,5	M=0,39	DP=0,49	M=0,5	DP=0,5	t(3533,901)=12,548;p<0,001 eta=0,069
Anos com trabalho por turnos		M=0,29	DP=0,78	M=0,28	DP=0,65	M=0,29	DP=0,77	t(3776,075)=1,218;p<0,001 eta=0,078
Último ano com trabalho por turnos		M=0,15	DP=0,36	M=0,19	DP=0,39	M=0,16	DP=0,36	t(3395,934)= -5,234;p<0,001 eta=0,032
Anos com trabalho noturno		M=0,29	DP=0,83	M=0,16	DP=0,57	M=0,28	DP=0,81	t(4233,018)=11,059;p<0,001 eta=0,058
Último ano com trabalho noturno		M=0,13	DP=0,33	M=0,09	DP=0,29	M=0,12	DP=0,33	t(3698,964)=5,822;p<0,001 eta=0,029

4.2. Modelos preditivos

De forma a responder ao segundo objetivo analítico desta investigação, são utilizados diferentes algoritmos de *machine learning*, baseados em árvores de decisão, que por serem parametrizados de

diferentes formas, dão origem a diversos modelos para a previsão do *turnover* voluntário, tal como evidenciado no capítulo anterior. Na metodologia, na Tabela 3.8, está representado o melhor modelo de cada algoritmo, ao nível das parametrizações, e os seus resultados são apresentados neste capítulo. Adicionalmente, no Anexo B, é possível verificar não só a parametrização, mas, também, os resultados obtidos de diferentes modelos por cada um dos cinco algoritmos em estudo.

É importante referir que para ser possível chegar aos modelos finais é necessário realizar um grande número de experiências para cada um dos cinco algoritmos, onde os modelos são analisados e as parametrizações e amostras alteradas de forma a melhorar a performance e evitar o sub e sobre ajustamento dos mesmos. Todos os modelos criados são comparados entre si para a identificação dos melhores, que alcançam os critérios de sucesso definidos para o segundo objetivo analítico.

4.2.1. Avaliação dos modelos e seleção do melhor

Os resultados dos melhores modelos, por algoritmo, são apresentados na Tabela 4.3 e respeitam as métricas de avaliação apresentadas na metodologia, sendo que estas são calculados para a amostra de treino (70%) e para a amostra de teste (30%). Os resultados mostram que a qualidade dos diferentes modelos é semelhante, uma vez que os valores são próximos para todas as métricas de avaliação. Ainda assim, ao nível do treino o modelo Y apresenta os valores mais altos na maioria das métricas, com uma precisão de 80,57% e PCCC de 81,40%. Já ao nível da performance na amostra de teste, não é o modelo Y que mais se destaca podendo o modelo mostrar algum sobreajustamento. O destaque é para o modelo J, uma vez que para a amostra de teste apresenta uma melhor qualidade do que os restantes (PCCC=78,61%, Sensibilidade=78,81, Especificidade=78,39%, precisão=80,20% e AUC=0,86).

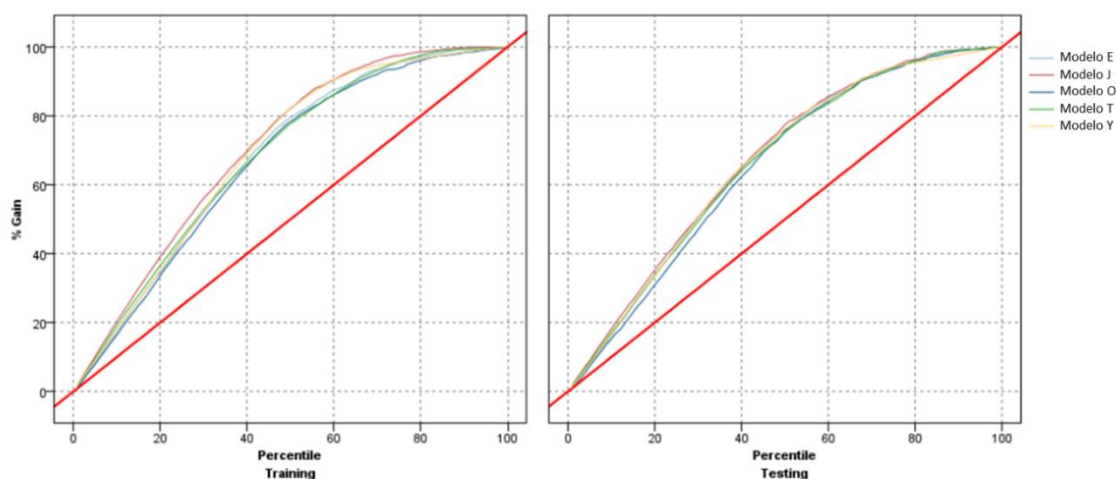
Pela comparação das métricas de sensibilidade e especificidade é possível concluir que os modelos apresentam uma capacidade semelhante no que toca à previsão quer dos colaboradores que se mantêm ativos, quer dos colaboradores que decidem sair voluntariamente das empresas do Grupo.

Tabela 4.3. Resultados dos melhores modelos por algoritmo

Modelo		E	J	O	T	Y
Algoritmo		CART	CHAID	CUEST	Tree-AS	C5.0
Treino (70%)	PCCC	78,80%	81,45%	77,98%	77,16%	81,40%
	Sensibilidade	77,08%	82,05%	78,04%	77,59%	81,59%
	Especificidade	80,44%	80,87%	77,92%	76,76%	81,21%
	Precisão	79,00%	80,38%	77,14%	76,12%	80,57%
	AUC	0,85	0,899	0,828	0,852	0,868
Teste (30%)	PCCC	78,51%	78,61%	76,96%	76,62%	77,36%
	Sensibilidade	76,11%	78,81%	77,08%	76,32%	77,30%
	Especificidade	78,51%	78,39%	76,83%	76,95%	77,43%
	Precisão	79,73%	80,20%	78,70%	78,62%	79,18%
	AUC	0,846	0,86	0,821	0,84	0,844

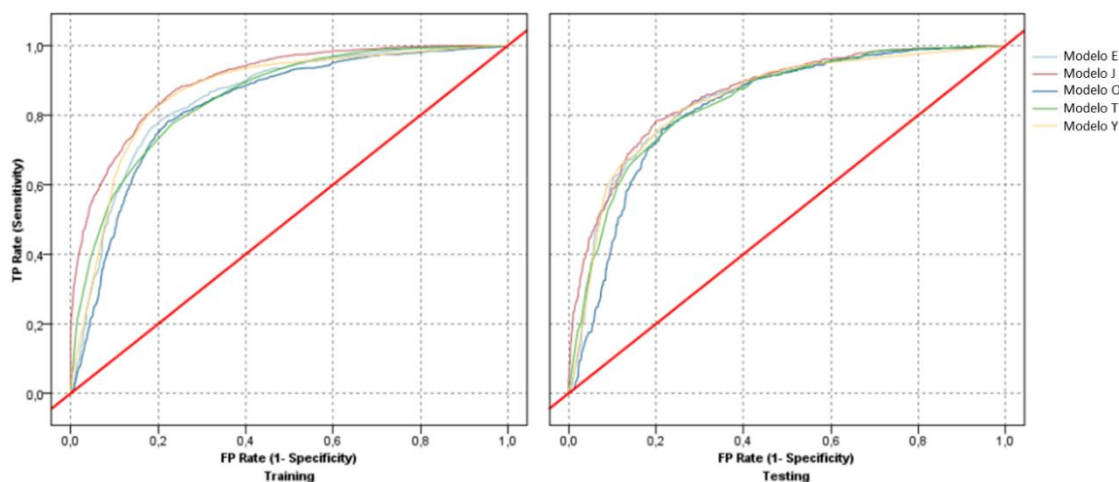
A suportar a tabela dos resultados, estão os gráficos de ganhos e a curva ROC, representados nas Figuras 4.1 e 4.2, respetivamente. Em ambas as representações, os modelos apresentam um comportamento bastante semelhante, principalmente no que concerne à performance sobre a amostra de teste. Destacam-se os modelos J (linha vermelha) e Y (linha amarela) quanto ao seu desempenho na amostra de treino e o modelo J destaca-se quanto à sua performance na amostra de teste. Enquanto a curva ROC é o reflexo gráfico dos valores do AUC, apresentados na tabela anterior, para cada modelo, o gráfico de ganhos representa a relação entre a percentagem de casos positivos em função do percentil de casos mais prováveis de serem positivos. Por exemplo, no modelo J, na amostra de teste, para os 40% de colaboradores mais prováveis de saírem voluntariamente, acerta-se em 65%.

Figura 4.1. Gráfico de ganhos dos melhores modelos por algoritmo



Fonte: *output* SPSS Modeler

Figura 4.2. Curva ROC dos melhores modelos por algoritmo



Fonte: *output* SPSS Modeler

Adicionalmente e de uma forma geral, verifica-se que as cinco variáveis mais importantes para a previsão do *turnover* voluntário são, para todos os modelos, variáveis da dimensão profissional. Além de serem todas do foro profissional, os modelos consideram exatamente as mesmas variáveis como as mais importantes, ainda que dependendo do modelo possam aparecer numa ordem diferente, e estas dizem respeito à antiguidade, à avaliação média, ao número de anos efetivo e ao tipo de contrato dos colaboradores (Tabela 4.4). O modelo Y destaca ainda o número de horas teóricas de trabalho por dia e o modelo T não identifica qualquer variável como importante. Tal situação decorre ou porque todas as variáveis apresentam o mesmo nível de importância para o modelo, ou porque nenhuma foi considerada relevante por este.

Tabela 4.4. Top cinco variáveis mais importantes por modelo

Modelo	Variáveis mais importantes (Top 5)				
E	Avaliação média	Tipo de contrato	Antiguidade	Anos efetivo	Prémios
J	Tipo de contrato	Avaliação média	Antiguidade	Anos efetivo	Prémios
O	Avaliação média	Antiguidade	Tipo de contrato	Anos efetivo	Prémios
Y	Antiguidade	Tipo de contrato	Avaliação média	Anos efetivo	Horas trab. por dia

Tendo em consideração todos os resultados obtidos por cada um dos melhores modelos por algoritmo, o modelo eleito como o melhor é o modelo J, pois destaca-se positivamente dos restantes. A escolha recai sobretudo pelos resultados do modelo sobre a amostra de teste, uma vez que esta reflete o desempenho dos modelos quando aplicados sobre novos colaboradores.

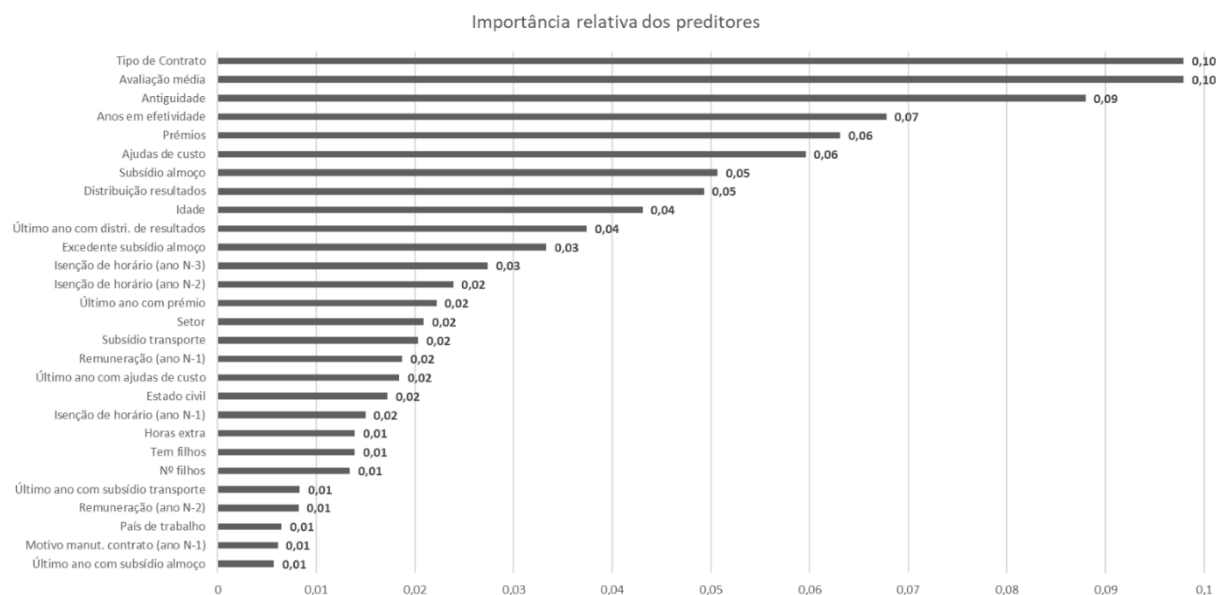
4.2.1.1 Avaliação dos modelos e seleção do melhor

De forma a ser possível identificar quais as características mais explicativas do *turnover* voluntário, é analisado o gráfico de importância relativa, atribuída pelo modelo escolhido, das variáveis que mais contribuem para a explicabilidade desse mesmo modelo. A Figura 4.3 baseia-se num dos *outputs* do modelo J e apresenta a importância relativa das variáveis, enquanto preditoras, mais importantes na estimativa do modelo. Os valores refletem a importância que cada fator tem na previsão e, como se trata de valores relativos, a sua soma é igual a um.

Através da figura, é possível verificar que as características que apresentam uma maior importância relativa para o modelo J se enquadram na dimensão profissional, seguindo-se as variáveis da dimensão salarial dos colaboradores. Apenas a variável idade, que pertence à dimensão pessoal, surge entre os dez fatores com maior importância relativa. As seguintes características pessoais aparecem entre o 19º lugar e o 22º lugar e referem-se ao estado civil e ao número de filhos dos

colaboradores. Assim sendo, é notório que as variáveis do foro profissional e salarial são aquelas que mais influenciam o modelo preditivo para o *turnover* voluntário.

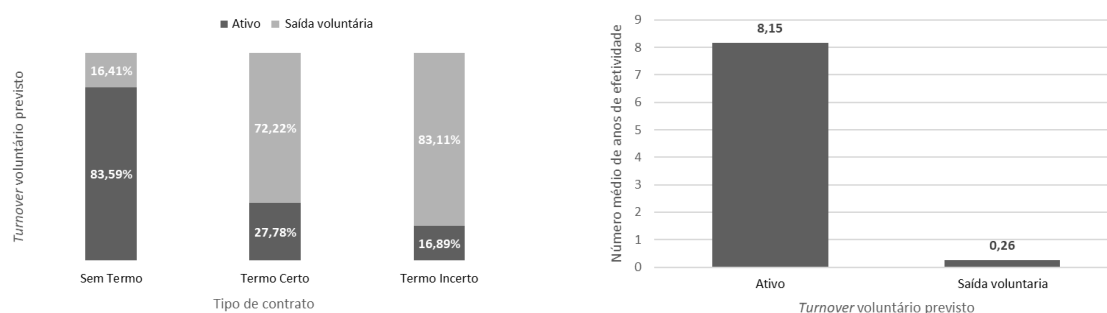
Figura 4.3. Gráfico de importância do modelo J



Fonte: adaptado do *Output* SPSS Modeler

O *turnover* voluntário quando comparado com as quatro variáveis com maior importância relativa para o modelo J, apresenta para todas elas uma distribuição bastante diferente. A partir da Figura 4.4 constata-se que os colaboradores com contratos a termo, seja este a termo certo ou incerto, são mais prováveis de sair voluntariamente. Por outro lado, os colaboradores com contrato sem termo são mais prováveis de ficar no Grupo. Da mesma forma, os colaboradores mais prováveis de se manterem ativos estão, em média, num contrato de efetivo (sem termo) há cerca de 8 anos, enquanto os colaboradores mais prováveis de apresentarem a demissão estão, em média, há menos de meio ano num contrato de efetivo.

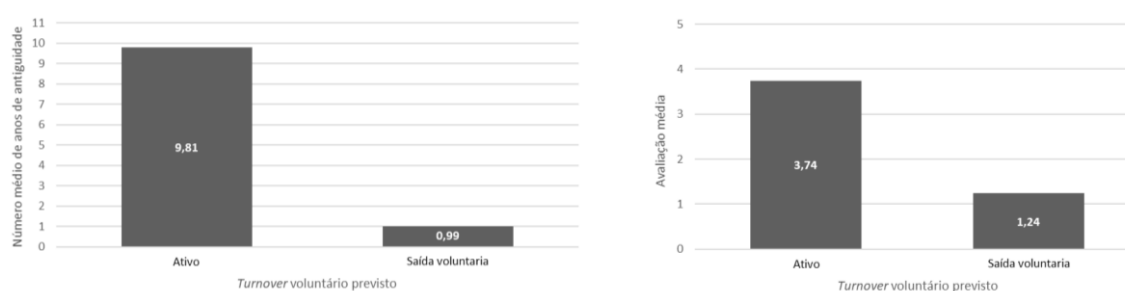
Figura 4.4. Distribuição do *turnover* voluntario pelo tipo de contrato e anos de efetividade



Fonte: adaptado do *output* SPSS Modeler

Adicionalmente, os colaboradores com uma antiguidade média de cerca de 9 anos apresentam uma maior probabilidade de continuarem no Grupo e colaboradores com uma antiguidade média de cerca de um ano são mais prováveis de saírem voluntariamente. Por fim, colaboradores com uma avaliação média de 3,74, numa escala de 1 a 5, são mais prováveis de continuarem ativos. Já os colaboradores com uma avaliação média avaliação mais baixa, de 1,24, apresentam maior probabilidade de se demitirem (Figura 4.5).

Figura 4.5. Distribuição do *turnover* voluntário pela antiguidade e avaliação média



Fonte: adaptado do *output* SPSS Modeler

4.2.1.2 Perfis do *turnover* voluntário

A identificação dos perfis do *turnover* voluntário tem como base a interpretação das regras da árvore de decisão do modelo escolhido. Contudo, como o modelo escolhido, modelo J, é obtido por uma combinação de modelos (*bagging*), recorre-se à aplicação do mesmo algoritmo sem a utilização de *bagging*. Desta forma, é possível identificar quais as regras que explicam a tomada de decisão do modelo.

Para os perfis, são selecionadas as regras com maior número de confiança e maior número de registos de suporte, considerando os critérios de sucesso definidos na metodologia para o terceiro objetivo analítico. Desta forma, são apresentados os perfis associados quer aos colaboradores que se mantêm ativos, quer para os colaboradores que saem voluntariamente, tendo por base regras com um grau de confiança acima dos 90% e um suporte superior a 100 colaboradores.

Os perfis dos colaboradores que tendem a manter-se ativos (*turnover* voluntário igual a zero):

- Colaboradores sem avaliação e que desempenham funções há mais de 10 anos;
- Colaboradores jovens adultos ou mais velhos (mais de 25 anos), que desempenham funções sem ser por turnos na Europa (incluindo Portugal) ou América e que obtêm uma boa avaliação no único ano de trabalho no Grupo;
- Colaboradores com uma boa avaliação média ao longo do tempo e que desempenham funções há mais de 4 anos;

- Colaboradores com uma avaliação média, ao longo dos anos de atividade, entre o nível bom e o nível elevado e que desempenham funções há mais de uma década;
- Colaboradores adultos (mais de 38 anos) e efetivos com uma elevada avaliação média ao longo dos anos.

Perfis dos ex-colaboradores que saem voluntariamente (*turnover* voluntário igual a um):

- Colaboradores não efetivos sem avaliação;
- Colaboradores não efetivos com avaliação baixa no decorrer do contrato.

4.3. Discussão e implicações

O estudo do *turnover* mantém-se como um tópico relevante quer para o tecido empresarial, quer para a comunidade académica, que nos últimos anos tem conduzido várias investigações que têm em vista entender este acontecimento através da aplicação de *machine learning* (Jayaratne & Jayatilleke, 2022). Nesse sentido, também a presente investigação se propõe ao estudo e previsão, não do *turnover* geral, mas sim do *turnover* voluntário, já estudado por autores como Rombaut e Guerry (2021), recorrendo igualmente à aplicação de técnicas *machine learning*, mas não em contexto português.

Para a realização da investigação é apenas utilizado um conjunto de dados pessoais e profissionais de colaboradores, sem a adição de informações provenientes de outras fontes, onde se inclui a realização de questionários (Anand *et al.*, 2022), seguindo as recomendações dos autores Rombaut e Guerry (2018). Já no que toca à metodologia, o presente estudo segue o CRISP-DM que é considerado adequado a projetos de *data mining* focados no negócio, e que não foi aplicado por nenhum dos artigos incluídos na revisão de literatura. O sucesso da investigação demonstra, em parte, o sucesso da aplicação de uma metodologia rigorosa e replicável.

Através das fases de compreensão e preparação dos dados, a investigação chegou a uma base de dados final, utilizada para a modelação, com mais de 31000 colaboradores que trabalham ou trabalharam para diferentes empresas portuguesas, incluídas num Grupo Económico Português, dos mais variados setores de negócio e com atividade em diversos países. Tal diversidade contrasta com grande parte da literatura, que se foca em colaboradores de profissão comum ou colaboradores pertencentes a empresas com atividades semelhantes (Matsuo *et al.*, 2023; Ogbeibu *et al.*, 2022). Nas fases seguintes, e em concordância com a literatura, são aplicados estatística descritiva, testes estatísticos e técnicas de análise preditiva com aprendizagem supervisionada (Vanegas, Wine & Drasgow, 2022), de forma a alcançar todos os objetivos analíticos e, consequentemente, os de negócio propostos. No entanto, a investigação destaca-se da literatura pela utilização de mais de 60 variáveis, todas elas relacionadas com a vida pessoal e profissional dos colaboradores (Fallucchi *et al.*, 2020).

Considerando os resultados apresentados neste capítulo, é possível verificar que o modelo baseado no algoritmo CHAID, não aplicado pela literatura apesar de a mesma utilizar vários algoritmos baseados em árvores de decisão (Gao, Wen & Zhang, 2019), é o que apresenta o melhor desempenho. As métricas de avaliação, baseadas na matriz de confusão, alcançam valores acima de 78%. A performance do modelo acompanha a performance dos modelos identificados na literatura, chegando mesmo a registar melhores resultados que alguns dos estudos (Matsuo *et al.*, 2023; Rombaut & Guerry, 2018). O facto de se ter dividido a amostra em 10 grupos aleatórios permite, ainda, concluir que a inclusão de mais casos, ou seja, considerar mais do que uma amostra, não melhora a performance dos modelos.

Adicionalmente, os resultados mostram que é a situação profissional dos colaboradores que mais influencia o *turnover* voluntário no Grupo Económico Português, o que vai de encontro à literatura (Nandialath *et al.*, 2018). À semelhança de alguns estudos apresentados na literatura, também o modelo escolhido, considera a antiguidade como uma das principais características que influencia a saída voluntária, ou não, dos colaboradores (Rombaut & Guerry, 2021). Contudo, identifica outras variáveis, não identificadas pela literatura, como principais preditores: o tipo de contrato, a avaliação média e o número de anos em efetividade. Em adição, o presente estudo não identifica a proximidade entre o local de residência e o local de trabalho dos colaboradores como um fator determinante para o *turnover*, ao contrário de autores como Atef, Elzanfaly e Ouf (2022) e Rombaut e Guerry (2021).

No que toca às características da dimensão pessoal dos colaboradores, a idade é o fator com maior importância relativa para o modelo J (o escolhido), fator este que é, também, identificado na literatura como uma das principais variáveis significativas na explicação do *turnover* (Atef, Elzanfaly & Ouf, 2022). Adicionalmente, a análise estatística demonstrou que entre as variáveis consideradas significativas para o *turnover* voluntário, as variáveis relacionadas com a remuneração e a sua variação são as que apresentam maior relação com este, conclusão que também vários estudos identificam (Gao, Wen & Zhang, 2019).

A influência destas características na previsão do *turnover* voluntário é bastante evidente nos principais perfis identificados a partir das regras do modelo. De uma forma generalizada, é possível concluir que os colaboradores com contratos a termo, certo ou incerto, são mais propensos a pedir a demissão antes do fim do seu contrato. O facto de os colaboradores, mais prováveis de saírem voluntariamente, estarem associados a contratos que duram entre três meses e quatro anos estão naturalmente associados a: antiguidades baixas e a avaliações baixas, pois se saírem antes de completarem um ano de casa não chegam a ser avaliados. Possíveis causas para estes colaboradores pedirem a demissão antes do final do contrato passam pela própria tipologia de contrato, que não incentiva o vínculo a longo prazo entre o colaborador e a empresa (têm uma data de fim), sendo um cenário mais propício para a procura de outras oportunidades fora da empresa. Além disso, os

colaboradores com contratos a termo estão, no Grupo, associados a menos benefícios, quer a nível monetário, quer a nível não monetário, como planos de benefícios flexíveis ou seguros de vida e saúde.

Por outro lado, o modelo identifica que os colaboradores categorizados como quadros superiores pelo Grupo, onde se encontram sobretudo cargos de chefia ou de direção, tendem a manter-se ativos. Este grupo de colaboradores é identificado pela primeira regra do modelo (colaboradores sem avaliação e antiguidade superior a dez anos) e está associado a remunerações mais elevadas e acesso a diversos benefícios. Adicionalmente e de forma geral, colaboradores com maior antiguidade e avaliações médias mais altas apresentam uma maior propensão em ficar no Grupo.

Por fim, enfatiza-se que os quatro objetivos analíticos, propostos na metodologia, são alcançados com sucesso e, atendendo a toda a discussão, são discriminadas na Tabela 4.5 recomendações para as principais partes interessadas da presente investigação.

Tabela 4.5. Recomendações aos principais *stakeholders*

Recomendação	Stakeholder
Garantir a qualidade os dados armazenados, evitando ao máximo erros e incoerências que podem comprometer os resultados e as conclusões de projetos de <i>data mining</i> .	Grupo Económico Português, Profissionais de RH e Profissionais de <i>Analytics</i>
Utilização da investigação como exemplo real da aplicação de técnicas preditivas com aprendizagem supervisionada à área de Recursos Humanos, em contexto português, promovendo o interesse e o contacto com esta área da Gestão.	Professores de <i>Business Analytics</i>
Utilização da investigação como exemplo real de uma alternativa capaz de prever o <i>turnover</i> voluntário em empresas portuguesas, identificando os fatores com maior influência para este acontecimento.	Professores e profissionais de RH
Utilização da investigação como exemplo do potencial da conjugação entre as áreas de <i>Business Analytics</i> e de Recursos Humanos, nomeadamente em Portugal – <i>People Analytics</i> .	Profissionais, professores e investigadores de RH e <i>Business Analytics</i>
Utilização da investigação como exemplo real da aplicação da metodologia CRISP-DM.	Investigadores de <i>Business Analytics</i>
Utilização da RSL, que sumariza o conhecimento existente sobre a previsão do <i>turnover</i> , para futuras investigações neste âmbito.	Investigadores de RH
Aproximar os benefícios oferecidos aos colaboradores efetivos, aos colaboradores contratados a termo, de forma a criar melhores condições para que estes colaboradores se mantenham ativos durante todo o período do contrato.	Grupo Económico Português
Incentivar desde logo a contratação sem termo, que pela duração do período experimental salvaguarda a empresa no caso de o colaborador contratado não cumprir com as expectativas.	Grupo Económico Português
Propor aos colaboradores contratados a termo, nos casos em que se aplique, o contrato sem termo com a maior antecedência possível e não esperar pelo fim do contrato a termo.	Grupo Económico Português
Optar por centralizar num único <i>software</i> todos os dados criados e geridos pelo departamento de RH, de forma a ser possível incluir e analisar o máximo de informação, que pode contribuir para gerar mais e melhor conhecimento sobre os colaboradores.	Grupo Económico Português e Profissionais de RH

Capítulo 5 – Conclusão

A existência de elevadas taxas de *turnover* nas empresas é um problema capaz de pôr em risco a continuidade destas no mercado, dada a sua associação a mais despesas, perda de conhecimento e de produtividade. No entanto, é sobre o *turnover* voluntário que as empresas devem prestar maior atenção, visto que acarreta elevados custos e a decisão está do lado dos colaboradores.

Assim, perceber como é que as empresas podem conseguir antecipar a saída voluntária dos seus colaboradores é um tema relevante e, dessa forma, a presente investigação procurou responder à questão de como prever a saída voluntária de colaboradores, tendo como objetivos caracterizar demograficamente e profissionalmente o *turnover* voluntário, construir um modelo preditivo de classificação para o *turnover* voluntário, identificar os perfis de colaboradores com elevada propensão para o *turnover* voluntário e identificar os fatores mais associados ao *turnover* voluntário. Para o efeito, foi, inicialmente, efetuada uma RSL para a identificar o conhecimento já existente neste âmbito, nomeadamente ao nível da contextualização dos estudos, das metodologias aplicadas, dos resultados obtidos e dos seus contributos.

Após a análise dos 24 artigos incluídos na RSL e seguindo a metodologia CRISP-DM, foram analisados dados pessoais e profissionais de colaboradores e ex-colaboradores, armazenados num *software* de cadastro, recolhidos pelo departamento de RH de um Grupo Económico Português, garantindo sempre a sua anonimização. O conjunto de dados passou, de forma exaustiva, pelas fases de compreensão, de tratamento e de modelação, tendo sempre em vista o cumprimento dos objetivos analíticos, baseados nos objetivos propostos pelo estudo.

De forma a caracterizar o *turnover voluntário*, foram aplicadas técnicas de estatística descritiva bivariada e testes estatísticos para identificar quais as características, pessoais e profissionais, impactam significativamente o *turnover* voluntário, numa amostra superior a 31000 colaboradores. A análise concluiu que entre 69 variáveis, apenas quatro não impactam significativamente o *turnover* voluntário, evidenciando o potencial dos dados armazenados pelo departamento de RH. Destacando-se, ainda, a relação do *turnover* voluntário com a idade, antiguidade, avaliação média, remuneração (de cada um dos anos), e o tipo de contrato, sendo nestas características que se identificam maiores diferenças entre os colaboradores que abandonam a empresa e os que nela permanecem.

Para a construção do modelo preditivo foram utilizados diferentes algoritmos, baseados em árvores de decisão, dado o seu carácter interpretativo adequado ao negócio. O melhor modelo resultou da utilização do algoritmo CHAID, que obteve a melhor avaliação na amostra de teste, com PCCC igual a 78,61%, sensibilidade igual a 78,81, especificidade igual a 78,39%, precisão igual a 80,20% e AUC igual a 0,86. Dado que todas as métricas apresentam valores superiores a 75%, o segundo objetivo analítico é alcançado com sucesso.

Através das regras proposicionais do modelo escolhido (árvore de decisão com o algoritmo CHAID), cujo suporte é de 100 colaboradores e confiança acima de 90%, permitem a identificação de perfis tanto para os colaboradores que tendem a manter-se no Grupo, associados a uma maior antiguidade e melhores avaliações, quer para os colaboradores que tendem a sair voluntariamente, associados a contratos a termo (certo ou incerto) e com baixas avaliações. O cumprimento do terceiro objetivo foi, também, realizado com sucesso.

A partir dos resultados dos modelos de classificação, consideram-se como os fatores mais associados ao *turnover* voluntário, o tipo de contrato, a avaliação média, a antiguidade e os anos de efetividade, uma vez que são considerados pelo modelo como os fatores de maior importância relativa para a previsão.

Cumpridos todos os objetivos do estudo, verifica-se que o mesmo é capaz de responder à questão da investigação “Como prever a saída voluntária de um colaborador?”. De facto, o estudo revelou que se pode usar uma árvore de classificação com o algoritmo CHAID para prever com sucesso (principais métricas refletem taxas de acerto perto de 80%) o *turnover* voluntário, tendo como variáveis preditores as características demográficas, profissionais e de remuneração, e uma parametrização em que a profundidade máxima é de 6, o número mínimo de casos no nós pais é de 50 e filho de 25, com *ensemble bagging*, levando a uma árvore com algum grau de complexidade, com mais de 100 nós.

5.1. Contributos

Perante o problema identificado, a investigação, ao responder à questão de investigação, tem contributos, quer ao nível académico, quer ao nível profissional e empresarial. Para os futuros investigadores interessados em estudar e analisar o *turnover*, o presente estudo é um bom ponto de partida, pois apresenta uma RSL que sumariza os principais estudos publicados em artigos científicos nesse mesmo âmbito e identifica os melhores em cada uma das dimensões consideradas (contexto, metodologia, resultados e contributos). A aplicação com sucesso da metodologia CRISP-DM contribui, também, para os investigadores da área de *Business Analytics*, uma vez que esta pode ser replicada em investigações de *data mining / machine learning*.

Os resultados da investigação contribuem para a literatura que, até quanto se sabe, não identifica a tipologia de contratos, os anos de efetividade e a avaliação média dos colaboradores como principais características na explicação do *turnover* voluntário. Ao conduzir a investigação em empresas portuguesas, deu-se a conhecer mais um contexto de aplicação de técnicas preditivas para previsão o *turnover* voluntário. Adicionalmente, o processo adotado de modelação revela que a performance dos modelos não melhora com a adição de uma maior amostra, mas sim com o balanceamento entre o número de colaboradores ativos e o número de colaboradores que saíram voluntariamente, e a

utilização de um grande número de variáveis não implica um pior desempenho dos modelos. De uma forma geral, o estudo demonstra que as técnicas de *machine learning* podem ser aplicadas com sucesso na área de RH, reforçando que são uma boa alternativa aos métodos estatísticos mais tradicionais.

Para o Grupo Económico Português, a identificação de perfis de colaboradores e dos fatores mais associados ao *turnover* voluntário contribui para que este possa ajustar as suas estratégias, definindo novas ações que possam não só prevenir as saídas voluntárias, como agilizar os processos que um despedimento deste tipo desencadeia, como o recrutamento de um novo colaborador. Além disso, o modelo desenvolvido pode ser utilizado pelos profissionais do Grupo para a identificar, entre os colaboradores ativos, os mais prováveis de vir a pedir a demissão e, assim, tomarem medidas preventivas.

Por fim, as ilações da investigação têm potencial interesse para outros profissionais, fora do Grupo que é alvo deste estudo. A aplicação de uma abordagem analítica na área de RH, pode ser capaz de gerar novo conhecimento útil de ser aplicado a outras realidades e, academicamente, apresentar novas perspetivas para futuras investigações.

5.2. Limitações e pistas futuras de investigação

Ao nível das limitações, identifica-se que o facto de os dados geridos pelo departamento de RH estarem armazenados em diferentes plataformas limita a informação passível de incluir no estudo o que, consequentemente, pode limitar as suas conclusões. Por outro lado, verifica-se que a qualidade dos dados recolhidos pode ser melhorada, existindo registos errados, incoerentes ou até mesmo omissos. Tal situação, além de implicar uma grande alocação a nível de tempo num projeto desta natureza, pode comprometer os resultados e as conclusões do estudo. É, ainda, importante referir que apesar de o estudo se basear em colaboradores que trabalham para diferentes empresas, com atividades bastante diferenciadas e com atividade a nível mundial, todas elas pertencem ao mesmo Grupo e, por isso, seguem e aplicam estratégias e práticas comuns, ao nível da gestão de RH, o que pode comprometer a generalização dos resultados a outras realidades.

Como trabalho futuro sugere-se aprofundar e dar continuidade à análise desenvolvida através da: i) inclusão de outras variáveis enquadradas na atividade dos departamentos de RH como recrutamento e seleção, formação, saúde ocupacional ou benefícios; ii) análise de quais as características mais associadas ao *turnover* voluntário nos colaboradores com contrato efetivo e com contrato a termo, de forma a concluir as potenciais diferenças entre o que motiva cada um destes grupos de colaboradores a saírem das empresas; e analisar o modelo escolhido numa perspetiva financeira, associando os custos do *turnover* voluntário para as empresas consideradas.

Referências bibliográficas

- Anand, V., Verma, L., Santhanam, N., & Grover, A. (2022). Turnover intention among Indian police: Do organizational and community stressors matter? *Journal of Criminal Justice*, 82, 101969. <https://doi.org/10.1016/j.jcrimjus.2022.101969>
- Atef, M., Elzanfaly, D., & Ouf, S. (2022). Early Prediction of Employee Turnover Using Machine Learning Algorithms. *International Journal of Electrical and Computer Engineering Systems*, 13(2), 135–144. <https://doi.org/10.32985/ijeces.13.2.6>
- Ayele, W. Y. (2020). Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset. *International Journal of Advanced Computer Science and Applications*, 11(6), 20–32. <https://doi.org/10.14569/IJACSA.2020.0110603>
- Berengueres, J., Duran, G., & Castro, D. (2017). Happiness, an inside job? Turnover prediction using employee likeability, engagement and relative happiness. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017* (pp. 509–516). Association for Computing Machinery, Inc. <https://doi.org/10.1145/3110025.3110132>
- Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F., & Silipo, R. (2020). Guide to Intelligent Data Science - How to Intelligently Make Use of Real Data (2nd ed.). Springer. <https://doi.org/10.1007/978-3-030-45574-3>
- Bhanu Prakash, K., Adishesha Reddy, A., & Siva Reddy, P. (2019). Human capital talentum analytics-a focus study on schools of business (Sobs) in Telangana & Karnataka. *International Journal of Recent Technology and Engineering*, 7(6), 1949–1952.
- Boroş, S. & Curşeu, L. (2013). Is it here where I belong? An integrative model of turnover intentions. *Journal of Applied Social Psychology*, 43(8), 1553-1562. <https://doi.org/10.1111/jasp.12104>
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Classification and Regression Trees. Chapman and Hall, Wadsworth, New York.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. USA: SPSS Inc.
- Chen, M.-L., Su, Z.-Y., Lo, C.-L., Chiu, C.-H., Hu, Y.-H. & Shieh, T.-Y. (2014). An empirical study on the factors influencing the turnover intention of dentists in hospitals in Taiwan. *Journal of Dental Sciences*, 9(4), 332-344. <https://doi.org/10.1016/j.jds.2013.01.003>
- Constantino, J. (2020). Gestão de Resultados como Preditor de Insolvências: Evidência nas Empresas Portuguesas, Dissertação de Mestrado, ISCTE-IUL, Lisboa.
- Denvir, A., & McMahon, F. (1992). Labour turnover in London hotels and the cost effectiveness of preventative measures. *International Journal of Hospitality Management*, 11(2), 143–154. [https://doi.org/10.1016/0278-4319\(92\)90007-I](https://doi.org/10.1016/0278-4319(92)90007-I)
- Dickter, D. N., Roznowski, M., & Harrison, D. A. (1996). Temporal tempering: An event history analysis of the process of voluntary turnover. *Journal of Applied Psychology*, 81(6), 705–716. <https://doi.org/10.1037/0021-9010.81.6.705>
- Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, 9(4), 86. <https://doi.org/10.3390/computers9040086>
- Felps, W., Mitchell, T., Hekman, D., Lee, T., Holtom, B. & Harman, W. (2009). Turnover contagion: how coworker's job embeddedness and job search behaviors influence quitting. *Academy of Management Journal*, 52(3), 545-561. <https://doi.org/10.5465/amj.2009.41331075>
- Frederiksen, A. (2017). Job satisfaction and employee turnover: A firm-level perspective. *German Journal of Human Resource Management: Zeitschrift Für Personalforschung*, 31(2), 132–161. <https://doi.org/10.1177/2397002216683885>
- Gao, X., Wen, J., & Zhang, C. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, 2019, 1–12. <https://doi.org/10.1155/2019/4140707>

- Guz, A. N., & Rushchitsky, J. J. (2009). Scopus: A system for the evaluation of scientific journals. *International Applied Mechanics*, 45(4), 351–362. <https://doi.org/10.1007/s10778-009-0189-4>
- Harris, G., Craig, E. & Light, A. (2011). Talent and analytics: new approaches, higher ROI. *Journal of Business Strategy*, 32(6), 4-13. <https://doi.org/10.1108/02756661111180087>
- Jayaratne, M., & Jayatilleke, B. (2021). Predicting job-hopping motive of candidates using answers to open-ended interview questions. *Journal of Computational Social Science*. <https://doi.org/10.1007/s42001-021-00138-4>
- Kang, I. G., Croft, B., & Bichelmeyer, B. A. (2020). Predictors of Turnover Intention in U.S. Federal Government Workforce: Machine Learning Evidence That Perceived Comprehensive HR Practices Predict Turnover Intention. *Public Personnel Management*, 50(4), 538-558. <https://doi.org/10.1177/0091026020977562>
- Karande, S., Shelake, A., M, S., & Sophia, S. (2019). Prediction of Employee Retention using Cassandra and Ensemble Learning. *Journal of Information Organization*, 9(4), 134. <https://doi.org/10.6025/jio/2019/9/4/134-140>
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29 (2), 119–127.
- Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. In *Keele University Technical Report TR/SE-0401*. <https://doi.org/10.1145/3328905.3332505>
- Kuzior, A., Kettler, K., & Rąb, Ł. (2022). Great Resignation—Ethical, Cultural, Relational, and Personal Dimensions of Generation Y and Z Employees' Engagement. *Sustainability*, 14(11), 6764. <https://doi.org/10.3390/su14116764>
- Lake, C. J., Highhouse, S., & Shrift, A. G. (2018). Validation of the Job-Hopping Motives Scale. *Journal of Career Assessment*, 26(3), 531–548. <https://doi.org/10.1177/1069072717722765>
- Larose, D., & Larose, C. (2015). *Data Mining and Predictive Analytics (Second Edi)*. John Wiley & Sons, Inc.
- Lee, J. L., & Kim, Y. (2023). Research Topic Trends on Turnover Intention among Korean Registered Nurses: An Analysis Using Topic Modeling. *Healthcare*, 11(8), 1139. <https://doi.org/10.3390/healthcare11081139>
- Loh, W.-Y. and Y.-S. Shih (1997). Split selection methods for classification trees. *Statistica Sinica* 7, 815–840.
- Ma, Z., Li, R., Li, T. et al. (2020). A data-driven risk measurement model of software developer turnover. *Soft Comput*, 24, 825–842. <https://doi.org/10.1007/s00500-019-04540-z>
- Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., ... Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. <https://doi.org/10.1109/TKDE.2019.2962680>
- Matsuo, M., Matsumoto, K., Misako Higashijima, Susumu Shirabe, Tanaka, G., Yoshida, Y., Higashi, T., Hiroya Miyabara, Komatsu, Y., & Iwanaga, R. (2023). Diagnostic model for preschool workers' unwillingness to continue working: Developed using machine-learning techniques. *Medicine*, 102(2), e32630–e32630. <https://doi.org/10.1097/md.00000000000032630>
- Mu, G. M., Gordon, D., Xu, J., Cayas, A., & Madesi, S. (2023). Benefits and limitations of partnerships amongst families, schools and universities: A systematic literature review. *International Journal of Educational Research*, 120. <https://doi.org/10.1016/j.ijer.2023.102205>
- Nandialath, A. M., David, E., Das, D., & Mohan, R. (2018). Modeling the determinants of turnover intentions: a Bayesian approach. *Evidence-Based HRM: A Global Forum for Empirical Scholarship*, 6(1), 2–24. <https://doi.org/10.1108/ebhrm-10-2016-0025>
- Ngo-Hencha, P. E. (2017). A Review of Existing Turnover Intention Theories. *International Scholarly and Scientific Research & Innovation*, 11(11), 2760–2767.
- Nijjer, S., Singh, J., & Raj, S. (2019). Developing HRIS for predictive attrition and retention management of indian it engineers- using ANN, ANOVA and smart PLS. *International Journal of Innovative Technology and Exploring Engineering*, 8(9 Special Issue), 710–715. <https://doi.org/10.35940/ijitee.I1114.0789S19>

- O'Connell, M., & Mei-Chuan, K. (2007). The Cost of Employee Turnover. *Industrial Management*, 49(1), 14–19.
- Ogbeibu, S., Chiappetta Jabbour, C. J., Burgess, J., Gaskin, J., & Renwick, D. W. S. (2021). Green talent management and turnover intention: the roles of leader STARA competence and digital task interdependence. *Journal of Intellectual Capital*, 23(1), 27–55. <https://doi.org/10.1108/jic-01-2021-0016>
- Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, 11, 169–198. <https://doi.org/10.1613/jair.614>
- Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M.-M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, 372. <https://doi.org/10.1136/bmj.n160>
- PRISMA. (2020). PRISMA. Prisma-Statement.org. <http://www.prisma-statement.org/>
- Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.
- Park, J., Kwon, S. & Jeong, SP. (2023) A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks. *Journal of Big Data*, 10(1), 36. <https://doi.org/10.1186/s40537-023-00715-6>
- Raman, R., Bhattacharya, S., & Pramod, D. (2019). Predict employee attrition by using predictive analytics. *Benchmarking: An International Journal*, 26(1), 2–18. <https://doi.org/10.1108/bij-03-2018-0083>
- Rombaut, E., & Guerry, M.-A. (2018). Predicting voluntary turnover through human resources database analysis. *Management Research Review*, 41(1), 96–112. <https://doi.org/10.1108/mrr-04-2017-0098>
- Rombaut, E., & Guerry, M.-A. (2020). The effectiveness of employee retention through an uplift modeling approach. *International Journal of Manpower*, 41(8), 1199–1220. <https://doi.org/10.1108/ijm-04-2019-0184>
- Rombaut, E., & Guerry, M.-A. (2021). Determinants of voluntary turnover: a data-driven analysis for blue and white collar workers. *Work: A Journal of Prevention, Assessment & Rehabilitation*, 69(3), 1083–1101. <https://doi.org/10.3233/WOR-213538>
- Ryan, J. C. (2020). Retaining, resigning and firing: bibliometrics as a people analytics tool for examining research performance outcomes and faculty turnover. *Personnel Review*, 50(5), 1316–1335. <https://doi.org/10.1108/pr-12-2019-0676>
- Sagie, A., Birati, A. & Tziner, A. (2002). Assessing the costs of behavioral and psychological withdrawal: a new model and an empirical illustration. *Applied Psychology: An International Review*, 51(1), 67–89. <https://doi.org/10.1111/1464-0597.00079>
- Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, 104(10), 1207–1225. <https://doi.org/10.1037/apl0000405>
- Salehi, G., Díaz, E., & Redondo, R. (2023). Forty-five years of research on vegetarianism and veganism: A systematic and comprehensive literature review of quantitative studies. *Heliyon*, 9(5). <https://doi.org/10.1016/j.heliyon.2023.e16091>
- Schröer C., Kruse F., Gómez J.M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Comput. Sci.*, 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Sexton, S., McMurtrey, S., Michalopoulos, O. & Smith, M. (2005). Employee turnover: a neural network solution. *Computers & Operations Research*, 32(10), 2635–2651. <https://doi.org/10.1016/j.cor.2004.06.022>
- Shaw, D., Duffy, K., Johnson, L., & Lockhart, E. (2005). Turnover, social capital losses, and performance. *Academy of Management Journal*, 48(4), 594–606. <https://doi.org/10.5465/amj.2005.17843940>
- Solano, J. A., Lancheros Cuesta, D. J., Umaña Ibáñez, S. F., & Coronado-Hernández, J. R. (2021). Predictive models assessment based on CRISP-DM methodology for students performance in

- Colombia - Saber 11 Test. In *Procedia Computer Science*, 198, 512–517. <https://doi.org/10.1016/j.procs.2021.12.278>
- Stark, S., Chernyshenko, O., Drasgow, F., Lee, W., White, L., & Young, M. (2011). Optimizing prediction of attrition with the U.S. Army's Assessment of Individual Motivation (AIM). *Military Psychology*, 23(2), 180–201. <https://doi.org/10.1080/08995605.2011.550234>
- Thompson, S. C., Holmgren, A. J., & Ford, E. W. (2020). Information system use antecedents of nursing employee turnover in a hospital setting. *Health Care Management Review*, Publish Ahead of Print. <https://doi.org/10.1097/hmr.0000000000000308>
- Tóth, Á., Suta, A., Pimentel, J., & Argoti, A. (2023). A comprehensive, semi-automated systematic literature review (SLR) design: Application to P-graph research with a focus on sustainability. *Journal of Cleaner Production*, 415. <https://doi.org/10.1016/j.jclepro.2023.137741>
- Vanegas, J. M., Wine, W., & Drasgow, F. (2021). Predictions of attrition among US Marine Corps: Comparison of four predictive methods. *Military Psychology*, 34(2), 147–166. <https://doi.org/10.1080/08995605.2021.1978754>
- Vera-Baceta, M. A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*. Springer Netherlands. <https://doi.org/10.1007/s11192-019-03264-z>
- Vrigazova, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. *Business Systems Research*, 12(1), 228–242. <https://doi.org/10.2478/bsrj-2021-0015>
- Wasmuth, J., & Davis, W. (1983). Strategies for Managing Employee Turnover. *Cornell Hotel and Restaurant Administration Quarterly*, 24(2), 65–75. <https://doi.org/10.1177/001088048302400216>
- Younis, S., & Ahsan, A. (2021). Know Your Stars Before They Fall Apart: A Social Network Analysis of Telecom Industry to Foster Employee Retention using Data Mining Technique. *IEEE Access*, 9. <https://doi.org/10.1109/access.2021.3050327>
- Yuan, S., Kroon, B., & Kramer, A. (2021). Building prediction models with grouped data: A case study on the prediction of turnover intention. *Human Resource Management Journal*. <https://doi.org/10.1111/1748-8583.12396>

Anexos

Anexo A - Fórmula da variável Nº mudanças categoria profissional

Nº mudanças categoria profissional = SOMA(SE(INCLUIR(FILTRAR(Categ. Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-2)));"NA");-1)="NA";0;SE (INCLUIR(FILTRAR(Categ.Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-2)));"NA");-1)=INCLUIR(FILTRAR(Categ.Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-1)));"NA");-1;0;1);SE(INCLUIR(FILTRAR(Categ.Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-3)));"NA");-1)="NA";0; SE(INCLUIR(FILTRAR(Categ.Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-3)));"NA");-1)=INCLUIR(FILTRAR(Categ.Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-2)));"NA");-1;0; SE(INCLUIR(FILTRAR(Categ. Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-3)));"NA");-1)=INCLUIR(FILTRAR(Categ. Prof.:(Nºcolab=Nºcolab)*(data<=data_análise(n-1)));"NA");-1;0;1)))

Anexo B - Parametrizações e resultados dos diferentes modelos de cada algoritmo

Parametrizações e resultados dos modelos com algoritmo CART

Modelos - CART					
Parameterizações					
	A	B	C	D	E
Ensemble	Não	Não	Não	Não	Boosting
Profundidade máxima	5	5	5	7	5
Nº casos Nó Pai	2%	2%	2%	2	2%
Nº casos Nó Filho	1%	1%	1%	1	1%
Custos de má classificação	Não	Não	Não	Não	Não
Amostra em análise	0 e 1	0, 1 e 2	0 e 2	0 e 2	0 e 2
Avaliação - amostra teste					
	A	B	C	D	E
PCCC	75,43%	78,63%	76,11%	76,11%	78,51%
Sensibilidade	84,65%	54,92%	79,35%	68,22%	76,11%
Especificidade	65,19%	91,50%	72,51%	84,87%	78,51%
Precisão	72,97%	77,79%	76,22%	83,36%	79,73%

Parametrizações e resultados dos modelos com algoritmo CHAID

Modelos - CHAID					
Parameterizações					
	F	G	H	I	J
Ensemble	Não	Não	Boosting	Bagging	Bagging
Profundidade máxima	5	5	5	5	6
Nº casos Nó Pai	2%	2%	2%	2%	50
Nº casos Nó Filho	1%	1%	1%	1%	25
Custos de má classificação	Não	Sim	Sim	Sim	Sim
Amostra em análise	0 e 2	0 e 2	0 e 2	0 e 2	0 e 2
Avaliação - amostra teste					
	F	G	H	I	J
PCCC	77,25%	75,94%	76,68%	77,47%	78,61%
Sensibilidade	73,08%	80,00%	77,08%	77,41%	78,81%
Especificidade	81,87%	71,43%	76,23%	77,55%	78,39%
Precisão	81,74%	75,66%	78,27%	79,29%	80,20%

Parametrizações e resultados dos modelos com algoritmo QUEST

Modelos - QUEST					
Parameterizações					
	K	L	M	N	O
<i>Ensemble</i>	Não	Não	Boosting	Bagging	Boosting
Profundidade máxima	5	5	6	6	7
Nº casos Nó Pai	2%	2%	2%	2%	2%
Nº casos Nó Filho	1%	1%	1%	1%	1%
Custos de má classificação	Não	Sim	Sim	Sim	Sim
Amostra em análise	0 e 2	0 e 2	0 e 2	0 e 2	0 e 2
Avaliação - amostra teste					
	K	L	M	N	O
PCCC	76,11%	76,39%	76,34%	75,71%	76,96%
Sensibilidade	73,73%	77,41%	76,00%	75,78%	77,08%
Especificidade	78,75%	75,27%	76,71%	75,63%	76,83%
Precisão	79,39%	77,66%	78,37%	77,54%	78,70%

Parametrizações e resultados dos modelos com algoritmo Tree-AS

Modelos - Tree-AS					
Parameterizações					
	P	Q	R	S	T
<i>Ensemble</i>	Não	Não	Não	Não	Não
Profundidade máxima	5	5	6	7	6
Nº casos Nó Pai	2%	2%	50	2%	2%
Nº casos Nó Filho	1%	1%	25	1%	1%
Custos de má classificação	Não	Sim	Sim	Não	Não
Amostra em análise	0 e 2	0 e 2	0 e 2	0 e 2	0 e 2
Avaliação - amostra teste					
	P	Q	R	S	T
PCCC	76,34%	76,34%	76,73%	76,17%	76,62%
Sensibilidade	76,32%	76,32%	78,05%	76,43%	76,32%
Especificidade	76,35%	76,35%	75,27%	75,87%	76,95%
Precisão	78,18%	78,18%	77,80%	77,86%	78,62%

Parametrizações e resultados dos modelos com algoritmo C5.0

Modelos - C5.0					
Parameterizações					
	U	V	W	X	Y
<i>Ensemble</i>	Não	Não	Não	Boosting	Não
Profundidade máxima	5	5	5	5	5
Nº casos Nó Pai	2%	2%	2%	50%	2%
Nº casos Nó Filho	1%	1%	1%	25%	1%
Custos de má classificação	Não	Sim	Sim	Não	Sim
Amostra em análise	0 e 2	0 e 2	0 e 2	0 e 2	0 e 2
Avaliação - amostra teste					
	U	V	W	X	Y
PCCC	76,73%	76,96%	77,99%		77,36%
Sensibilidade	75,03%	79,57%	80,65%		77,30%
Especificidade	78,63%	74,07%	75,03%		77,43%
Precisão	79,59%	77,31%	78,20%		79,18%

Anexo C – Distribuição das características profissionais por *turnover* voluntário

Tabela: Distribuição das primeiras nove características por *turnover* voluntário

		Ativos (n=28940)		Saídas Voluntárias (n=2897)		Total (n=31837)		Teste estatístico
		Colaboradores	%	Colaboradores	%	Colaboradores	%	Medida de associação
Setor	Automóvel	197	0,68	10	0,35	207	0,65	$\chi^2(10)=408,129;p<0,001$ Cramer V=0,113
	Concessões e Serviços	4410	15,24	549	18,95	4959	15,58	
	Construção	13754	47,54	1169	40,35	14923	46,89	
	Construção - Obras ferroviárias	2318	8,01	120	4,14	2438	7,66	
	Construção - Obras marítimas	66	0,23	1	0,03	67	0,21	
	Construção - Obras subterrâneas	4338	14,99	535	18,47	4873	15,31	
	Construção - Reabilitação de estrut.	44	0,15	0	0	44	0,14	
	Corporativo	1398	4,83	44	1,52	1442	4,53	
	Distribuição	372	1,29	56	1,93	428	1,34	
	Hotelaria	1705	5,89	373	12,88	2078	6,53	
	Imobiliária	329	1,14	40	1,38	369	1,16	
Total		28931	100	2897	100	31828	100	
Área de RH	Açores	84	0,29	7	0,24	91	0,29	$\chi^2(3)=44,937;p<0,001$ Cramer V=0,038
	Continente	23953	82,79	2279	78,67	26232	82,42	
	Estrangeiro	4621	15,97	596	20,57	5217	16,39	
	Madeira	273	0,94	15	0,52	288	0,9	
	Total	28931	100	2897	100	31828	100	
País trabalho	África	3312	11,47	533	18,42	3845	12,1	$\chi^2(4)=153,541;p<0,001$ Cramer V=0,070
	América	840	2,91	26	0,9	866	2,73	
	Europa	294	1,02	26	0,9	320	1,01	
	Médio Oriente	125	0,43	7	0,24	132	0,42	
	Portugal	24310	84,17	2301	79,54	26611	83,75	
	Total	28881	100	2893	100	31774	100	
Relação de emprego	Destacado	3843	13,28	444	15,33	4287	13,47	$\chi^2(3)=75,845;p<0,001$ Cramer V=0,049
	Expatriado	767	2,65	150	5,18	917	2,88	
	Lic.s/Venciment	11	0,04	3	0,1	14	0,04	
	Padrão	24310	84,03	2300	79,39	26610	83,61	
	Total	28931	100	2897	100	31828	100	
Concelho igual	Não	25431	88,44	2509	87,57	27940	88,36	$\chi^2(1)=1,914;p=0,169$ Cramer V=0,008
	Sim	3323	11,56	356	12,43	3679	11,64	
	Total	28754	100	2865	100	31619	100	
Distrito igual	Não	14931	51,93	1423	49,67	16354	51,72	$\chi^2(1)=5,321;p=0,022$ Cramer V=0,013
	Sim	13823	48,07	1442	50,33	15265	48,28	
	Total	28754	100	2865	100	31619	100	
Nº mudanças categoria profissional		M=0,13	DP=0,37	M=0,14	DP=0,41	M=0,13	DP=0,37	t(3369,178)=-1,684;p=0,092 eta=0,036
Motivo alteração catg. Prof. (ano N-1)	Não aplicável	25169	86,97	2596	89,61	27765	87,21	$\chi^2(3)=102,127;p<0,001$ Cramer V=0,057
	Promoção mérito	2287	7,9	117	4,04	2404	7,55	
	Promoção legal	670	2,32	125	4,31	795	2,5	
	Reclassificação	814	2,81	59	2,04	873	2,74	
	Total	28940	100	2897	100	31837	100	
Motivo alteração catg. Prof. (ano N-2)	Não aplicável	27377	94,6	2784	96,1	30161	94,74	$\chi^2(3)=13,237;p<0,001$ Cramer V=0,020
	Promoção mérito	1011	3,49	73	2,52	1084	3,4	
	Promoção legal	217	0,75	20	0,69	237	0,74	
	Reclassificação	335	1,16	20	0,69	355	1,12	
	Total	28940	100	2897	100	31837	100	

Tabela: Distribuição das seguintes catorze características por *turnover* voluntário

		Ativos (n=28940)		Saídas Voluntárias (n=2897)		Total (n=31837)		Teste estatístico
		Colaboradores	%	Colaboradores	%	Colaboradores	%	Medida de associação
Motivo alteração catg. Prof. (ano N-3)	Não aplicável	27614	95,42	2830	97,69	30444	95,62	$\chi^2(3)=34,183$; $p<0,001$ Cramer V=0,033
	Promoção mérito	838	2,9	48	1,66	886	2,78	
	Promoção legal	187	0,65	10	0,35	197	0,62	
	Reclassificação	301	1,04	9	0,31	310	0,97	
	Total	28940	100	2897	100	31837	100	
Motivo transf. no Grupo (ano N-1)	Não aplicável	26791	92,57	2851	98,41	29642	93,11	$\chi^2(3)=140,012$; $p<0,001$ Cramer V=0,066
	Início de destacamento	740	2,56	13	0,45	753	2,37	
	Fim de destacamento	531	1,83	13	0,45	544	1,71	
	Mudança de empresa	878	3,03	20	0,69	898	2,82	
	Total	28940	100	2897	100	31837	100	
Motivo transf. no Grupo (ano N-2)	Não aplicável	27931	96,51	2867	98,96	30798	96,74	$\chi^2(3)=51,018$; $p<0,001$ Cramer V=0,040
	Início de destacamento	336	1,16	7	0,24	343	1,08	
	Fim de destacamento	282	0,97	7	0,24	289	0,91	
	Mudança de empresa	391	1,35	16	0,55	407	1,28	
	Total	28940	100	2897	100	31837	100	
Motivo transf. no Grupo (ano N-3)	Não aplicável	28082	97,04	2870	99,07	30952	97,22	$\chi^2(3)=40,628$; $p<0,001$ Cramer V=0,036
	Início de destacamento	305	1,05	12	0,41	317	1	
	Fim de destacamento	210	0,73	5	0,17	215	0,68	
	Mudança de empresa	343	1,19	10	0,35	353	1,11	
	Total	28940	100	2897	100	31837	100	
Motivo manut. Contrato (ano N-1)	Não aplicável	22542	77,89	2016	69,59	24558	77,14	$\chi^2(2)=175,778$; $p<0,001$ Cramer V=0,074
	Passagem a efetivo	1268	4,38	82	2,83	1350	4,24	
	Renovação	5130	17,73	799	27,58	5929	18,62	
	Total	28940	100	2897	100	31837	100	
Motivo manut. Contrato (ano N-2)	Não aplicável	26216	90,59	2507	86,54	28723	90,22	$\chi^2(2)=62,308$; $p<0,001$ Cramer V=0,044
	Passagem a efetivo	659	2,28	66	2,28	725	2,28	
	Renovação	2065	7,14	324	11,18	2389	7,5	
	Total	28940	100	2897	100	31837	100	
Motivo manut. Contrato (ano N-3)	Não aplicável	26793	92,58	2662	91,89	29455	92,52	$\chi^2(2)=16,030$; $p<0,001$ Cramer V=0,022
	Passagem a efetivo	614	2,12	41	1,42	655	2,06	
	Renovação	1533	5,3	194	6,7	1727	5,42	
	Total	28940	100	2897	100	31837	100	
Tipo de contrato	Sem Termo	18419	63,65	539	18,61	18958	59,55	$\chi^2(2)=2220,863$; $p<0,001$ Cramer V=0,264
	Termo Certo	6862	23,71	1504	51,92	8366	26,28	
	Termo Incerto	3659	12,64	854	29,48	4513	14,18	
	Total	28940	100	2897	100	31837	100	
Anos efetivo		M=7,8	DP=10,87	M=0,75	DP=2,83	M=7,16	DP=10,59	t(14561,442)=85,102; $p<0,001$ eta=0,224
Regresso baixa prolong.	0	28840	99,65	2890	99,76	31730	99,66	$\chi^2(1)=0,849$; $p=0,406$ Cramer V=0,005
	1	100	0,35	7	0,24	107	0,34	
	Total	28940	100	2897	100	31837	100	
Antiguidade		M=9,21	DP=10,39	M=1,65	DP=3,45	M=8,52	DP=10,19	t(9714,759)=85,33; $p<0,001$ eta=0,246
Horas de trabalho dia		M=8,02	DP=0,57	M=7,89	DP=0,58	M=8,01	DP=0,58	t(31835)=11,572; $p<0,001$ eta=0,113
Dias de trabalho por semana		M=5,01	DP=0,35	M=5,05	DP=0,41	M=5,01	DP=0,36	t(3337,852)=-5,143; $p<0,001$ eta=0,067
Avaliação média		M=3,33	DP=1,67	M=1,61	DP=1,99	M=3,17	DP=1,77	t(3315,927)=44,965; $p<0,001$ eta=0,357