

UNIVERSIDADE FEDERAL FLUMINENSE

ELAINE FLAVIO RANGEL SEIXAS

**UMA ABORDAGEM BASEADA EM
TECNOLOGIAS DE INTELIGÊNCIA
ARTIFICIAL PARA RETENÇÃO DE TALENTOS**

NITERÓI

2025

ELAINE FLAVIO RANGEL SEIXAS

**UMA ABORDAGEM BASEADA EM
TECNOLOGIAS DE INTELIGÊNCIA
ARTIFICIAL PARA RETENÇÃO DE TALENTOS**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação.

Orientador:
Prof. Dr. José Viterbo Filho

Coorientadora:
Profa. Dra. Flavia Cristina Bernardini

NITERÓI

2025

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S457a Seixas, Elaine Flavio Rangel
UMA ABORDAGEM BASEADA EM TECNOLOGIAS DE INTELIGÊNCIA
ARTIFICIAL PARA RETENÇÃO DE TALENTOS / Elaine Flavio Rangel
Seixas. - 2025.
140 f.: il.

Orientador: José Viterbo Filho.
Coorientador: Flavia Cristina Bernardini.
Tese (doutorado)-Universidade Federal Fluminense, Instituto
de Computação, Niterói, 2025.

1. Retenção de talentos. 2. Rotatividade de funcionários.
3. Aprendizado de máquina. 4. Sistema de recomendação. 5.
Produção intelectual. I. Filho, José Viterbo, orientador.
II. Bernardini, Flavia Cristina, coorientadora. III.
Universidade Federal Fluminense. Instituto de Computação.
IV. Título.

CDD - XXX

ELAINE FLAVIO RANGEL SEIXAS

**UMA ABORDAGEM BASEADA EM TECNOLOGIAS DE INTELIGÊNCIA ARTIFICIAL
PARA RETENÇÃO DE TALENTOS**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito para a obtenção do Grau de Doutor em Computação. Área de concentração: Ciência da Computação.

Aprovada em 04 de julho de 2025.

BANCA EXAMINADORA

Prof. Dr. José Viterbo Filho - Orientador, UFF

Prof. Dra. Flavia Cristina Bernardini, Coorientadora, UFF

Prof. Dr. Cristiano Maciel, UFMT

Prof. Dra. Raissa dos Santos Barcellos, UERJ

Prof. Dr. Daniel Oliveira, UFF

Prof. Dra. Luciana Cardoso de Castro Salgado, UFF

Niterói

2025

“A parte que ignoramos é muito maior que tudo quanto sabemos.” (Platão)

Agradecimentos

Agradecimentos especiais aos meus pais pelo alicerce de minha educação e formação de caráter, sempre presentes e confiantes em meu sucesso; pela amizade e respeito mútuo.

Agradeço ao meu marido Flávio, por estar sempre presente em todos os momentos da minha vida, pela cumplicidade e apoio incondicional.

Aos meus irmãos e cunhadas, Vinícius, Elisa, Felipe e Gleize agradeço pelo apoio e confiança. Aos meus sobrinhos, Pedro, João e Isabela, obrigada por fazerem parte da minha vida e tornarem tudo mais leve.

À família Seixas, Maria Helena, Mozart e Letícia, obrigada pelo incentivo e apoio. Sempre próximos, mesmo a distância.

Aos meus orientadores, José Viterbo e Flavia Bernardini, que me mostraram os caminhos a serem seguidos e pela confiança depositada, principalmente quando mudanças de rumo foram necessárias.

Aos meus amigos e colegas da UFF, em especial Mônica da Silva, pelas inúmeras discussões, que proporcionaram necessários ajustes de rota, além das parcerias em pesquisas e publicações.

Agradeço aos membros da banca que concordaram em participar da avaliação desta tese: Cristiano Maciel, Raissa Barcellos, Daniel Oliveira e Luciana Salgado.

Agradeço às agências CAPES e CNPQ pelos recursos essenciais para manutenção da pesquisa e publicações. Adicionalmente agradeço à empresa parceira Prime Up pela confiança, em especial ao José Blaschek, sempre disponível, com sua fundamental participação para a entrega deste trabalho.

Agradeço a todos que de alguma forma contribuíram para o desenvolvimento desta tese, mesmo que sem saber. Obrigada!

Resumo

O setor de Tecnologia da Informação (TI) é caracterizado por dinamicidade e complexidade, exigindo de empresas e profissionais uma cultura contínua de desenvolvimento. Um dos principais desafios enfrentados é a alta rotatividade de funcionários, que impacta negativamente a gestão de equipes e o capital intelectual das organizações. Para mitigar esse problema, a literatura propõe modelos baseados em aprendizado de máquina voltados à previsão do risco de desligamento voluntário. Além da previsão, estratégias de retenção de talentos surgem como complementares e essenciais para a redução da rotatividade. Neste contexto, esta tese propõe ferramentas computacionais para apoiar gestores na identificação de risco de desligamento e na retenção de profissionais em empresas de TI. A pesquisa foi conduzida com base na metodologia *Design Science Research* (DSR), que assegura rigor, relevância e inovação na construção de artefatos. Como etapa inicial, realizamos um mapeamento e uma revisão sistemática da literatura, com o objetivo de compreender o estado da arte e identificar lacunas no uso de tecnologias de inteligência artificial aplicadas à gestão de pessoas. Com base nesses achados, desenvolvemos um *framework* composto por duas soluções integradas: (i) um modelo preditivo de desligamento voluntário, baseado em aprendizado de máquina, sendo o classificador *Random Forest* o que apresentou melhor desempenho na prova de conceito (acurácia de 92% e *F1-score* de 90%); e (ii) uma proposta de retenção que realiza o *matching* entre os perfis dos colaboradores e os projetos disponíveis, auxiliando o gestor na alocação adequada de talentos. Essa abordagem também favorece o desenvolvimento profissional, funcionando como uma ferramenta complementar à gestão estratégica de pessoas.

Palavras-chave: Retenção de talentos, Rotatividade de funcionários, Aprendizado de máquina, Sistema de recomendação, *Design Science Research*, Inteligência artificial, Gestão de pessoas.

Abstract

The Information Technology (IT) sector is marked by dynamism and complexity, requiring companies and professionals to adopt a culture of continuous development. One of the main challenges faced in this context is the high employee turnover, which negatively impacts team management and organizational knowledge. To address this issue, the literature proposes machine learning-based models aimed at predicting voluntary turnover risk. In addition to prediction, talent retention strategies emerge as essential and complementary to reducing turnover. In this context, this thesis proposes computational tools to support managers in identifying the risk of voluntary departure and retaining professionals in IT companies. The research is grounded in the Design Science Research (DSR) methodology, which ensures rigor, relevance, and innovation in artifact development. As a preliminary step, we conducted a systematic mapping and a systematic literature review to understand the state of the art and identify gaps in the application of artificial intelligence technologies to human resource management. Based on these findings, we developed a framework composed of two integrated solutions: (i) a voluntary turnover prediction model based on machine learning, with the Random Forest classifier achieving the best performance (92% accuracy and 90% F1-score); and (ii) a retention proposal that matches employee profiles with available projects, assisting managers in allocating talent appropriately. This approach also promotes professional development, serving as a complementary tool for strategic talent management.

Keywords: Talent retention, Employee turnover, Machine learning, Recommendation system, Design Science Research, Artificial intelligence, People management.

Listas de Figuras

Fig. 1	Visão geral dos ciclos da DSR aplicados ao desenvolvimento da tese	16
Fig. 2	Critérios de Seleção de Estudos	35
Fig. 3	Processo de busca e seleção de estudos primários	36
Fig. 4	Cronologia das publicações na Fase 2.	38
Fig. 5	Cronologia das publicações na Fase 3.	38
Fig. 6	Cronologia das publicações aceitas	39
Fig. 7	Distribuição dos estudos primários aceitos por questão de pesquisa	40
Fig. 8	Aspectos relevantes para migração ou retenção de talentos	42
Fig. 9	Critérios de Seleção de Estudos	52
Fig. 10	Processo de busca e seleção de estudos primários	53
Fig. 11	Total de trabalhos publicados entre 2014 e maio de 2024	54
Fig. 12	Total de trabalhos aceitos por base entre 2014 e maio de 2024	55
Fig. 13	Distribuição de estudos por tipo de publicação e ano	55
Fig. 14	Tipos de proposta de modelo dos estudos	56
Fig. 15	Classificadores utilizados nos estudos primários selecionados	57
Fig. 16	Classificadores utilizados nos estudos selecionados agrupados por categoria	57
Fig. 17	Utilização de técnicas de balanceamento de dados nos estudos	58
Fig. 18	<i>Macroprocesso da proposta para retenção de talentos</i>	64
Fig. 19	<i>Framework</i> proposto: Processo de aprendizado de máquina para predição de risco de desligamento voluntário	65
Fig. 20	<i>Framework</i> proposto para retenção de talentos: Indicação do projeto certo para o funcionário	70
Fig. 21	Etapas para realização do experimento	76
Fig. 22	Matriz de Correlação dos atributos	78
Fig. 23	Processo de aprendizado de máquina para predição de risco de desligamento voluntário	82
Fig. 24	Matriz de Confusão para classificador <i>Random Forest</i>	86
Fig. 25	Curva ROC para classificador <i>Random Forest</i>	87
Fig. 26	Matriz de Confusão para classificador <i>Random Forest</i> com <i>Imputer</i>	88
Fig. 27	Curva ROC para classificador <i>Random Forest</i> com <i>Imputer</i>	88

Fig. 28	Matriz de Confusão para classificador <i>Logistic Regression</i>	89
Fig. 29	Curva ROC para classificador <i>Logistic Regression</i>	90
Fig. 30	Matriz de Confusão para classificador SVM	91
Fig. 31	Curva ROC para classificador SVM	91
Fig. 32	Matriz de Confusão para classificador KNN	92
Fig. 33	Curva ROC para classificador KNN	93
Fig. 34	Matriz de Confusão para classificador <i>Gradient Boosting</i>	94
Fig. 35	Curva ROC para classificador <i>Gradient Boosting</i>	94
Fig. 36	Matriz de Confusão para classificador XGBoost	95
Fig. 37	Curva ROC para classificador XGBoost	96
Fig. 38	Top 10 atributos do <i>Random Forest</i>	98
Fig. 39	Processo do sistema de recomendação para retenção de talentos	100
Fig. 40	Rankeamento dos projetos por funcionário com base na cobertura de competências e grau médio de habilidade	103

Lista de Tabelas

Tabela1	Objetivo do Mapeamento Sistemático segundo paradigma GQM	32
Tabela2	Distribuição dos artigos selecionados nas fases do processo de busca e seleção de estudos primários 37	
Tabela3	Estudos relevantes selecionados	41
Tabela4	<i>String</i> de busca	50
Tabela5	Descrição dos Atributos da Base de Funcionários	79
Tabela6	Descrição dos Atributos da Base de Projetos	81
Tabela7	Relatório de classificação do <i>Random Forest</i>	86
Tabela8	Relatório de classificação do <i>Random Forest</i> com <i>imputer</i>	87
Tabela9	Relatório de classificação do <i>Logistic Regression</i>	89
Tabela10	Relatório de classificação do <i>Support Vector Machine</i>	90
Tabela11	Relatório de classificação do KNN	92
Tabela12	Relatório de classificação do <i>Gradient Boosting</i>	93
Tabela13	Relatório de classificação do XGBoost	95
Tabela14	Comparação de Modelos Classificadores	97
Tabela15	Descrição dos Hiperparâmetros por Modelo Classificador	97
Tabela16	Aspectos relevantes para migração ou retenção de talentos verificados no MSL	128
Tabela17	Estudos relevantes selecionados	130
Tabela18	Formulário de avaliação de qualidade da RSL	134

Lista de Abreviaturas e Siglas

Natural Language Generation - NLG Geração de Linguagem Natural

ABES Associação Brasileira das Empresas de Software

CEP Comitê de Ética em Pesquisa

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

DAI Doutorado Acadêmico para Inovação

DL *Deep Learning*

DSR *Design Science Research*

DTC *Decision Tree Classifier*

ENN *Edited Nearest Neighbor*

FN *False Negative*

FP *False Positive*

GB *Gradient Boosting*

GBM *Gradient Boosting Machine*

GERH Gestão Estratégica de Recursos Humanos

GNB *Gaussian Naive Bayes*

GRH Gestão de Recursos Humanos

GT Gestão de Talentos

IA Inteligência Artificial

IDC *International Data Corporation*

KNN *k-Nearest Neighbor*

LGPD Lei Geral de Proteção de Dados

LR *Logistic Regression*

ML *Machine Learning*

MSL Mapeamento Sistemático da Literatura

PLN Processamento de Linguagem Natural

PLS-SEM *Variance based Structural Equation Modeling*

RF *Random Forest*

RH Recursos Humanos

ROC *Receiver Operating Characteristic*

RSL Revisão Sistemática da Literatura

SBC Sociedade Brasileira de Computação

SHAP *SHapley Additive exPlanations*

SMOTE *Synthetic Minority Over-sampling Technique*

SOL SBC-OpenLib

SVM *Support Vector Machine*

TI Tecnologia da Informação

TIC Tecnologia da Informação e Comunicação

TN *True Negative*

TP *True Positive*

UFF Universidade Federal Fluminense

XGBoost *Extreme-Gradient Boosting*

SUMÁRIO

1	Introdução	12
1.1	Motivação	14
1.2	Objetivos	15
1.3	Metodologia	15
1.4	Contribuições	18
1.5	Trabalhos publicados	19
1.5.1	Trabalhos relacionados com esta tese	19
1.5.2	Outras publicações	19
1.6	Organização do trabalho	20
2	Referencial teórico	22
2.1	Aspectos da gestão de recursos humanos	22
2.1.1	Rotatividade	23
2.1.2	Retenção de talentos	24
2.2	Inteligência artificial	25
2.2.1	Aprendizado de máquina	26
2.2.2	Sistemas de recomendação	28
2.3	Inteligência artificial em recursos humanos	29
3	Mapeamento sistemático	31
3.1	Protocolo do mapeamento sistemático	32
3.1.1	Objetivo e questões de pesquisa	32
3.1.2	Estratégia para busca dos estudos primários	33
3.1.3	Critérios para seleção dos estudos	34
3.1.4	Estratégia para extração de dados	35
3.2	Metodologia para busca e seleção de estudos primários	36
3.3	Resultados	37
3.3.1	Resultados quantitativos	37
3.3.2	Contribuições dos estudos	42
3.4	Discussão	46

3.5	Ameaças à validade	47
4	Revisão sistemática da literatura	48
4.1	Protocolo da revisão sistemática	49
4.1.1	Objetivo e questões de pesquisa	49
4.1.2	Estratégia para busca dos estudos primários	49
4.1.3	Critérios para seleção de estudos	50
4.1.4	Avaliação da qualidade	52
4.1.5	Estratégia para extração de dados	53
4.2	Resultados	53
4.2.1	Metodologia para busca e seleção de estudos primários	53
4.2.2	Resultados quantitativos	54
4.2.3	Contribuições dos estudos	56
4.3	Discussão	60
4.4	Ameaças à validade	61
5	Abordagem proposta	62
5.1	Introdução	62
5.2	Visão geral do <i>framework</i> proposto	64
5.3	<i>Framework</i> para previsão de risco de desligamento voluntário	64
5.4	Modelo para Indicação de Projetos	70
5.5	Discussão	71
6	Experimento	75
6.1	Metodologia	75
6.2	Base de dados	76
6.2.1	Matriz de correlação	77
6.2.2	Descrição dos atributos das bases de dados	79
6.3	Modelo para previsão de risco de desligamento voluntário	81
6.3.1	Construção do modelo	81
6.3.2	Resultados	85
6.3.2.1	Resultados dos treinamentos dos modelos	85
6.3.2.2	Comparação entre classificadores	96
6.4	Modelo para retenção de talentos	100
6.5	Discussão	104
7	Conclusão	105

7.1	Considerações finais	105
7.2	Limitações	107
7.3	Perspectivas futuras	108
Referências		109
Apêndice A – Tabela de aspectos relevantes para migração ou retenção de talentos		128
Apêndice B – Relação de estudos primários selecionados na RSL		130
Apêndice C – Formulário de avaliação de qualidade da RSL		134

1 Introdução

Os constantes avanços tecnológicos, econômicos e culturais representam um desafio em diferentes setores da sociedade. Para as empresas que atuam no setor de serviços de Tecnologia da Informação (TI), espera-se atingir uma estrutura organizacional e de desenvolvimento sustentável. Para os jovens funcionários, espera-se um plano de desenvolvimento que proporcione oportunidades de crescimento profissional, incluindo habilidades técnicas e comportamentais. As demandas de mercado em TI são bastante dinâmicas e complexas, o que exige tanto dos funcionários quanto das empresas uma cultura de desenvolvimento constante. Por parte dos funcionários de suas habilidades e conhecimentos e por parte da empresa de seus processos, produtos e serviços.

Sendo largamente conhecida e utilizada, a TI é amplamente entendida como a tecnologia capacitadora do século 21 ([CORTEZ; BATISTA; GOES, 2022](#); [JORGENSON; VU, 2007](#); [SCHMIDT; ALT; SHIRAZI, 2012](#)). Isto ficou mais evidente com a pandemia de COVID-19, com a manifestação da ubiquidade e importância da TI em todos os setores da indústria. Devido à pandemia, por intermédio dos serviços de TI, houve uma rápida mudança para o trabalho remoto, gestão de logística de forma remota, telemedicina, educação on-line, aumento do *e-commerce*, entre outros exemplos, evidenciando a importância do setor de TI em diversos setores da economia ([SCIENCES ENGINEERING; MEDICINE, 2020](#)).

O setor de TI desempenha um importante papel na maioria das empresas. Um problema conhecido neste setor é a elevada rotatividade das empresas, resultando em desafios para a gestão do conhecimento, o que pode causar um grande impacto na eficiência destas empresas, sem mencionar o custo envolvido. De acordo com dados do Novo CAGED, Cadastro Geral de Empregados e Desempregados do Governo Federal Brasileiro, a rotatividade no setor de TI em 2022 no Brasil foi de aproximadamente 40%. Considerando apenas os postos com curso superior completo, este índice é de aproximadamente 19%, um índice alto de acordo com especialistas do setor de Recursos Humanos (RH) ([BRASIL, 2022](#); [TOTVS, 2022](#)). Já em 2024, o índice de rotatividade ficou em torno de 35%, segundo dados do Novo CAGED, um índice considerado alto ([BRASIL, 2024](#)).

A atração e retenção de talentos representam um grande desafio para empresas que necessitam de mão de obra qualificada de tecnologia. A Brasscom (Associação das Empresas de Tecnologia da Informação e Comunicação (TIC) e de Tecnologias Digitais) divulgou o estudo intitulado “Demanda de Talentos em TIC e Estratégia ΣTCEM”. Realizado em dezembro de 2021, este estudo apresenta uma projeção da demanda de 797 mil novos taletos para o setor de Tecnologia entre 2021 e 2025, com uma média simples de 159 mil novos talentos demandados por ano ([BRASSCOM, 2021](#)). Outro dado importante deste estudo é a quantidade de formados com perfil tecnológico ao ano no Ensino Superior, que é de 53 mil. Isto representa um *deficit* significativo na formação de profissionais na área, acirrando a competição por profissionais no setor, o que impacta ainda mais a retenção de talentos e taxas de rotatividade.

De acordo com a prévia do Estudo Mercado Brasileiro de Software - Panorama e Tendências 2025, realizada por meio de *webinar* apresentado pela Associação Brasileira das Empresas de Software (ABES) e a *International Data Corporation* (IDC), o Brasil subiu da 12^a posição do ranking global de investimentos em TI em 2022 para 10^a posição em 2024. Sendo responsável por 34,7% dos investimentos da América Latina, o Brasil ocupa a 1^a posição em investimentos no setor para a região. O crescimento do mercado de TI no Brasil em 2024, que foi de 13,9%, superou o mundial que foi de 10,8%. Para 2025, há uma expectativa de crescimento de 9,5%, ainda superior à mundial, com expectativa de 8,9%. Com um total de 58,6 bilhões de dólares em investimentos no setor em 2024, a mão de obra do setor de TI é bastante disputada, acirrando a busca por talentos e, consequentemente, a rotatividade das empresas ([NETO, 2023, 2024, 2025](#); [MASSONI et al., 2019](#)).

Por sua vez, abordagens de Inteligência Artificial (IA) são amplamente utilizadas em ecossistemas de RH ([STROHMEIER; PIAZZA, 2015](#); [QAMAR et al., 2021](#); [BANKINS, 2021](#)). Trabalhos relatam a utilização de ferramentas de IA em RH para recrutamento e seleção de candidatos para as vagas em aberto das empresas, sendo este método de contratação amplamente utilizado em empresas de tecnologia ([HMOUD; LASZLO et al., 2019](#)). Outros trabalhos relatam o uso de ferramentas de IA para verificação do nível de satisfação dos funcionários, rotatividade, avaliação de desempenho de RH, entre outras aplicações ([QAMAR et al., 2021](#)). Estes trabalhos refletem principalmente a realidade fora do Brasil.

Uma pesquisa realizada no Brasil durante o período do terceiro trimestre de 2024, com mais de 110 profissionais de RH de diversos setores, sendo 32% do setor de tecnologia, e com 65% de empresas entre com até 499 funcionários, foi divulgada no relatório intitulado “O futuro do RH no Brasil” ([DEEL., 2024](#)). De acordo com dados desta pesquisa, mais de 70% dos respondentes informam que já automatizam processos de RH, sendo os mais automatizados: folha de paga-

mento, seleção de currículos, agendamento de entrevistas e avaliação de desempenho. Mais de 57% dos respondentes afirmam que não utilizam IA na rotina das estruturas de RH. Ainda segundo a pesquisa, aproximadamente 61% dos respondentes consideram importante a aplicação de IA no RH, contra aproximadamente 19% que dispensa a necessidade dessa tecnologia nas atividades do RH. Mais de 42% dos respondentes afirmam que não há adoção de *big data* e análises preditivas em processos de RH na empresa, embora 89% dos respondentes afirmem interesse em investir em tais tecnologias em seus processos de RH em até 2 anos, principalmente considerando gestão de talentos. Um dado importante da pesquisa foi o relato de aproximadamente 53% dos respondentes sobre a alta rotatividade associada à geração Z (profissionais entre 18 e 28 anos), o que reforça o desafio em relação à retenção de talentos.

Pesquisas do Gartner, uma empresa global de pesquisa e consultoria, apontam que gestores de RH estão buscando formas para retenção de talentos que não estejam necessariamente associadas à remuneração direta ou indireta ([WILES, 2022](#)). Entre as abordagens possíveis para retenção de talentos, está a utilização de ferramentas baseadas em tecnologias de IA, como aprendizado de máquina e sistemas de recomendação, por exemplo, que busquem formas de verificar o risco de desligamento voluntário do talento e recomendar formas alternativas de medidas de retenção. Uma forma ainda não explorada, de acordo com as pesquisas realizadas, seria a alocação do talento no projeto que ofereça oportunidade de desenvolvimento de conhecimentos e habilidades com desafios interessantes.

1.1 Motivação

A motivação desta tese surge da necessidade das empresas em reter seus profissionais altamente qualificados, considerados estratégicos e mapeados como talentos ([COSTA; DIAS et al., 2024](#)). Adicionalmente, tenta cobrir lacunas em relação a ferramentas para apoiar o gestor na retenção de talentos, buscando soluções com foco em desenvolvimento e baixo investimento financeiro para a empresa ([OWENS; KHAZANCHI, 2011](#)).

Para orientar esta tese, formulamos a seguinte questão de pesquisa:

QP. *Como a utilização de técnicas computacionais pode auxiliar o gestor a reter os talentos de sua equipe?*

Esta questão de pesquisa pode ser subdividida nas seguintes subquestões de pesquisa complementares:

- **SQP1.** *Quais aspectos são determinantes para retenção ou migração de talentos?*
- **SQP2.** *Que técnicas computacionais podem ser empregadas para promover retenção de talentos?*
- **SQP3.** *Como identificar o risco de um talento pedir para sair da empresa?*
- **SQP4.** *Uma vez identificado o risco de saída do talento, como promover a retenção deste talento?*

Buscando responder à questão de pesquisa, considera-se a utilização de técnicas computacionais baseadas em tecnologias de inteligência artificial.

1.2 Objetivos

O principal objetivo deste trabalho é propor ferramentas computacionais que ofereçam suporte aos gestores na retenção de talentos em empresas prestadoras de serviços de TI.

Para alcançar este objetivo, são definidos os seguintes objetivos específicos:

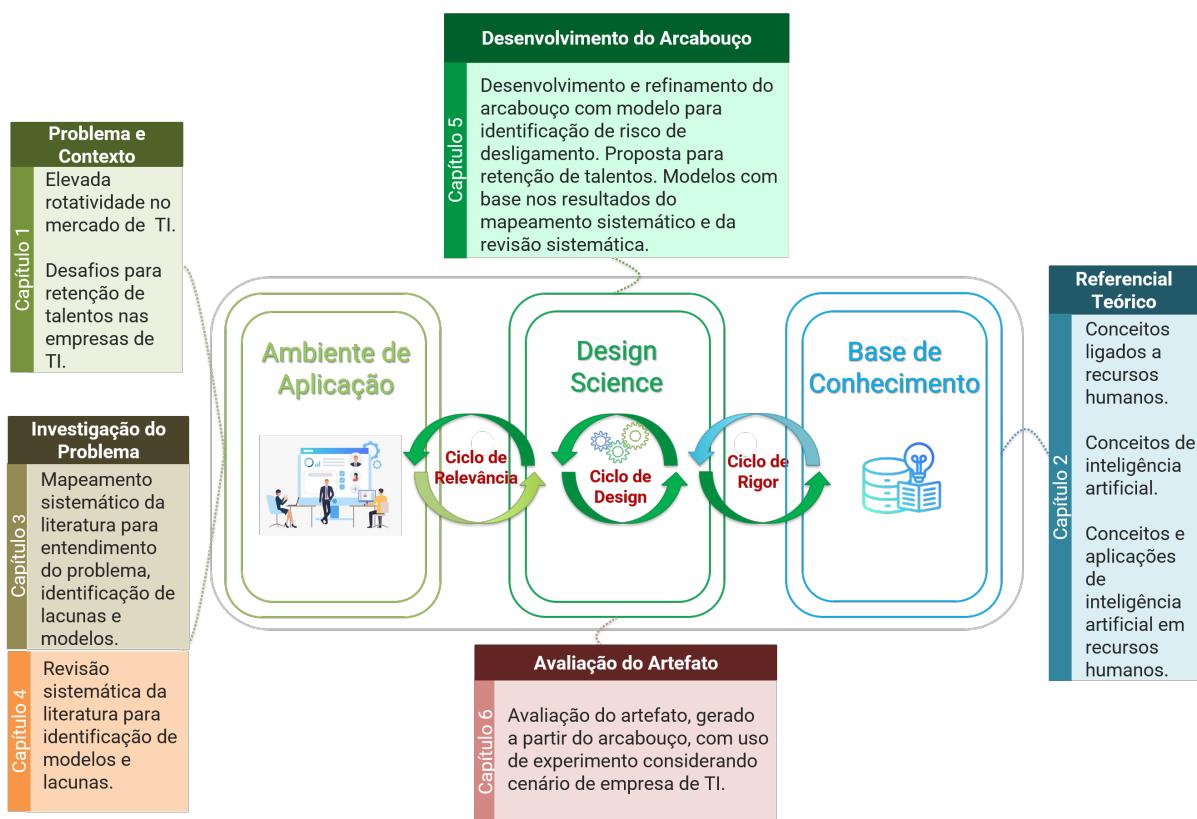
1. Identificar o estado da arte de estudos que abordem questões de previsão de desligamento voluntário dos profissionais e estudos sobre retenção de talentos. Desta forma, identificar melhores práticas e lacunas existentes;
2. Propor *framework* ou arcabouço que integre abordagens apoiadas em tecnologias de IA visando previsão de desligamento voluntário, adicionado de solução visando a retenção de talentos;
3. Realizar experimento para validação do arcabouço considerando cenário de uma empresa de TI.

1.3 Metodologia

Para atingir os objetivos supracitados, adotou-se pesquisa aplicada, por meio de *Design Science Research* (DSR). Sendo caracterizada por uma abordagem metodológica voltada para o *design* e a investigação de artefatos inovadores que solucionam problemas práticos, a DSR contribui simultaneamente para o avanço do conhecimento científico, desta forma, combinando inovação prática e rigor metodológico (HEVNER; MARCH et al., 2004; PEFFERS et al., 2007; DRESCH; LACERDA; ANTUNES JR, 2015; WIERINGA, 2014).

[Hevner e Chatterjee \(2010\)](#) propõem três ciclos interdependentes que estruturam a DSR: o ciclo de relevância, o ciclo de rigor e o ciclo de *design*. Essa visão destaca o equilíbrio necessário entre prática e teoria, assegurando que os artefatos sejam aprimorados iterativamente de acordo com as necessidades das partes interessadas para resolver um problema, ao mesmo tempo que sejam úteis e cientificamente válidos. A seguir são apresentados o conceito acerca de cada ciclo e uma visão geral das etapas realizadas durante cada ciclo, conforme ilustrado na Figura 1.

Figura 1: Visão geral dos ciclos da DSR aplicados ao desenvolvimento da tese



Fonte: Adaptado de [Hevner e Chatterjee \(2010\)](#) e [Nakamura \(2022\)](#)

Ainda de acordo com o modelo proposto nesta tese e com a proposta de [Hevner e Chatterjee \(2010\)](#), no ciclo de relevância, são identificados o problema e oportunidades de pesquisa para um determinado ambiente de aplicação, no qual o problema é observado, fornecendo requisitos e critérios para avaliação. Este ciclo conecta a pesquisa ao ambiente ambiente prático, garantindo que o problema abordado seja relevante.

Para uma maior investigação do problema nesta tese, foi realizado um Mapeamento Sistemático da Literatura (MSL), visando a identificação do referencial teórico, incluindo conceitos de retenção de talentos, informações sobre sistemas existentes para previsão de desligamento

voluntário e propostas de modelos para retenção de talentos utilizando inteligência artificial. Os achados no MSL suportaram os pressupostos iniciais, porém foram encontrados poucos estudos sobre previsão de desligamento voluntário, o que motivou uma maior investigação sobre este tema. Em seguida foi realizada uma Revisão Sistemática da Literatura (RSL), com o objetivo de buscar modelos de referência e entender como os modelos para previsão de risco de desligamento voluntário e retenção de talentos estavam sendo propostos. Desta forma, buscamos criar uma base sólida de conhecimento, ajudando a categorizar as tecnologias existentes e buscar lacunas a serem preenchidas.

O ciclo de rigor assegura que a pesquisa seja fundamentada em teorias e métodos científicos sólidos, ou seja, assegura a base científica da pesquisa. Desta forma, garante a validade teórica e metodológica (HEVNER; CHATTERJEE, 2010). Na seção de referencial teórico, foram apresentados conceitos ligados a RH. Ainda foram investigados conceitos de IA, além de conceitos de aplicação de IA em RH.

O ciclo central da DSR é o ciclo de *design*, no qual o artefato é projetado e aperfeiçoado iterativamente. Ele envolve a construção e avaliação do artefato, sendo alimentado pelos requisitos do ciclo de relevância e validado com base nos critérios definidos no ciclo de rigor (HEVNER; CHATTERJEE, 2010; WIERINGA, 2014).

A definição da arquitetura do arcabouço, que visa promover a identificação de risco de desligamento voluntário de um funcionário e retenção de talentos, foi baseada nos modelos encontrados nos estudos do MSL e nos estudos da RSL. A proposta de arcabouço foi sendo refinada a partir dos estudos da RSL, alcançando um modelo com utilização de IA para previsão de desligamento voluntário, com proposta de teste com diversos classificadores de aprendizado de máquina. Adicionalmente, foi apresentada uma proposta para retenção de talentos, no qual são mapeados os projetos mais adequados ao perfil do funcionário, visando a verificação e possível alocação pelo gestor do funcionário certo no projeto certo, por meio de *match* entre perfil do funcionário e competências necessárias para desenvolvimento do projeto.

Uma vez definido o arcabouço, foi realizado um experimento, para a aplicação, validação e avaliação da proposta, tendo como base um cenário real, com utilização de bases de dados cedidas por uma empresa que atua com prestação de serviços de TI. No experimento são apresentados os resultados da aplicação do modelo e verificado o classificador de aprendizado de máquina com melhor desempenho para utilização no artefato desenvolvido, considerando a utilização daquela base de dados. Constatamos com uso de métodos quantitativos que, para essa aplicação, o classificador com melhor desempenho foi o *Random Forest*.

Desta forma, os três ciclos se interligam para assegurar que a pesquisa em DSR seja re-

levante, respondendo a problemas reais do ambiente, rigorosa, baseada em conhecimento científico validado e criativa e eficaz, por meio do desenvolvimento iterativo de artefatos úteis ([HEVNER; CHATTERJEE, 2010](#)).

1.4 Contribuições

Entre as contribuições diretas, está o desenvolvimento de um arcabouço para tratamento de questões de retenção de talentos, com proposta para identificação do risco de desligamento voluntário de um funcionário identificado como um talento e indicação de ação para desenvolvimento visando retenção do talento em questão. Adicionalmente, é esperada contribuição direta para identificação de parâmetros a serem utilizados em modelos de aprendizado de máquina para pesquisas com identificação de risco de saída de um funcionário.

De forma indireta, esta proposta contribui para a qualificação e desenvolvimento de funcionários de alto desempenho, endereçando questões de rotatividade e gestão do conhecimento. Adicionalmente, ainda de forma indireta, a proposta favorece um ambiente de trabalho estimulante, saudável e produtivo a partir de um olhar para oportunidades de desenvolvimento.

Este trabalho faz parte do programa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) de Doutorado Acadêmico para Inovação (DAI), sendo uma parceria entre a Universidade Federal Fluminense (UFF), o CNPq e a empresa parceira Prime Up, processo N° 142182/2019-2. No programa DAI é estabelecido um convênio entre a universidade, o CNPq e a empresa parceira, visando, por intermédio da pesquisa de doutorado, o desenvolvimento de um produto ou pesquisa sobre alguma questão definida pela empresa. O financiamento desta tese é feito em maior parte pelo CNPq e em parte pela empresa parceira por meio do convênio estabelecido com a UFF. O convênio foi aprovado para fase de execução no dia 28 de dezembro de 2022. Desta forma, a pesquisa se insere de forma prática no mercado, lançando luz sobre soluções para o tema escolhido pela empresa.

Entre as contribuições para a empresa parceira, podemos citar o desenvolvimento de uma tese em tema de sua escolha, que poderá ser utilizada para concessão de benefícios fiscais por intermédio da Lei do Bem ([BRASIL, 2023](#)). O produto desta tese poderá ser utilizado internamente pela empresa parceira, de forma a tentar evitar a perda de seus profissionais de alto desempenho. Adicionalmente, caso seja de interesse da empresa, o produto desta tese poderá ser transformado em um produto comercial, que poderá oferecer para seus clientes.

1.5 Trabalhos publicados

Nesta seção são apresentadas as publicações realizadas no decorrer do doutorado. Primeiramente são apresentadas as publicações relacionadas à tese. As demais publicações, decorrentes de parcerias com colegas e professores da Pós-Graduação em Computação da UFF, são apresentadas na sequência.

1.5.1 Trabalhos relacionados com esta tese

- SEIXAS, E. F. R.; SEIXAS, F.; VITERBO, J.; BERNARDINI, F.; FREITAS, K.; FERNANDES, G. An Artificial Intelligence Technologies Approach for Talent Retention. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY & SYSTEMS, 2024, Cham: Springer Nature Switzerland, 2024. p. 412–421.
- SEIXAS, E. F. R.; VITERBO, J.; BERNARDINI, F.; SEIXAS, F.; PANTOJA, C. Applying artificial intelligence for talent retention: a systematic literature review. In: 18th IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 2023. [S.l.]: IEEE, 2023. p. 1–6.

1.5.2 Outras publicações

- SEIXAS, F. L.; SEIXAS, E. R.; FREITAS, A. A. Enhancing dementia prediction models: Leveraging temporal patterns and class-balancing methods. *Applied Soft Computing*, [S.l.], v. 171, 2025, Artigo n. 112754.
- SEIXAS, E. F. R.; RAMOS, B.; SEIXAS, F.; SILVA, M.; BERNARDINI, F.; VITERBO, J. An accessibility review: Google Classroom. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS, 2024, Cham: Springer Nature Switzerland, 2024. p. 171–180.
- SILVA, M.; FERRO, M.; MOURÃO, E.; SEIXAS, E. F. R.; VITERBO, J.; SALGADO, L. C. Ethics and AI in higher education: a study on students' perceptions. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS, 2024, Cham: Springer Nature Switzerland, 2024. p. 149–158.
- SILVA, M.; SEIXAS, E. F. R.; FERRO, M.; VITERBO, J.; SEIXAS, F.; SALGADO, L. C. Ética e responsabilidade na era da inteligência artificial: um survey com estudantes

de computação. In: WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 2024. [S.I.]: SBC, 2024. p. 854–865.

- HENRIQUES, F.; SOUZA, M.; VASCONCELOS, L.; PAIVA, M.; SEIXAS, E.; TREVISAN, D.; VITERBO, J. The use of design thinking for reflection in recruitment and selection of IT professionals. In: 26th INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK IN DESIGN (CSCWD), 2023. [S.I.]: IEEE, 2023. p. 1354–1359.
- LIMA, A. S.; SEIXAS, E. F. R.; SILVA, M.; SEIXAS, F.; SALGADO, L.; VITERBO, J. Automating risk stratification processes in obstetric emergency: a case study. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY & SYSTEMS, 2023, Cham: Springer International Publishing, 2023. p. 519–528.
- RAMOS, B. A.; SEIXAS, E. R.; SEIXAS, F. L. Uma análise de acessibilidade: o Google Sala de Aula. In: WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 2022. [S.I.]: SBC, 2022. p. 381–391.
- RANGEL, M.; BERNARDINI, F.; VITERBO, J.; MONTEIRO, R.; SEIXAS, E.; SANTOS PINTO, H. Uso de aprendizado de máquina para categorização automática de conjuntos de dados de portais de dados abertos. In: WORKSHOP DE COMPUTAÇÃO APLICADA EM GOVERNO ELETRÔNICO (WCGE), 2020. [S.I.]: SBC, 2020. p. 120–131.

1.6 Organização do trabalho

Além desta introdução, este trabalho está organizado em mais seis seções. Na Seção 2 apresentamos conceitos importantes para o entendimento deste estudo, tais como conceitos de rotatividade, retenção de talentos e tecnologias de IA. O mapeamento sistemático da literatura, detalhado na Seção 3, foi conduzido para obter um maior entendimento do cenário de uso de tecnologias de Inteligência Artificial em RH e para embasar os conceitos da pesquisa, oferecendo uma visão geral do tópico. Na Seção 4 apresentamos a revisão sistemática da literatura realizada com o objetivo de entender melhor as propostas de modelos para previsão de desligamento voluntário e retenção de talentos e, a partir dos achados e lacunas identificadas nos estudos primários, desenvolver a proposta de *framework*. Na Seção 5 é apresentada a proposta de um *framework* visando o atendimento dos objetivos apresentados nesta tese. Na Seção 6 é proposto um experimento com aplicação do *framework* com utilização de base de dados forne-

cido por uma empresa de TI. Finalmente, na Seção 7 são apresentadas as considerações finais sobre este trabalho e indicados trabalhos futuros.

2 Referencial teórico

Compondo o ciclo de rigor de nossa DSR, a seção de referencial teórico constitui a base de conhecimento que apoiará a validação do ciclo de *design*. Nesta seção, são apresentados os principais conceitos relativos à gestão de recursos humanos, como conceito de rotatividade e retenção de talentos. Em sequência, são apresentados conceitos de tecnologias de inteligência artificial, tais como aprendizado de máquina e sistemas de recomendação. Finalmente, são apresentados trabalhos que relacionam tecnologias de inteligência artificial com aplicações em recursos humanos. Desta forma, buscamos uma base científica para a tese, formando uma base de conhecimento e garantindo validade teórica e metodológica para nossa DSR.

2.1 Aspectos da gestão de recursos humanos

O conceito de Gestão de Recursos Humanos (GRH) engloba práticas específicas de RH, como recrutamento, seleção, treinamento, desenvolvimento, avaliação de desempenho e recompensas, sendo uma perspectiva tradicional, que avalia práticas de RH de forma isolada. A GRH é o elo que liga os funcionários à organização, garantindo alinhamento e melhor desempenho para ambos (BOON et al., 2018; DE BRITO; OLIVEIRA, 2016). Para Ivancevich (2009), a GRH estuda o que pode ou deve ser feito para tornar o trabalho mais produtivo ou satisfeito, facilitando o aproveitamento eficaz das pessoas para atingir metas individuais e organizacionais e inclui as atividades: cumprimento da legislação; análise de cargos; planejamento de recursos humanos; recrutamento, seleção, motivação e orientação de funcionários; avaliação de desempenho e compensação; treinamento e desenvolvimento; relações trabalhistas; segurança, saúde e bem-estar.

No entanto, a Gestão Estratégica de Recursos Humanos (GERH) difere significativamente da gestão tradicional de RH. Na GERH as práticas de RH ajudam as organizações a atingir metas estratégicas e melhorar o desempenho da empresa, através do desenvolvimento e implementação de processos de RH que visam aprimoramento e facilitação do cumprimento dos objetivos estratégicos da organização. Desta forma, a GERH contribui para aumentar a eficácia

organizacional, por exemplo, aumentando ao máximo a satisfação e auto-realização do funcionário, desenvolvendo e mantendo a qualidade de vida no trabalho, de modo que o emprego na organização seja algo desejável (IVANCEVICH, 2009; BOON et al., 2018; DE BRITO; OLIVEIRA, 2016).

A Gestão de Talentos (GT) começou a ganhar visibilidade dentro da GERH, tendo grande impacto no desempenho das organizações. Diferente da GRH, que tem como foco todos os funcionários da organização, sem diferenciação quanto às suas habilidades, competências e experiência, a GT foca em um grupo de indivíduos de acordo com seu talento e produtividade. Entre as atividades de GT relacionadas à GERH podemos citar o acompanhamento de indicadores que possam estar causando rotatividade na empresa e desenvolvimento de políticas e ações para retenção de talentos (AL-DALAHMEH; HÉDER; DAJNOKI, 2020).

2.1.1 Rotatividade

A rotatividade de empregados, ou do termo em inglês *turnover*, é um tema amplamente pesquisado, com grande foco das pesquisas nos motivos que levam os funcionários a deixarem as empresas. (KUMAR M; GOVINDARAJO, 2014).

A rotatividade pode ser definida pela porcentagem de empregados que deixam a organização, sendo repostos por novos funcionários. A rotatividade, de forma geral, considera tanto saídas por decisão dos funcionários quanto demissões, transferências, término de contratos e mesmo aposentadorias. Os dois principais tipos de rotatividade descritos na literatura são (AL-DALAHMEH; HÉDER; DAJNOKI, 2020; LAZZARI; ALVAREZ; RUGGIERI, 2022; TOTVS, 2022):

- **Rotatividade involuntária:** quando a decisão de saída do funcionário parte da empresa, ou seja, quando é demitido.
- **Rotatividade voluntária:** quando o funcionário pede demissão. Pode ser classificada como:
 - **Rotatividade voluntária funcional:** quando um funcionário que está com produtividade baixa solicita demissão, considera-se que o desligamento gerará benefícios para a organização;
 - **Rotatividade voluntária disfuncional:** quando um funcionário qualificado decide deixar a organização, configurando situação desfavorável para a empresa.

Para calcular a rotatividade é utilizada a fórmula (RÊGO et al., 2022):

$$Rotatividade = \frac{\frac{A+D}{2}}{EM} \cdot 100 \quad (2.1)$$

Onde:

A = admissões de pessoal dentro do período considerado (entradas);

D = desligamentos de pessoal(tanto por iniciativa da empresa como por iniciativa dos empregados) dentro do período considerado (saídas);

EM = efetivo médio dentro do período considerado. Pode ser obtido pela soma dos efetivos existentes no início e no final do período, dividido por dois.

A rotatividade representa um custo elevado para as empresas na maior parte dos casos. Por exemplo, no caso de um desligamento voluntário, haverá necessidade de uma nova contratação, treinamento e adaptação do novo funcionário, o que representa custos para a empresa até que o funcionário comece a performar. Quanto mais alta a rotatividade, menor a eficiência cultural e dos processos da empresa. Entre os desafios enfrentados pelas empresas está prever o risco de saída voluntária do funcionário, que passa a ser uma questão estratégica para a empresa. O grande desafio para se prever o risco de saída voluntária de um funcionário está na definição dos parâmetros a serem utilizados. Algumas pesquisas utilizam variáveis como absenteísmo, atrasos e indiferença, por exemplo ([CHIAVENATO, 2014](#); [CHAKRABORTY et al., 2021](#); [TOTVS, 2022](#)). Outros estudos utilizam variáveis como satisfação no trabalho ([CHANG et al., 2023](#); [SEELAM et al., 2022](#); [PORKODI; SRIHARI; VIJAYAKUMAR, 2022](#)).

De acordo com pesquisas do Gartner, trabalhadores de TI têm intenção 10,6% menor de ficar na empresa do que trabalhadores de outros setores, sendo a taxa mais baixa entre todas as funções corporativas. Adicionalmente, 49% dos candidatos que receberam uma oferta de emprego estão considerando pelo menos outras duas ofertas ao mesmo tempo e 65% dos funcionários de TI dizem que a possibilidade de trabalhar de forma flexível, ou não, impactará sua decisão de permanecer no trabalho ([GARTNER, 2022](#)).

2.1.2 Retenção de talentos

Segundo [Pereira; \(2013\)](#), “Talento é alguém que se identifica com a missão, a visão e os valores da empresa, que possui potencial para se desenvolver e crescer com o negócio, que consegue perceber o que deve ser feito e tem disposição para realizar suas tarefas cada vez melhor. Talentos não são gênios, mas sim pessoas normais, com a habilidade de aplicar corretamente o que sabem. São capazes de compreender o ambiente, antever problemas e identificar soluções.” Já

segundo [Al-Dalahmeh, Héder e Dajnoki \(2020\)](#), “Talentos, ou funcionários-chave, são aqueles indivíduos que costumam surpreender seus gestores com seu nível de desempenho, mostrando comportamentos preferenciais e seguindo o código de conduta da organização.”

Segundo [Veloso \(2019\)](#), retenção de talentos é um conjunto de medidas e ações estruturadas que visam evitar a rotatividade de pessoas que apresentam alto desempenho. Uma das razões mais óbvias para investir na retenção de talentos é reduzir o prejuízo de um alto *turnover* nas empresas, uma vez que demissões custam caro para os empregadores.

Um número crescente de organizações está competindo ferozmente para atrair e reter o talento necessário para acelerar o crescimento e o desempenho. Em uma pesquisa realizada pelo Gartner com líderes de RH, em junho de 2021, sobre o que as empresas estão fazendo ou planejando fazer para endereçar o *turnover* de empregados, podendo selecionar mais de uma alternativa, 74% dos participantes da pesquisa informaram que “Oferecer mais opções de flexibilidade de trabalho”, 44% irão aumentar as oportunidades de desenvolvimento para os empregados, 41% aumentar a habilidade do gestor em conversar sobre carreira e retenção, 37% investir mais no ambiente de trabalho e intervenções de engajamento, 25% aproveitar mais o *talent analytics* para entender melhor os riscos de retenção, 25% aumentar recompensas e reconhecimentos, 24% aumentar compensação ou benefícios e, finalmente, 23% aumentar comunicação sobre benefícios dos empregados. Com esta pesquisa é possível observar que as empresas estão mais interessadas em investir em outras alternativas para tratar retenção de talentos que não sejam apenas alternativas financeiras ([WILES, 2022](#)).

De acordo com [Holtom et al. \(2021\)](#), “a pandemia mudou a forma como trabalhamos e como vemos o trabalho. Os líderes precisam concentrar a atenção e a energia nas condições que ajudarão os funcionários com entusiasmo, em vez de relutância, a permanecer e prosperar - ou seja, um senso de adequação e propósito, um sistema de apoio no trabalho e pacotes personalizados que seriam difíceis de encontrar em qualquer outro lugar.”

2.2 Inteligência artificial

A Inteligência Artificial (IA) é a área da ciência da computação voltada para o estudo e projeto de sistemas computacionais inteligentes. Consideram-se sistemas inteligentes os que possuem características observadas na cognição e comportamento humano, por exemplo, compreensão da linguagem, habilidades de aprendizado, raciocínio e de resolução de problemas ([RUSSELL, 1997](#)).

Para muitos pesquisadores, o objetivo da IA é emular a cognição humana, enquanto para

alguns pesquisadores, é a criação de inteligência sem considerar qualquer característica humana ou noção abstrata de inteligência. Essa diversidade de objetivos não é necessariamente um problema, pois cada abordagem revela novas ideias e fornece uma base para prosseguir a pesquisa em IA. De acordo com o relatório de [West e Allen \(2018\)](#), sistemas inteligentes são dotados de pelo menos três aspectos: inteligência, intencionalidade e adaptabilidade. Um sistema inteligente pode ser construído usando técnicas de IA.

A IA constitui um campo robusto da ciência da computação, abarcando áreas como aprendizado de máquina, visão computacional e processamento de linguagem natural. Destas, o aprendizado de máquina, mais conhecido pelo seu termo em inglês, *Machine Learning* (ML), se destaca por permitir que algoritmos aprendam padrões diretamente dos dados, habilitando previsões e decisões sem programação explícita ([MUKHAMEDIEV et al., 2022](#); [BISHOP; NASRABADI, 2006](#)). Uma aplicação amplamente difundida dessa tecnologia são os sistemas de recomendação, os quais aplicam filtros colaborativos, baseados em conteúdo ou técnicas híbridas para sugerir itens personalizados conforme o perfil e comportamento do usuário ([LE-BLANC et al., 2024](#); [MAPHOSA; MAPHOSA, 2023](#); [SOLANO-BARLIZA et al., 2024](#)). Os sistemas de recomendação são bastante úteis na recuperação da informação, pois ajudam os usuários a descobrir resultados de uma busca que, de outra forma, demandaria um significativo esforço ([ISINKAYE; FOLAJIMI; OJOKOH, 2015](#)).

2.2.1 Aprendizado de máquina

Os algoritmos de ML encontram os padrões primeiro e depois executam a ação com base nesses padrões. O aprendizado de máquina pode ser classificado em quatro categorias: (a) aprendizado supervisionado, (b) aprendizado não supervisionado, (c) aprendizado semi-supervisionado e (d) aprendizado por reforço ([AYODELE, 2010](#)).

O aprendizado supervisionado pode ser definido como um tipo de aprendizado de máquina, em que tanto a entrada quanto a saída são fornecidas ao sistema. O sistema, então, busca um padrão, ou melhor ajuste, do modelo e seus parâmetros, considerando os dados de entrada e os dados rotulados na saída. Uma vez ajustado ou treinado, o modelo permite categorizar ou classificar um conjunto de dados de entrada. O aprendizado supervisionado pode ser dividido em duas formas: (a) classificação - o algoritmo de aprendizado de máquina supervisionado classifica os dados de entrada em classes pré-definidas. Esses algoritmos ajudam a prever a categoria de saída a partir dos dados rotulados. (b) Regressão - o algoritmo de aprendizado supervisionado encontra a relação entre os dados de entrada e a saída do modelo ([BELL, 2022](#)). Exemplos de algoritmos de aprendizado supervisionado incluem *k-Nearest Neighbor* (KNN)

(CUNNINGHAM; DELANY, 2021), *Support Vector Machine* (SVM) (CERVANTES et al., 2020), árvore de decisão (DE VILLE, 2013), *random forest* (CUTLER; CUTLER; STEVENS, 2012), regressão linear, regressão logística (ou *Logistic Regression* (LR)) (SPERANDEI, 2014), *Gradient Boosting* (GB) e *Extreme-Gradient Boosting* (XGBoost) (MITRAVINDA; SHETTY, 2022).

No aprendizado não supervisionado, são fornecidos os valores/dados de entrada e não há saída fixada pelo sistema. Os algoritmos de aprendizado não supervisionados não têm dados rotulados. Eles estimam a saída encontrando o padrões ocultos dos dados de entrada. Há dois tipos de aprendizado de máquina não supervisionado (DAYAN; SAHANI; DEBACK, 1999): (a) *clustering* - este método de aprendizado não supervisionado procura padrões para a criação de grupos com base em medidas de semelhanças nos dados de entrada apresentados. Esses agrupamentos de dados são chamados de *clusters*. (b) Associação - este método de aprendizado não supervisionado busca encontrar regras a partir dos dados de entrada e estimar relações entre os mesmos. Exemplos de algoritmos de aprendizado não supervisionado incluem *K-means*, Componente principal ou PCA e *hierarchical clustering*.

O aprendizado semi-supervisionado usa os dados de saída rotulados e não rotulados para construir um modelo de previsão (VAN ENGELEN; HOOS, 2020). A diferença entre os algoritmos acima e o método semi-supervisionado é que os dados não rotulados são mais numerosos do que os dados rotulados. Exemplos de algoritmos do aprendizado semi-supervisionado incluem heurísticas, modelos generativos e métodos baseados em grafos.

O aprendizado por reforço concentra-se em encontrar a melhor ação, dada uma situação ou contexto, buscando maximizar um resultado. As decisões são tomadas sequencialmente. Em cada passo, o algoritmo assume o caminho para resultados totais, que pode ter um resultado de saída positivo ou negativo. O resultado geral é, portanto, a soma de todos os resultados positivos e negativos ao longo o caminho. O objetivo do algoritmo é encontrar a melhor maneira de maximizar o resultado (GARCIA; FERNÁNDEZ, 2015). Exemplos de algoritmos aplicados em aprendizado por reforço incluem *Q-learning*, *policy interaction* e *Deep-Q network*.

Além das classificações, o aprendizado profundo (*DeepLearning* ou DL) é considerado uma subclasse de algoritmos de aprendizado de máquina. O DL utiliza redes neurais de várias multicamadas para extrair características/atributos dos dados de entrada para fazer previsões mais confiáveis (SHINDE; SHAH, 2018). Os métodos de aprendizado profundo são capazes de funcionar tanto em ambientes supervisionados quanto não supervisionados. As técnicas de DL incluem redes neurais artificiais, redes neurais convolucionais e regressão linear múltipla.

2.2.2 Sistemas de recomendação

Os sistemas de recomendação podem ser comparados em termos do domínio de problema onde são aplicados, as informações usadas e o tipo de recomendação produzido na saída. Há três tipos de sistema de recomendação (MOHANTY et al., 2020): (1) os baseados em conteúdo, (2) os que utilizam métodos de filtragem colaborativa, e (3) os que utilizam as características de ambos os sistemas.

Nos sistemas de recomendação baseados em conteúdo, o sistema aprende a fazer recomendações analisando a semelhança de características dos perfis de usuário entre as características dos itens buscados (WANG; LIANG et al., 2018). Por exemplo, com base na classificação de um usuário para diferentes gêneros de filmes, o sistema aprenderá a recomendar o gênero avaliado positivamente pelo usuário. Um sistema de recomendação baseado em conteúdo estima um perfil de usuário com base nos itens buscados anteriormente pelo usuário. Um perfil de usuário representa o conjunto de interesses do usuário, sendo também capaz de se adaptar a novos interesses. O processo de recomendação consiste em identificar a melhor correspondência do perfil do usuário e as características do objeto em foco. O sistema baseado em conteúdo pode ser decomposto por três módulos, representando etapas sequenciais: (a) módulo de análise de conteúdo - responsável por extrair as características mais revelantes e representar o conteúdo de forma adequada. Técnicas de extração de características podem ser usadas para essa tarefa. (b) módulo de aprendizado - este módulo estima o melhor perfil do usuário usando os dados obtidos na etapa anterior. Técnicas de aprendizado de máquina são usadas para na generalização e estimativa de um modelo baseado em preferências do usuário, tanto positivas quanto negativas. (c) módulo de filtro - este componente utiliza o perfil de usuário para derivar itens relacionados ao mesmo. Isso é realizado combinando o perfil de usuário com os itens a serem recomendados. Com base em medidas de similaridade, uma análise de preferência é produzida, que pode ter como resultado uma avaliação binária (por exemplo, sim ou não) ou contínua (por exemplo, *score* de similaridade).

Os métodos de filtragem colaborativa utilizam o comportamento do usuário para as recomendações (EKSTRAND; RIEDL; KONSTAN et al., 2011). Esta abordagem utiliza uma matriz de utilidade em vez das características correspondentes aos perfis de usuários. Algumas limitações do sistema de recomendação baseado em conteúdo podem ser superadas pela abordagem colaborativa. Por exemplo, para fazer previsões para aqueles itens aos quais o conteúdo não está disponível, ele usa a avaliação ou *feedback* de outros usuários. Esses sistemas avaliam a qualidade de um item com base na revisão por pares. Também pode sugerir itens com conteúdo diferenciado desde que outros usuários tenham demonstrado interesse no conteúdo. Há

duas categorias de filtragem colaborativa: (a) abordagem baseada em memória - a matriz de utilidade é aprendida e as sugestões são dadas com base em informações fornecidas pelo usuário. (b) Filtragem dos itens - neste método, para cada item apresentado a um usuário particular, é utilizada como critério de recomendação, a classificação desse item com base na vizinhança mais próxima de usuários.

2.3 Inteligência artificial em recursos humanos

O aumento do uso da IA nos mais variados tipos de aplicação já é uma realidade em nossa vida cotidiana. Assim como em inúmeras outras áreas, ferramentas com tecnologias baseadas em inteligência artificial são amplamente utilizadas em diversos processos de RH (STROHMEIER; PIAZZA, 2015; QAMAR et al., 2021; BANKINS, 2021), automatizando e suportando desde o recrutamento e seleção até processos como gestão de carreira, desenvolvimento organizacional e treinamento (HMOUD; LASZLO et al., 2019; VELUCHAMY; SANCHARI; GUPTA, 2021). Trabalhos relatam a utilização de ferramentas de IA em RH para recrutamento e seleção de candidatos para as vagas em aberto das empresas, sendo este método de contratação amplamente utilizado em empresas de tecnologia (HMOUD; LASZLO et al., 2019). Outros trabalhos relatam o uso de ferramentas de IA para verificação do nível de satisfação dos funcionários, *turnover*, avaliação de desempenho de RH, entre outras aplicações (QAMAR et al., 2021).

Acompanhando esta tendência, foi observado um crescente interesse acadêmico relacionado ao uso de IA nas práticas de negócios de RH. Votto et al. (2021) realizaram uma revisão sistemática da literatura para explorar a literatura de Sistemas de Informação de Recursos Humanos Táticos e entender quais componentes existem na literatura e como eles são representados posteriormente. Já Pereira et al. (2023) revisou sistematicamente a literatura sobre a ligação entre IA e resultados no local de trabalho. O trabalho de Köchling e Wehner (2020) revisa pesquisas sobre tomada de decisão algorítmica no contexto de Gestão de Recursos Humanos (GRH), destacando questões éticas relacionadas a algoritmos. Em Budhwar et al. (2022) os autores desenvolvem uma estrutura conceitual que integra pesquisas sobre aplicações de IA em gerenciamento de recursos humanos e oferece uma base coesa para futuros empreendimentos de pesquisa. O estudo de Kaushal et al. (2021) visa esclarecer como a IA foi integrada em diversas funções de RH e sua influência na direção das organizações, força de trabalho e RH.

Atualmente há uma crescente preocupação acerca das questões éticas envolvidas com o uso de tais ferramentas. Os trabalhos de Bankins (2021) e Veluchamy, Sanchari e Gupta (2021) abordam o tema. O primeiro trabalho constrói uma estrutura de tomada de decisão para apoiar

a implantação ética de IA para GRH e orientar as determinações da combinação ideal de envolvimento humano e de máquina para diferentes tarefas de GRH. Fazer isso apoia a implantação da IA para a melhoria do trabalho e dos trabalhadores e gera resultados acadêmicos e práticos. Já o trabalho de [Veluchamy, Sanchari e Gupta \(2021\)](#) discute a capacidade da eliminar vieses conscientes e inconscientes no processo de recrutamento e seleção.

3 Mapeamento sistemático

Nesta seção é apresentada a investigação do problema, compondo o ciclo de relevância da DSR, que permitirá a identificação de potenciais lacunas a serem pesquisadas. O conhecimento obtido nessa mapeamento também leva a uma relação com o ciclo de rigor, o que garante o rigor da pesquisa necessária para construção do arcabouço fundamentado em uma sólida base teórica ([HEVNER; CHATTERJEE, 2010](#)).

Visando melhor entendimento do estado da arte no contexto do problema estudado por esta tese, foi realizado um Mapeamento Sistemático da Literatura (MSL), com uma busca abrangente, respondendo a quatro questões, buscando compreensão sobre os principais motivos que levam à saída dos funcionários, os cenários de uso de IA em RH, ferramentas comerciais e modelos ou arquiteturas baseadas em IA para retenção de talentos. Este mapeamento foi publicado em 2023 na conferência internacional: *18th Iberian Conference on Information Systems and Technologies (CISTI)* ([SEIXAS et al., 2023](#)).

Um mapeamento sistemático é um tipo de estudo secundário com o objetivo de fornecer uma visão abrangente de uma área de pesquisa, por meio da classificação e contagem das publicações conforme categorias definidas. Ele é utilizado para identificar tendências, lacunas e distribuição das evidências científicas sobre determinado tema, sendo útil especialmente na fase inicial de investigação de um domínio científico ([PETERSEN; VAKKALANKA; KUZNIARZ, 2015](#); [FELIZARDO et al., 2020](#)).

Este mapeamento sistemático tem como base as diretrizes propostas por [Petersen, Vakkalanka e Kuzniarz \(2015\)](#), combinadas com protocolo sugerido por [Kitchenham e Chartes \(2007\)](#).

3.1 Protocolo do mapeamento sistemático

3.1.1 Objetivo e questões de pesquisa

Para definição do objetivo deste mapeamento sistemático, foi utilizado o Paradigma GQM (*Goal-Question-Metric*) (VAN SOLINGEN et al., 2002; BASILI; ROMBACH, 1988), conforme apresentado na Tabela 1.

Tabela 1: Objetivo do Mapeamento Sistemático segundo paradigma GQM

Analisar	publicações científicas
Com o propósito de	identificar e analisar
Em relação a	aspectos determinantes que envolvam retenção ou migração de talentos e se há pesquisas que utilizam inteligência artificial
Do ponto de vista dos	Pesquisadores
No contexto	de ecossistemas de Recursos Humanos, de ecossistemas de inovação, com foco em retenção de talentos

Para atingir o objetivo deste mapeamento sistemático, focamos nas seguintes questões de pesquisa:

- **QP1: Quais aspectos são relevantes para retenção ou migração de talentos?**

Por intermédio desta questão de pesquisa, buscamos identificar quais os principais motivos que levam um funcionário a pedir para sair da empresa. Desta forma também são mapeados o que poderiam ser atributos a serem utilizados no artefato.

- **QP2: Em que tipo de processos a IA está sendo aplicada em Recursos Humanos em empresas no cenário de inovação?**

Por meio desta questão buscamos entender o ambiente de aplicação, mapeando quais as áreas de RH onde são aplicadas soluções baseadas em IA de acordo com os estudos, também se os estudos apontam o uso de IA para retenção de talentos.

- **QP3: Quais ferramentas existem ou são utilizadas para retenção de talentos em empresas de tecnologia?**

Com esta questão procuramos verificar se existem informações sobre ferramentas comerciais para retenção de talentos, se há informações sobre custos e o que as ferramentas realmente fazem, buscando, assim, lacunas para aprofundamento da pesquisa.

- **QP4:** *Existem propostas de modelos ou arquiteturas utilizando IA para retenção de talentos?*

Finalmente, buscamos propostas de modelos que possam subsidiar nossa proposta de arcabouço. Por meio destas propostas, identificar as principais ferramentas utilizadas, melhores práticas e possibilidades de lacunas.

Por intermédio das questões de pesquisa buscou-se identificar os estudos alinhados ao tema de tese, com o objetivo de refinar a proposta da tese, além de obter respostas para questões relevantes para o entendimento do cenário do tema de tese proposto.

3.1.2 Estratégia para busca dos estudos primários

Estudos primários são estudos individuais que contribuem para um mapeamento sistemático, enquanto um mapeamento sistemático é um estudo secundário ([KITCHENHAM; CHARTES, 2007; SILVA; VALENTIM; CONTE, 2015; PETERSEN; VAKKALANKA; KUZNIARZ, 2015; NAPOLEÃO et al., 2017](#)). Uma vez elaboradas as questões de pesquisa, o próximo passo é buscar estudos relacionados a estas questões para nos ajudar a respondê-las. Para tal é estabelecida a estratégia para pesquisa, onde serão definidos os termos a serem utilizados. Estes termos compõem as sequências de palavras-chaves utilizados na *string* de busca, que tem a finalidade de buscar trabalhos relevantes para o escopo da pesquisa nas bibliotecas digitais selecionadas, que serão apresentadas na sequência. Utilizamos o PICOC (*Population, Intervention, Comparison, Outcome e Context*) para estruturar e melhorar a *string* de busca ([KITCHENHAM; CHARTES, 2007; PETERSEN; VAKKALANKA; KUZNIARZ, 2015](#)):

- **(P) Population:** *Profissionais mapeados como talentos;*
- **(I) Intervention:** *Artificial Intelligence;*
- **(C) Comparison:** Não se aplica, pois o objetivo não é realizar uma comparação;
- **(O) Outcome:** *Talent Retention;*
- **(C) Context:** *"Innovation Ecosystem" AND "Talent Retention".*

A *string* de busca utilizada para a pesquisa nas bases foi:

(("artificial intelligence" OR "AI") AND ("retaining talent" OR "talent retention" OR "personal development" OR "employee retention") AND ("innovation"))

A busca de artigos foi feita nas bibliotecas digitais que consideramos ser as mais relevantes: ACM Digital Library¹, Engineering Village², IEEE Digital Library³, ISI Web of Science⁴, Science Direct⁵, Scopus⁶, Springer⁷ e Taylor & Francis⁸.

Para escolha das bibliotecas digitais seguimos critérios propostos por Petersen, Vakkalanka e Kuzniarz (2015) e Kitchenham e Chartes (2007):

- Possuir mecanismo de busca simples, que permita utilização de *string* de busca, proporcionando a localização de publicações com análise de seus títulos e resumos;
- Possuir máquinas de busca abrangentes com um bom funcionamento;
- Suas bases devem conter publicações de diversas áreas de conhecimento.

O período da busca inclui artigos de periódicos científicos e textos de conferências publicados entre 2012 e 2022. A pesquisa considerou trabalhos divulgados até 10 de outubro de 2022, data da extração de dados nas bases.

3.1.3 Critérios para seleção dos estudos

Segundo Kitchenham e Chartes (2007) devemos definir critérios para inclusão e exclusão dos artigos identificados na busca baseados nas questões de pesquisa, considerando o contexto previamente definido. Desta forma, foram definidos os critérios apresentados na Figura 2.

Ainda segundo Kitchenham e Chartes (2007), uma das formas de minimizar questões como viés de pesquisas, que geralmente tendem a publicar um resultado positivo em detrimento de resultados negativos, seguimos a orientação de incluir estudos publicados em conferências.

Os critérios para seleção dos artigos foram aplicados em duas etapas. Inicialmente foi feita uma seleção preliminar de artigos. Nesta etapa foram analisadas as informações do título e do

¹<http://portal.acm.org>

²<https://www.engineeringvillage.com>

³<http://ieeexplore.ieee.org>

⁴<http://www.isiknowledge.com>

⁵<http://www.sciencedirect.com>

⁶<http://www.scopus.com>

⁷<http://link.springer.com>

⁸<https://www.tandfonline.com>

resumo (*abstract*) dos artigos para verificação se atendiam a algum dos critérios de inclusão. Em caso de dúvida o artigo era incluído ao grupo de artigos aceitos para uma análise posterior mais aprofundada. Para cada artigo foi definido um dos critérios.

Figura 2: Critérios de Seleção de Estudos



Fonte: Elaborado pela autora

Para a seleção final dos estudos primários foi feita a leitura completa de cada um dos artigos que atendiam a um dos critérios de inclusão na etapa preliminar. O objetivo desta etapa é realizar uma análise mais detalhada dos estudos, mantendo a classificação ou reclassificando de acordo com os critérios de seleção definidos.

3.1.4 Estratégia para extração de dados

A estratégia de extração de dados foi elaborada para garantir uniformidade na apresentação dos dados coletados e facilitar as respostas às perguntas de pesquisa estabelecidas no MSL. Com base nos critérios de qualidade definidos, foi desenvolvido um questionário de extração

de dados, contendo perguntas sobre informações relevantes a serem analisadas nos estudos primários selecionados.

3.2 Metodologia para busca e seleção de estudos primários

Para definição do processo de busca e seleção dos estudos primários, adaptamos o processo sugerido por [Dyba, Dingsoyr e Hanssen \(2007\)](#). A Figura 3 apresenta o processo seguido neste mapeamento sistemático.

Inicialmente foram feitas as buscas dos estudos primários nas bibliotecas digitais selecionadas, a partir da *string* de busca apresentada na Subseção 3.1.2, obtendo um resultado quantitativo de 1492 estudos. Em seguida são retirados os artigos duplicados e os artigos publicados antes de 2012, de acordo com os critérios de exclusão. Na sequência, com a leitura dos títulos e resumos (*abstracts*) foram aplicados os critérios de inclusão e exclusão previamente definidos, sendo selecionados os 113 estudos primários potencialmente relevantes. Destes estudos, foi feita a leitura completa dos estudos com acesso gratuito via Portal de periódicos da CAPES e aplicados novamente os critérios de inclusão e exclusão. Este MSL obteve como resultado quantitativo um total de 13 estudos primários relevantes selecionados.

Figura 3: Processo de busca e seleção de estudos primários



Fonte: Elaborado pela autora

3.3 Resultados

3.3.1 Resultados quantitativos

A Tabela 2 apresenta com detalhes a distribuição dos artigos de acordo com cada fase do processo de busca e seleção dos estudos primários.

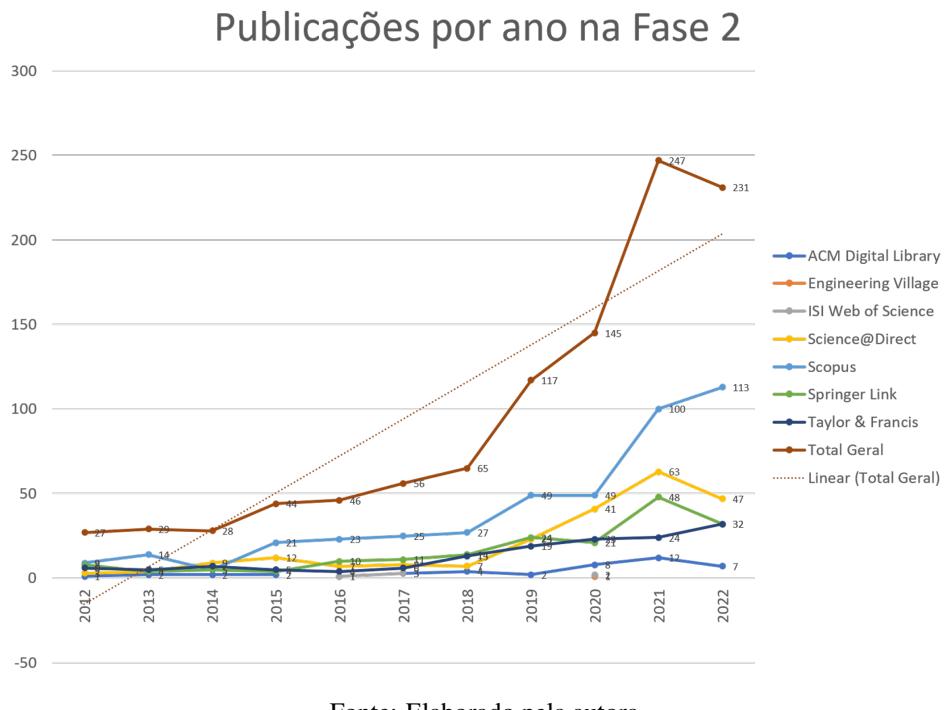
É importante destacar que na busca direta na base da biblioteca digital IEEE não obtivemos retorno de artigos com a *string* de busca utilizada. Adicionalmente, dos 113 trabalhos selecionados para leitura completa não obtivemos acesso gratuito ao texto completo de um total de 15 estudos primários, sendo 14 estudos da biblioteca Scopus e 1 estudo da biblioteca ACM.

Tabela 2: Distribuição dos artigos selecionados nas fases do processo de busca e seleção de estudos primários

Bibliotecas Digitais	Fase 1	Fase 2	Fase 3	Aprovados
ACM Digital Library	52	43	4	1
Engineering Village	5	2	0	0
ISI Web of Science	8	6	0	0
Science Direct	422	224	26	0
Scopus	480	435	73	12
Springer	254	181	7	0
Taylor & Francis	271	144	3	0
Total	1492	1035	113	13

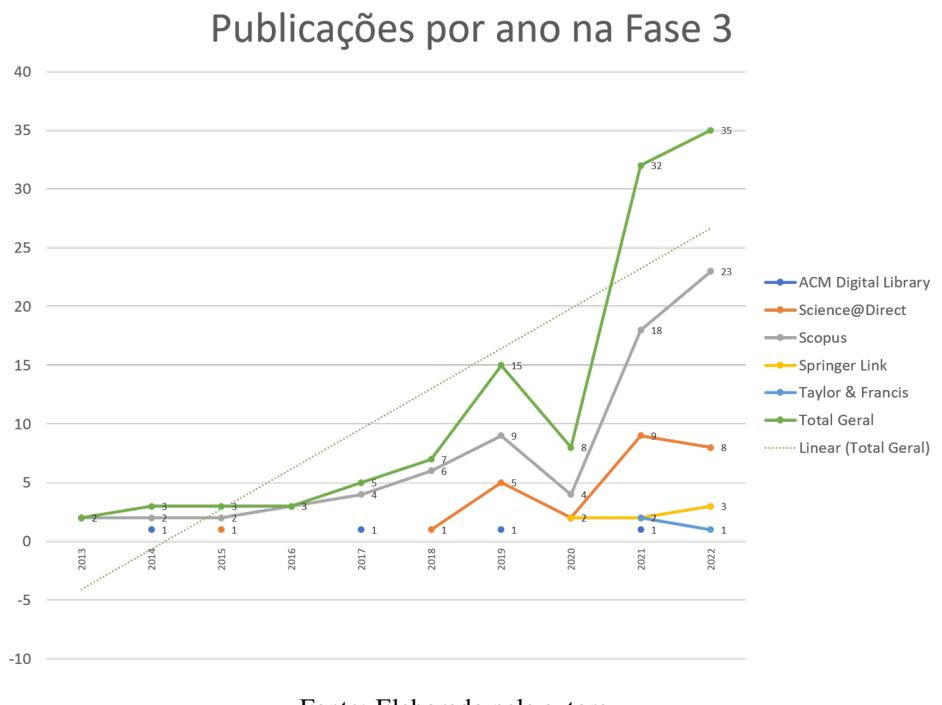
Considerando os critérios de exclusão previamente definidos, os artigos nas fases 2 e 3 foram publicados no período entre os anos de 2012 e 2022. Notamos que, a partir de 2019 houve um aumento significativo do número de publicações sobre o assunto pesquisado, com o período de 2019 a 2022 representando 71,5% do total de trabalhos publicados. Somente nos anos de 2021 e 2022 foram publicados 46,18% do total de trabalhos, o que levanta a hipótese, que pode ser futuramente testada, de que a pandemia de COVID-19 pode ter causado um aumento na demanda por pesquisas relacionadas a questões de RH com o aumento do trabalho remoto. Mesmo com a condução deste mapeamento sistemático sendo realizado em outubro de 2022, a tendência é de aumento no número de publicações sobre o tema pesquisado, conforme apresentado nas Figuras 4 e 5. Adicionalmente, notamos que o número de publicações na Fase 3 sofreu uma queda drástica em 2020, o que pode significar que os estudos publicados neste período não estavam aderentes ao que se buscava para responder às questões de pesquisa.

Figura 4: Cronologia das publicações na Fase 2.



Fonte: Elaborado pela autora

Figura 5: Cronologia das publicações na Fase 3.

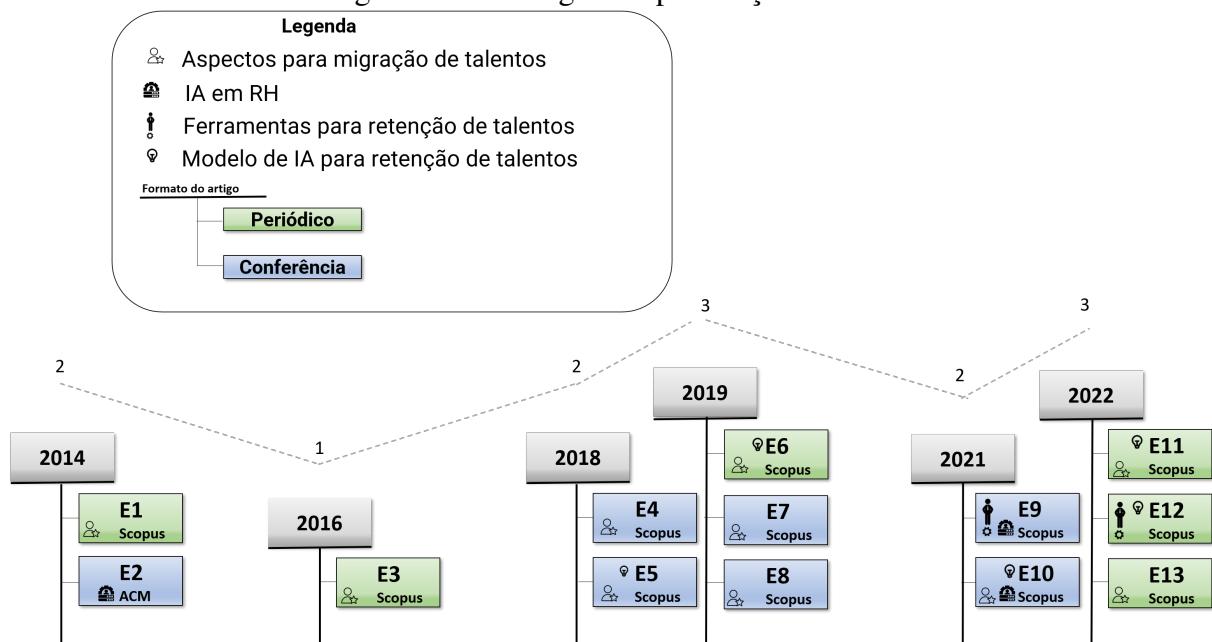


Fonte: Elaborado pela autora

Os estudo primários aprovados foram publicados entre 2014 e 2022. Considerando a cronologia das publicações, observamos que houve um aumento no número de publicações aceitas

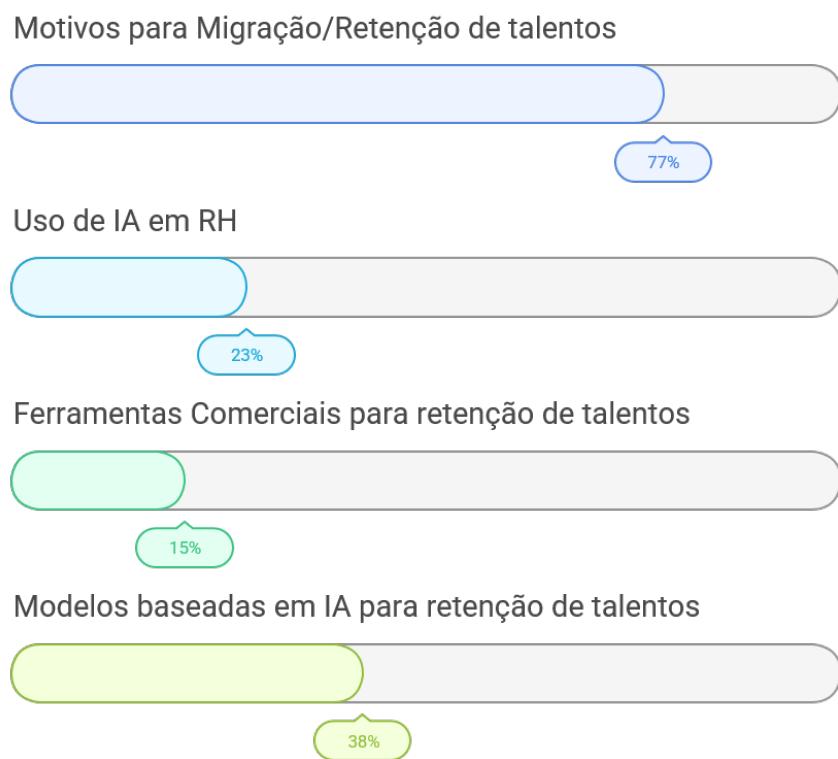
entre os anos de 2019 e 2022, correspondendo a 61,54% do total de publicações aceitas, seguindo a tendência geral de aumento de publicações sobre o tema. A Figura 6 apresenta a distribuição dos estudos primários aceitos. Nesta representação os estudos foram classificados quanto ao tipo de publicação, considerando se foi publicado em periódico (*journal*) ou se foi publicado em conferência (*conference*), com informação da biblioteca digital de onde o artigo foi extraído. Adicionalmente, classificamos os estudos de acordo com as questões de pesquisa: se a pesquisa trata de aspectos relevantes para migração de talentos, se trata do uso de IA em recursos humanos, se trata de ferramentas existentes utilizadas para retenção de talentos, ou se trata de propostas de modelos de IA para retenção de talentos. Verificamos que alguns trabalhos abordam mais de um desses temas. A maior parte dos trabalhos aceitos aborda os motivos para migração ou retenção de funcionários, representando 77% dos trabalhos. Já os trabalhos sobre o uso de IA em RH representam 23% dos estudos. Aproximadamente 15% dos estudos discorre sobre ferramentas comerciais para retenção de talentos, enquanto 38% dos trabalhos apresentam modelos ou arquiteturas baseadas em IA para retenção de talentos, conforme ilustrado na Figura 7.

Figura 6: Cronologia das publicações aceitas



Fonte: Elaborado pela autora

Figura 7: Distribuição dos estudos primários aceitos por questão de pesquisa



Fonte: Elaborado pela autora

A lista dos estudos selecionados é apresentada na Tabela 17.

Tabela 3: Estudos relevantes selecionados

ID	Título	Citação
$E_{MS} 1$	<i>Instrument development “intention to stay instrument” (ISI)</i>	(KUMAR M; GOVINDA-RAJO, 2014)
$E_{MS} 2$	<i>Predicting Employee Expertise for Talent Management in the Enterprise</i>	(VARSHNEY et al., 2014)
$E_{MS} 3$	<i>The right friends in the right places: Understanding network structure as a predictor of voluntary turnover</i>	(BALLINGER; CROSS; HOLTON, 2016)
$E_{MS} 4$	<i>Employee retention and turnover in global software development: Comparing in-house offshoring and offshore outsourcing</i>	(BASS et al., 2018)
$E_{MS} 5$	<i>A Study on Job Satisfaction Factors in Retention and Turnover Groups using Dominance Analysis and LDA Topic Modeling with Employee Reviews on Glassdoor.com</i>	(LEE; KANG, 2018)
$E_{MS} 6$	<i>Perceived Organizational Support, Alternative Job Opportunity, Organizational Commitment, Job Satisfaction and Turnover Intention: A Moderated-mediated Model</i>	(ALBALAWI et al., 2019)
$E_{MS} 7$	<i>When agile means staying: The relationship between agile development usage and individual IT professional outcomes</i>	(SETOR; JOSEPH, 2019)
$E_{MS} 8$	<i>Employee Retention in Agile Project Management</i>	(ISSA et al., 2019)
$E_{MS} 9$	<i>Employee Turnover Prediction System: With Special Reference to Apparel Industry in Sri Lanka</i>	(RODRIGO; RATNAYAKE, 2021)
$E_{MS} 10$	<i>A Machine Learning Approach for Employee Retention Prediction</i>	(MARVIN; JACKSON; ALAM, 2021)
$E_{MS} 11$	<i>Green talent management and turnover intention: the roles of leader STARA competence and digital task interdependence</i>	(OGBEIBU et al., 2022)
$E_{MS} 12$	<i>Predicting Employee Attrition Using Machine Learning Approaches</i>	(RAZA et al., 2022)
$E_{MS} 13$	<i>Factors Affecting Employee’s Retention: Integration of Situational Leadership With Social Exchange Theory</i>	(XUECHENG; IQBAL; SAINA, 2022)

3.3.2 Contribuições dos estudos

RQP1: Quais aspectos são relevantes para retenção ou migração de talentos?

A maior parte dos estudos primários selecionados aponta, de forma direta ou indireta, possíveis motivos para migração ou retenção de talentos nas empresas, correspondendo a aproximadamente 77% dos estudos. Alguns estudos como o de [Ballinger, Cross e Holtom \(2016\)](#) e de [Marvin, Jackson e Alam \(2021\)](#) utilizam dados demográficos dos sistemas de RH para montagem do modelo de IA para predição de retenção de funcionários.

A partir do mapeamento sistemático realizado, constatamos que os principais estudos para responder a esta questão de pesquisa são os trabalhos de [Kumar M e Govindarajo \(2014\)](#), [Bass et al. \(2018\)](#), [Setor e Joseph \(2019\)](#), [Issa et al. \(2019\)](#) e [Xuecheng, Iqbal e Saina \(2022\)](#). A Figura 8 apresenta os principais aspectos relevantes para retenção ou migração de talentos elencados nos estudos primários. O detalhamento dos itens encontrados nos estudos estão dispostos no Apêndice A.

Figura 8: Aspectos relevantes para migração ou retenção de talentos



Fonte: Elaborado pela autora

RQP2: Em que tipo de processos a IA está sendo aplicada em Recursos Humanos em empresas no cenário de inovação?

O propósito desta pergunta é entender como a Inteligência Artificial (IA) está sendo utilizada em Recursos Humanos (RH) e se há propostas associadas a retenção de talentos, por meio dos relatos dos estudos primários. Verificamos que três trabalhos aprovados apresentam informações sobre como a IA é aplicada nos diversos processos de RH. No trabalho de [Varsh-](#)

ney et al. (2014) foram utilizados modelos de predição, como Regressão Logística (*Logistic Regression - LR*) e Naive Bayes, para atualização de informações cadastrais dos funcionários sobre suas competências e conhecimentos. Os dados utilizados para a atualização são oriundos de informações de mídias sociais internas e sistemas da empresa, como de treinamento. Desta forma houve uma otimização do trabalho do RH, mais especificamente de equipes de Desenvolvimento Organizacional e Consultoria Interna de RH, mantendo atualizadas informações importantes para decisões gerenciais.

O trabalho de [Rodrigo e Ratnayake \(2021\)](#) contribui para responder a esta questão de pesquisa por meio de uma das contribuições de seu estudo, onde é feita uma compilação de trabalhos que utilizam IA para predição de *turnover*. Partes destes trabalhos tinham como foco a retenção de talentos. Estudos para redução de *turnover* afetam diretamente as equipes de Recrutamento e Seleção, além de todos os processos de RH envolvidos em contratação e desligamento de funcionários.

Outro estudo primário que aparece novamente é o estudo de [Marvin, Jackson e Alam \(2021\)](#). Neste trabalho são propostos modelos que utilizam IA para prever a probabilidade de retenção do candidato antes que ele seja selecionado para fazer parte de qualquer programa de treinamento da empresa. Desta forma, busca-se um melhor investimento em treinamento por parte da empresa, investindo nos funcionários com menor risco de sair, e, consequentemente, gestão do conhecimento, visto que serão treinados funcionários que manterão o conhecimento na empresa.

RQP3: *Quais ferramentas existem ou são utilizadas para retenção de talentos em empresas de tecnologia?*

Esta pergunta busca informações sobre quais são as ferramentas comerciais existentes para retenção de talentos que tenham chamado a atenção dos pesquisadores. Dois trabalhos apresentaram esta informação, sendo um de forma direta e outro de forma indireta. Embora o trabalho de [Rodrigo e Ratnayake \(2021\)](#) esteja no contexto da indústria têxtil, a comparação das ferramentas comerciais apresentada foi um fator determinante para a seleção deste trabalho, com sua importante contribuição para a resposta da terceira questão de pesquisa deste mapeamento.

Em [Rodrigo e Ratnayake \(2021\)](#) é apresentada proposta de atributos para um modelo teórico. Para definição destes atributos é feita uma comparação entre as ferramentas comerciais e o modelo a ser desenvolvido, com as principais funcionalidades que o modelo deveria ter, se-

gundo os pesquisadores. Entre as ferramentas comerciais citadas estão: Visier⁹; Cultureamp¹⁰; Qualtrics¹¹; IBM Watson Analytics¹²; HIQlabs¹³. Segundo análise apresentada neste estudo, o Visier, o Qualtrics e o HIQlabs apresentam a visualização da taxa de rotatividade de cada funcionário, adicionalmente, estes sistemas fazem a identificação dos *drivers* de rotatividade de cada funcionário. O Culture Amp, o Qualtrics e o hiQLabs apresentam sugestões automatizadas de estratégias de retenção para cada *driver* de rotatividade e somente o Qualtrics apresenta alertas para funcionários de alto risco.

Já o estudo de Raza et al. (2022) tinha o propósito de apresentar um modelo para predição de desligamento de funcionários. Na fase inicial o time de pesquisa utiliza os dados do relatório gerado pelo sistema “IBM HR Employee Attrition” buscando inspiração para a definição dos dados a serem utilizados em seu modelo.

No momento da escrita deste trabalho, não foi possível identificar o *site* da hiQLabs, nem maiores informações sobre a empresa. Verificamos no site da Culture Amp que trata-se de uma plataforma comercial bastante ampla, que interliga informações de pesquisa de clima organizacional com indicadores importantes de RH, como a taxa de *turnover* da empresa e os riscos associados, entre outras funções. Já a Qualtrics é focada em avaliações de experiências, seja de clientes ou de funcionários, aplicando suas soluções para serviços ao cliente, recursos humanos e pesquisadores. A Visier apresenta uma plataforma para *People Analytics*, onde oferece *insights* para entender a rotatividade e seu impacto na organização. A IBM apresenta em seu *site* uma série de estudos de casos e ferramentas desenvolvidas visando o uso de IA para soluções de RH, com propostas aplicadas a diversas áreas de recursos humanos.

RQP4: Existem propostas de modelos ou arquiteturas utilizando IA para retenção de talentos?

Um total de cinco estudos primários selecionados neste MSL sugerem modelos que utilizam ferramentas de IA para tratar de forma direta ou indireta a retenção de talentos. Esta seleção tem como base trabalhos que apresentam modelos que foram testados a partir de dados e técnicas descritos. Não foram selecionados para responder a esta questão de pesquisa trabalhos que apresentem propostas de modelos que não foram testados ou que não estejam relacionados de forma direta ou indireta à retenção de talentos. No trabalho de Lee e Kang (2018) são

⁹<https://www.visier.com/solutions/outcome/talent-retention/>

¹⁰<https://www.cultureamp.com>

¹¹<https://www.qualtrics.com/pt-br/discover/employee-xm/>

¹²<https://www.ibm.com/watson/uk-en/talent/insights/>

¹³<https://www.hiqlabs.com/>

analisados 217.379 dados de avaliação de funcionários publicados no site Glassdoor. O foco do trabalho é, por meio de mineração de texto de domínio do RH, identificar a importância relativa dos fatores que afetam a satisfação no trabalho, classificando os funcionários de acordo com a retenção e a rotatividade. Os autores afirmam que o trabalho oferece duas contribuições: aplicação do raciocínio indutivo à pesquisa organizacional usando *big data*, com realização de várias análises para diversos grupos usando mineração de texto e análise de dominância. A segunda contribuição foi a identificação de pontos-chave, por meio dessas análises, para ajudar a melhorar a retenção de funcionários.

No estudo de [Albalawi et al. \(2019\)](#) foi testado um modelo estrutural que examina o papel mediador do comprometimento organizacional na ligação entre suporte organizacional percebido, oportunidades alternativas de trabalho percebidas e intenção de rotatividade, além do papel moderador da satisfação no trabalho nas relações propostas. Por meio de quatro hipóteses foi utilizado o método de Modelagem de Equação Estrutural baseada em variância (*Variance based Structural Equation Modeling* (PLS-SEM)) para mapear se as variáveis ‘Percepção de suporte organizacional’, ‘Percepção de oportunidades alternativas de trabalho’, ‘Satisfação no trabalho’ e ‘Comprometimento organizacional’ têm correlação com intenção de rotatividade.

[Marvin, Jackson e Alam \(2021\)](#) apresentam proposta de modelo para previsão de probabilidade de retenção de candidato antes de seleção para participação em programa de treinamento da empresa, evitando possíveis custos com profissionais com intenção de deixar a empresa. São citadas como contribuições principais:

- Os modelos ilustram as características que afetam as decisões dos funcionários para a retenção da empresa;
- Avaliação de desempenho e precisão dos modelos com conjuntos de dados sintetizados e reais obtidos para fins de análise de RH;
- Teste de nove modelos de classificação, indicando o melhor classificador para aplicação preditiva relacionada a Recursos Humanos. Foram testados os modelos: Regressão logística; Naive Bayes (*Gaussian Naive Bayes*); k-NN (*k-Nearest Neighbors* - k-vizinhos mais próximos); Árvore de Decisão; *Random Forest*; Algoritmo de Árvores Extras (); Algoritmos de aumento de gradiente (*GBM - Gradient Boosting Algorithms*); XGBoost; LightGBM. O classificador Random Forest foi o que apresentou melhores resultados.

Por outro lado [Ogbeibu et al. \(2022\)](#) testa hipóteses com uso de PLS-SEM em relação à intenção de turnover, utilizando conceitos de Gestão de Talentos Sustentável (*Green Talent Management - GTM*).

Conforme citado anteriormente, [Raza et al. \(2022\)](#) realizaram estudo para previsão de desgaste dos funcionários e, consequentemente, risco de *turnover*. Foram utilizadas quatro técnicas avançadas baseadas em aprendizado de máquina: *Extra Trees Classifier (ETC)*, *Support vector machine (SVM)*, Regressão Logística (*Logistic Regression (LR)*) e Árvore de Decisão (*Decision Tree Classifier (DTC)*). Após comparação, os autores concluíram que ETC foi a melhor técnica testada.

3.4 Discussão

Compreender os motivos que levam à rotatividade é essencial para a compreensão dos modelos apresentados nos estudos aprovados neste MSL. Alguns estudos, que trabalham com dados de sistemas de recursos humanos, utilizam dados considerados demográficos (como gênero, idade, escolaridade, local de residência, etc.) para a previsão do risco de rotatividade. Outros estudos utilizam dados de pesquisas de opinião para verificar o impacto de um determinado aspecto na previsão do risco de *turnover*, como, por exemplo, a satisfação no trabalho. Os estudos que apresentam dados de pesquisa representam 54% do total de estudos primários selecionados neste mapeamento, enquanto os estudos que utilizaram dados de sistema correspondem a 31%.

Os motivos que levam à saída voluntária de funcionários estão diretamente ligados aos atributos que fazem parte do modelo. Quando são utilizados dados de sistema para alimentação do modelo, os atributos que retratam os motivos para saída geralmente fazem parte dos dados dos sistemas de RH. A definição de parâmetros é fundamental para o sucesso do modelo e representa um desafio para os modelos de previsão de *turnover*.

O baixo número de estudos encontrados que tratam sobre retenção de talento pode ter sido causado pela restrição de cenário, considerando apenas o cenário de inovação, com o uso da expressão “*innovation*” na *string* de busca. É importante ressaltar que no momento que realizamos este MSL estávamos considerando o contexto de empresas do cenário de inovação. Após a realização deste mapeamento, entendemos que seria mais adequada a utilização de cenário mais restrito ao qual faz parte a empresa parceira, que é o de empresas de Tecnologia da Informação (TI).

Embora a parte do conteúdo deste mapeamento tenha sido publicado, tal publicação foi feita na língua inglesa e em versão reduzida. Desta forma, optamos por manter o texto completo nesta tese, visando oferecer maior acesso aos leitores.

3.5 Ameaças à validade

Entre as ameaças à validade deste MSL podemos citar:

- Seleção dos estudos conduzida por apenas um dos pesquisadores. Por um lado são mitigadas questões como tratamento de divergência de opiniões, mas por outro lado a discussão sobre a qualidade dos estudos é penalizada. Ainda assim, o protocolo utilizado foi elaborado em conjunto;
- A busca foi realizada em oito bibliotecas digitais, porém a *string* de busca não foi adaptada para cada biblioteca, o que pode ter resultado em perda de estudos;
- Não foi aplicado *snowballing* nem incluídos estudos primários manualmente, o que pode ter resultado em perda de estudos;
- Nenhum estudo realizado no cenário brasileiro foi encontrado nas bases com utilização da *string* de busca utilizada, este fato exclui a possibilidade de maior entendimento do cenário brasileiro, que pode ser diferente das realidades dos locais de aplicação dos estudos selecionados;
- No momento da realização das buscas, estava-se considerando empresas do cenário de inovação; a limitação de cenário pode ter excluído alguns trabalhos do resultado das buscas;
- A falta de acesso a alguns artigos pode ter excluído estudos importantes para responder às questões de pesquisa.

4 Revisão sistemática da literatura

Embora o mapeamento sistemático tenha apresentado esclarecimentos importantes acerca dos atributos e outras questões, foram identificados poucos modelos que utilizam IA para previsão de desligamento voluntário, ou para retenção de talentos, que atendessem aos rigorosos critérios de qualidade estabelecidos. Por este motivo, para um maior entendimento dos modelos propostos na literatura e melhor identificação de lacunas, foi realizada uma Revisão Sistemática da Literatura (RSL) focada exclusivamente em trabalhos que apresentem modelos baseados em IA para previsão de desligamento voluntário ou para retenção de talentos. Assim como o mapeamento sistemático, a RSL compõe o ciclo de relevância da DSR.

Uma Revisão Sistemática da Literatura fornece processos metodológicos e estruturados para identificar e agregar evidências de pesquisa. A RSL identifica, avalia e interpreta todas as pesquisas disponíveis relevantes para uma questão de pesquisa específica, área de tópico ou fenômeno de interesse. Portanto, é um método de categorizar e resumir imparcialmente as informações existentes sobre uma questão de pesquisa. Estudos primários são estudos individuais que contribuem para uma revisão sistemática, enquanto uma revisão sistemática é um estudo secundário ([KITCHENHAM; CHARTES, 2007](#); [SILVA; VALENTIM; CONTE, 2015](#); [PETERSEN; VAKKALANKA; KUZNIARZ, 2015](#); [NAPOLEÃO et al., 2017](#)).

Mapeamentos sistemáticos e revisões sistemáticas são complementares. Ambos são revisões rigorosas, porém o mapeamento tem um escopo mais amplo e exploratório, provendo uma visão geral da área ou tópico de pesquisa, enquanto a revisão sistemática é focada em uma questão específica ([FELIZARDO et al., 2020](#); [PETERSEN; VAKKALANKA; KUZNIARZ, 2015](#)).

4.1 Protocolo da revisão sistemática

4.1.1 Objetivo e questões de pesquisa

Esta Revisão Sistemática da Literatura (RSL) tem como objetivo analisar publicações científicas com o propósito de identificar e analisar propostas de modelos ou arquiteturas utilizando tecnologias de IA para previsão de risco de desligamento voluntário, ou para retenção de talentos, buscando melhores práticas e lacunas. Desta forma focamos em uma das questões do mapeamento previamente apresentado, buscando responder à questão: “*Existem propostas de modelos ou arquiteturas utilizando IA para retenção de talentos ou previsão de desligamento voluntário?*”

Para atingir o objetivo desta revisão, também buscamos responder às seguintes subquestões de pesquisa:

- **SQP1:** *O estudo apresenta proposta para previsão de desligamento voluntário ou para retenção de talentos?* Por meio desta pergunta procuramos identificar se o estudo trata de previsão de desligamento voluntário, ou de retenção, das duas questões combinadas, ou se tem alguma outra aplicação.
- **SQP2:** *Quais as ferramentas ou técnicas de IA utilizadas?* Com esta pergunta buscamos mais informações sobre as técnicas de IA utilizadas e melhores práticas para proposta de modelo.
- **SQP3:** *Quais as principais técnicas aplicadas no preprocessamento dos dados?* Procuramos identificar a partir desta pergunta se os estudos apresentam técnicas para preprocessamento dos dados e como estão sendo aplicadas em caso positivo.
- **SQP4:** *Quais as propostas de modelos para retenção de talentos?* Finalmente, por meio desta subquestão buscamos identificar propostas de modelos para retenção de talentos, investigando possíveis lacunas de pesquisa.

4.1.2 Estratégia para busca dos estudos primários

Assim como no MSL realizado, utilizamos o PICOC (*Population, Intervention, Comparison, Outcome e Context*) para estruturar e melhorar a *string* de busca:

- **(P) Population:** *Talents* (Profissionais mapeados como talentos);

- (I) **Intervention:** *Artificial Intelligence*;
- (C) **Comparison:** Não se aplica, pois o objetivo não é realizar uma comparação;
- (O) **Outcome:** *Talent Retention*;
- (C) **Context:** *IT Industry AND "Human Resource Management"*.

A busca de artigos foi feita nas bibliotecas digitais que consideramos ser as mais relevantes, seguimos os critérios propostos por Petersen, Vakkalanka e Kuzniarz (2015) e Kitchenham e Charters (2007) para seleção das bibliotecas digitais: ACM Digital Library, Engineering Village, IEEE Digital Library, Web of Science, Scopus e Taylor & Francis.

Para construir a *string* de busca, utilizamos o operador booleano "*OR*" entre as palavras com ideias ou assuntos semelhantes para cada parâmetro e o operador booleano "*AND*" para unir os parâmetros. A *string* de busca utilizada para pesquisa nas bases está descrita na Tabela 4:

Tabela 4: *String* de busca

(("artificial intelligence"OR "AI"OR "machine learning"OR "deep learning"OR "neural networks"OR "intelligent systems") AND ("talent retention"OR "retaining talent"OR "employee retention"OR "turnover"OR "personnel retention"OR "migration of talent"OR "talent migration"OR "employee turnover") AND ("human resource management"OR "HR"OR "human capital management"OR "people analytics"OR "workforce management"OR "talent management"OR "succession planning"))
--

O período da busca inclui artigos de periódicos científicos e textos de conferências publicados entre 2014 e 2024. A pesquisa considerou trabalhos divulgados até maio de 2024.

4.1.3 Critérios para seleção de estudos

Os critérios para inclusão e exclusão dos estudos são apresentados na Figura 9. Assim como no MSL, os critérios pra seleção dos estudos foram aplicados em duas etapas. Inicialmente foram retirados os artigos duplicados. Na sequência foram analisados títulos e resumos dos artigos e verificado se os estudos atendiam a algum dos critérios de seleção. Em caso de dúvida o estudo era incluído para análise posterior do texto completo. Para cada um dos estudos foi aplicado um dos critérios de seleção. Para seleção final dos artigos foi feita uma análise do texto completo e aplicados novamente os critérios de inclusão e exclusão.

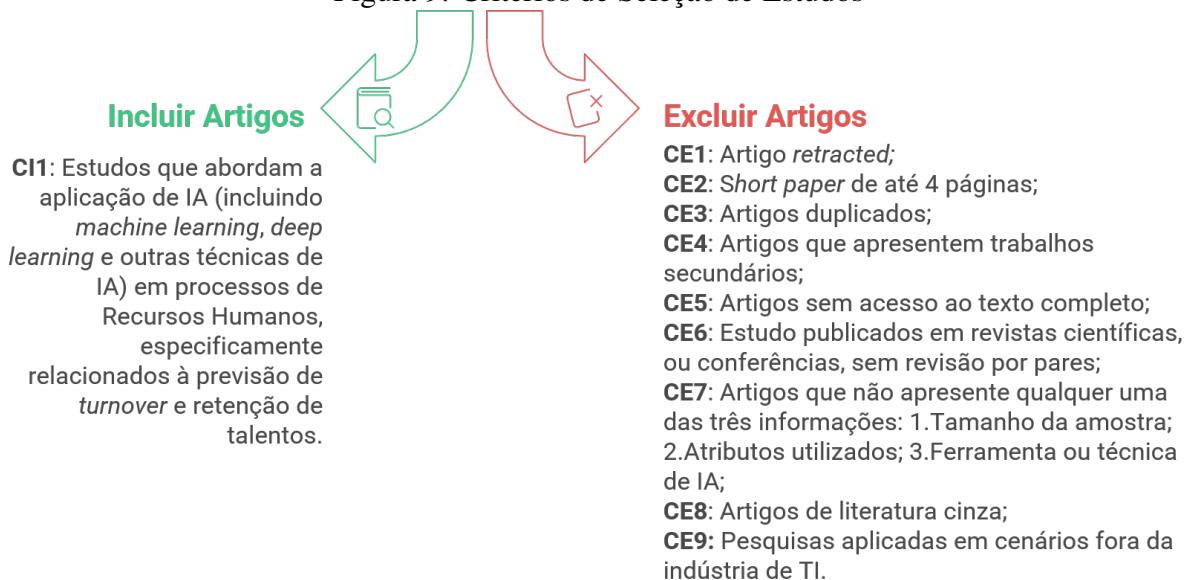
Em relação ao idioma de publicação dos estudos, diferente do MSL, onde focamos em estudos publicados apenas na língua inglesa, para esta RSL, embora a maior parte das revistas

científicas e conferências internacionais adotem o inglês como idioma principal, estudos publicados em outros idiomas devem ser considerados, seguindo as recomendações de [Kitchenham e Chartes \(2007\)](#).

Entre os critérios para exclusão dos artigos, ressaltamos que não serão incluídos artigos que contenham:

- Literatura cinza (teses, dissertações, relatórios técnicos, leis, capítulos de livros, patentes, blogs, *white papers*, artigos de opinião, entre outros) por se tratarem de documentos não revisados por pares. Adicionalmente somente foram consideradas conferências e periódicos onde há um processo de revisão por pares documentado, pois, em caso contrário, não há garantias da qualidade e credibilidade da pesquisa.
- Artigos retratados (*retracted*), visto que artigos retirados pelos autores ou pelos periódicos continuam disponíveis nas bases;
- Artigos classificados como *short paper* de até 4 páginas, pois costumam ser de trabalhos em andamento, sem resultados conclusivos;
- Artigos que aparecem duplicados em diferentes bases de dados, pois bases como a Scopus, por exemplo, indexam artigos de várias editoras (springer, Taylor&Francis, Elsevier);
- Artigos que apresentem trabalhos secundários (revisões de literatura) que não apresentam nenhuma solução nova diferente das verificadas na revisão;
- Artigos sem acesso ao texto completo, pois não permitirão uma avaliação completa dos métodos e resultados;
- Artigos que não apresentem qualquer uma das três informações: 1.Tamanho da amostra; 2.Atributos utilizados; 3.Ferramenta ou técnica de IA;
- Pesquisas aplicadas em cenários fora da indústria de TI, pois a pesquisa está no contexto desta indústria.

Figura 9: Critérios de Seleção de Estudos



Fonte: Elaborado pela autora

4.1.4 Avaliação da qualidade

Uma vez selecionados os artigos aderentes às questões de pesquisa definidas após leitura dos textos completos, na última fase da revisão são aplicados os Critérios de Qualidade (CQ) para garantir que sejam selecionados artigos relevantes. Para definição dos critérios de qualidade utilizamos como base os trabalhos de [Dyba, Dingsoyr e Hanssen \(2007\)](#), [Dybå e Dingsøyr \(2008\)](#) e [Kitchenham e Brereton \(2013\)](#), que ressaltam a importância de comparar a qualidade de artigos que usam diferentes métodos de pesquisa. Ainda neste contexto, assim como definido por [Dyba, Dingsoyr e Hanssen \(2007\)](#), consideramos necessário que os critérios de qualidade contemplam questões tais como rigor, credibilidade, relevância e clareza:

- Se a abordagem aplicada aos métodos de pesquisa foi minuciosa e adequada, apresentando o rigor esperado;
- Se os estudos têm credibilidade, com resultados significativos e bem apresentados;
- Se os estudos são relevantes e úteis, ajudando a responder às questões de pesquisa;
- Se os estudos apresentam metas e objetivos claramente relatados, além de uma descrição adequada do contexto em que a pesquisa foi realizada.

Desta forma, adotamos para o critério de qualidade questões que englobassem as quatro subquestões geradas nesta RSL. Foram atribuídos os seguintes pesos para as respostas: (1) Peso

2,0 para resposta "Sim"; (2) Peso 1,0 para resposta "Parcialmente"; (3) Peso 0,0 para resposta "Não". A pontuação atribuída à qualidade do artigo pode variar de 14,0 a 28,0 pontos. Artigos com pontuação de qualidade menor que 14,0 pontos são retirados da seleção por não atenderem aos critérios de qualidade. No Apêndice C estão detalhados os critérios aplicados a cada um dos estudos selecionados.

4.1.5 Estratégia para extração de dados

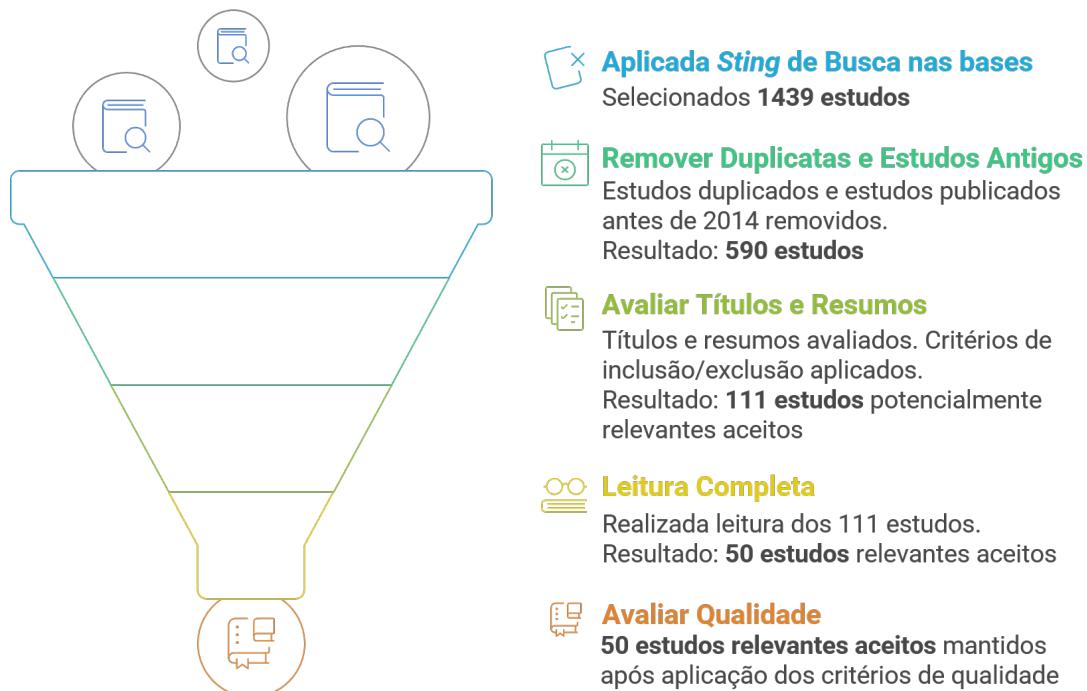
A estratégia de extração de dados visa garantir que os dados dos estudos sejam coletados de forma uniforme, favorecendo o processo de responder às subquestões desta RSL.

4.2 Resultados

4.2.1 Metodologia para busca e seleção de estudos primários

A metodologia para busca dos estudos primários relevantes seguiu processo semelhante ao da metodologia utilizada no MSL, conforme apresentado na Figura 10.

Figura 10: Processo de busca e seleção de estudos primários



Fonte: Elaborado pela autora

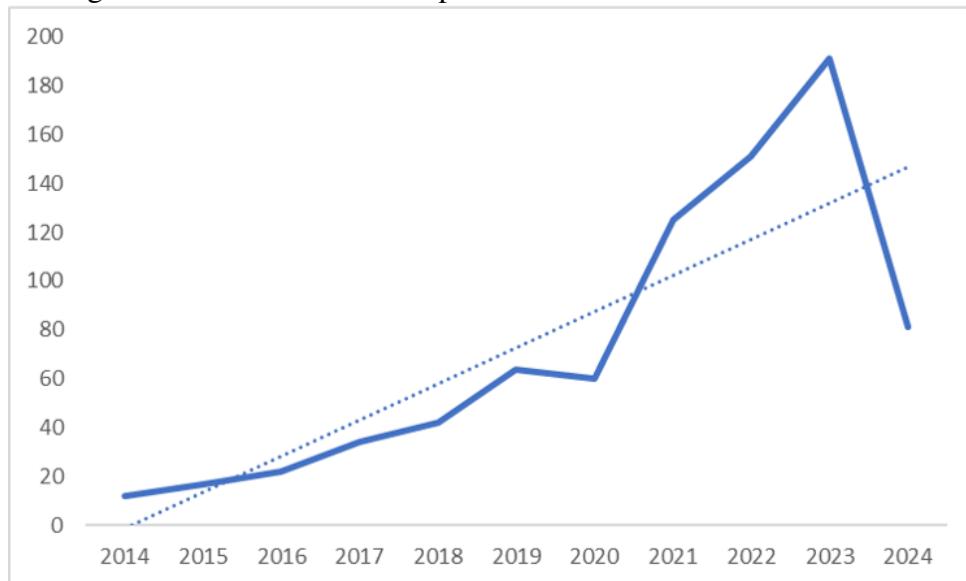
Inicialmente foram feitas as buscas dos estudos primários nas bibliotecas digitais supracita-

das, a partir da *string* de busca, obtendo um resultado quantitativo de 1439 estudos. Em seguida foram retirados os artigos duplicados e os artigos publicados antes de 2014, resultando em 590 estudos. Posteriormente, foi feita a leitura dos títulos e resumos (*abstract*) e aplicados os critérios de inclusão e exclusão previamente definidos, sendo selecionados os 111 estudos primários potencialmente relevantes. Destes estudos, foi feita a leitura completa dos estudos com acesso gratuito e aplicados novamente os critérios de inclusão e exclusão. Ao final foram selecionados um total de 50 estudos primários relevantes. Foram aplicados os critérios de qualidade para cada um dos estudos selecionados e todos os estudos foram mantidos. A relação dos estudos primários selecionados está disponível no Apêndice B.

4.2.2 Resultados quantitativos

O total de trabalhos encontrados no período de 2014 a maio de 2024, após utilização da *string* de busca nas bases, é apresentado na Figura 11. É possível observar um aumento significativo nas publicações sobre este tema, apresentando uma curva de tendência ascendente. Somente entre 2014 e maio de 2024 foram identificados 590 trabalhos publicados nesta temática, sem contar os 209 trabalhos duplicados.

Figura 11: Total de trabalhos publicados entre 2014 e maio de 2024

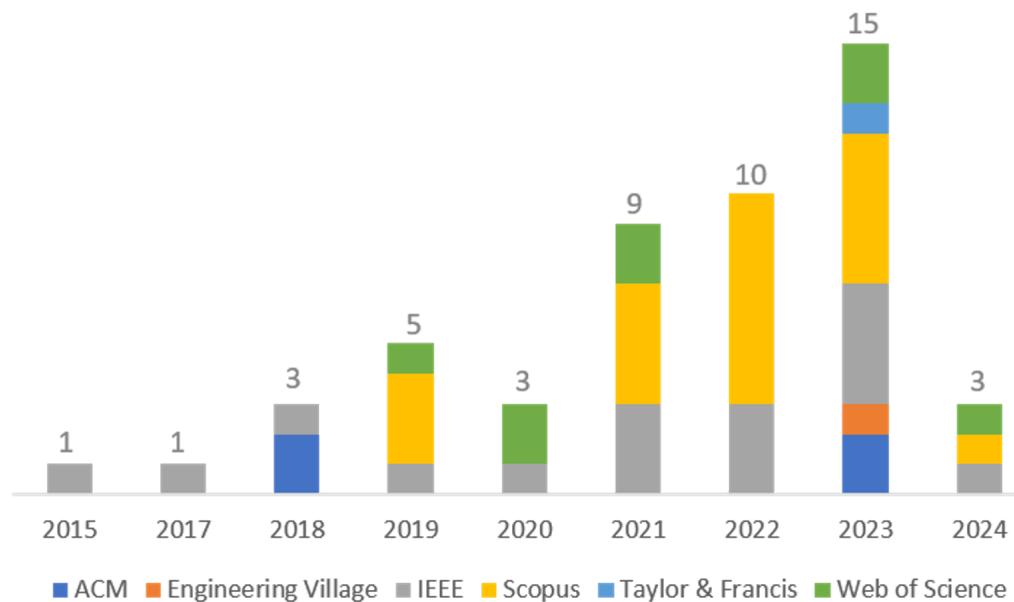


Fonte: Elaborado pela autora

A distribuição dos artigos aceitos ao final do processo de busca e seleção dos estudos primários é apresentada na Figura 12. Notamos que todas as bases consultadas apresentaram trabalhos aceitos, sendo alguns trabalhos incluídos em apenas uma das bases, sem estarem duplicados. As bases Scopus e IEEE lideram a quantidade de artigos aceitos, somando 72% dos estudos, sendo

40 % da Scopus e 32% da IEEE. Os estudos selecionados refletem a mesma tendência de aumento verificada no total de estudos, com 30% dos estudos selecionados tendo sido publicados em 2023.

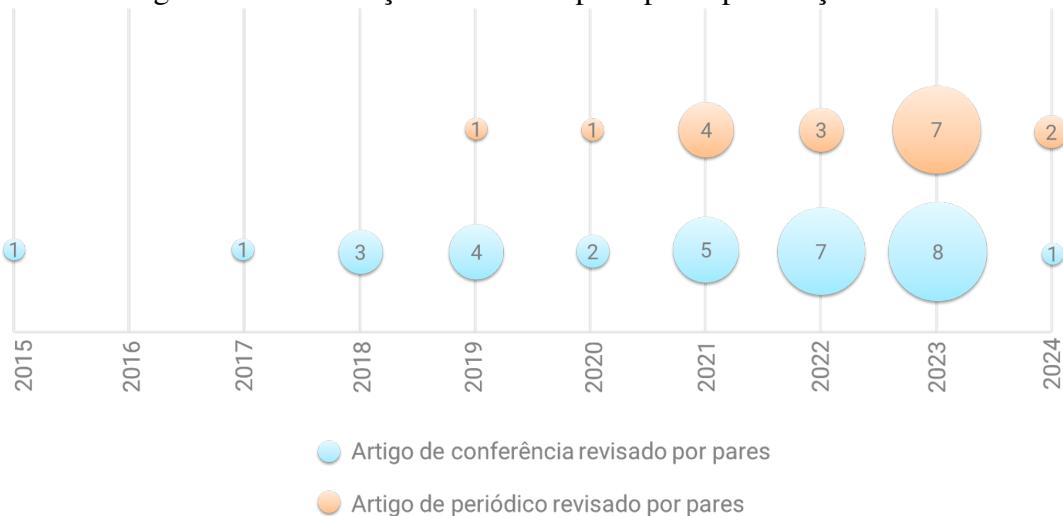
Figura 12: Total de trabalhos aceitos por base entre 2014 e maio de 2024



Fonte: Elaborado pela autora

Dos 50 artigos aceitos, 18 são artigos publicados em periódicos e 32 publicados em anais de conferências, conforme ilustrado na Figura 13.

Figura 13: Distribuição de estudos por tipo de publicação e ano



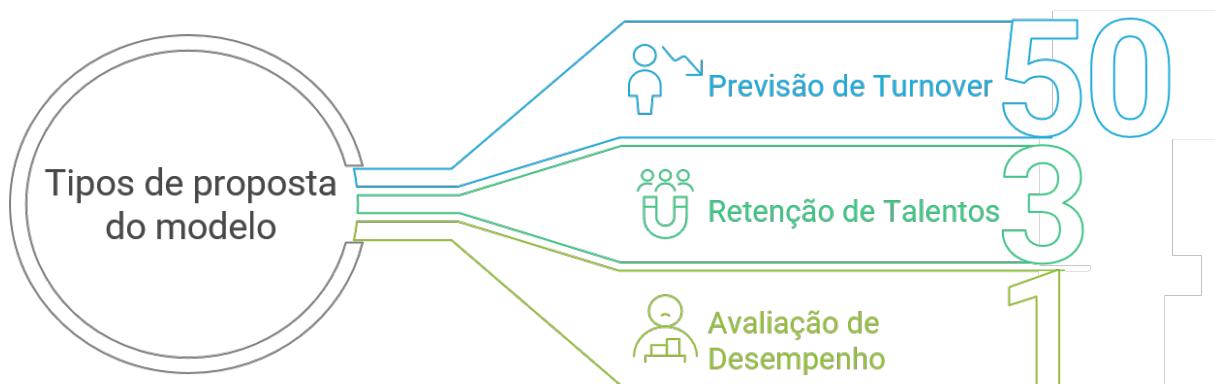
Fonte: Elaborado pela autora

4.2.3 Contribuições dos estudos

RSQP1: *O estudo apresenta proposta para previsão de desligamento voluntário ou para retenção de talentos?*

Todos os estudos selecionados apresenta modelo para previsão de desligamento voluntário. Além do modelo de previsão de desligamento voluntário, 3 estudos apresentam modelos para retenção de talentos (MITRAVINDA; SHETTY, 2022; MOHIUDDIN et al., 2023; RAMAMURTHY et al., 2015) e 1 estudo apresenta modelo para avaliação de desempenho (SUN et al., 2021), conforme ilustrado na Figura 14.

Figura 14: Tipos de proposta de modelo dos estudos



Fonte: Elaborado pela autora

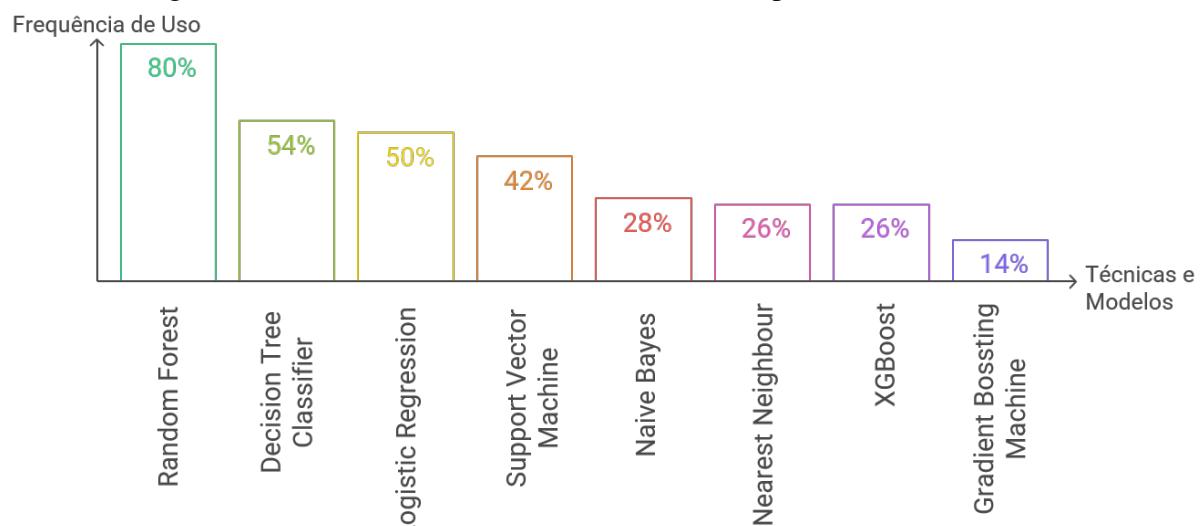
RSQP2: *Quais as ferramentas ou técnicas de IA utilizada?*

Todos os 50 estudos primários selecionados utilizam técnicas de *Machine Learning* (ML), destes, 7 trabalhos utilizam técnicas de *Deep Learning* (DL) (PAIGUDE et al., 2023; WANG; ZHI, 2021; TENG et al., 2019; SRIVASTAVA; EACHEMPATI, 2021; SUN et al., 2021; CHUNG et al., 2023; SARDAR; CHOURASIYA; VIJYALAKSHMI, 2023) e 2, além de ML, utilizam também técnicas de Processamento de Linguagem Natural (PLN) (MOHIUDDIN et al., 2023; MA et al., 2020).

Entre os tipos de classificadores utilizados, o mais aplicado foi o *Random Forest* (RF), constando em 80% dos estudos, em seguida *Decision Tree Classifier* (DTC), utilizado em 54% dos estudos, *Logistic Regression* (LR), constando em 50% dos estudos, *Support Vector Machine* (SVM), em 42% dos estudos. *Naive Bayes* é empregado em 28%, enquanto *k-Nearest Neighbor* (KNN) e *Extreme-Gradient Boosting* (XGBoost), em 26% dos estudos cada, e *Gradient Boosting Machine* (GBM) é utilizado em 14% dos estudos. A Figura 15 apresenta esta distribuição

da utilização dos classificadores nos estudos.

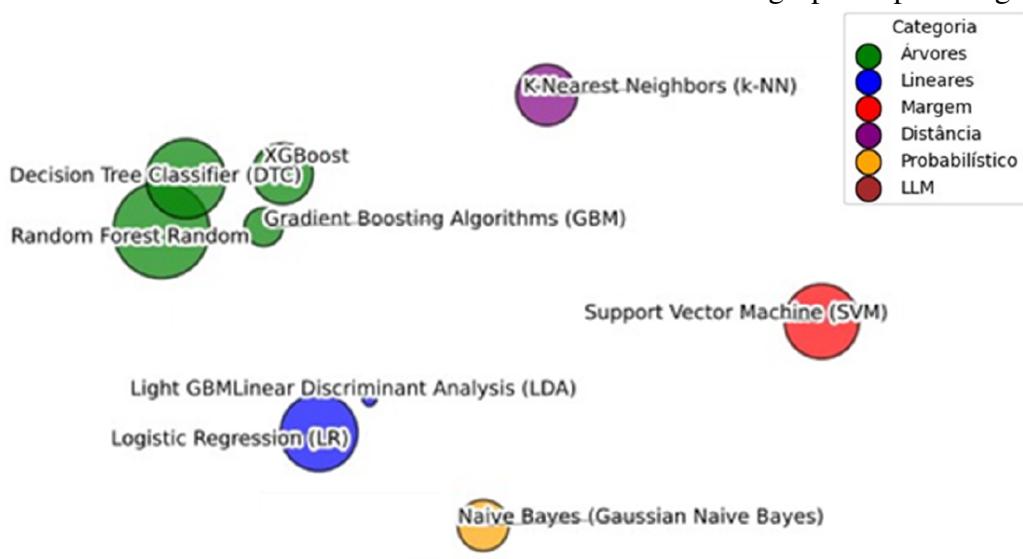
Figura 15: Classificadores utilizados nos estudos primários selecionados



Fonte: Elaborado pela autora

A Figura 16 apresenta a distribuição dos diferentes tipos de classificadores agrupados por categorias.

Figura 16: Classificadores utilizados nos estudos selecionados agrupados por categoria



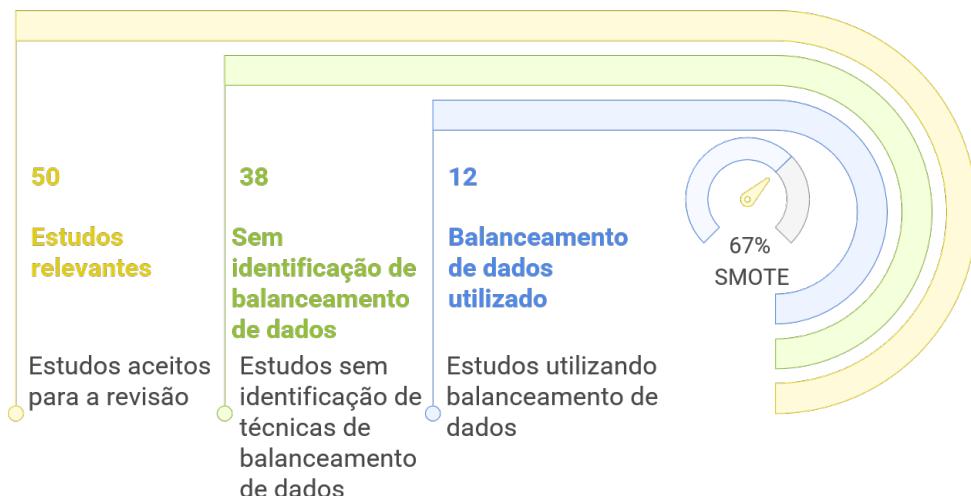
Fonte: Elaborado pela autora

RSQP3: Quais as principais técnicas aplicadas no preprocessamento dos dados?

Dos 50 estudos primários relevantes aceitos, em 38 estudos não foi possível identificar

a utilização de técnicas para balanceamento dos dados. Em 12 estudos, os autores declaram utilização de técnicas para balanceamento dos dados. Destes, aproximadamente 67% (8 estudos - *ERSL 24, ERSL 48, ERSL 28, ERSL 2, ERSL 4, ERSL 8, ERSL 43 e ERSL 5*) utiliza *Synthetic Minority Over-sampling Technique* (SMOTE), conforme ilustrado na Figura 17.

Figura 17: Utilização de técnicas de balanceamento de dados nos estudos



Fonte: Elaborado pela autora

O SMOTE é uma técnica de *oversampling* usada para lidar com conjuntos de dados desbalanceados, especialmente quando há uma minoria de exemplos pertencentes a uma determinada classe. Diferente da simples replicação de instâncias da classe minoritária, o SMOTE gera novas amostras sintéticas ao invés de duplicar exemplos existentes. A geração dessas amostras ocorre no espaço das características (*feature space*) por meio de uma interpolação entre uma instância da classe minoritária e seus vizinhos mais próximos. O algoritmo seleciona aleatoriamente um dos k vizinhos mais próximos, calcula a diferença entre os vetores de características e multiplica essa diferença por um valor aleatório entre 0 e 1. O resultado é então somado à instância original para formar uma nova amostra sintética. Esse processo cria maior variação entre as instâncias minoritárias, contribuindo para um treinamento mais robusto dos modelos de aprendizado de máquina e reduzindo o risco de *overfitting* que pode ocorrer com técnicas de replicação direta (GOORBERGH et al., 2022; FERNÁNDEZ et al., 2018; CHUNG et al., 2023).

RSQP4: Quais as propostas de modelos para retenção de talentos?

Três estudos apresentam propostas para retenção de talentos:

- Ramamurthy et al. (2015) desenvolveram e implementaram um sistema para retenção de

talentos baseado em análise preditiva da propensão de saída (*PTL – Propensity to Leave*), onde explicam os fatores que contribuem para essa probabilidade e geram recomendações personalizadas de retenção acionáveis por gestores. Para tal foram utilizados modelos de ML interpretáveis, principalmente C5.0 *rule sets* e regressão logística. Foi desenvolvido um *Dashboard* interativo para uso de os gestores, RH, altos executivos e os próprios funcionários. O sistema apresenta Categoria de risco (alto, médio, baixo) baseada no PTL *score*, Fatores observados que explicam o risco, Ações recomendadas específicas e contextualizadas (recomendações como: Conversas com o funcionário, revisão de plano de carreira, promoção ou realocação, aumento de salário quando compatível com a estratégia, etc). As ações são priorizadas com base na ação mais impactante para o perfil do funcionário. Apresenta espaço *feedback* do gestor, que pode ajustar manualmente o risco percebido.

- [Mitavinda e Shetty \(2022\)](#) propuseram e desenvolveram um modelo preditivo com foco em identificar causas de desligamento e recomendar ações específicas para retenção de talentos. Além de prever a probabilidade de saída de um funcionário, identifica o principal fator de risco para cada caso individual e gera recomendações personalizadas para retenção com base em registros semelhantes (ex: reduzir horas extras, revisar promoções, melhorar salário). Cada recomendação está associada a um atributo do modelo. Utiliza *SHapley Additive exPlanations* (SHAP) para identificar a variável mais influente para a predição de cada funcionário, por exemplo: "*OverTime*" foi o fator mais frequente, seguido por "*MonthlyIncome*", "*StockOptionLevel*" e "*Age*", neste caso o risco de saída é associado a *OverTime* e o sistema recomendará que "Compense de forma justa a hora extra de trabalho". Índice SHAP é uma abordagem matemática usada para descrever o papel de cada característica em uma previsão feita por um modelo de aprendizado de máquina.
- [Mohiuddin et al. \(2023\)](#) propuseram e implementaram um modelo explicável de apoio à decisão (*Human Resource Decision Support System*(HR-DSS)) voltado à retenção de talentos, com foco em predição e explicação de risco de saída de funcionários. Na proposta foram testados 8 modelos de aprendizado de máquina (*XGBoost* (XGB), *LightGBM*, *Random Forest*, *Decision Tree*, *Logistic Regression*, SVM, MLP e *Gaussian Naive Bayes* (GNB)), sendo o XGB o com melhor desempenho. Criaram um painel baseado em *SHapley Additive exPlanations* (SHAP) para análise visual das predições por funcionário, utilizando Geração de Linguagem Natural (*Natural Language Generation* - NLG) com GPT-3.5 para transformar SHAP em explicações textuais personalizadas para o RH.

4.3 Discussão

Dentre os diversos fatores que limitam o acesso a bases de RH, dois fatores são determinantes: se tratar de dados sensíveis e leis para proteção de dados pessoais. Alguns pesquisadores recorreram ao uso de bases sintéticas para tentar contornar a falta de acesso a bases reais de RH. Foi observado que mais da metade dos estudos primários aceitos (56%) utilizam bases sintéticas disponíveis no Kaggle¹. Destes, 21 estudos utilizam uma mesma base com 1470 registros e 7 estudos utilizam mesma base com 14.999 registros. Dois destes estudos utilizam as duas bases em seus experimentos. Mesmo utilizando as mesmas bases, há uma variedade de resultados, até mesmo para os trabalhos que utilizaram os mesmos tipos de classificadores para treinar os modelos. Esse fato enfatiza a importância do tratamento dos dados, mostrando como este pode influenciar o resultado. Adicionalmente, a utilização de dados sintéticos pode acarretar em viés, por isso a importância do uso de dados reais para tratamento de questões sensíveis como as ligadas à área da saúde e de RH.

Em relação à discussão ou tratamento de vieses, identificamos que 28% dos estudos apresentam alguma informação sobre este tema, mesmo que seja apenas a admissão da possibilidade da existência de viés no estudo. Metade destes estudos, ou seja, 7 estudos (*E_{RSL}* 35, *E_{RSL}* 4, *E_{RSL}* 8, *E_{RSL}* 20, *E_{RSL}* 32, *E_{RSL}* 40 e *E_{RSL}* 42) utilizam a base de dados com 1470 registros supracitada e admitem a possibilidade de viés nos dados. Os demais estudos (*E_{RSL}* 24, *E_{RSL}* 39, *E_{RSL}* 25, *E_{RSL}* 46, *E_{RSL}* 6, *E_{RSL}* 15 e *E_{RSL}* 31) utilizam bases de dados reais, mas ainda assim apresentam alguma discussão acerca de possíveis vieses. [Jesus, Júnior e Brandão \(2018\)](#) e [Chang et al. \(2023\)](#) indicam a possibilidade de existência de viés algorítmico, uma vez que ambos utilizam dados reais de plataformas de mídias sociais relacionadas a questões profissionais, sendo os dados do primeiro estudo provenientes do LinkedIn² e do segundo estudo oriundos do Glassdoor³.

Esta RSL não apresenta uma subquestão de pesquisa para avaliar os principais atributos utilizados nos sistemas, pois o foco da revisão está nas tecnologias de IA utilizadas. Ainda assim, pudemos observar que do total dos estudos, 80% apresenta informações sobre os principais atributos ligados ao motivo para pedido de desligamento por parte do funcionário. Destes motivos apresentados, o com maior ocorrência é "Satisfação no trabalho", que aparece em 22 estudos, seguido de "Salário"(19 estudos), "tempo de serviço"(13 estudos) e "Equilíbrio entre trabalho e vida pessoal"(7 estudos).

¹<https://www.kaggle.com/>

²<https://www.linkedin.com/>

³<https://www.glassdoor.com>

Verificamos que apenas um estudo brasileiro (JESUS; JÚNIOR; BRANDÃO, 2018) foi incluído na RSL, sendo o único estudo brasileiro que retornou nas busca atendendo aos critérios de seleção. Visando tentar mitigar esta questão, e buscar mais estudos sobre utilização de ferramentas de IA para retenção de talentos ou predição de desligamento voluntário realizados no Brasil, foram realizadas buscas na plataforma SBC-*OpenLib* (SOL) da Sociedade Brasileira de Computação (SBC). Utilizamos a *string* de busca em inglês e traduzida para o português, porém não houve retorno de estudos.

Entre as propostas apresentadas para modelos para retenção de talentos, foi possível observar que, mesmo indicando ações a serem seguidas pelo gestor, tais ações são diretamente associadas apenas aos atributos do sistema (RAMAMURTHY et al., 2015; MOHIUDDIN et al., 2023; MITRAVINDA; SHETTY, 2022). As propostas apresentadas associam as ações a serem seguidas pelo gestor de forma estática ao atributo indicado como principal motivo, ou seja, para cada atributo é pré-cadastrada uma ação, que será informada ao gestor em forma de recomendação. Não foi possível identificar qualquer proposta para retenção que vá além da identificação dos principais atributos e indicar ações para o gestor alinhadas ao atributo.

Os achados desta RSL foram fundamentais para identificação de características dos modelos a serem aplicados para previsão de desligamento voluntário, assim como para identificação de lacunas em propostas para retenção de talentos.

4.4 Ameaças à validade

Verificamos a possibilidade das seguintes ameaças à validade desta revisão sistemática da literatura:

- Embora a seleção dos estudos tenha sido conduzida por dois pesquisadores, o que proporcionou uma discussão acerca da qualidade dos trabalhos e quais deveriam ser incluídos, ainda pode ter havido perda de estudos pela limitação de cenário ao de empresas de TI;
- A busca foi realizada em seis bibliotecas digitais, porém a *string* de busca não foi adaptada para cada biblioteca, o que pode ter resultado em perda de estudos;
- Não foi aplicado *Snowballing* nem incluídos estudos primários manualmente, o que pode ter resultado em perda de estudos;
- A falta de acesso a alguns artigos pode ter excluído estudos importantes para responder às questões de pesquisa.

5 Abordagem proposta

5.1 Introdução

O mapeamento sistemático e a revisão sistemática da literatura forneceram a base teórica necessária para o desenvolvimento da proposta. Correspondendo ao ciclo de relevância da DSR, as informações extraídas tanto do MSL, quanto da RSL, foram fundamentais para a investigação do problema, e seu subsequente entendimento, dado o ambiente de aplicação. Outrossim, pudemos identificar algumas lacunas a partir da revisão da literatura. A seguir, apresentamos nossa proposta para preenchimento destas lacunas encontradas e aplicação do conhecimento obtido para desenvolver nossa proposta a partir dos achados dos estudos.

Pressupostos para o modelo, definidos a partir dos achados da investigação do problema (Capítulos 3 e 4):

- **Utilização de técnicas para balanceamento de dados:** Observamos que alguns estudos utilizam técnicas para balanceamento dos dados na etapa de pré-processamento dos dados. Embora técnicas para balanceamento dos dados, como SMOTE, por exemplo, possam reduzir o risco de *overfitting*, tais técnicas inserem dados sintéticos, aumentando o risco de inserção de vieses. Talvez esta questão não represente problemas em determinadas aplicações. Porém, sua utilização em bases de dados sensíveis, como bases médicas, bases de RH, entre outras, pode significar um risco.
 - Para minimizar este risco, sugerimos a utilização de técnicas mais "justas" para balanceamento de dados, que atenuem o risco de inserção de viés.
- **Como o pré-processamento dos dados e a forma de utilização da base podem influenciar o resultado da aplicação do modelo:** Verificamos por meio da revisão sistemática da literatura a utilização de mesma base de dados por diversos estudos diferentes, obtendo resultados diferentes. Conforme apresentado na subseção 4.3, foram identificados:

21 estudos¹ que utilizam mesma base com 1.490 registros, 7 estudos² com mesma base de 14.999 registros, sendo 2 estudos³ utilizando estas duas bases.

- Esta observação foi importante para concluirmos que a aplicação de nosso modelo depende da base de dados que será utilizada, dificultando a replicabilidade direta. Isto significa dizer que para cada nova base de dados, ou nova empresa, onde o modelo seja aplicado, será necessário novo treinamento dos classificadores, podendo haver resultado totalmente diferente do que será apresentado no experimento da próxima seção.
- **Classificadores de aprendizado de máquina:** Entre as ferramentas de IA utilizadas nos estudos primários selecionados, as mais utilizadas são os classificadores de aprendizado de máquina (ou *machine learning*). Todos os estudos primários selecionados na RSL utilizam tais técnicas, apresentando, em muitos casos, uma comparação entre os classificadores, destacando o classificador que obteve melhor desempenho.
 - Por intermédio dos estudos primários selecionados pudemos verificar os classificadores que apresentaram melhor desempenho, servindo como exemplo para nossos testes.
- **Propostas para retenção de talentos:** Alguns estudos aceitos na RSL apresentam proposta para retenção de talentos com utilização de sistema de recomendações. Verificamos que todas as propostas utilizam os atributos do sistema para indicar risco de desligamento voluntário e ações que podem ser tomadas pelos gestores, geralmente diretamente ligadas aos tipos de atributos.
 - Nossa proposta para retenção de talentos apresenta uma novidade em relação aos demais estudos. Estamos propondo a verificação de competências para garantir que os funcionários estejam nos projetos corretos, visando não só adequação de perfil, mas também de desafios.

Nesta seção apresentamos a primeira etapa do Ciclo de *Design* da DSR, que consiste em desenvolver e avaliar o arcabouço. Desta forma, apresentamos o desenvolvimento do arcabouço conceitual, ou *framework* conceitual, proposto.

¹Estudos que utilizam base sintética com 1.490 registros: *E_{RSL} 4, E_{RSL} 8, E_{RSL} 9, E_{RSL} 13, E_{RSL} 17, E_{RSL} 18, E_{RSL} 20, E_{RSL} 21, E_{RSL} 27, E_{RSL} 28, E_{RSL} 32, E_{RSL} 35, E_{RSL} 36, E_{RSL} 37, E_{RSL} 40, E_{RSL} 42, E_{RSL} 43, E_{RSL} 45, E_{RSL} 47, E_{RSL} 48 e E_{RSL} 49.*

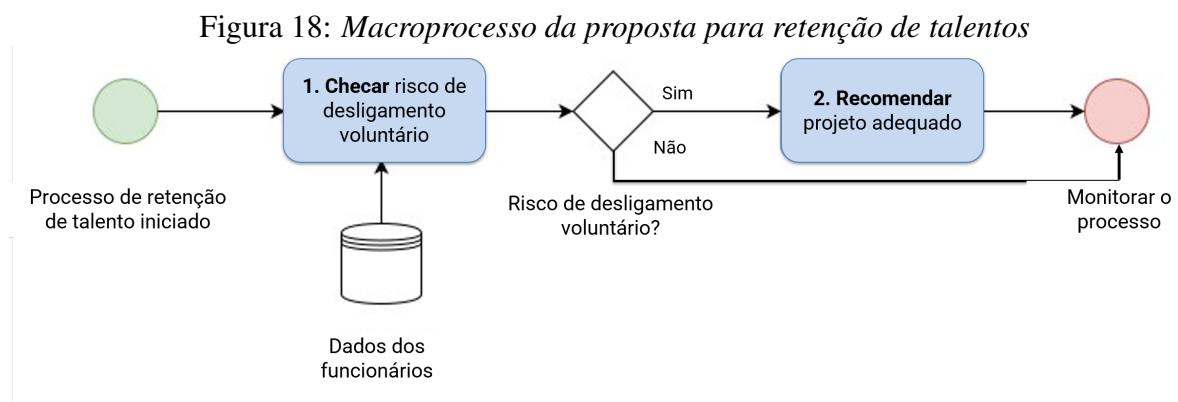
²Estudos que utilizam base sintética com 14.999 registros: *E_{RSL} 4, E_{RSL} 5, E_{RSL} 14, E_{RSL} 16, E_{RSL} 22, E_{RSL} 27 e E_{RSL} 50.*

³Estudos que utilizam as duas bases sintéticas com 1.490 e 14.999 registros: *E_{RSL} 4 e E_{RSL} 27*.

5.2 Visão geral do *framework* proposto

As etapas de processamento de dados em aprendizado de máquina variam de acordo com a técnica específica em uso, mas, geralmente, ocorrem em um ciclo de vida bem definido, que pode ser representado por um fluxo de trabalho. O ciclo de vida completo de um processo de aprendizado envolve uma série de etapas, desde a aquisição e preparação de dados até a implantação do modelo em um ambiente de produção. A construção de modelos está no centro de qualquer técnica de aprendizado de máquina (FILHO et al., 2022).

Nossa proposta para tratar o problema de rotatividade em empresas de TI utilizando tecnologias de inteligência artificial se divide em duas etapas ou dois *frameworks*. O primeiro (1. Checar risco de desligamento voluntário) tem como objetivo identificar risco do funcionário pedir demissão. No segundo (2. Recomendar projeto adequado), é proposto um modelo conceitual para ranqueamento dos projetos mais adequados ao perfil do funcionário com risco de saída, caracterizando uma proposta para retenção de talentos, que pode ser aplicada pelo gestor em conjunto com outras diretrizes que não serão abordadas neste trabalho. Desta forma, a sugestão de uso é para gestores de equipes e gerentes de RH, visando a retenção dos talentos da empresa, garantindo, assim, a gestão do conhecimento e sustentabilidade do negócio. A Figura 18 ilustra o macro processo da proposta para retenção de talento.



Fonte: Elaborado pela autora

5.3 Framework para previsão de risco de desligamento voluntário

Questões de pesquisa do MSL e da RSL analisam a existência de propostas de modelos ou arquiteturas na literatura que utilizam técnicas de IA para retenção de talentos, seja de forma direta ou indireta, bem como suas principais características. Foi verificada a existência de diversos

modelos que direta ou indiretamente tratavam este assunto. A partir do estudo destes modelos, chegamos à proposta do *framework*, que tem como foco a previsão de risco de desligamento voluntário de um funcionário, ou seja, do profissional pedir demissão. Também foi possível observar a partir destes estudos quais os principais classificadores de aprendizado de máquina são empregados no desenvolvimento destas propostas. A Figura 19 apresenta o processo proposto para previsão do risco de desligamento voluntário.

Figura 19: *Framework* proposto: Processo de aprendizado de máquina para predição de risco de desligamento voluntário



Fonte: Elaborado pela autora

A seguir são descritas as fases do processo para detalhamento do modelo:

- A primeira etapa (1) do processo de aprendizado de máquina supervisionado é o carregamento da base de dados.
- A próxima etapa (2) é o pré-processamento dos dados, onde ocorrem a limpeza de dados e/ou as conversões de tipos de atributos da base dados, conforme necessário.
- Neste momento a base de dados é dividida entre "dados para treinamento e validação" e "dados para teste".
- Na sequência, etapa (3) é realizado o balanceamento dos dados. A classe majoritária é representada por casos em que o funcionário permanece na posição de trabalho, e a classe minoritária os casos de desligamento voluntário do funcionário.
- Em seguida (4), é executado o algoritmo de treinamento utilizando vários modelos de classificação. Serão utilizadas medidas de desempenho apropriadas para classificadores binários, como sensibilidade ou *recall*, especificidade, valor predictivo positivo, acurácia, *F1-score*, e a área sob a curva ROC. O algoritmo de treinamento ainda utiliza validação cruzada de *n-folds*.

- Na última etapa do treinamento (5), os resultados de desempenho dos classificadores serão comparados e será selecionado aquele classificador que obtiver melhor desempenho.

Entre os principais classificadores utilizados nos estudos primários selecionados no MSL e na RSL, destacamos:

- ***Decision Tree Classifier (DTC)***: O classificador *Decision Tree*, ou árvore de decisão, é um método supervisionado amplamente utilizado para tarefas de classificação. Sua estrutura é hierárquica, formada por um conjunto de nós de decisão, ramos e folhas, onde cada nó interno representa uma condição baseada em um atributo do conjunto de dados, cada ramo indica o resultado de um teste, e as folhas correspondem às classes finais preditas (ZHU et al., 2019). O processo de construção da árvore é feito por meio de uma divisão recursiva dos dados em subconjuntos, com base nos atributos que mais reduzem a incerteza ou a impureza na classificação (YAHIA; HLEL; COLOMO-PALACIOS, 2021).
- ***Random Forest (RF)***: O *Random Forest* é um algoritmo de aprendizado de máquina do tipo *ensemble* baseado em múltiplas árvores de decisão. Ele combina os resultados de várias árvores independentes para produzir uma previsão mais robusta e precisa. É mencionado como uma ferramenta para classificação e regressão de compostos. O *Random Forest* é um algoritmo robusto que pode lidar com grandes volumes de dados e é eficaz em lidar com *overfitting*, um problema comum em modelos de aprendizado de máquina (CHUNG et al., 2023; ZHU et al., 2019; MITRAVINDA; SHETTY, 2022).
- ***Gradient Boosting (GB)***: O classificador *Gradient Boosting* é uma técnica de aprendizado de máquina baseada no conceito de ensemble, onde múltiplos modelos fracos, geralmente árvores de decisão, são combinados de forma sequencial para formar um modelo forte, com maior poder preditivo. Em cada iteração, o modelo tenta corrigir os erros cometidos nas etapas anteriores, ajustando-se aos resíduos ou aos gradientes negativos da função de perda, o que caracteriza o método como uma otimização baseada em gradiente. Uma das principais vantagens dessa abordagem é sua flexibilidade para diferentes tipos de funções de perda, bem como a possibilidade de incorporar mecanismos de regularização, como a redução da taxa de aprendizado (*shrinkage*) e o controle da profundidade das árvores, para evitar *overfitting* (BENTÉJAC; MORVAN; MOUTARLIER, 2020; FRIEDMAN, 2001).
- ***Extreme-Gradient Boosting (XGBoost)***: O XGBoost, é uma técnica de *ensemble* escalável que se destacou como uma solução eficiente e confiável em desafios de aprendizado de máquina. Ele se insere na família de algoritmos de *gradient boosting* e é reconhecido

por sua capacidade de oferecer melhor desempenho, precisão e uso eficiente da memória em comparação com outros modelos. Além disso, o XGBoost utiliza um algoritmo aproximado de aprendizado de árvore para lidar de forma eficaz com dados esparsos e incorpora parâmetros de regularização adicionais, oferecendo maior proteção contra o *overfitting* (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021; OSMAN et al., 2021).

- ***k-Nearest Neighbor (KNN)***: O classificador KNN é um algoritmo de mineração de dados amplamente utilizado para classificação supervisionada, conhecido por sua simplicidade e eficiência em diferentes domínios de aplicação. Ele opera sob o princípio de "*lazy learning*", onde a computação só ocorre no momento da previsão. Para classificar um novo ponto de dados, o KNN identifica os K vizinhos mais próximos no conjunto de treinamento e atribui a ele a classe mais comum entre esses vizinhos. Apesar de sua popularidade, o KNN enfrenta desafios como a seleção do valor ideal de K, a busca pelos vizinhos mais próximos e a sensibilidade a conjuntos de dados grandes ou com ruído, o que exige métodos eficientes para lidar com grandes volumes de dados e a imperfeição das amostras (SHOKRZADE et al., 2021; ZHANG, 2021; ZHANG; LI, 2021).
- ***Logistic Regression (LR)***: A Regressão Logística, ou *Logistic Regression*, é um método estatístico amplamente utilizado para tarefas de classificação binária em aprendizado de máquina, sendo especialmente valorizado por sua simplicidade, interpretabilidade e robustez. Ele estima a probabilidade de um evento ocorrer, transformando uma combinação linear de variáveis preditoras por meio da função logística. Esta função garante que as saídas estejam no intervalo de 0 a 1, interpretáveis como probabilidades, o que a torna adequada para prever resultados dicotômicos (opostos ou binários, por exemplo, sim/não, 0/1) (SARAN; NAR, 2025; OMAR et al., 2024).
- ***Support Vector Machine (SVM)***: O classificador SVM é amplamente aplicado em tarefas de classificação e regressão, sendo uma técnica de aprendizado supervisionado que se destaca pela sua capacidade de encontrar o hiperplano ótimo que melhor separa as classes de um conjunto de dados, maximizando a margem entre os pontos de diferentes categorias. Uma característica chave do SVM é sua capacidade de evitar o *overfitting*, mesmo com limites de decisão simples. A eficácia do SVM é influenciada pela otimização de seus hiperparâmetros e pela escolha de diferentes funções de *kernel*, que permitem mapear os dados para espaços de maior dimensão, tornando-o versátil para lidar com complexas relações não lineares (KURANI et al., 2023; OMAR et al., 2024; ROY; CHAKRABORTY, 2023).

Cabe ressaltar que os modelos de classificadores fornecem um resultado numérico no do-

mínio entre 0 e 1. Esse resultado numérico poderá ser associado ao risco de desligamento voluntário, quanto mais próximo a 1, maior o risco de desligamento voluntário.

Entre as métricas para avaliação dos classificadores, destacamos:

- **Acurácia:** A acurácia é uma das métricas mais tradicionais na avaliação de classificadores e representa a proporção de previsões corretas em relação ao total de instâncias avaliadas. Ela fornece uma medida geral da capacidade de um classificador em fazer previsões certas. Seu cálculo se dá pela razão entre a soma dos verdadeiros positivos (TP - *True Positive*) e verdadeiros negativos (TN - *True Negative*) pelo total de amostras, incluindo também os falsos positivos (FP - *False Positive*) e falsos negativos (FN - *False Negative*).

$$Acuracia = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

No entanto, é importante notar que a acurácia pode ser uma métrica enganosa em cenários onde as classes são desbalanceadas, pois um modelo que prevê majoritariamente a classe mais comum ainda pode apresentar uma alta acurácia, mascarando um desempenho fraco para as classes minoritárias ([SOKOLOVA; LAPALME, 2009; THARWAT, 2021](#)).

- **Recall:** O *recall*, também chamado de sensibilidade ou taxa de verdadeiros positivos (TPR – *True Positive Rate*), avalia a capacidade de um modelo em identificar corretamente todas as instâncias pertencentes à classe positiva. Ele calcula a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos negativos.

$$Recall = \frac{TP}{TP + FN} \quad (5.2)$$

Em outras palavras, o *recall* mede quão bem o modelo consegue detectar todos os casos positivos reais presentes no conjunto de dados, sendo especialmente relevante em aplicações onde a detecção de todas as ocorrências da classe de interesse é prioritária, mesmo que isso leve a alguns falsos positivos. O *recall* é particularmente indicado para cenários com dados desbalanceados, pois considera apenas a coluna correspondente à classe positiva da matriz de confusão, reduzindo os efeitos de distribuição desigual entre classes ([SOKOLOVA; LAPALME, 2009; THARWAT, 2021](#)).

- **Precisão:** A precisão, ou valor preditivo positivo (PPV - *Positive Predictive Value*), indica a proporção de exemplos classificados como positivos que realmente pertencem à classe positiva ou seja, avalia a exatidão das previsões positivas de um classificador. Ela

é calculada como a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos positivos.

$$Preciso = \frac{TP}{TP + FP} \quad (5.3)$$

Em termos práticos, a precisão responde à pergunta: "Das instâncias que o modelo previu como positivas, quantas realmente eram positivas?". Essa métrica é essencial quando o custo de falsos positivos é alto, ou seja, quando classificar erroneamente uma instância negativa como positiva traz impactos significativos. No entanto, assim como outras métricas dependentes da matriz de confusão, a precisão pode ser afetada por distribuições desbalanceadas de classe, tornando necessário o uso combinado com o *recall* para uma avaliação mais equilibrada ([SOKOLOVA; LAPALME, 2009; THARWAT, 2021](#)).

- ***F1-score***: O *F1-score* é uma métrica que busca um equilíbrio entre a precisão e o *recall*, especialmente quando há *trade-offs* entre elas, sendo a média harmônica dessas duas medidas. Seu valor varia de 0 a 1, sendo mais elevado quando tanto a precisão quanto o *recall* são altos. Ele é particularmente útil em situações onde há um desequilíbrio entre as classes, pois considera tanto os falsos positivos quanto os falsos negativos. Um *F1-score* alto indica que o classificador possui tanto alta precisão (poucos falsos positivos) quanto alto *recall* (poucos falsos negativos), sendo uma medida mais robusta para avaliar o desempenho geral do modelo do que a acurácia em muitos contextos. Um aspecto importante destacado por [Sokolova e Lapalme \(2009\)](#) é que o *F1-score* foca na capacidade do modelo em identificar corretamente as instâncias da classe positiva, o que o torna particularmente relevante em tarefas de classificação com foco na recuperação de itens relevantes ([SOKOLOVA; LAPALME, 2009; THARWAT, 2021](#)).
- ***Receiver Operating Characteristic (ROC)***: A curva ROC é uma representação gráfica que avalia o desempenho de um classificador ao longo de diferentes limiares de decisão. Ela plota a taxa de verdadeiros positivos (sensibilidade) no eixo Y contra a taxa de falsos positivos (1 - especificidade) no eixo X, para todos os possíveis limiares de classificação, permitindo visualizar o *trade-off* entre detectar corretamente os positivos e evitar falsas detecções. Uma curva ROC que se aproxima do canto superior esquerdo do gráfico indica um melhor desempenho, pois representa uma alta taxa de verdadeiros positivos com uma baixa taxa de falsos positivos. Um dos principais benefícios da curva ROC é sua independência em relação à distribuição das classes, o que a torna útil mesmo em situações de grande desequilíbrio entre classes. Além disso, a Área Sob a Curva (AUC - *Area Under the Curve*) fornece uma medida resumida de desempenho, com valores mais próximos de

1 indicando maior capacidade discriminativa do modelo. A interpretação da ROC também permite comparar múltiplos classificadores de forma intuitiva, mesmo quando suas configurações de limiar são diferentes (FAWCETT, 2006; THARWAT, 2021).

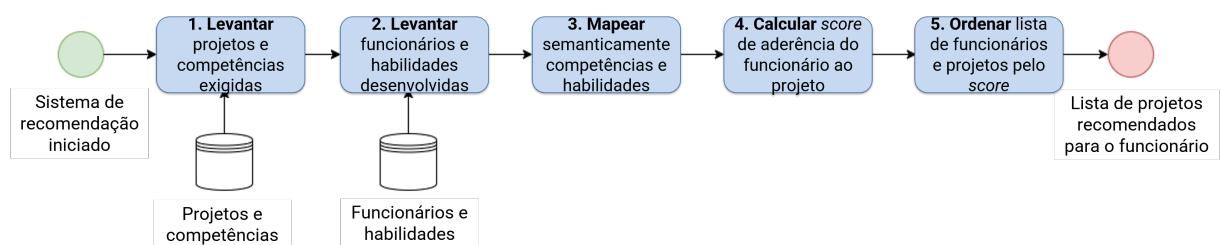
A definição dos parâmetros a serem utilizados no modelo de previsão de risco de desligamento voluntário é um grande desafio, uma vez que estes afetam diretamente o desempenho do modelo.

5.4 Modelo para Indicação de Projetos

Uma outra questão investigada do MSL foi a existência de ferramentas utilizadas para retenção de talentos em empresas de tecnologia. A partir dos estudos primários foi verificada a existência de 5 ferramentas comerciais com esta finalidade, apresentadas no Capítulo 3.

Já na RSL, foram identificados 3 estudos que apresentam propostas para retenção de talento. Entre as lacunas observadas, embora tenham sido encontrados estudos que propõem utilização de sistemas de recomendação para retenção de talentos, não encontramos estudos que buscassem associar os funcionários aos projetos corretos, de acordo com seu perfil. Esta proposta visa colocar a pessoa certa no projeto certo, promovendo, desta forma, um ambiente estimulante para os funcionários, por meio dos desafios dos projetos. A Figura 20 apresenta o processo de nossa proposta para retenção de talentos.

Figura 20: *Framework* proposto para retenção de talentos: Indicação do projeto certo para o funcionário



Fonte: Elaborado pela autora

A seguir serão descritas as fases do processo para detalhamento do *framework*:

- O fluxo inicia com (1) a coleta e a extração dos dados dos funcionários dos sistemas de gestão de pessoas e gestão de projetos da empresa. Também podem ser coletados dados

de pesquisas de interesse e entrevistas semi-estruturadas. Todos esses dados irão compor a coleção de documentos ou informações do funcionário, que pode seguir um formato descritivo ou semi-estruturado.

- Da mesma forma (2) são coletados ou extraídos dados de projeto existentes da empresa, nos quais são levantados os tipos de projeto, bem como as competências necessárias para na sua execução. Todos esses dados irão compor a coleção de documentos ou informações de projetos, que pode seguir um formato descritivo ou semi-estruturado.
- As competências e habilidades são mapeadas semanticamente (3). O objetivo nesta etapa é extrair tópicos significativos e distintos de forma supervisionada ou não supervisionada, utilizando parte de um conjunto de nomes de categorias fornecidos pela empresa. Essa tarefa ajuda a reduzir a dimensionalidade do problema, a entender de forma mais clara e distinta quais tópicos o funcionário estaria interessado, e também beneficia tarefas de classificação baseadas em palavras-chave. O algoritmo de recomendação é executado utilizando tanto a base de documentos de informações do funcionário, como os dados e informações do projeto.
- A próxima tarefa (4) é calcular o *score* de aderência do funcionário ao projeto usando os dados do funcionário e as informações dos projetos disponíveis na empresa. A combinação representa o alinhamento do interesse do funcionário e necessidades de desenvolvimento, com as competências necessárias nos projetos da empresa. O resultado dessa combinação é traduzido em um *score*.
- A lista de funcionários e projetos é ordenada de acordo com o *score* (5).
- Na saída do modelo temos a lista com os projetos recomendados para o funcionário de acordo com seu perfil.

5.5 Discussão

No MSL e na RSL foram verificados motivos para o desligamento voluntário de funcionários. Estes motivos se traduzem em atributos do sistema. A existência destes atributos depende das informações que o RH da empresa dispõe e de sua capacidade de geração e atualização de dados. Pode-se indicar a importância de determinados atributos para o modelo, mas não há garantias de disponibilidade de dados para compor este atributo.

Conforme apontado na RSL, verificamos que todos os estudos primários aceitos apresentam modelo para previsão de desligamento voluntário, com utilização dos mais diversos tipos

de classificadores e ferramentas de IA. Porém, observamos que algumas propostas verificam a influência de parâmetros tais como "satisfação no trabalho", "equilíbrio entre trabalho e vida pessoal", "suporte organizacional percebido", "gestão de talento" e "papel do gestor na decisão do funcionário de deixar a empresa", como os trabalhos de Živković, Šebalj e Franjković (2024), Monisaa Tharani e Vivek Raj (2020), Shen et al. (2023), Veglio, Romanello e Pedersen (2025), Seelam et al. (2022), Rajeswari et al. (2022) e Marvin, Jackson e Alam (2021), entre outros. Tais parâmetros qualitativos usualmente são medidos a partir de ferramentas como pesquisa de clima organizacional. Geralmente, os relatórios de retorno de pesquisas de clima associam fatores às áreas e não ao funcionário, visto que os fatores medidos não podem ser individualizados por questões de confidencialidade e proteção dos funcionários por meio de anonimato. A apresentação de resultados de pesquisa de clima de forma agrupada e anônima não é uma escolha, mas uma condição *sine qua non* para a pesquisa. A garantia de confidencialidade e anonimato é associada aos motivos de aderência do funcionário à pesquisa de clima e, consequente, resposta às questões da pesquisa (BISPO, 2006; CHIAVENATO, 2014). Por este motivo, tais atributos não são comumente encontrados em bases de RH e possivelmente não serão utilizados em modelos que trabalham com bases reais de dados de RH. Tais dados como "satisfação no trabalho" podem surgir de conversas de *feedback* 1:1 entre gestores e funcionários, mas geralmente não são cadastrados em sistemas de dados de RH. Também, podem aparecer em entrevistas de desligamento, mas, neste caso, a informação não estaria cadastrada no sistema antes do funcionário pedir desligamento, impossibilitando seu uso para previsão de desligamento voluntário.

A etapa de pré-processamento dos dados vai depender dos tipos de dados que estiverem disponíveis para treinamento dos classificadores, da qualidade e forma de apresentação destes dados. Portanto, esta etapa pode variar de acordo com as bases de dados disponibilizadas por cada empresa. Normalmente, as empresas costumam manter dados demográficos de seus funcionários, mas, ainda assim, pode haver variação da forma de armazenamento destes dados, visto que não há um padrão para todas as empresas. Alguns dados que podem estar em bases de RH dependerão dos processos internos da empresa, como utilização de ferramentas de avaliação de desempenho, por exemplo.

A etapa de balanceamento dos dados é importante para garantir um melhor desempenho do modelo. Embora seja possível o emprego de métodos de sobreamostragem (*oversampling*) e subamostragem (*undersampling*) de classe, como o *Synthetic Minority Over-sampling Technique* (SMOTE) (FERNÁNDEZ et al., 2018) ou *Edited Nearest Neighbor* (ENN) (BATISTA; PRATI; MONARD, 2004), tais métodos acarretam em imputação de dados sintéticos, com possível imputação de viés. Como a proposta está direcionada ao uso de dados sensíveis de RH, a

imputação de dados sintéticos deve ser evitada. Para tal, sugerimos a utilização de técnicas como *class-weight*.

A técnica de balanceamento de dados conhecida como *class-weight* é utilizada em aprendizado de máquina para lidar com o problema de desbalanceamento de classes, situação em que uma ou mais classes possuem significativamente menos amostras em comparação com as demais. Trata-se de uma estratégia de nível algorítmico, que altera a função de perda do modelo atribuindo pesos diferenciados a cada classe, com o objetivo de penalizar mais fortemente os erros nas classes minoritárias e, assim, equilibrar a influência de cada classe no processo de aprendizado. Ao contrário de métodos de reamostragem, como *oversampling* ou *undersampling*, o *class weight* não modifica o conjunto de dados original, o que evita riscos como *overfitting* (associado ao *oversampling*) ou perda de informação (relacionada ao *undersampling*). Essa abordagem tem sido empregada com sucesso em diversos algoritmos, incluindo regressão logística, SVM e *random forest*, nos quais os pesos podem ser diretamente incorporados à função de custo do algoritmo. Essa técnica é vantajosa por sua simplicidade e eficácia, especialmente em cenários onde a modificação do conjunto de dados não é desejável ([CARVALHO; PINHO; BRÁS, 2025](#)).

Considerando os requisitos encontrados no ciclo de relevância da DSR, baseados nos estudos primários aceitos na RSL, sugerimos a adoção de classificadores que tenham algoritmos de aprendizado supervisionado, que representem diferentes paradigmas de aprendizado de máquina. O desempenho dos classificadores treinados dependerá da quantidade, tipo e qualidade dos dados disponíveis. Por este motivo, para cada nova empresa ou ambiente de aplicação deste *framework* conceitual, será necessário novo treinamento dos diferentes classificadores para determinação do classificador mais adequado para utilização, considerando o desempenho. Sugerimos o uso de diferentes tipos de classificadores, tais como: árvore de decisão, *random forest*, *gradient boosting*, *XGBoost*, KNN, regressão logística e SVM.

Para treinamento dos classificadores, as etapas de preprocessamento, limpeza e normalização dos dados e treinamento de classes, além do tratamento de dados faltantes, são fundamentais. Uma vez que dados de RH são considerados dados sensíveis, assim como dados médicos, que envolvem leis de proteção de dados e possíveis questões trabalhistas, sugerimos adoção de técnicas para balanceamento de dados como *class-weight*. Adicionalmente, para tratamento de questões como dados faltantes, considerando casos com menos de 50% de dados faltantes, sugerimos a adoção de *Imputer KNN*, com k=1.

Para o bom funcionamentos dos *frameworks* propostos, é indispensável que a base de dados contenha atributos que não se limitem dados demográficos. Atributos como nota de avaliação

de desempenho, benefícios, tempo de empresa, posição na faixa salarial e dados de capacitações e certificação, e outros dados com atualização frequente, são relevantes para o bom funcionamento do modelo para previsão de desligamento voluntário.

Foi observado que nas propostas dos estudos identificados na RSL são feitas recomendações aos gestores a partir da identificação de risco de saída de um funcionário. As recomendações estão diretamente associadas ao tipo de risco identificado. Em nossa proposta, verificamos a existência dos projetos que são mais adequados ao perfil do funcionário e ranqueamos este projeto. Desta forma, é possível a verificação pelo gestor se os projetos nos quais os funcionários estão alocados são os mais adequados para seu perfil, ou até mesmo compatíveis com as aspirações de desenvolvimento do funcionário, uma vez que são utilizadas também informações de treinamentos e certificações. Assim, é possível mitigar o risco de desligamento voluntário, mantendo o funcionário em projetos que sejam estimulantes e desafiadores na medida. Para realização desta proposta de recomendação dos projetos, é fundamental que a base de dados de funcionários da empresa tenha informações sobre treinamentos e certificações realizados, ou de suas principais competências, assim como os projetos estejam associados às competências necessárias para sua execução.

A utilização e tratamento de dados sensíveis, como dados de RH, tornam imprescindível a atenção em relação às questões éticas e uso responsável da IA. Desta forma, o cuidado com a anonimização dos dados é fundamental. Por este motivo, para treinamento dos modelos, é necessário que os dados recebidos sejam anonimizados pelo RH e que o tratamento destes dados seja feito de forma a garantir sua anonimidade. Adicionalmente, faz-se necessária atenção especial ao acesso aos dados completos, que devem ser disponibilizado de forma que os gestores só tenham acesso aos dados de seus subordinados. Assim, é responsabilidade do RH, que tem acesso a todos os dados, informar quais dados devem ser visualizados por quais gestores.

6 Experimento

Nesta seção é apresentado um experimento, no qual é descrita a aplicação dos modelos conceituais propostos, com utilização de bases de dados disponibilizadas por uma empresa de serviços de TI. Este experimento integra a segunda etapa do Ciclo de *Design* da DSR, em que é feita a avaliação da abordagem proposta. Após desenvolvimento do modelo, esta etapa contempla a validação do modelo proposto, sendo alimentada pelos requisitos do ciclo de relevância e validada com base nos critérios definidos no ciclo de rigor. Este experimento constitui uma prova de conceito para o modelo conceitual proposto.

6.1 Metodologia

A Resolução nº 510/2016, do Ministério da Saúde dispõe sobre normas e diretrizes brasileiras aplicáveis a pesquisas cujos procedimentos metodológicos envolvam a utilização de dados diretamente obtidos com os participantes ou de informações identificáveis, ou, ainda, que possam acarretar riscos maiores do que os existentes na vida cotidiana. Esta resolução estabelece de forma clara que não serão registradas nem avaliadas pelo sistema CEP/CONEP (sistema dos Comitês de Ética em Pesquisa e da Comissão Nacional de Ética em Pesquisa) pesquisas com bancos de dados, cujas informações são agregadas, sem possibilidade de identificação individual ([BRASIL, 2016](#)). Por este motivo não submetemos projeto ao Comitê de Ética em Pesquisa (CEP). Não obstante, seguimos todas as recomendações da resolução em relação à confidencialidade dos dados e tratamento dos mesmos. Assim sendo, enfatizamos junto à empresa parceira a necessidade de recebimento das bases de dados anonimizadas e com os dados dispostos de forma a impedir a identificação dos funcionários.

A presente investigação busca prever a saída voluntária de funcionários a partir da aplicação de métodos quantitativos de ciência de dados sobre registros administrativos e históricos de recursos humanos, além de procurar compreender os fatores associados ao desligamento voluntário. O estudo adota uma abordagem sistemática, iniciando pela consolidação e padronização dos dados, seguida da engenharia de atributos e, por fim, da aplicação de algoritmos de

classificação para identificar padrões e fatores preditivos.

Para entendimento dos atributos e tratamento dos dados, reforçando sua anonimidade, foram realizadas reuniões com o RH da empresa que cedeu as bases de dados. As reuniões ocorreram em duas etapas, antes e depois do recebimento das bases de dados. Estas etapas foram fundamentais para o correto tratamento dos dados e entendimento do cenário da empresa. Além das reuniões, foi feito um tratamento inicial dos dados, com avaliação da matriz de correlação. Na sequência, foram aplicados os *frameworks* para previsão de desligamento voluntário e para retenção de talentos. A Figura 21 apresenta as etapas seguidas neste experimento.

Figura 21: Etapas para realização do experimento



Fonte: Elaborado pela autora

6.2 Base de dados

Os dados utilizados nesta tese são provenientes de 2 bases de dados, uma de funcionários e outra de projetos. As bases foram disponibilizadas por uma empresa de serviços de TI, que tem 3 filiais, sendo duas no estado do Rio de Janeiro e uma no estado de São Paulo. A empresa é classificada como empresa de grande porte, de acordo com anuário do DIEESE ([SERVIÇO BRASILEIRO DE APOIO ÀS MICRO E PEQUENAS EMPRESAS; DEPARTAMENTO INTERSINDICAL DE ESTATÍSTICA E ESTUDOS SOCIOECONÔMICOS, 2020](#)). Ao disponibilizar as bases, a empresa autorizou seu uso para esta pesquisa. Todos os cuidados para tratamento, anonimização e não exposição dos dados indicados pelo CEP foram tomados.

As bases de dados foram disponibilizadas totalmente anonimizadas, de forma a atender à Lei Geral de Proteção de Dados (LGPD) ([BRASIL, 2018](#)) e aos preceitos éticos de pesquisa. Para garantir a anonimidade dos dados, foi solicitado à empresa que um rigoroso processo fosse seguido antes de sua disponibilização. Informações como, por exemplo, “salários” foram

alteradas para “faixas salariais”, “idade” para “faixa de idade”. Adicionalmente foram criadas chaves criptografadas para cada funcionário e para cada projeto.

A base de dados de funcionários reúne informações de diversos sistemas da empresa, além de informações imputadas manualmente, contendo registros administrativos e históricos de recursos humanos, totalizando 302 linhas com registros únicos e 42 duas colunas de atributos. A base contém informações tanto de funcionários ativos quanto de funcionários que não fazem mais parte do quadro da empresa, apresentando data de contratação e data de desligamento. Entre as informações imputadas manualmente está o motivo do desligamento, coletadas em entrevistas de desligamento.

A base com dados de projeto apresenta as principais competências necessárias para atuação nos projetos. Contém 227 linhas e 25 colunas, registrando até 21 competências por projeto, contendo chave única de identificação do projeto.

6.2.1 Matriz de correlação

A matriz de correlação é uma ferramenta estatística que expressa a intensidade e direção da relação linear entre variáveis numéricas. Cada valor da matriz varia entre -1 e +1 ([SCHOBER; BOER; SCHWARTE, 2018](#); [ASLAM; RAMLI; SHAH, 2019](#)):

- **+1** indica correlação perfeitamente positiva: quando uma variável aumenta, a outra também aumenta.
- **-1** indica correlação perfeitamente negativa: quando uma variável aumenta, a outra diminui.
- **0** significa ausência de correlação linear.

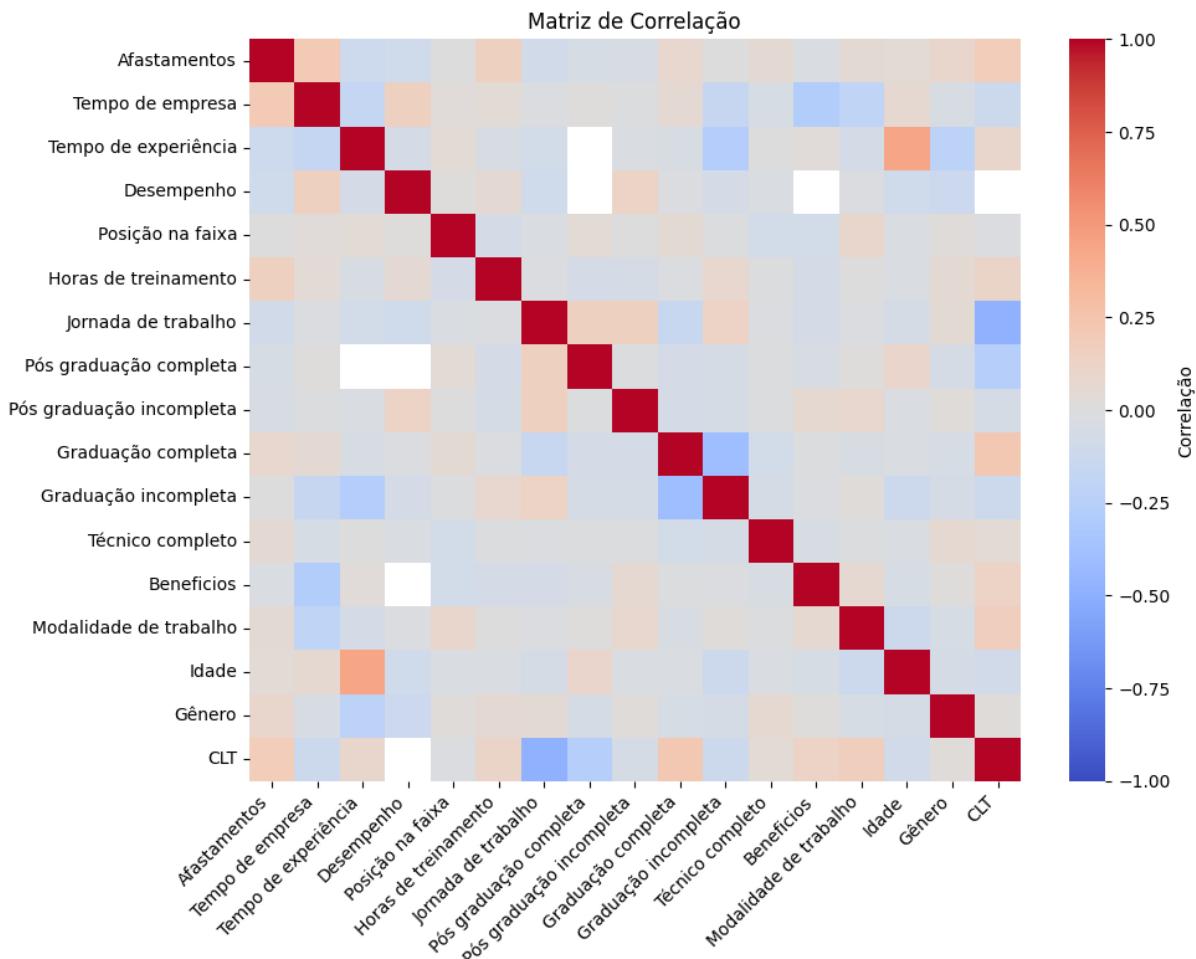
Essa análise é útil para: (1) identificar redundâncias entre variáveis (multicolinearidade); (2) descobrir padrões subjacentes; (3) selecionar variáveis para modelagem preditiva (*feature selection*).

Durante o processo de análise exploratória dos dados, foi realizada a construção de uma matriz de correlação com o objetivo de identificar relações lineares entre as variáveis independentes. Essa matriz revelou elevadas correlações entre algumas variáveis categóricas transformadas em binárias, o que poderia introduzir redundância e multicolinearidade no modelo preditivo, comprometendo sua interpretabilidade e robustez.

Avaliando a matriz de correlação, removemos variáveis altamente correlacionadas, como *Seguro saúde* e *Seguro de Vida*, que apresentavam correlações próximas de 1 entre si, sugerindo sobreposição de informação. A variável *Não*, por sua vez, era redundante por representar ausência de benefícios, cuja informação já estava implicitamente capturada pelas variáveis de presença de benefício, e portanto foi descartada para evitar viés artificial no modelo.

Além disso, alguns campos relacionados a benefícios corporativos apresentavam forte correlação positiva entre si (por exemplo, correlação de 0,63 entre *Folga Anual*, *Reembolso de Certificações*, *Programa de Indicações*, *PLR*, *plano odontológico*), indicando que esses pacotes tendem a ser oferecidos conjuntamente. Para reduzir a dimensionalidade sem perder o conteúdo informativo, optou-se por agregá-los em uma variável representativa única, renomeando a coluna *Ticket* como *Benefício*, de forma a consolidar o pacote de vantagens oferecido ao funcionário. A Figura 22 apresenta a matriz de correlação dos atributos presentes na base de funcionários, após tratamento inicial com combinação das variáveis identificadas com alta correlação.

Figura 22: Matriz de Correlação dos atributos



Fonte: Elaborado pela autora

6.2.2 Descrição dos atributos das bases de dados

Após tratamento dos dados a partir da avaliação da matriz de correlação, a base de dados de funcionários ficou com 19 atributos. A Tabela 5 mostra a relação de atributos da base de dados de funcionários, a descrição e o respectivo domínio numérico.

Tabela 5: Descrição dos Atributos da Base de Funcionários

Atributo	Descrição	Domínio
Pessoa	Identificador hexadecimal único representando o funcionário.	Exemplo: xyMhZ05n0F/Zc
Afastamentos (dias/ano)	Número de afastamentos por ano.	Numérico: 0 – 24
Tempo de empresa	Tempo de empresas em meses.	Numérico: 0 – 257
Tempo de experiência	Tempo de experiência em anos.	Numérico: 0 – 26
Benefícios	Ticket, seguro saúde, seguro de vida, plano odontológico, PLR, folga anual, reembolso de certificações, programa de indicações.	Binário para cada tipo de benefício
Desempenho	Nota de avaliação do desempenho em 2023.	Numérico: 2.2 – 4.94
Posição na faixa	Posição do funcionário na faixa: -1 significa no início da faixa, 0 meio e 1 final da faixa.	Numérico: -1 – 1
Horas de treinamento	Horas de treinamento.	Numérico: 0 – 2000
Jornada de trabalho	Jornada de 20 horas, 30 horas ou 40 horas semanais.	0 ou 1 para cada tipo de jornada
Pós-graduação completa	Formação nível doutorado, mestrado.	0 ou 1 para cada nível de pós-graduação
Pós-graduação incompleta	Formação nível doutorado, mestrado incompleta.	0 ou 1 para cada nível
Graduação completa	Educação superior completa. Valor 1 representa indivíduos com graduação completa; 0, caso contrário.	0 ou 1

Continua na próxima página...

Tabela 5: Descrição dos Atributos da Base de Funcionários

Atributo	Descrição	Domínio
Graduação incompleta	Educação superior incompleta. Valor 1 representa indivíduos com graduação incompleta; 0, caso contrário.	0 ou 1
Ensino médio completo	Ensino médio completo. Valor 1 representa indivíduos com ensino médio completo; 0, caso contrário.	0 ou 1
Ensino médio incompleto	Ensino médio completo. Valor 1 representa indivíduos com ensino médio incompleto; 0, caso contrário.	0 ou 1
Modalidade de trabalho	A distância ou híbrido.	0 ou 1 para cada tipo de modalidade de trabalho.
Faixa etária	entre 18 e 29 anos, entre 30 e 39 anos, entre 40 e 49 anos, 50 anos ou mais.	0 ou 1 para cada faixa
Gênero	Feminino ou masculino.	0 ou 1 para cada gênero
Contrato de trabalho	CLT, ou contrato por tempo determinado, ou estágio, ou pessoa jurídica.	0 ou 1 para cada tipo de contrato de trabalho

A Tabela 6 mostra a relação de atributos da base de dados de projeto, a descrição e o respectivo domínio numérico.

Tabela 6: Descrição dos Atributos da Base de Projetos

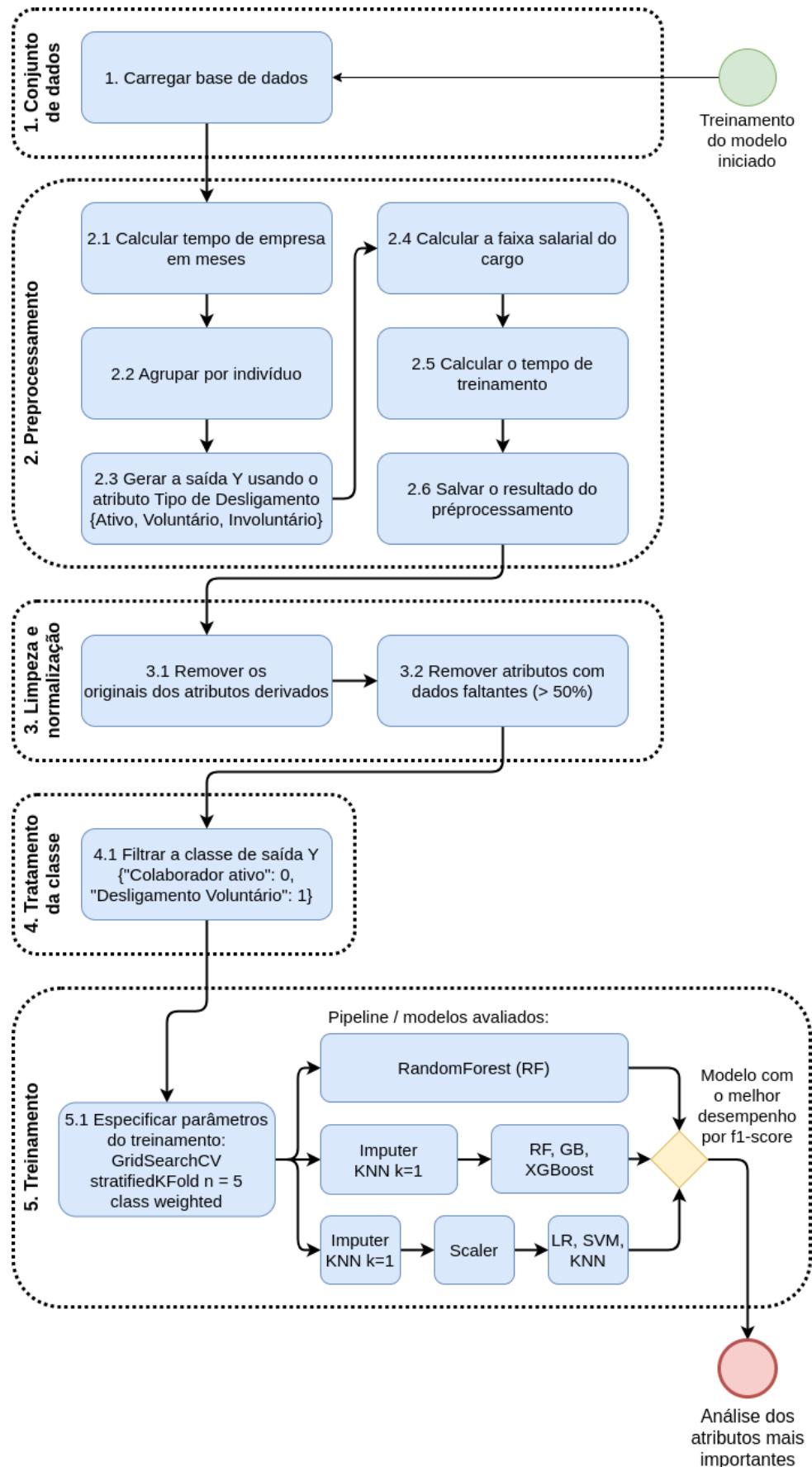
Atributo	Descrição	Domínio
Projeto	Identificador hexadecimal único representando o projeto	Exemplo: xyZvsMphBN.Ns
Pessoas	Identificadores hexadecimais representando os funcionários alocados naquele projeto, separados por vírgula.	Exemplo: xyMhZ05n0F/Zc, xyGM5Xx2JY90Q.
Primeira Alocação Projeto	Data da primeira alocação no projeto	Data
Última Alocação Projeto	Data da última alocação no projeto	Data
Competências 1 – 21	Designação das competências exigidas no projeto	Exemplo: Arquitetura de sistemas, Segurança, Desenvolvimento Mobile, Gerenciamento de teste, issue tracking, integração contínua, container, banco de dados, APIs...

6.3 Modelo para previsão de risco de desligamento voluntário

6.3.1 Construção do modelo

A Figura 23 apresenta o *framework* proposto de forma detalhada, com o processo de aprendizado de máquina para previsão de risco de desligamento voluntário. Cada etapa do processo será especificada a seguir.

Figura 23: Processo de aprendizado de máquina para predição de risco de desligamento voluntário



Fonte: Elaborado pela autora

1. Conjunto de dados

- **1.1. Carregar bases de dados:** A primeira etapa ao iniciar o treinamento do modelo é carregar base de dados.

2. Preprocessamento

- **2.1. Calcular tempo de empresa em meses:** A primeira etapa envolveu a padronização de registros temporais, especialmente datas de contratação e desligamento, com o objetivo de calcular o tempo de permanência na organização. A abordagem adotada priorizou a transformação dessas informações em unidades mensuráveis e comparáveis (meses), permitindo análises longitudinais coerentes.
- **2.2. Agrupar por indivíduo:** Concomitantemente, foi realizada a consolidação de múltiplos registros associados a uma mesma identidade funcional. Essa etapa implicou o agrupamento de atributos por funcionário, com concatenação de valores distintos em campos categóricos, e retenção da primeira ocorrência para atributos estáveis. Tal estratégia permitiu a construção de um perfil único e integrado por funcionário, etapa fundamental para reduzir viés por duplicidade e garantir coesão na modelagem preditiva.
- **2.3. Gerar saídas Y usando o atributo Tipo de Desligamento:** A seguir, aplicou-se uma reconstrução semântica de variáveis categóricas, com destaque para o tipo de desligamento. Por meio de mapeamentos supervisionados, converteu-se a nomenclatura original em rótulos (ativo, voluntário e involuntário). Este refinamento semântico permitiu a definição clara da variável alvo, fundamental para tarefas de classificação supervisionada.
- **2.4. Calcular faixa salarial do cargo:** Na dimensão salarial, foi necessário lidar com variações de valores de faixas salariais para o mesmo cargo. Foi adotado uma faixa salarial única por cargo, pegando o menor e maior valores das faixas existentes.
- **2.5. Calcular o tempo de treinamento:** No caso do funcionário apresentar mais de um treinamento, os valores de tempo dos diversos treinamentos são somados. Adicionalmente, estipulou-se um valor máximo para tempo de treinamento, uma vez que alguns valores cadastrados ultrapassavam duas mil horas de treinamento.
- **2.6. Salvar o resultado do preprocessamento:** Após a realização de todos os passos da etapa de preprocessamentos, os dados são salvos em nova planilha.

3. Limpeza e normalização

- **3.1. Remover os originais dos atributos derivados:** Após preprocessamento, é realizada limpeza removendo os atributos substituídos.
- **3.2. Remover atributos com dados faltantes (>50%):** Todos os atributos com mais de 50% de dados faltantes foram removidos.

4. Tratamento de classe

- **4.1. Filtrar a classe de saída Y:** Definido filtro de classe com “funcionário ativo” 0 e “Desligamento voluntário” 1.

5. Treinamento

- **5.1. Especificar parâmetros de treinamento:** Nesta etapa foram especificados os parâmetros *GridSearchCV*, *stratifiedKFold* e *class weighted*. Na sequência foi realizado treinamento e avaliação dos modelos *random forest*; *random forest*, *Gradient Boosting* e *XGBoost* com Imputer KNN (k=1); regressão logística, SVM e KNN com Imputer KNN (k=1).
- Após treinamento de todos os modelos, verificamos o classificador que apresenta melhor desempenho é o *random forest*, com base no *F1-score*.

Na etapa de treinamento, o conjunto de dados foi dividido em duas partes: 70% para o treinamento dos modelos e 30% para teste, com o objetivo de avaliar o desempenho preditivo em dados não vistos anteriormente. A divisão foi feita de forma estratificada, ou seja, mantendo a proporção entre as classes da variável alvo tanto no conjunto de treino quanto no conjunto de teste. Essa abordagem é especialmente importante em contextos de desbalanceamento de classes, como é o caso deste estudo, pois evita que o modelo seja treinado ou avaliado com uma distribuição artificial de exemplos.

Para a seleção dos melhores hiperparâmetros de cada modelo, foi utilizada a técnica de validação cruzada estratificada com 5 dobras (*StratifiedKFold*). Essa técnica divide o conjunto de treinamento em cinco subconjuntos diferentes, garantindo que cada modelo seja treinado e validado em diferentes partições dos dados. Isso reduz a variabilidade nos resultados e aumenta a robustez da avaliação. A busca pelos melhores hiperparâmetros foi feita por meio do *GridSearchCV*, que realiza uma busca exaustiva por todas as combinações possíveis dos valores definidos para cada parâmetro. Essa abordagem permite identificar a configuração mais eficiente para cada algoritmo, com base em métricas de desempenho consistentes ao longo das dobras da validação cruzada.

Além disso, foi utilizado o parâmetro *class_weight=balanced* nos modelos que suportam essa funcionalidade, como Random Forest, Regressão Logística e SVM. Essa técnica ajusta os pesos das classes de forma inversamente proporcional à sua frequência, permitindo que o modelo dê a devida atenção à classe minoritária, que normalmente tende a ser sub-representada no aprendizado supervisionado. Isso é fundamental para evitar viés de predição para a classe majoritária, contribuindo para um melhor equilíbrio entre precisão e recall.

Foram treinados diversos modelos com e sem imputação de dados via KNN ($k=1$), incluindo *Random Forest*, *Gradient Boosting*, XGBoost, Regressão Logística, SVM e o próprio KNN como classificador (TROYANSKAYA et al., 2001). A escolha desses algoritmos buscou representar diferentes paradigmas de aprendizado de máquina, como modelos de árvore, modelos lineares, métodos de margem máxima e métodos baseados em distância, permitindo uma análise comparativa ampla sobre qual técnica se adequa melhor aos dados disponíveis.

Por fim, o desempenho de cada modelo foi avaliado com base no F1 score, que oferece uma média balanceada do desempenho entre as classes, sendo particularmente apropriada em situações de desbalanceamento. O modelo que apresentou melhor desempenho foi selecionado para análise de importância de variáveis e aplicação final no conjunto de teste. Essa abordagem estruturada e justificada permitiu não apenas otimizar os resultados preditivos, mas também garantir interpretabilidade e validade dos achados.

6.3.2 Resultados

Nesta seção são apresentados os resultados da aplicação do *framework*, com utilização da base de dados disponibilizada pela empresa de serviços de TI, e sua respectiva avaliação.

6.3.2.1 Resultados dos treinamentos dos modelos

A. *Random Forest* (RF)

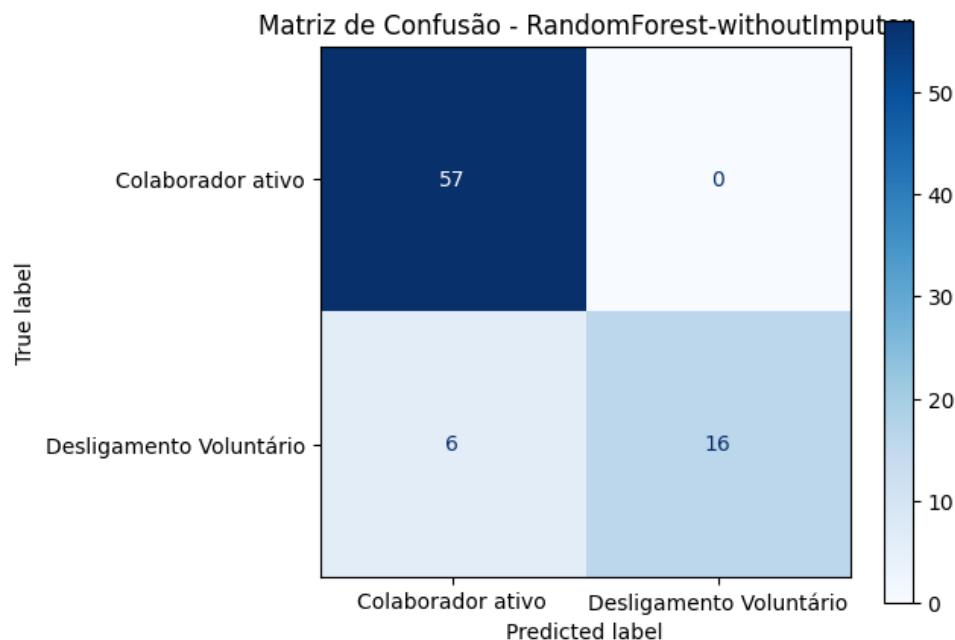
O *Random Forest* é um método *ensemble* baseado em árvores de decisão que utiliza o método de *bagging* para construir múltiplas árvores treinadas sobre subconjuntos dos dados e combina suas previsões por votação majoritária. A configuração ótima encontrada envolveu 100 árvores, critério de impureza gini e seleção de até 12 atributos por divisão. Com essa configuração, o modelo alcançou acurácia geral de 92% e um F1 score de 0,90, indicando bom equilíbrio entre as classes. Destaca-se também sua precisão de 1,00 para a classe minoritária, com um *recall* de 0,73, evidenciando uma leve tendência a falsos negativos, porém ainda superior em comparação aos demais modelos.

A seguir apresentamos a tabela de resultados (Tabela 7), a matriz confusão, apresentada na Figura 24, e a curva ROC apresentada na Figura 25.

Tabela 7: Relatório de classificação do *Random Forest*

Classe	Precision	Recall	F1-score	Support
0.0	0.90	1.00	0.95	57
1.0	1.00	0.73	0.84	22
Acurácia	—	—	0.92	79
Macro avg	0.95	0.86	0.90	79
Weighted avg	0.93	0.92	0.92	79

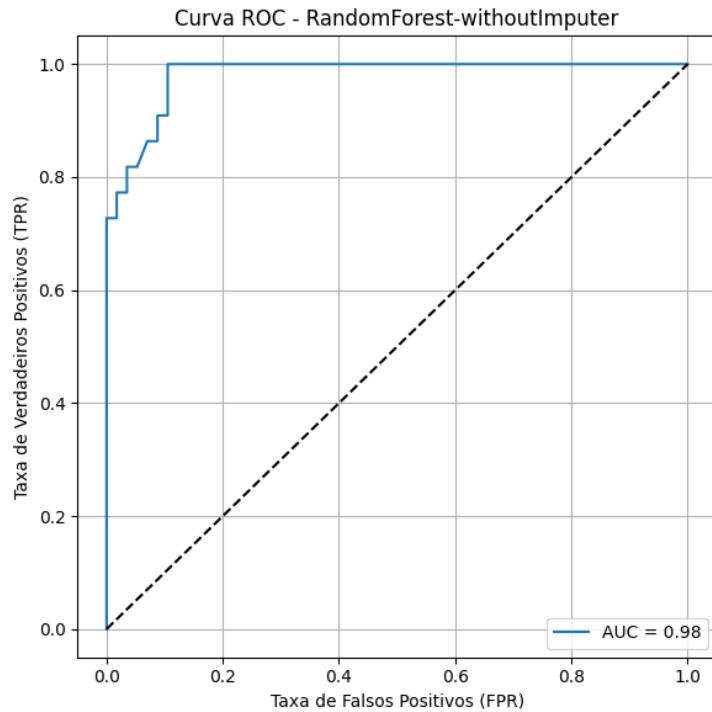
Figura 24: Matriz de Confusão para classificador *Random Forest*



Fonte: Elaborado pela autora

A matriz de confusão é uma ferramenta estatística fundamental para avaliar o desempenho de modelos de classificação, tanto em problemas binários quanto multiclasse. Ela organiza os resultados em uma tabela $N \times N$, no qual N representa o número de classes, permitindo identificar de forma clara os acertos e erros de classificação de um modelo. No caso de classificações binárias, a matriz é composta por quatro categorias principais: Verdadeiros Positivos (*True Positive* (TP)), Falsos Positivos (*False Positive* (FP)), Verdadeiros Negativos (*True Negative* (TN)) e Falsos Negativos (*False Negative* (FN)), o que permite visualizar quantas vezes o modelo acertou ou errou cada classe. A partir dessas informações, é possível derivar diversas métricas de desempenho, como acurácia, sensibilidade (*recall*), precisão, *F1-score*, entre outras (THARWAT, 2021).

Figura 25: Curva ROC para classificador *Random Forest*



Fonte: Elaborado pela autora

B. Random Forest (RF) com Imputer

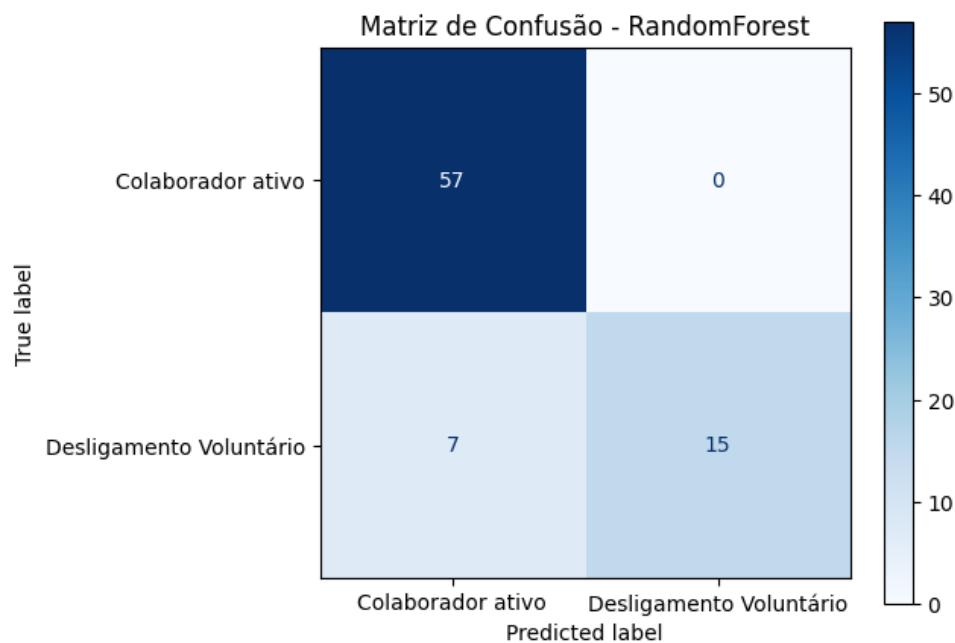
O *Random Forest* com imputação apresentou desempenho ligeiramente inferior, com acurácia de 91% e F1 score de 0,88. Embora mantenha excelente precisão na classe minoritária (1,00), seu *recall* foi de 0,68, indicando uma tendência mais forte a deixar de identificar casos dessa classe. Isso sugere que o processo de imputação pode ter interferido na representação de padrões relevantes para os casos menos frequentes.

Na sequência, apresentamos a tabela de resultados do treinamento com classificador RF com *Imputer* (Tabela 8), matriz confusão, ilustrada na Figura 26 e a curva ROC ilustrada na Figura 27.

Tabela 8: Relatório de classificação do *Random Forest* com *imputer*

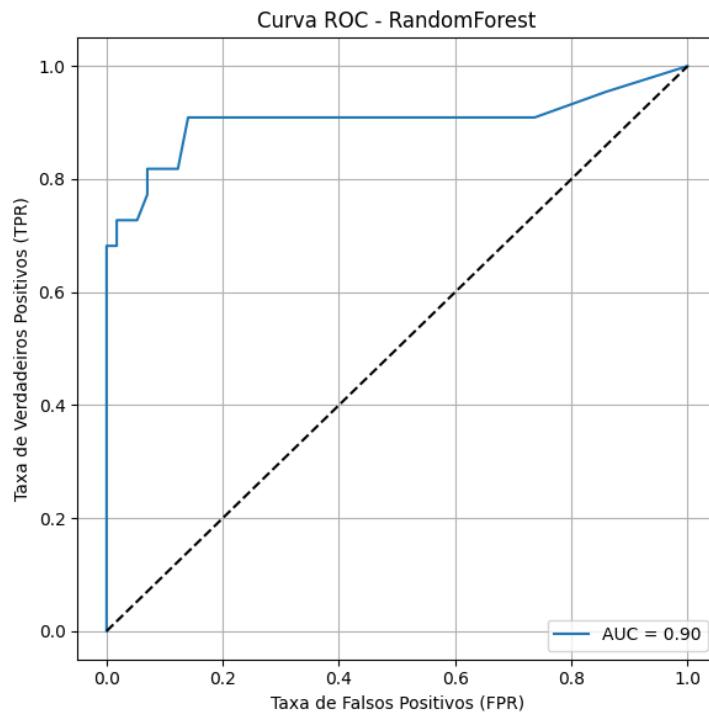
Classe	Precision	Recall	F1-score	Support
0.0	0.89	1.00	0.94	57
1.0	1.00	0.68	0.81	22
Acurácia	–	–	0.91	79
Macro avg	0.95	0.84	0.88	79
Weighted avg	0.92	0.91	0.91	79

Figura 26: Matriz de Confusão para classificador *Random Forest* com *Imputer*



Fonte: Elaborado pela autora

Figura 27: Curva ROC para classificador *Random Forest* com *Imputer*



Fonte: Elaborado pela autora

C. Logistic Regression (LR)

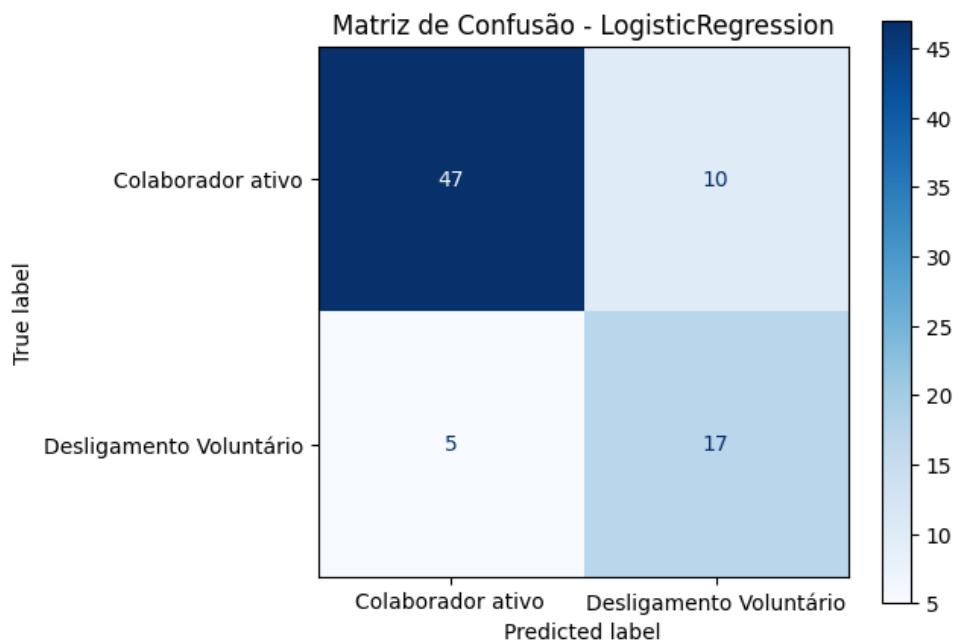
O *Logistic Regression*, por sua vez, apresentou acurácia de 81% e F1 score de 0,78, evidenciando limitações em capturar padrões não lineares. A precisão de 0,63 e o *recall* de 0,77 na classe minoritária mostram que o modelo errou mais na identificação correta dos casos dessa classe, ainda que tenha sido razoavelmente sensível.

A partir do treinamento com classificador LR foi obtida a seguinte a tabela de resultados do treinamento (Tabela 9), a matriz confusão, apresentada na Figura 28, e a curva ROC apresentada na Figura 29.

Tabela 9: Relatório de classificação do *Logistic Regression*

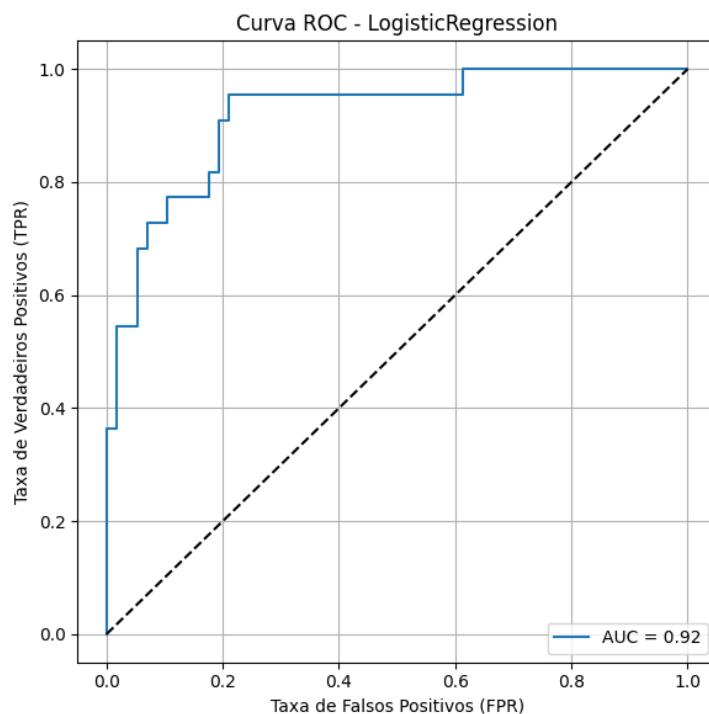
Classe	Precision	Recall	F1-score	Support
0.0	0.90	0.82	0.86	57
1.0	0.63	0.77	0.69	22
Acurácia	—	—	0.81	79
Macro avg	0.77	0.80	0.78	79
Weighted avg	0.83	0.81	0.82	79

Figura 28: Matriz de Confusão para classificador *Logistic Regression*



Fonte: Elaborado pela autora

Figura 29: Curva ROC para classificador *Logistic Regression*



Fonte: Elaborado pela autora

D. Support Vector Machine (SVM)

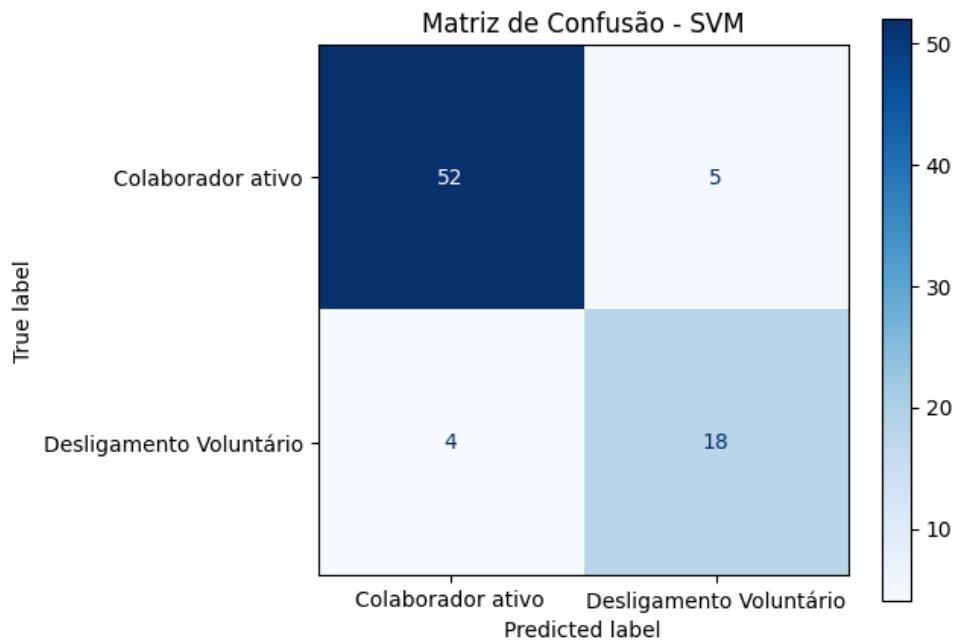
O modelo SVM, com *kernel* linear, mostrou-se competitivo, com acurácia de 89% e F1 score de 0,86. Apesar de ligeiramente inferior aos modelos de árvore, o SVM apresentou melhor equilíbrio entre precisão (0,78) e recall (0,82) na classe minoritária, o que pode ser vantajoso em aplicações que valorizam a redução simultânea de falsos positivos e falsos negativos.

Apresentamos a seguir a Tabela 10, com os resultados do treinamento do classificador SVM, a matriz confusão, apresentada na Figura 30, e a curva ROC apresentada na Figura 31.

Tabela 10: Relatório de classificação do *Support Vector Machine*

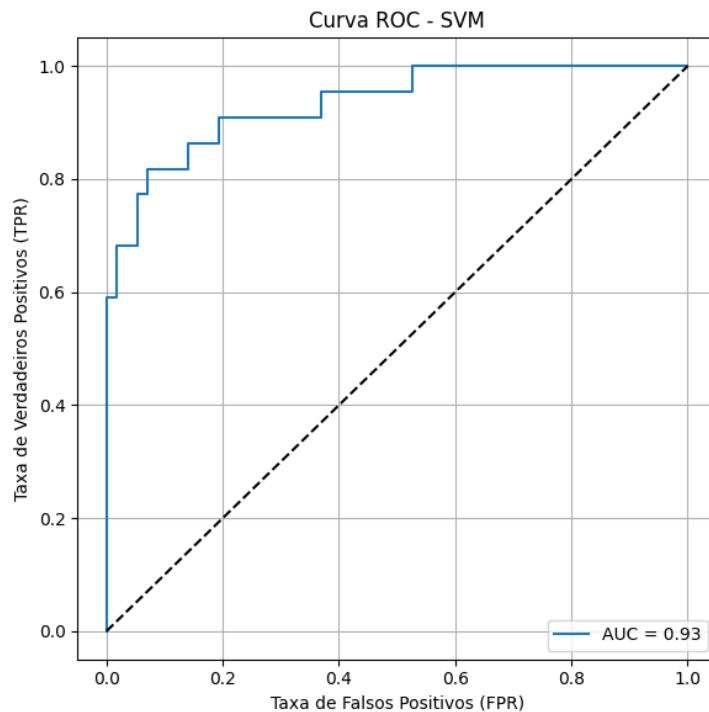
Classe	Precision	Recall	F1-score	Support
0.0	0.93	0.91	0.92	57
1.0	0.78	0.82	0.80	22
Acurácia	–	–	0.89	79
Macro avg	0.86	0.87	0.86	79
Weighted avg	0.89	0.89	0.89	79

Figura 30: Matriz de Confusão para classificador SVM



Fonte: Elaborado pela autora

Figura 31: Curva ROC para classificador SVM



Fonte: Elaborado pela autora

E. *k-Nearest Neighbor* (KNN)

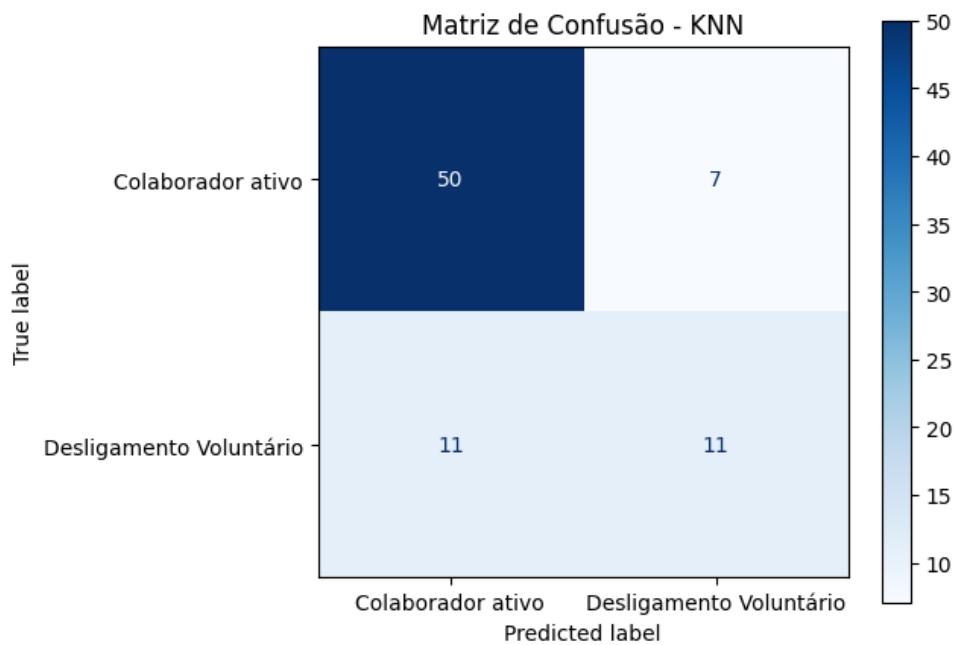
O classificador KNN apresentou o pior desempenho geral, com acurácia de 77% e F1 score de 0,70, sendo particularmente prejudicado pelo baixo *recall* (0,50) na classe minoritária. Isso evidencia que o KNN teve dificuldades em identificar corretamente os exemplos menos representados, provavelmente por sua sensibilidade a ruídos e à distribuição dos dados no espaço amostral.

A partir do treinamento com classificador KNN obtivemos a Tabela 11, com os resultados do treinamento do classificador, a matriz confusão, ilustrada na Figura 32, e a curva ROC apresentada na Figura 33.

Tabela 11: Relatório de classificação do KNN

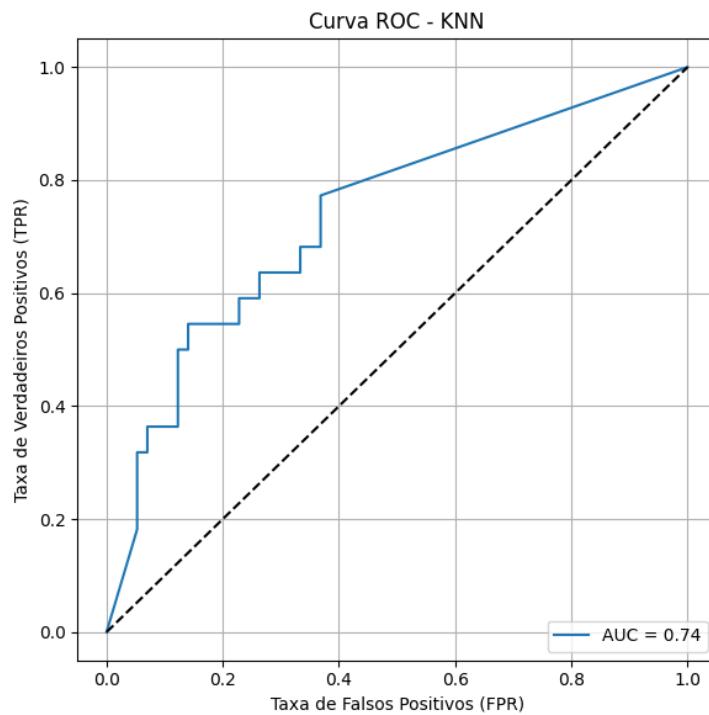
Classe	Precision	Recall	F1-score	Support
0.0	0.82	0.88	0.85	57
1.0	0.61	0.50	0.55	22
Acurácia	–	–	0.77	79
Macro avg	0.72	0.69	0.70	79
Weighted avg	0.76	0.77	0.76	79

Figura 32: Matriz de Confusão para classificador KNN



Fonte: Elaborado pela autora

Figura 33: Curva ROC para classificador KNN



Fonte: Elaborado pela autora

F. Gradient Boosting (GB)

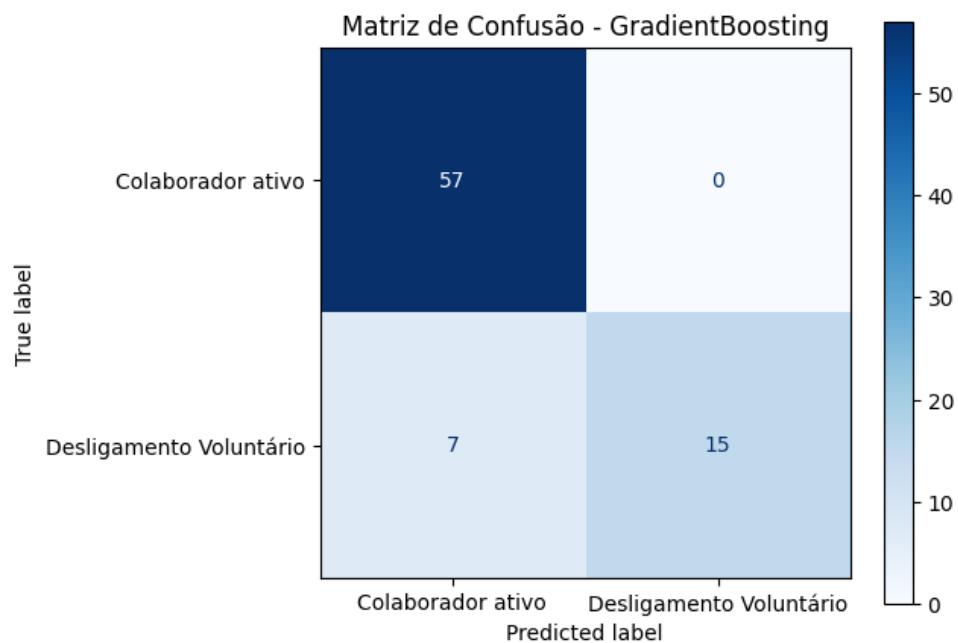
O *Gradient Boosting*, outro modelo baseado em *ensemble* de árvores além do *Random Forest*, alcançou acurácia de 91% e F1 score de 0,88, com desempenho equilibrado entre precisão e *recall*, reforçando a eficácia de abordagens baseadas em *boosting* na modelagem de relações não lineares.

Na sequência, apresentamos a Tabela 12, com os resultados do treinamento com classificador GB, assim como a matriz confusão, apresentada na Figura 34, e a curva ROC apresentada na Figura 35.

Tabela 12: Relatório de classificação do *Gradient Boosting*

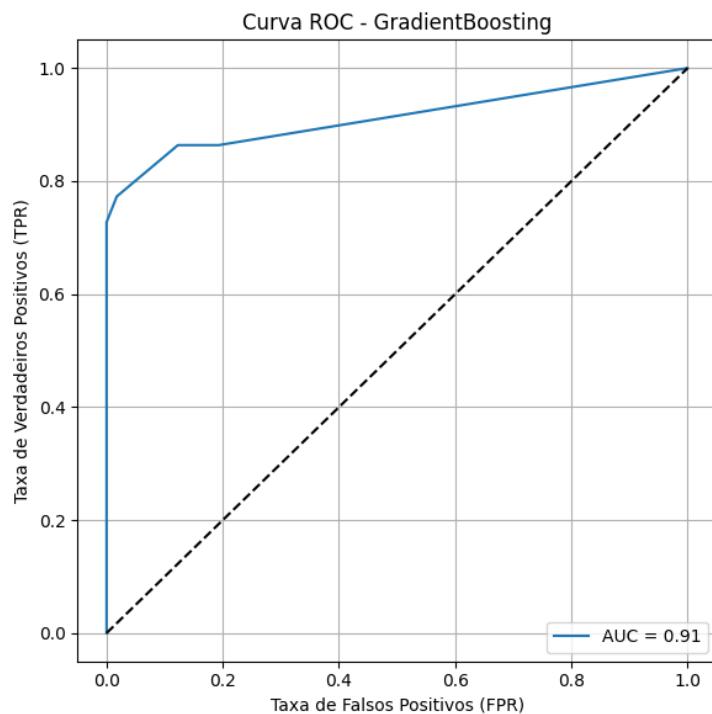
Classe	Precision	Recall	F1-score	Support
0.0	0.89	1.00	0.94	57
1.0	1.00	0.68	0.81	22
Acurácia	—	—	0.91	79
Macro avg	0.95	0.84	0.88	79
Weighted avg	0.92	0.91	0.91	79

Figura 34: Matriz de Confusão para classificador *Gradient Boosting*



Fonte: Elaborado pela autora

Figura 35: Curva ROC para classificador *Gradient Boosting*



Fonte: Elaborado pela autora

G. Extreme-Gradient Boosting (XGBoost)

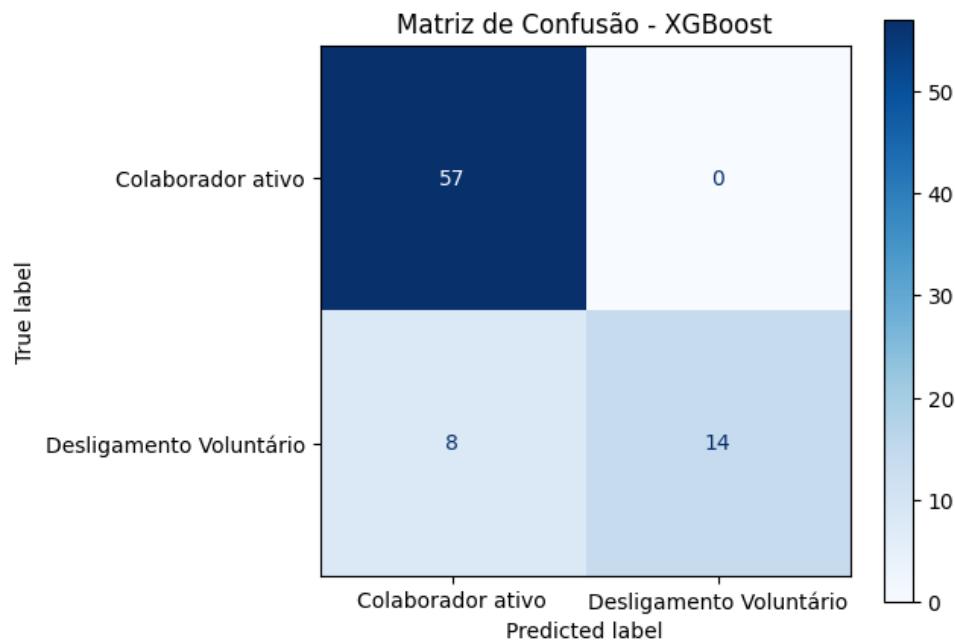
O XGBoost obteve acurácia de 90% e F1 score de 0,86, destacando-se pela alta precisão nas duas classes, mas penalizado por um *recall* de apenas 0,64 na classe minoritária, o que pode indicar uma maior rigidez nos critérios de decisão.

A partir do treinamento com classificador XGBoost foi obtida a Tabela 13, com o resultado do treinamento realizado, a matriz confusão, exibida na Figura 36, e a curva ROC apresentada na Figura 37.

Tabela 13: Relatório de classificação do XGBoost

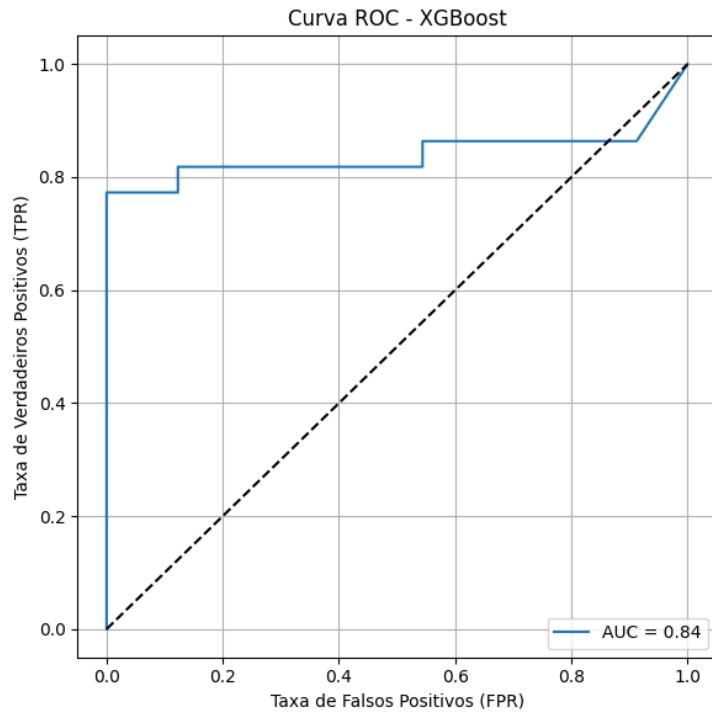
Classe	Precision	Recall	F1-score	Support
0.0	0.88	1.00	0.93	57
1.0	1.00	0.64	0.78	22
Acurácia	–	–	0.90	79
Macro avg	0.94	0.82	0.86	79
Weighted avg	0.91	0.90	0.89	79

Figura 36: Matriz de Confusão para classificador XGBoost



Fonte: Elaborado pela autora

Figura 37: Curva ROC para classificador XGBoost



Fonte: Elaborado pela autora

6.3.2.2 Comparação entre classificadores

A Tabela 14 apresenta a comparação entre os classificadores treinados. Constatamos que o *Random Forest* foi o classificador que apresentou melhor desempenho. Os resultados reforçam a superioridade dos modelos baseados em *ensembles* de árvores, especialmente o *Random Forest* sem imputação, que não apenas obteve os melhores valores globais de acurácia e F1 score, mas também demonstrou robustez na detecção da classe minoritária sem comprometer a precisão geral. Esse equilíbrio entre sensibilidade e especificidade, aliado à capacidade de lidar com conjuntos de dados possivelmente incompletos, justifica sua escolha como o modelo mais adequado para o problema em análise. A Tabela 15 apresenta a descrição de cada hiperparâmetro utilizado nos classificadores.

Tabela 14: Comparação de Modelos Classificadores

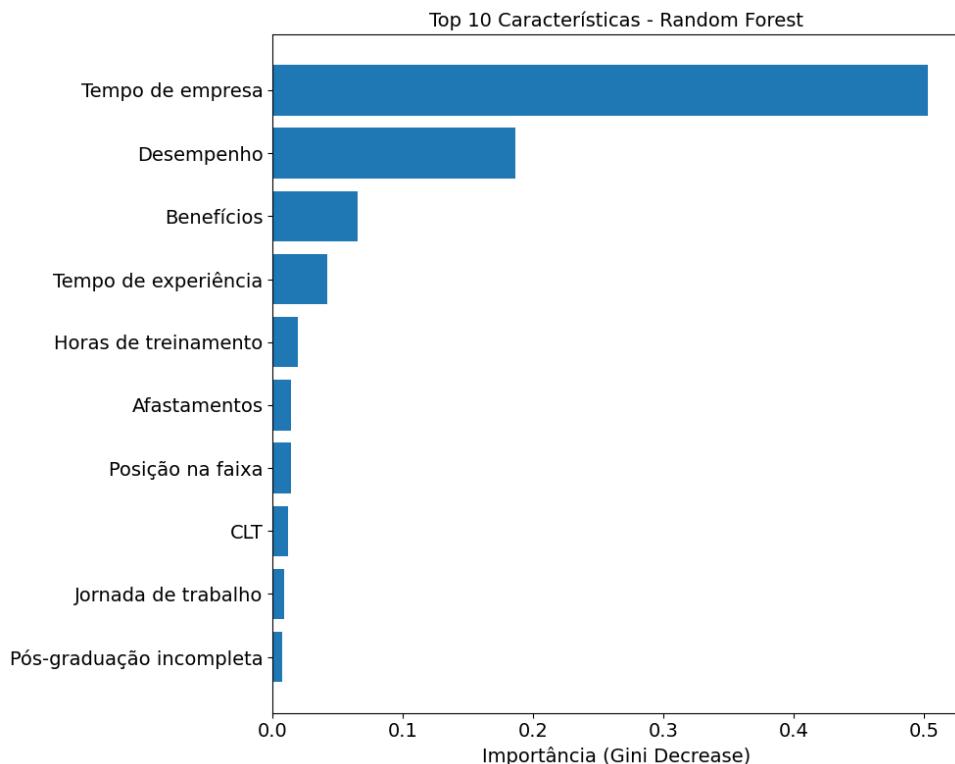
Modelo	Melhores Hiperparâmetros	Acurácia	F1 Macro
RF	{n_estimators: 100, max_depth: None}	0.92	0.90
RF com Imputer	{n_estimators: 100, max_depth: None}	0.91	0.88
LR	{C: 10.0, solver: lbfgs}	0.81	0.78
SVM	{C: 1.0, gamma: scale}	0.89	0.86
KNN	{n_neighbors: 3, p: 2}	0.77	0.70
GB	{n_estimators: 100, learning_rate: 0.01, max_depth: 3}	0.91	0.88
XGBoost	{n_estimators: 100, learning_rate: 0.01, max_depth: 3}	0.90	0.86

Tabela 15: Descrição dos Hiperparâmetros por Modelo Classificador

Modelo	Hiperparâmetro	Descrição
RF RF com Imputer	n_estimators	Número de árvores na floresta. Maior valor pode melhorar o desempenho, mas aumenta o custo computacional.
	max_depth	Profundidade máxima das árvores. None permite crescimento total até as folhas puras.
LR	C	Inverso da força de regularização. Valores maiores significam menos regularização.
	solver	Algoritmo para otimização. lbfgs é eficiente e apropriado para problemas de pequena a média escala.
SVM	C	Parâmetro de regularização. Controla o <i>trade-off</i> entre margem de separação e erro de classificação.
	gamma	Coeficiente do kernel RBF. O valor scale usa $1/(n_features \cdot X.var())$.
KNN	n_neighbors	Número de vizinhos mais próximos usados na classificação.
	p	Parâmetro da métrica de distância: p=2 representa a distância Euclidiana.
GB XGBoost	n_estimators	Número de árvores adicionadas sequencialmente.
	learning_rate	Taxa de aprendizado que controla a contribuição de cada árvore.
	max_depth	Profundidade máxima das árvores individuais.

Após definido o melhor classificador, foram verificados os 10 atributos que mais influenciavam o risco de um funcionário pedir para sair, considerando o classificador *Random Forest*, conforme ilustrado na Figura 38.

Figura 38: Top 10 atributos do *Random Forest*



Fonte: Elaborado pela autora

A análise da importância das variáveis foi conduzida utilizando o modelo *Random Forest* com hiperparâmetros previamente otimizados e definidos com base no melhor desempenho alcançado durante os testes. O modelo foi treinado com todos os dados disponíveis, sem aplicar imputação de valores faltantes. Após o treinamento, a importância de cada variável foi extraída com base no critério de redução da impureza de Gini acumulada ao longo das árvores da floresta, refletindo o quanto cada variável contribui para a correta classificação dos exemplos. As 10 variáveis mais importantes identificadas pelo modelo foram:

1. *Tempo de empresa*: esta variável teve de longe a maior importância. Representa o tempo total de permanência do funcionário na empresa em meses. Esse forte peso sugere que a longevidade na organização é um forte preditor de comportamento, possivelmente relacionado à retenção ou à propensão de desligamento voluntário.
2. *Desempenho*: a nota da avaliação de desempenho recente também teve alta relevância. Isso indica que funcionários com avaliações mais baixas ou inconsistentes podem estar

mais propensos a sair ou serem considerados de maior risco, sendo um reflexo do alinhamento (ou desalinhamento) entre funcionário e empresa.

3. *Benefícios*: esse grupo de benefícios foi tratado como uma variável composta. Sua importância reforça a percepção de que recompensas e incentivos extracurriculares impactam diretamente na satisfação e permanência dos funcionários.
4. *Tempo de experiência*: anos de experiência profissional total indicam senioridade. A interação entre tempo de casa e tempo de mercado parece ser relevante para o modelo, talvez indicando perfis com maior mobilidade ou estabilidade no mercado.
5. *Horas de treinamento*: representa o investimento da empresa em capacitação. A presença dessa variável entre as mais importantes sugere que o acesso a treinamentos pode influenciar na motivação e perspectiva de futuro do funcionário.
6. *Afastamentos*: um número elevado de afastamentos pode estar relacionado a problemas de saúde, insatisfação ou desengajamento, o que pode sinalizar risco de desligamento.
7. *Posição na Faixa*: Representa a posição na faixa salarial de cargo. Sua importância sugere que o posicionamento na faixa salarial influenciam diretamente o comportamento dos funcionários.
8. *CLT*: O tipo de vínculo empregatício também apareceu entre as variáveis mais importantes. Estar sob regime CLT, em oposição a contrato temporário ou terceirizado, pode indicar maior estabilidade ou expectativas diferentes em relação à permanência.
9. *Jornada de trabalho*: expressa em cargas horárias semanais de 20, 30 ou 40 horas, a jornada de trabalho é um atributo relevante na análise de fatores que influenciam a decisão de permanência ou desligamento do colaborador. Em geral, jornadas mais longas podem estar associadas a maior desgaste físico e emocional, impactando negativamente o equilíbrio entre vida pessoal e profissional e contribuindo para o aumento da rotatividade. Neste estudo, a matriz de correlação revela uma correlação moderada e negativa (cerca de -0,50) entre a jornada semanal e o tipo de contrato de trabalho, sugerindo que contratos com menor carga horária estão mais frequentemente associados a modalidades contratuais específicas, como o trabalho temporário ou parcial. Esse achado indica que a jornada de trabalho não atua de forma isolada, mas se relaciona com o regime contratual e possivelmente com condições de estabilidade, benefícios e perspectivas de crescimento, influenciando de maneira significativa a decisão do funcionário em permanecer ou deixar a organização.

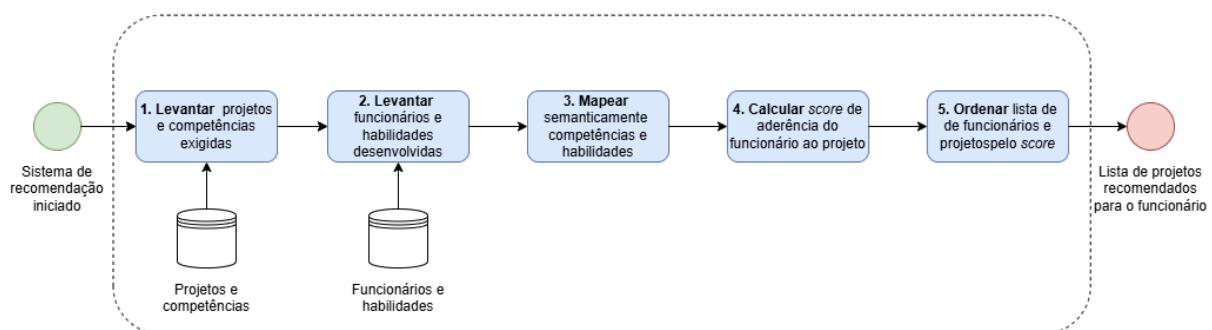
10. *Pós-graduação incompleta*: ter uma pós em andamento pode sinalizar busca por crescimento profissional ou insatisfação com a posição atual, sendo um possível indicador de transição iminente.

Esses resultados permitem não apenas interpretar as decisões do modelo, mas também oferecem *insights* para a área de gestão de pessoas, ajudando na criação de estratégias de retenção baseadas em dados objetivos.

6.4 Modelo para retenção de talentos

Um sistema de recomendação foi elaborado para prover o gestor com informações sobre os projetos mais adequados ao perfil do funcionário com risco de desligamento voluntário. Este dado pode levar o gestor a refletir, por exemplo, se o fato de o projeto no qual o funcionário está alocado não ser adequado para seu perfil interfere em sua decisão de sair da empresa, funcionando como uma ferramenta de retenção. O processo do sistema está ilustrado na Figura 39 e descrito de forma detalhada no Capítulo 5.

Figura 39: Processo do sistema de recomendação para retenção de talentos



Fonte: Elaborado pela autora

A base de dados dos funcionários apresenta atributos como habilidades e grau de habilidade. Já a base de projetos apresenta competências de 1 a 21. As competências foram categorizadas, em ambas as bases seguindo a mesma categoria. O sistema trabalha com um indicador de cobertura, indicando quantas competências associadas aos projetos são cobertas pelos funcionários. Além da categoria da competência, é calculada uma média dos graus de competência para cada categoria, a partir do grau de habilidade cadastrado. Então é calculado um índice chamado "score" que é obtido por meio do produto da cobertura das competências com o grau médio das

categorias que estão mapeadas nas competências cobertas. Para cada projeto é calculado um *score* de afinidade e realizado o ranqueamento deste *score*.

Para definir os projetos mais indicados para cada funcionário, foi desenvolvido um método baseado no alinhamento entre as competências demandadas pelos projetos e as habilidades apresentadas pelos funcionários. A abordagem combinou técnicas de Processamento de Linguagem Natural (PLN), representações semânticas de textos com *embeddings*, e uma pontuação composta que avalia a adequação entre o perfil do funcionário e os requisitos do projeto.

O primeiro passo consistiu na construção de dois conjuntos de dados.

1. As competências demandadas pelos projetos, extraídas das descrições técnicas dos projetos e estruturadas em um campo de lista.
2. As habilidades dos funcionários, mapeadas a partir dos treinamentos realizados, cursos concluídos e experiências prévias, categorizadas segundo um modelo de classificação interno.

Como as terminologias utilizadas para descrever competências e habilidades nem sempre coincidem literalmente, foi necessário realizar um mapeamento semântico entre essas duas fontes de informação. Para isso, utilizou-se o modelo pré-treinado SBERT, que permite gerar representações vetoriais ou *embeddings* de frases e termos em múltiplos idiomas, otimizadas para tarefas de similaridade semântica ([REIMERS; GUREVYCH, novembro 2019](#)).

Os *embeddings* das competências e das categorias de habilidades foram comparados entre si utilizando a similaridade do cosseno. Esse processo permitiu identificar, para cada competência descrita nos projetos, a categoria de habilidade mais semanticamente próxima no vocabulário dos funcionários. Esse mapeamento resultou em uma tabela de equivalência semântica entre competências e habilidades.

Uma vez estabelecido o mapeamento, cada funcionário foi avaliado em relação a cada projeto com base em dois critérios principais:

- Cobertura: número de competências demandadas pelo projeto que são cobertas pelas habilidades do funcionário. Essa medida reflete o grau de compatibilidade entre o perfil do funcionário e os requisitos do projeto.
- Média dos graus de habilidade: cada funcionário possui um grau associado a cada uma de suas habilidades, derivado de seu histórico de capacitação. Para as categorias relevantes

ao projeto, é computada a média desses graus, representando a profundidade de domínio técnico.

- A pontuação final (*score*) que determina a adequação do funcionário ao projeto é dada pelo produto dessas duas medidas: $score = cobertura \times grau$. Essa multiplicação garante que um bom candidato seja aquele que não apenas cobre a maior parte das competências, mas também apresenta níveis elevados de habilidade técnica nas áreas relevantes.

A abordagem utilizada fundamenta-se em princípios de Sistemas de Recomendação Baseados em Conteúdo, aplicados ao contexto de alocação de recursos humanos. O uso de *embeddings* semânticos para emparelhar competências e habilidades evita limitações impostas por métodos lexicais exatos (ex.: *string matching*), sendo especialmente útil em ambientes nos quais a terminologia pode variar entre diferentes fontes de informação. A métrica composta cobertura × média de grau representa uma forma equilibrada de considerar tanto a extensão quanto a intensidade do alinhamento entre o funcionário e o projeto, incorporando aspectos quantitativos e qualitativos da formação profissional.

A Figura 40 apresenta os resultados do processo de ranqueamento que associa cada funcionário aos três projetos mais compatíveis com seu perfil de habilidades. Cada linha representa uma combinação entre um funcionário (*_Funcionario_id*) e um projeto (*_Projeto_id*), acompanhada de três métricas principais. A cobertura indica o número de competências exigidas pelo projeto que são atendidas pelas habilidades do funcionário. A média do grau corresponde ao valor médio de proficiência do funcionário nas categorias cobertas, refletindo o nível de domínio técnico. Por fim, o *score* é calculado pelo produto entre cobertura e média do grau, sendo utilizado como critério de priorização: quanto maior o *score*, mais indicado o funcionário está para aquele projeto. Por exemplo, projeto com ID *xyyuIbZRssxF6* apresenta uma cobertura de 19 competências, média de grau 3,58 e *score* final de 68,08, o que o posiciona, entre os projetos listados, como o projeto como o mais aderente ao funcionário com ID *09b209332534f4436b8ed6455771df9f*.

Figura 40: Ranqueamento dos projetos por funcionário com base na cobertura de competências e grau médio de habilidade

<u>Funcionario_id</u>	<u>Projeto_id</u>	<u>Cobertura</u>	<u>Média_grau</u>	<u>Score</u>
09b209332534f4436b8ed6455771df9f	xyyulbZRssxF6	19	3,58	68,08
	xygNlrxVwsdPo	18	3,57	64,25
	xyCstjmS9iwkA	12	3,57	42,86
8f2f8e45abfa14e2977aeb92009706ce	xyyulbZRssxF6	14	3,97	55,62
	xygNlrxVwsdPo	13	3,92	50,90
	xyngK/Lzqwdjc	9	3,96	35,62
34eb23d6c393685f155dc77e999ea11a	xyyulbZRssxF6	17	3,13	53,26
	xygNlrxVwsdPo	16	3,15	50,34
	xyngK/Lzqwdjc	12	2,92	35,10
146961436751e1895036212ee9866d27	xyyulbZRssxF6	14	3,52	49,33
	xygNlrxVwsdPo	13	3,56	46,33
	xyngK/Lzqwdjc	10	3,00	30,00
ae1c9dc27dea12267360bd2f79b7c5a6	xyyulbZRssxF6	17	2,79	47,35
	xygNlrxVwsdPo	16	2,82	45,10
	xyngK/Lzqwdjc	12	2,38	28,59
1ec16ce1b7c9d10134965e7cef978718	xygNlrxVwsdPo	10	4,61	46,10
	xyyulbZRssxF6	10	4,61	46,10
	xyngK/Lzqwdjc	6	4,67	28,00
0acabed25ef5549abe56cb9c3778a173	xygNlrxVwsdPo	15	3,00	45,04
	xyyulbZRssxF6	15	3,00	45,04
	xyOXurDR0yqGw	8	2,95	23,63
2afba3c599bcf75ae87bd938f8441c74	xyyulbZRssxF6	16	2,80	44,85
	xygNlrxVwsdPo	15	2,86	42,85
	xyCstjmS9iwkA	10	2,55	25,50
9a75843adafce1e697381dc6dfd7e894	xyyulbZRssxF6	18	2,29	41,20
	xygNlrxVwsdPo	17	2,26	38,45
	xyCstjmS9iwkA	10	2,46	24,58
36fb6f75654fb27c0338fbaddaae115c	xyyulbZRssxF6	18	2,24	40,38
	xygNlrxVwsdPo	17	2,32	39,38
	xyCstjmS9iwkA	12	2,23	26,80
6948405993a8a045a4016d6c59ca27d5	xygNlrxVwsdPo	12	3,29	39,44
	xyyulbZRssxF6	12	3,29	39,44
	xyngK/Lzqwdjc	6	3,24	19,44
64e6fd3ae45f99c652ca129a880be146	xyyulbZRssxF6	21	1,86	39,09
	xygNlrxVwsdPo	20	1,85	37,09
	xyCstjmS9iwkA	13	1,62	21,11
38f412b3eab2b96b9446955e5ef855fd	xyyulbZRssxF6	12	3,24	38,91
	xygNlrxVwsdPo	11	3,45	37,91
	xyngK/Lzqwdjc	8	3,18	25,41
23a6f950805bb55ccdc0ffd4611286f	xygNlrxVwsdPo	11	3,33	36,62
	xyyulbZRssxF6	11	3,33	36,62
	xyjINvrmxRUXsk	6	3,70	22,20
ff10ce3d9ec7a13af760236f99ef468a	xygNlrxVwsdPo	14	2,60	36,42
	xyyulbZRssxF6	14	2,60	36,42
	xyjINvrmxRUXsk	8	2,99	23,95
d5f0c53ff690edf768ed3c099d42dc36	xyyulbZRssxF6	19	1,91	36,26
	xygNlrxVwsdPo	18	1,95	35,11
	xyngK/Lzqwdjc	13	2,16	28,03
7be9f84b17a092d2d73da6d505ab55a1	xygNlrxVwsdPo	12	3,01	36,12

Fonte: Elaborado pela autora

6.5 Discussão

Para um melhor funcionamento do modelos ao ser implementado na empresa, é importante que haja uma categorização dos motivos de desligamento, visando a melhoraria do desempenho do sistema. Adicionalmente, é importante que o máximo possível de campos da base de dados esteja preenchido, buscando, desta forma, uma maior robustez do modelo.

Tivemos que lidar com cenário de dados faltantes e desbalanceamento de classes em uma base de dados sensíveis, no qual há necessidade de atenção especial com risco de imputação de viés. Por este motivo optamos pelo uso de *class weight*, nos classificadores que o suportam, visando minimização de desbalanceamento de classes. Adicionalmente, utilizamos imputação de dados via KNN, com *imputer k=1*, para tratamento de dados faltantes.

O classificador RF tem todas as características para lidar com dados faltantes, funcionando também de forma adequada nestes casos (TANG; ISHWARAN, 2017). Porém, havia a dúvida se a imputação de dados com utilização de *imputer* poderia maximizar o desempenho do classificador. Foi possível observar que o classificador com utilização de *imputer* obteve um desempenho ligeiramente menor do que sem a utilização da técnica de imputação de dados faltantes.

Este experimento se caracteriza em uma prova de conceito para nossa proposta de *framework*. Considerando o tempo necessário para testes em ambientes reais de propostas desta magnitude, não foi possível a realização de testes diretos no ambiente de aplicação.

A base de dados de projeto apresenta as principais competências necessárias para atuação em cada projeto. Também deveria apresentar a mesma chave para funcionários que a base de dados de RH, associando-os aos projetos, porém as bases apresentavam chaves para funcionários diferentes. Desta forma, não foi possível associar os funcionários mapeados com risco de desligamento voluntário aos projetos, impedindo a análise da proposta inicial que seria verificar se o projeto está correto para o funcionário. Quando implementado na empresa, será possível também a apresentação desta informação.

7 Conclusão

Este capítulo apresenta as principais conclusões desta tese. Em seguida são apresentadas propostas recomendações para trabalhos futuros, considerando o potencial de aplicação e expansão do método desenvolvido.

7.1 Considerações finais

Esta tese apresenta proposta de arcabouços para previsão de desligamento voluntário e retenção de talentos. As propostas foram testadas considerando o cenário de uma empresa de serviços de TI.

Para desenvolvimento da proposta, foi utilizada a metodologia *Design Science Research*. No ciclo de rigor, por meio do referencial teórico, foram apresentados conceitos sobre RH, IA e o uso de IA para RH, organizando a base de conhecimento. Já no ciclo de relevância, além da definição do contexto e do problema, verificamos, a partir do mapeamento sistemático e da revisão sistemática da literatura, as lacunas e oportunidades, com a investigação do problema. Desta forma, foi delineado o ambiente de aplicação, com identificação de melhores práticas e discussões orientadoras para o melhor desenvolvimento da proposta. Alimentado pelo ciclo de rigor e pelo ciclo de relevância, o modelo foi desenvolvido, avaliado e validado, compondo o ciclo de *design*.

Bases de dados de RH apresentam obstáculos consideráveis para desenvolvimento de modelos para previsão de desligamento voluntário, entre eles podemos citar o acesso aos dados, o tratamento dos dados pelas equipes de RH, principalmente os dados obtidos de forma manual. A mera existência de dados imputados manualmente já colocam em risco o modelo, devido à incerteza em relação à continuidade de disponibilização de tais dados e maior possibilidade de erros nos dados. Somado a esses fatores, estão os motivos de desligamento que independem de informações capturadas pelo RH, como motivos pessoais/familiares e motivos não documentados relacionados ao trabalho. Adicionalmente, a maior parte dos atributos não sofre mudanças

constantes, o que pode limitar o desempenho do sistema ([RAMAMURTHY et al., 2015](#)).

A presença de atributos como a "Nota de Avaliação de Desempenho", atualizada anualmente, e o registro de horas de treinamento, com atualizações mais frequentes, representa uma potencial vantagem competitiva para o modelo preditivo, ao incorporar indicadores relevantes de desempenho e capacitação dos funcionários, considerando a base disponibilizada pela empresa de serviços de TI para realização do experimento.

Diversos estudos da literatura utilizam pesquisas de formulário com questões qualitativas para geração de dados para treinamento de modelos, como "Satisfação no trabalho", porém a obtenção e utilização destes tipos de dados não é trivial. Informações qualitativas tendem a não fazer parte do *hall* de informações cadastradas pelo RH. Mesmo quando a empresa realiza pesquisas de clima, onde tais questões são abordadas, a confidencialidade das respostas deve ser garantida pela empresa. Geralmente este tipo de pesquisa é realizado por empresas de consultoria que entregam uma compilação dos dados para a empresa contratante, de forma a garantir que as respostas sejam sigilosas, dando maior transparência ao processo. Frequentemente os dados compilados não chegam a nível individual, mas sim de uma equipe completa, visando a confidencialidade e proteção do funcionário com anonimato. Desta forma, não é possível associar este tipo de atributo a um funcionário, tornando incerta sua utilização em um sistema que se apoia em disponibilização de dados pelo RH.

Foi realizado experimento, no qual, após tratamento dos dados, foram treinados diversos classificadores, com e sem imputação de dados via KNN ($k=1$). O classificador que apresentou melhor desempenho foi o *random forest*, com acurácia de 92%. Entre os atributos com maior influência no risco de desligamento voluntário estão: 'tempo de empresa', 'desempenho', 'benefícios', 'tempo de experiência', 'horas de treinamento', 'afastamentos', 'posição na faixa', 'CLT', 'jornada de trabalho' e 'pós-graduação incompleta'.

Esta solução passou por uma prova de conceito, considerando o cenário de uma empresa prestadora de serviços de TI. Para aplicação em outras empresas, a menos que a base de dados tenha os mesmos tipos de atributos, será necessário novo treino dos classificadores, podendo haver resultado diferente do alcançado nesta tese.

O contato com o especialista no domínio, ou sua participação na pesquisa, neste caso um profissional de RH, é fundamental tanto para validação dos dados como entendimento dos processos. Adicionalmente, o tratamento de questões de confidencialidade e gestão de dados sensíveis é primordial. Para realização desta tese, tivemos reuniões com o RH da empresa de serviços de TI que disponibilizou as bases de dados. Estas reuniões foram importantes para entendimento das bases disponibilizadas e processos que pudessem afetar os dados.

Em relação às contribuições esperadas, as contribuições diretas foram alcançadas, com desenvolvimento e avaliação de um modelo para previsão de risco de desligamento voluntário e modelo para retenção de talentos, com identificação de projetos mais adequados ao perfil do funcionário. Por meio das bases de dados utilizadas, foi possível a identificação de parâmetros importantes para o sucesso do modelo, como tempo de empresa, desempenho do funcionário e benefícios, por exemplo. Entre as contribuições alcançadas para a empresa parceira, de forma imediata, podemos citar o uso da Lei do Bem para benefícios fiscais, já comprovada com utilização das publicações que são produtos desta tese. Ainda considerando as contribuições para a empresa, podemos apontar o sucesso do desenvolvimento do tema de sua escolha e o estreitamento da relação com a academia. Adicionalmente, uma contribuição observada, é a explicação sobre conceitos de RH para área de TI.

A solução desenvolvida pode ser utilizada para uma gestão pró-ativa, ajudando o gestor a alocar o funcionário de forma mais assertiva nos novos projetos. Esta solução combinada a outras iniciativas pode aumentar a capacidade de retenção de talentos na empresa.

7.2 Limitações

Alguns desafios foram encontrados no decorrer do desenvolvimento da tese. Em especial em relação às bases de dados. Embora as bases de dados apresentem atributos interessantes para o treinamento dos modelos propostos, foi enfrentada a existência de muitos dados faltantes. Adicionalmente, foi necessária a realização de conversões dos dados para melhor adequação ao treinamento do modelo, o que pode ter introduzido viés ao processo.

Não foi possível a aplicação do modelo conceitual da forma como foi proposto. Isto ocorreu devido ao fato da base de funcionários e da base de projetos não estarem com a mesma chave para identificação de funcionários. Para que o modelo funcionasse conforme proposto no modelo conceitual, após identificação do risco de desligamento voluntário para determinado funcionário, por meio da chave de identificação criada, deveria ser possível identificar os projetos nos quais aquele funcionário está alocado. Como as chaves das bases estavam diferentes, não foi possível a associação direta, e consequente verificação se o funcionário estava alocado no projeto mais adequado ao seu perfil.

7.3 Perspectivas futuras

Para um melhor funcionamento do modelo para previsão de desligamento voluntário, ao ser implementado na empresa, é importante que haja uma categorização dos motivos de desligamento, visando a melhoraria do desempenho do sistema. Desta forma, sugerimos a digitalização do formulário de entrevista de desligamento, com motivos pré definidos, baseados em experiências passadas e *benchmarking* com outras empresas do setor. O formulário pode também ter a opção de cadastro de novo motivo de desligamento, enriquecendo o mesmo.

Entre os possíveis desdobramentos futuros desta proposta, pode haver desenvolvimento de uma ferramenta ou sistema, a partir dos *frameworks* apresentados nesta tese, com *dashboard* que indique ao gestor o risco de desligamento voluntário, as variáveis associadas ao risco, com indicação de ações para retenção associadas às variáveis. Além disso, a indicação se o funcionário está no projeto mais aderente ao seu perfil, somada da indicação de *ranking* dos projetos mais pertinentes para o perfil do funcionário.

Um trabalho futuro sugerido, seria a ampliação do *framework*, com integração de dados exógenos, como dados de pesquisas de mercado sobre remuneração (tais como faixa salarial média para determinado cargo ou atividade), dados de mercado de trabalho (como índices de desemprego e rotatividade), entre outros. Assim, seria possível a verificação, por exemplo, do posicionamento na faixa salarial de acordo com o mercado, enriquecendo o modelo.

A base de dados utilizada contém informações de diversos sistemas diferentes, além de dados com imputação manual. As informações estão distribuídas por estes sistemas. Desta forma, seria interessante que a ferramenta fosse integrada com as ferramentas existentes, como Convenia¹ e Qulture.Rocks², por exemplo, com dados alimentados pelas gerências e pelos próprios funcionários. Adicionalmente, com desenvolvimento e posterior integração com ferramentas para entrevista de desligamento e treinamento, além de integração com ferramenta existente para gestão de projetos, automatizando todo o processo de obtenção dos dados para previsão de risco de desligamento. Também há possibilidade de desenvolvimento de *dashboards*, com diversas informações sobre as equipes e funcionários, identificando sua progressão, com atualização frequente do perfil e dos dados, funcionando como ferramenta de apoio a decisão para gestores e RH.

¹<https://www.convenia.com.br/>

²<https://www.qulture.rocks/>

Referências

- ALBALAWI, Abdulmajeed Saad; NAUGTON, Shahnaz; ELAYAN, Malek Bakheet; SLEIMI, Mohammad Tahseen. Perceived organizational support, alternative job opportunity, organizational commitment, job satisfaction and turnover intention: A moderated-mediated model. **Organizacija**, volume 52, número 4, páginas 310–324, 2019.
- ARROMRIT, Trirat; SRISAKAEW, Korrapin; ROSWHAN, Napatsakorn; MAHIKUL, Wiriya. A Supervised Machine Learning Method for Predicting the Employees Turnover Decisions. In: IEEE. 2023 IEEE 8th International Conference On Software Engineering and Computer Systems (ICSECS). s.l.: IEEE, 2023. Páginas 122–127.
- ASLAM, M. S.; RAMLI, M. F.; SHAH, S. A. A. A Brief on Correlation and Co-integration. In: 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET). s.l.: s.n., 2019. Páginas 121–125. DOI: [10.1109/ICSET.2019.8878278](https://doi.org/10.1109/ICSET.2019.8878278).
- AYODELE, Taiwo Oladipupo. Types of machine learning algorithms. **New advances in machine learning**, InTech Rijeka, Croatia, volume 3, páginas 19–48, 2010.
- BALLINGER, Gary A; CROSS, Rob; HOLTON, Brooks C. The right friends in the right places: Understanding network structure as a predictor of voluntary turnover. **Journal of Applied Psychology**, American Psychological Association, volume 101, número 4, página 535, 2016.
- BANKINS, Sarah. The ethical use of artificial intelligence in human resource management: a decision-making framework. **Ethics and Information Technology**, Springer, volume 23, número 4, páginas 841–854, 2021.
- BASHA, Md Shaik Amzad; VARIKUNTA, Obulesu; DEVI, A Uma; RAJA, S. Machine Learning in HR Analytics: A Comparative Study on the Predictive Accuracy of Attrition Models. In: IEEE. 2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT). s.l.: IEEE, 2024. Páginas 475–480.
- BASILI, Victor R; ROMBACH, H Dieter. The TAME project: Towards improvement-oriented software environments. **IEEE Transactions on software engineering**, IEEE, volume 14, número 6, páginas 758–773, 1988.

- BASS, Julian M; BEECHAM, Sarah; RAZZAK, Mohammed Abdur; NOLL, John. Employee retention and turnover in global software development: Comparing in-house offshoring and offshore outsourcing. In: PROCEEDINGS of the 13th International Conference on Global Software Engineering. s.l.: s.n., 2018. Páginas 82–91.
- BATISTA, Gustavo EAPA; PRATI, Ronaldo C; MONARD, Maria Carolina. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, volume 6, número 1, páginas 20–29, 2004.
- BELL, Jason. What is machine learning? **Machine Learning and the City: Applications in Architecture and Urban Design**, Wiley Online Library, páginas 207–216, 2022.
- BENTÉJAC, Candice; CSÖRGŐ, Anna; MARTÍNEZ-MUÑOZ, Gonzalo. A comparative analysis of gradient boosting algorithms. **Artificial Intelligence Review**, Springer US, volume 54, número 3, páginas 1937–1967, 2021. DOI: [10.1007/s10462-020-09896-5](https://doi.org/10.1007/s10462-020-09896-5).
- BENTÉJAC, Olivier; MORVAN, Franck; MOUTARLIER, Fabrice. Gradient Boosting Decision Trees: A survey on algorithms and applications. **Expert Systems with Applications**, volume 136, páginas 216–235, 2020. DOI: [10.1016/j.eswa.2019.06.012](https://doi.org/10.1016/j.eswa.2019.06.012).
- BHUPATHI, Priyadharshini; PRABU, S; GOH, Alexis PI. Artificial intelligence-enabled knowledge management using a multidimensional analytical framework of visualizations. **International Journal of Cognitive Computing in Engineering**, Elsevier, volume 4, páginas 240–247, 2023.
- BISHOP, Christopher M; NASRABADI, Nasser M. **Pattern recognition and machine learning**. s.l.: Springer, 2006. volume 4.
- BISPO, Carlos Alberto Ferreira. Um novo modelo de pesquisa de clima organizacional. **Production**, SciELO Brasil, s.l., volume 16, páginas 258–273, 2006.
- BONGALE, Anupkumar M; DHARRAO, Deepak; UROLAGIN, Siddhaling. Exploratory data analysis and classification of employee retention based on logistic regression model. In: IEEE. 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS). s.l.: IEEE, 2023. volume 1, páginas 1929–1933.
- BOON, Corine; ECKARDT, Rory; LEPAK, David P; BOSELIE, Paul. Integrating strategic human capital and strategic human resource management. **The International Journal of Human Resource Management**, Taylor & Francis, volume 29, número 1, páginas 34–67, 2018.

BRASIL. O que é a Lei do Bem. Brasil: Ministério da Ciência, Tecnologia e Inovação, 2023. Disponível em: <https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/lei-do-bem/paginas/o-que-e-a-lei-do-bem>. Acesso em: 22 de julho de 2025.

_____. **Painel de Informações do Novo CADEG.** Brasil: Ministério do Trabalho, 2022. Disponível em: <http://pdet.mte.gov.br/novo-caged?view=default>. Acesso em: 22 de março de 2023.

_____. **Painel de Informações do Novo CADEG - Maio 2025.** Brasil: Ministério do Trabalho, 2024. Disponível em: <https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/estatisticas-trabalho/novo-caged/2025/maio/pagina-inicial>. Acesso em: 15 de maio de 2025.

BRASIL. Lei nº13.709, de 14 de agosto de 2018 — Dispõe sobre o tratamento de dados pessoais, e dá outras providências. s.l., 2018. Sanção em 14-agosto-2018; vigência a partir de 16-agosto-2020. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

_____. **Resolução nº.510, de 7 de abril de 2016: dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais.** Brasília, DF, 2016. Publicado no Diário Oficial da União em 24 de maio de 2016. Disponível em: <https://bvsms.saude.gov.br/bvs/saudelegis/cns/2016/res0510_07_04_2016.html>.

BRASSCOM, Associação de Empresas de Tecnologia da Informação e Comunicação.

Demanda de Talentos em TIC e Estratégia ΣTCEM: Relatório de Inteligência e Informação BRI2-007-v112. São Paulo, Brasil: s.n., 2021. <https://brasscom.org.br/pdfs/demandade-talentos-em-tic-e-estrategia-tcem/>. Acesso em: 22 de março de 2023.

BUDHWAR, Pawan; MALIK, Ashish; DE SILVA, MT Thedushika; THEVISUTHAN, Praveena. Artificial intelligence—challenges and opportunities for international HRM: a review and research agenda. **The International Journal of human resource management**, Taylor & Francis, volume 33, número 6, páginas 1065–1097, 2022.

CARVALHO, Miguel; PINHO, Armando J; BRÁS, Susana. Resampling approaches to handle class imbalance: a review from a data perspective. **Journal of Big Data**, Springer, s.l., volume 12, número 1, página 71, 2025.

CERVANTES, Jair; GARCIA-LAMONT, Farid; RODRÍGUEZ-MAZAHUA, Lisbeth; LOPEZ, Asdrubal. A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, Elsevier, volume 408, páginas 189–215, 2020.

- CHAKRABORTY, Raj; MRIDHA, Krishna; SHAW, Rabindra Nath; GHOSH, Ankush. Study and prediction analysis of the employee turnover using machine learning approaches. In: IEEE. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). s.l.: IEEE, 2021. Páginas 1–6.
- CHANG, Victor; MOU, Yeqing; XU, Qianwen Ariel; XU, Yue. Job satisfaction and turnover decision of employees in the Internet sector in the US. **Enterprise Information Systems**, Taylor & Francis, volume 17, número 8, página 2130013, 2023.
- CHIAVENATO, Idalberto. **Gestão de pessoas: o novo papel dos recursos humanos nas organizações**. 4. edição. Barueri, SP: Manole, 2014.
- CHUNG, Doohee; YUN, Jinseop; LEE, Jeha; JEON, Yeram. Predictive model of employee attrition based on stacking ensemble learning. **Expert Systems with Applications**, Elsevier, volume 215, página 119364, 2023.
- CORTEZ, Rafael M.; BATISTA, Jose A. N.; GOES, Paulo B. Rethinking the digital transformation in knowledge-intensive services. **Technological Forecasting and Social Change**, Elsevier, s.l., volume 178, página 121563, 2022. DOI: [10.1016/j.techfore.2022.121563](https://doi.org/10.1016/j.techfore.2022.121563).
- COSTA, Luiz Alexandre; DIAS, Edson; RIBEIRO, Danilo; FONTÃO, Awdren; PINTO, Gustavo; SANTOS, Rodrigo Pereira Dos; SEREBRENIK, Alexander. An Actionable Framework for Understanding and Improving Talent Retention as a Competitive Advantage in IT Organizations. In: PROCEEDINGS of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings. s.l.: [sine nomine], 2024. Páginas 290–291.
- COSTA, R; CARVALHO, P. Machine learning predictive models for preventing employee turnover costs. In: IEEE. 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). s.l.: IEEE, 2022. Páginas 1–6.
- CUNNINGHAM, Padraig; DELANY, Sarah Jane. k-Nearest neighbour classifiers-A Tutorial. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, volume 54, número 6, páginas 1–25, 2021.
- CUTLER, Adele; CUTLER, D Richard; STEVENS, John R. Random forests. **Ensemble machine learning: Methods and applications**, Springer, páginas 157–175, 2012.
- AL-DALAHMEH, Maha Lufti; HÉDER, Mária; DAJNOKI, Krisztina. The effect of talent management practices on employee turnover intention in the information and communication technologies (ICTs) sector: Case of Jordan. **Problems and Perspectives in Management**, 2020.

- DAYAN, Peter; SAHANI, Maneesh; DEBACK, Grégoire. Unsupervised learning. **The MIT encyclopedia of the cognitive sciences**, MIT Press, páginas 857–859, 1999.
- DE BRITO, Renata Peregrino; OLIVEIRA, Lucia Barbosa de. A relação entre gestão de recursos humanos e desempenho organizacional. **Journal Brazilian Business Review**, volume 13, número 3, páginas 94–115, 2016.
- DE VILLE, Barry. Decision trees. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, volume 5, número 6, páginas 448–455, 2013.
- DEEL. **O futuro do RH no Brasil - o que esperar do setor em 2025**. São Paulo, Brasil: s.n., 2024. <https://www.deel.com/pt/recursos/relatorio-o-futuro-do-rh/>. Acesso em: 22 de maio de 2025.
- DRESCH, Aline; LACERDA, Daniel Pacheco; ANTUNES JR, José Antônio Valle. **Design science research**. s.l.: Springer, 2015.
- DYBA, Tore; DINGSOYR, Torgeir; HANSSEN, Geir K. Applying systematic reviews to diverse study types: An experience report. In: IEEE. FIRST international symposium on empirical software engineering and measurement (ESEM 2007). s.l.: IEEE, 2007. Páginas 225–234.
- DYBÅ, Tore; DINGSØYR, Torgeir. Empirical studies of agile software development: A systematic review. **Information and software technology**, Elsevier, volume 50, número 9-10, páginas 833–859, 2008.
- EKSTRAND, Michael D; RIEDL, John T; KONSTAN, Joseph A et al. Collaborative filtering recommender systems. **Foundations and Trends® in Human–Computer Interaction**, Now Publishers, Inc., volume 4, número 2, páginas 81–173, 2011.
- FALLUCCHI, Francesca; COLADANGELO, Marco; GIULIANO, Romeo; DE LUCA, Ernesto William. Predicting employee attrition using machine learning techniques. **Computers**, scopus, s.l., volume 9, número 4, páginas 1–17, 2020. Cited by: 140; All Open Access, Green Open Access. DOI: [10.3390/computers9040086](https://doi.org/10.3390/computers9040086). Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098559248&doi=10.3390%2fcomputers9040086&partnerID=40&md5=8d864d2c495c9aa0e596bdbba57e143df>>.
- FAWCETT, Tom. An introduction to ROC analysis. **Pattern Recognition Letters**, Elsevier, volume 27, número 8, páginas 861–874, 2006. DOI: [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010).

- FELIZARDO, Katia Romero; SOUZA, Érica Ferreira de; NAPOLEÃO, Bianca Minetto; VIJAYKUMAR, Nandamudi Lankalapalli; BALDASSARRE, Maria Teresa. Secondary studies in the academic context: A systematic mapping and survey. **Journal of Systems and Software**, Elsevier, s.l., volume 170, página 110734, 2020.
- FERNÁNDEZ, Alberto; GARCIA, Salvador; HERRERA, Francisco; CHAWLA, Nitesh V. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **Journal of artificial intelligence research**, volume 61, páginas 863–905, 2018.
- FILHO, Carlos Poncinelli; MARQUES JR, Elias; CHANG, Victor; DOS SANTOS, Leonardo; BERNARDINI, Flavia; PIRES, Paulo F; OCHI, Luiz; DELICATO, Flavia C. A systematic literature review on distributed machine learning in edge computing. **Sensors**, MDPI, volume 22, número 7, página 2665, 2022.
- FRIEDMAN, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine. **Annals of Statistics**, volume 29, número 5, páginas 1189–1232, 2001. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451).
- GARCIA, Javier; FERNÁNDEZ, Fernando. A comprehensive survey on safe reinforcement learning. **Journal of Machine Learning Research**, volume 16, número 1, páginas 1437–1480, 2015.
- GARTNER. **O futuro do trabalho se reinventa**. [Sine loco: sine nomine], 2022. Disponível em: <https://www.gartner.com.br/pt-br/insights/futuro-do-trabalho>. Acesso em: 13 de março de 2023.
- GOORBERGH, Ruben van den; SMEDEN, Maarten van; TIMMERMAN, Dirk; VAN CALSTER, Ben. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. **Journal of the American Medical Informatics Association**, volume 29, número 9, páginas 1525–1534, 2022. ISSN 1527-974X. DOI: [10.1093/jamia/ocac093](https://doi.org/10.1093/jamia/ocac093). eprint: <https://academic.oup.com/jamia/article-pdf/29/9/1525/45444435/ocac093.pdf>. Disponível em: <<https://doi.org/10.1093/jamia/ocac093>>.
- HABOUS, Amine; OUBENAALLA, Youness et al. Predicting employee attrition using supervised learning classification models. In: IEEE. 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS). s.l.: IEEE, 2021. Páginas 1–5.
- HEVNER, Alan; CHATTERJEE, Samir. **Design research in information systems: theory and practice**. s.l.: Springer Science & Business Media, 2010. volume 22.
- HEVNER, Alan R; MARCH, Salvatore T; PARK, Jinsoo; RAM, Sudha. Design science in information systems research. **MIS quarterly**, JSTOR, s.l., páginas 75–105, 2004.

- HMOUD, Bilal; LASZLO, Varallyai et al. Will artificial intelligence take over human resources recruitment and selection. **Network Intelligence Studies**, Romanian Foundation for Business Intelligence, Editorial Department, volume 7, número 13, páginas 21–30, 2019.
- HOLTOM, Brooks; LEI, Zhike; REEVES, Cody; DARABI, Tiffany. **Are You Trying to Retain the Right Employees?** s.l.: Harvard Business Review, 2021. Disponível em: <https://hbr.org/2021/10/are-you-trying-to-retain-the-right-employees?registration=success/>. Acesso em: 25 de março de 2023.
- ISINKAYE, Folasade Olubusola; FOLAJIMI, Yetunde O; OJOKOH, Bolande Adefowoke. Recommendation systems: Principles, methods and evaluation. **Egyptian informatics journal**, Elsevier, volume 16, número 3, páginas 261–273, 2015.
- ISSA, Lana; ALKHATIB, Mahdi; AL-BADARNEH, Aalaa; QUSEF, Abdallah. Employee Retention in Agile Project Management. In: IEEE. 2019 10th International Conference on Information and Communication Systems (ICICS). [Sine loco: sine nomine], 2019. Páginas 160–165.
- IVANCEVICH, John M. **Gestão de recursos humanos**. [Sine loco]: AMGH Editora, 2009.
- JESUS, Ana Carolina C. de; JÚNIOR, Márcio Enio G. D.; BRANDÃO, Wladimir C. Exploiting linkedin to predict employee resignation likelihood. In: PROCEEDINGS of the 33rd Annual ACM Symposium on Applied Computing. Pau, France: Association for Computing Machinery, 2018. (SAC '18), páginas 1764–1771. ISBN 9781450351911. DOI: [10.1145/3167132.3167320](https://doi.org/10.1145/3167132.3167320). Disponível em: <<https://doi.org/10.1145/3167132.3167320>>.
- JIN, Ziwei; SHANG, Jiaxing; ZHU, Qianwen; LING, Chen; XIE, Wu; QIANG, Baohua. RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Scopus, s.l., 12343 LNCS, páginas 503–515, 2020. DOI: [10.1007/978-3-030-62008-0_35](https://doi.org/10.1007/978-3-030-62008-0_35). Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096609229&doi=10.1007%2f978-3-030-62008-0_35&partnerID=40&md5=b54b17366877aab4caf41c44689b46b>.
- JORGENSEN, Dale W.; VU, Khuong M. Information Technology and the World Growth Resurgence. **The World Economy**, Wiley, s.l., volume 30, número 6, páginas 763–782, 2007. DOI: [10.1111/j.1467-9701.2007.01019.x](https://doi.org/10.1111/j.1467-9701.2007.01019.x).

- KAUSHAL, Neelam; KAURAV, Rahul Pratap Singh; SIVATHANU, Brijesh; KAUSHIK, Neeraj. Artificial intelligence and HRM: identifying future research Agenda using systematic literature review and bibliometric analysis. **Management Review Quarterly**, Springer, páginas 1–39, 2021.
- KITCHENHAM, Barbara; BRERETON, Pearl. A systematic review of systematic review process research in software engineering. **Information and software technology**, Elsevier, volume 55, número 12, páginas 2049–2075, 2013.
- KITCHENHAM, Barbara; CHARTES, Sue. **Guidelines for performing systematic literature reviews in software engineering**. [Sine loco], 2007.
- KÖCHLING, Alina; WEHNER, Marius Claus. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. **Business Research**, Springer, volume 13, número 3, páginas 795–848, 2020.
- KRISHNA, Shobhanam; DWIVEDI, Rohit; MURTI, Ashutosh. Predictive Analytics for Employee Attrition: A Comparative Study of Machine Learning Algorithms. In: IEEE. 2023 Third International Conference on Digital Data Processing (DDP). s.l.: IEEE, 2023. Páginas 180–187.
- KRISHNA, Shobhanam; SIDHARTH, Sumati. Analyzing employee attrition using machine learning: the new AI approach. In: IEEE. 2022 IEEE 7th International conference for Convergence in Technology (I2CT). s.l.: IEEE, 2022. Páginas 1–14.
- _____. HR analytics: Employee attrition analysis using random forest. **International Journal of Performability Engineering**, volume 18, número 4, página 275, 2022.
- KUMAR M, Dileep; GOVINDARAJO, Normala S. Instrument development “intention to stay instrument” (ISI). **Asian Social Science**, Canadian Center of Science e Education, volume 10, número 12, páginas 1–21, 2014.
- KURANI, Akshit; DOSHI, Pavan; VAKHARIA, Aarya; SHAH, Manan. A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting. **Annals of Data Science**, Springer Singapore, volume 10, número 1, páginas 183–208, 2023. DOI: [10.1007/s40745-021-00344-x](https://doi.org/10.1007/s40745-021-00344-x).
- LAZZARI, Matilde; ALVAREZ, Jose M; RUGGIERI, Salvatore. Predicting and explaining employee turnover intention. **International Journal of Data Science and Analytics**, Springer, volume 14, número 3, páginas 279–292, 2022.

- LEBLANC, Patrick M; BANKS, David; FU, Linhui; LI, Mingyan; TANG, Zhengyu; WU, Qiuyi. Recommender systems: a review. **Journal of the American Statistical Association**, Taylor & Francis, s.l., volume 119, número 545, páginas 773–785, 2024.
- LEE, Jongseo; KANG, Juyoung. A study on job satisfaction factors in retention and turnover groups using dominance analysis and LDA topic modeling with employee reviews on Glassdoor. In: 38TH International Conference on Information Systems: Transforming Society with Digital Innovation. s.l.: s.n., 2018.
- MA, Zifei; LI, Ruiyin; LI, Tong; ZHU, Rui; JIANG, Rong; YANG, Juan; TANG, Mingjing; ZHENG, Ming. A data-driven risk measurement model of software developer turnover. **Soft Computing**, Springer, volume 24, páginas 825–842, 2020.
- MAPHOSA, Vusumuzi; MAPHOSA, Mfowabo. Fifteen years of recommender systems research in higher education: Current trends and future direction. **Applied Artificial Intelligence**, Taylor & Francis, s.l., volume 37, número 1, página 2175106, 2023.
- MARÍN DÍAZ, Gabriel; GALÁN HERNÁNDEZ, José Javier; GALDÓN SALVADOR, José Luis. Analyzing employee attrition using explainable AI for strategic HR decision-making. **Mathematics**, MDPI, volume 11, número 22, página 4677, 2023.
- MARVIN, Ggaliwango; JACKSON, Majwega; ALAM, Md Golam Rabiul. A machine learning approach for employee retention prediction. In: IEEE. 2021 IEEE Region 10 Symposium (TENSYMP). s.l.: IEEE, 2021. Páginas 1–8.
- MASSONI, Tiago; GINANI, Nilton; SILVA, Wallison; BARROS, Zeus; MOURA, Georgia. Relating voluntary turnover with job characteristics, satisfaction and work exhaustion-An initial study with Brazilian developers. In: IEEE. 2019 IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE). s.l.: IEEE, 2019. Páginas 85–88.
- MITRAVINDA, KM; SHETTY, Sakshi. Employee attrition: Prediction, analysis of contributory factors and recommendations for employee retention. In: IEEE. 2022 IEEE International conference for women in innovation, technology & entrepreneurship (ICWITE). s.l.: IEEE, 2022. Páginas 1–6.
- MOHAMAD, Mas Rahayu; NASARUDDIN, Fariza Hanum; HAMID, Suraya; BUKHARI, Sarah; IJAB, Mohamad Taha. Predicting employees' turnover in IT industry using classification method with feature selection. In: IEEE. 2021 International conference on computer science and engineering (IC2SE). s.l.: IEEE, 2021. volume 1, páginas 1–7.

MOHANTY, Sachi Nandan; CHATTERJEE, Jyotir Moy; JAIN, Sarika; ELNGAR, Ahmed A; GUPTA, Priya. **Recommender system with machine learning and artificial intelligence: Practical tools and applications in medical, agricultural and other industries.** s.l.: John Wiley & Sons, 2020.

MOHIUDDIN, Karishma; ALAM, Mirza Ariful; ALAM, Mirza Mohtashim; WELKE, Pascal; MARTIN, Michael; LEHMANN, Jens; VAHDATI, Sahar. Retention is all you need. In: PROCEEDINGS of the 32nd ACM International Conference on Information and Knowledge Management. s.l.: s.n., 2023. Páginas 4752–4758.

MONISAA THARANI, S K; VIVEK RAJ, S N. Predicting employee turnover intention in ITITeS industry using machine learning algorithms. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). s.l.: s.n., 2020. Páginas 508–513. DOI: [10.1109/I-SMAC49090.2020.9243552](https://doi.org/10.1109/I-SMAC49090.2020.9243552).

MUKHAMEDIEV, Ravil I; POPOVA, Yelena; KUCHIN, Yan; ZAITSEVA, Elena; KALIMOLDAYEV, Almas; SYMAGULOV, Adilkhan; LEVASHENKO, Vitaly; ABDOLDINA, Farida; GOPEJENKO, Viktors; YAKUNIN, Kirill et al. Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. **Mathematics**, MDPI, s.l., volume 10, número 15, página 2552, 2022.

NAKAMURA, Walter Takashi. **UX-MAPPER: A User Experience Method to Analyze App Store Reviews.** 2022. Tese (Doutorado) – Programa de Pós-graduação em Informática, Universidade Federal do Amazonas, Manaus, AM, Brasil.

NAPOLEÃO, Bianca; FELIZARDO, Katia Romero; SOUZA, Érica Ferreira de; VIJAYKUMAR, Nandamudi L. Practical similarities and differences between Systematic Literature Reviews and Systematic Mappings: a tertiary study. In: SEKE. s.l.: s.n., 2017. volume 2017, páginas 85–90.

NETO, Jorge Sukarie. **Brazilian Software Market | 2023 - Scenario & Trends.** s.l.: s.n., 2023. Disponível em: <https://abes.com.br/dados-do-setor/>. Acesso em: 22 de março de 2023.

_____. **Brazilian Software Market | 2024 - Scenario & Trends.** s.l.: s.n., 2024. Disponível em: <https://abes.com.br/dados-do-setor/>. Acesso em: 07 de abril de 2025.

_____. **Brazilian Software Market | 2025 - Scenario & Trends.** s.l.: s.n., 2025. Disponível em: <https://abes.com.br/dados-do-setor/>. Acesso em: 20 de maio de 2025.

- OGBEIBU, Samuel; CHIAPPETTA JABBOUR, Charbel Jose; BURGESS, John; GASKIN, James; RENWICK, Douglas WS. Green talent management and turnover intention: the roles of leader STARA competence and digital task interdependence. **Journal of Intellectual Capital**, Emerald Publishing Limited, volume 23, número 1, páginas 27–55, 2022.
- OMAR, Evi Diana; MAT, Hasnah; KARIM, Ainil Zafirah Abd; SANAUDI, Ridwan; IBRAHIM, Fairol H; OMAR, Mohd Azahadi; ISMAIL, Muhd Zulfadli Hafiz; JAYARAJ, Vivek Jason; GOH, Bak Leong. Comparative analysis of logistic regression, gradient boosted trees, SVM, and random forest algorithms for prediction of acute kidney injury requiring dialysis after cardiac surgery. **International Journal of Nephrology and Renovascular Disease**, Dove Medical Press, volume 17, páginas 197–204, 2024. DOI: [10.2147/IJNRD.S461028](https://doi.org/10.2147/IJNRD.S461028).
- OSMAN, Ahmedbahaaldin Ibrahem Ahmed; AHMED, Ali Najah; CHOW, Ming Fai; HUANG, Yuk Feng; EL-SHAFIE, Ahmed. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. **Ain Shams Engineering Journal**, Elsevier, volume 12, número 2, páginas 1545–1556, 2021. DOI: [10.1016/j.asej.2020.11.014](https://doi.org/10.1016/j.asej.2020.11.014).
- OWENS, Dawn; KHAZANCHI, Deepak. Best practices for retaining global IT talent. In: IEEE. 2011 44th Hawaii International Conference on System Sciences. s.l.: [sine nomine], 2011. Páginas 1–12.
- PAIGUDE, Supriya; PANGARKAR, Smita C; HUNDEKARI, Sheela; MALI, Manisha; WANJALE, Kirti; DONGRE, Yashwant. Potential of artificial intelligence in boosting employee retention in the human resource industry. **International Journal on Recent and Innovation Trends in Computing and Communication**, volume 11, 3s, páginas 01–10, 2023.
- PANDEY, Shweta; MEHTA, Jalpa. HRPA: Human resource prediction analytics. In: IEEE. 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT). s.l.: IEEE, 2022. Páginas 721–726.
- PAPINENI, Swarajya Lakshmi V; MALLIKARJUNA REDDY, A.; YARLAGADDA, Sudeepti; YARLAGADDA, Snigdha; AKKINENI, Haritha. An extensive analytical approach on human resources using random forest algorithm. **International Journal of Engineering Trends and Technology**, s.n., s.l., volume 69, número 5, páginas 119–127, 2021. Cited by: 24; All Open Access, Green Open Access. DOI: [10.14445/22315381/IJETT-V69I5P217](https://doi.org/10.14445/22315381/IJETT-V69I5P217). Disponível em:

<<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106929543&doi=10.14445%2f22315381%2fIJETT-V69I5P217&partnerID=40&md5=c5ab282bfae6cb84c9f047c3cff9df59>>.

PATIL, Aaditya; SARDA, Tejas; SHETYE, Snehal; BACHCHAN, Srishti; GUTTE, Vitthal S. Comparing Logistic Regression and Tree Models on HR Data. In: IEEE. 2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech). s.l.: IEEE, 2023. Páginas 671–677.

PEFFERS, Ken; TUUNANEN, Tuure; ROTHENBERGER, Marcus A; CHATTERJEE, Samir. A design science research methodology for information systems research. **Journal of management information systems**, Taylor & Francis, volume 24, número 3, páginas 45–77, 2007.

PEREIRA, Vijay; HADJIELIAS, Elias; CHRISTOFI, Michael; VRONTIS, Demetris. A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective. **Human Resource Management Review**, Elsevier, volume 33, número 1, página 100857, 2023.

PEREIRA; Daniel César. **Retenção de talentos: manual do participante**. s.l., 2013.

PETERSEN, Kai; VAKKALANKA, Sairam; KUZNIARZ, Ludwik. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and software technology**, Elsevier, volume 64, páginas 1–18, 2015.

PORKODI, S; SRIHARI, S; VIJAYAKUMAR, N. Talent management by predicting employee attrition using enhanced weighted forest optimization algorithm with improved random forest classifier. **International Journal of Advanced Technology and Engineering Exploration**, Accent Social e Welfare Society, s.l., volume 9, número 90, página 563, 2022.

_____. **International Journal of Advanced Technology and Engineering Exploration**, Scopus, s.l., volume 9, número 90, páginas 563–582, 2022. DOI:

[10.19101/IJATEE.2021.875340](https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136813409&doi=10.19101%2fIJATEE.2021.875340&partnerID=40&md5=b6cc71a1bfe7cf96a249f69fa1fc0). Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136813409&doi=10.19101%2fIJATEE.2021.875340&partnerID=40&md5=b6cc71a1bfe7cf96a249f69fa1fc0>>.

POURKHODABAKHSH, Nima; MAMOUDAN, Mobina Mousapour; BOZORGI-AMIRI, Ali. Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover. **Applied Intelligence**, Springer, s.l., volume 53, número 12, páginas 16309–16331, 2023.

- QAMAR, Yusra; AGRAWAL, Rakesh Kumar; SAMAD, Taab Ahmad; JABBOUR, Charbel Jose Chiappetta. When technology meets people: the interplay of artificial intelligence and human resource management. **Journal of Enterprise Information Management**, Emerald Publishing Limited, volume 34, número 5, páginas 1339–1370, 2021.
- RAJESWARI, G Raja; MURUGESAN, Rajkumar; ARUNA, R; JAYAKRISHNAN, Balaji; NILAVATHY, K. Predicting employee attrition through machine learning. In: IEEE. 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC). s.l.: IEEE, 2022. Páginas 1370–1379.
- RAMAMURTHY, Karthikeyan Natesan; SINGH, Moninder; YU, Yichong; ASPIS, Jessica; IAMES, Matthew; PERAN, Michael; HELD, Qin S. A talent management tool using propensity to leave analytics. In: IEEE. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). s.l.: IEEE, 2015. Páginas 1–10.
- RATTAN, Vikas; MITTAL, Ruchi; MALIK, Varun; SINGH, Jaiteg. Supervised machine learning approach for employee attrition analysis. In: AIP PUBLISHING, 1. AIP Conference Proceedings. s.l.: AIP Publishing, 2023. volume 2916.
- RAZA, Ali; MUNIR, Kashif; ALMUTAIRI, Mubarak; YOUNAS, Faizan; FAREED, Mian Muhammad Sadiq. Predicting Employee Attrition Using Machine Learning Approaches. **Applied Sciences**, MDPI, volume 12, número 13, página 6424, 2022.
- RÊGO, Gildygleide Cruz de Brito; LANZARINI, Ricardo; CLAUDINO, Adson de Lima; BARROS, Aline Gisele Azevedo Lima de. Índice de turnover em empresas organizadoras de eventos como diferencial competitivo de mercado. pt-BR. **Marketing & Tourism Review**, s.l., volume 8, número 2, 2022. DOI: [10.29149/mtr.v8i2.7359](https://doi.org/10.29149/mtr.v8i2.7359). Disponível em: <<https://revistas.face.ufmg.br/index.php/mtr/article/view/7359>>. Acesso em: 21 julho 2025.
- REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In _____. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, novembro 2019. Páginas 3982–3992. DOI: [10.18653/v1/D19-1410](https://doi.org/10.18653/v1/D19-1410). Disponível em: <<https://aclanthology.org/D19-1410/>>.
- RODRIGO, Dishala Sandamini; RATNAYAKE, Gayashini Shyanka. Employee Turnover Prediction System: With Special Reference to Apparel Industry in Sri Lanka. In: IEEE. 2021 6th International Conference for Convergence in Technology (I2CT). s.l.: IEEE, 2021. Páginas 1–9.

- ROY, Atin; CHAKRABORTY, Subrata. Support vector machine in structural reliability analysis: A review. **Reliability Engineering System Safety**, Elsevier, volume 233, página 109126, 2023. DOI: [10.1016/j.ress.2023.109126](https://doi.org/10.1016/j.ress.2023.109126).
- RUSSELL, Stuart J. Rationality and intelligence. **Artificial intelligence**, Elsevier, volume 94, número 1-2, páginas 57–77, 1997.
- SARAN, Nurdan Ayse; NAR, Fatih. Fast binary logistic regression. **PeerJ Computer Science**, PeerJ, volume 11, e2579, 2025. DOI: [10.7717/peerj-cs.2579](https://doi.org/10.7717/peerj-cs.2579).
- SARDAR, Devadarshan K; CHOURASIYA, Sidhant; VIJYALAKSHMI, V. Employee Turnover Prediction by Machine Learning Techniques. In: IEEE. 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT). s.l.: IEEE, 2023. Páginas 265–272.
- SCHMIDT, Albrecht; ALT, Florian; SHIRAZI, Alireza Sahami. Interacting with 21st-Century Computers. **IEEE Pervasive Computing**, IEEE, s.l., volume 11, número 1, páginas 22–31, 2012. DOI: [10.1109/MPRV.2012.3](https://doi.org/10.1109/MPRV.2012.3).
- SCHOBER, Patrick; BOER, Christa; SCHWARTE, Lothar A. Correlation Coefficients: Appropriate Use and Interpretation. **Anesthesia Analgesia**, Lippincott Williams Wilkins, volume 126, número 5, páginas 1763–1768, 2018. DOI: [10.1213/ANE.0000000000002864](https://doi.org/10.1213/ANE.0000000000002864).
- SCIENCES ENGINEERING, National Academies of; MEDICINE. **Information Technology Innovation: Resurgence, Confluence, and Continuing Impact**. s.l.: National Academies Press, 2020.
- SEELAM, Srinivasa Reddy; KUMAR, Kandula Hemanth; SUPRITHA, Mutyala Sai; GNANESWAR, Gaddam; REDDY, Vuyyuru Venu Madhav. Comparative study of predictive models to estimate employee attrition. In: IEEE. 2022 7th International conference on communication and electronics systems (ICCES). s.l.: IEEE, 2022. Páginas 1602–1607.
- SEIXAS, Elaine F Rangel; VITERBO, José; BERNARDINI, Flavia; SEIXAS, Flávio; PANTOJA, Carlos. Applying artificial intelligence for talent retention: a systematic literature review. In: IEEE. 2023 18th Iberian Conference on Information Systems and Technologies (CISTI). Aveiro, Portugal: [sine nomine], 2023. Páginas 1–6.
- SERVIÇO BRASILEIRO DE APOIO ÀS MICRO E PEQUENAS EMPRESAS; DEPARTAMENTO INTERSINDICAL DE ESTATÍSTICA E ESTUDOS SOCIOECONÔMICOS. **Anuário do trabalho nos Pequenos Negócios: 2018**. 11. edição. Brasília, DF: DIEESE, 2020. Publicação institucional.

- SETOR, Tenace; JOSEPH, Damien. When agile means staying: The relationship between agile development usage and individual IT professional outcomes. In: PROCEEDINGS of the 2019 on Computers and People Research Conference. s.l.: s.n., 2019. Páginas 168–175.
- SHANKAR, Shardul; VYAS, Ranjana; TEWARI, Vijayshri. Applying machine learning algorithms to determine and predict the reasons and models for employee turnover. **International Journal of Information Technology and Management**, Inderscience Publishers (IEL), s.l., volume 23, número 1, páginas 48–63, 2024.
- SHEN, Dazhong; ZHU, Hengshu; XIAO, Keli; ZHANG, Xi; XIONG, Hui. Exploiting connections among personality, job position, and work behavior: Evidence from joint Bayesian learning. **ACM Transactions on Management Information Systems**, ACM New York, NY, volume 14, número 3, páginas 1–20, 2023.
- SHI, Yong. A Machine Learning Study on the Model Performance of Human Resources Predictive Algorithms. In: IEEE. 2022 4th International Conference on Applied Machine Learning (ICAML). s.l.: IEEE, 2022. Páginas 405–409.
- SHINDE, Pramila P; SHAH, Seema. A review of machine learning and deep learning applications. In: IEEE. 2018 Fourth international conference on computing communication control and automation (ICCUBEA). s.l.: IEEE, 2018. Páginas 1–6.
- SHOKRZADE, Amin; RAMEZANI, Mohsen; TAB, Fardin Akhlaghian; MOHAMMAD, Mahmud Abdulla. A novel extreme learning machine based kNN classification method for dealing with big data. **Expert Systems with Applications**, Elsevier, volume 183, página 115293, 2021.
- SILVA, Williamson; VALENTIM, NM Costa; CONTE, Tayana. Integrating the Usability into the Software Development Process. In: PROCEEDINGS of the 17th International Conference on Enterprise Information Systems. s.l.: s.n., 2015. volume 3, páginas 105–113.
- SISODIA, Dilip Singh; VISHWAKARMA, Somdutta; PUJAHARI, Abinash. Evaluation of machine learning models for employee churn prediction. In: 2017 International Conference on Inventive Computing and Informatics (ICICI). s.l.: s.n., 2017. Páginas 1016–1020.
- SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, Elsevier, volume 45, número 4, páginas 427–437, 2009. DOI: [10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002).
- SOLANO-BARLIZA, Andrés; ARREGOCÉS-JULIO, Isabel; AARÓN-GONZALVEZ, Marlin; ZAMORA-MUSA, Ronald; DE-LA-HOZ-FRANCO, Emiro; ESCORCIA-GUTIERREZ, José; ACOSTA-COLL, Melisa.

Recommender systems applied to the tourism industry: a literature review. **Cogent Business & Management**, Taylor & Francis, s.l., volume 11, número 1, página 2367088, 2024.

SPERANDEI, Sandro. Understanding logistic regression analysis. **Biochimia medica**, Medicinska naklada, volume 24, número 1, páginas 12–18, 2014.

SRIVASTAVA, Praveen Ranjan; EACHEMPATI, Prajwal. Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach. **Journal of Global Information Management (JGIM)**, IGI Global, volume 29, número 6, páginas 1–29, 2021.

STROHMEIER, Stefan; PIAZZA, Franca. Artificial intelligence techniques in human resource management—a conceptual exploration. **Intelligent Techniques in Engineering Management: Theory and Applications**, Springer, páginas 149–172, 2015.

SUN, Ying; ZHUANG, Fuzhen; ZHU, Hengshu; SONG, Xin; HE, Qing; XIONG, Hui. Modeling the impact of person-organization fit on talent management with structure-aware attentive neural networks. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, volume 35, número 3, páginas 2809–2822, 2021.

TANG, Fei; ISHWARAN, Hemant. Random forest missing data algorithms. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, Wiley Online Library, volume 10, número 6, páginas 363–377, 2017.

TENG, Mingfei; ZHU, Hengshu; LIU, Chuanren; ZHU, Chen; XIONG, Hui. Exploiting the contagious effect for employee turnover prediction. In: 01. PROCEEDINGS of the AAAI conference on artificial intelligence. s.l.: s.n., 2019. volume 33, páginas 1166–1173.

THARWAT, Alaa. Classification assessment methods. **Applied Computing and Informatics**, Emerald Publishing Limited, volume 17, número 1, páginas 168–192, 2021. DOI: [10.1016/j.aci.2018.08.003](https://doi.org/10.1016/j.aci.2018.08.003).

TOTVS, Equipe. **O que é turnover? Guia completo sobre o índice de rotatividade, suas causas e como reduzi-lo!** s.l.: s.n., 2022. Disponível em: <https://www.totvs.com/blog/gestao-para-recursos-humanos/o-que-e-turnover/>. Acesso em: 22 de março de 2023.

TROYANSKAYA, Olga; CANTOR, Michael; SHERLOCK, Gavin; BROWN, Patrick; HASTIE, Trevor; TIBSHIRANI, Robert; BOTSTEIN, David; ALTMAN, Russ B. Missing value estimation methods for DNA microarrays. **Bioinformatics**, Oxford University Press, volume 17, número 6, páginas 520–525, 2001.

- VAN ENGELEN, Jesper E; HOOS, Holger H. A survey on semi-supervised learning. **Machine learning**, Springer, volume 109, número 2, páginas 373–440, 2020.
- VAN SOLINGEN, Rini; BASILI, Vic; CALDIERA, Gianluigi; ROMBACH, H Dieter. Goal question metric (gqm) approach. **Encyclopedia of software engineering**, Wiley Online Library, 2002.
- VARSHNEY, Kush R; CHENTHAMARAKSHAN, Vijil; FANCHER, Scott W; WANG, Jun; FANG, Dongping; MOJSILOVIĆ, Aleksandra. Predicting employee expertise for talent management in the enterprise. In: PROCEEDINGS of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. s.l.: s.n., 2014. Páginas 1729–1738.
- VEGLIO, Valerio; ROMANELLO, Rubina; PEDERSEN, Torben. Employee turnover in multinational corporations: a supervised machine learning approach. **Review of Managerial Science**, Springer, volume 19, número 3, páginas 687–728, 2025.
- VELOSO, Elza Fatima Rosa. **Retenção de talentos: O que é, Importância e Dicas**. [Sine loco: sine nomine], 2019. Disponível em: <https://fia.com.br/blog/retencao-de-talentos/>. Acesso em: 13 de março de 2023.
- VELUCHAMY, Ramar; SANCHARI, C; GUPTA, S. Artificial intelligence within recruitment: eliminating biases in human resource management. **Artificial intelligence**, volume 8, número 3, 2021.
- VOTTO, Alexis Megan; VALECHA, Rohit; NAJAFIRAD, Peyman; RAO, H Raghav. Artificial intelligence in tactical human resource management: A systematic literature review. **International Journal of Information Management Data Insights**, Elsevier, volume 1, número 2, página 100047, 2021.
- WANG, Donghui; LIANG, Yanchun; XU, Dong; FENG, Xiaoyue; GUAN, Renchu. A content-based recommender system for computer science publications. **Knowledge-Based Systems**, Elsevier, volume 157, páginas 1–9, 2018.
- WANG, T; CHEN, G. Analyzing employee turnover based on job skills. In: ZNPROCEEDINGS of the International Conference on Data Processing and Applications, i6. [Sine loco: sine nomine], 2018.
- WANG, Xinlei; ZHI, Jianing. A machine learning-based analytical framework for employee turnover prediction. **Journal of Management Analytics**, Taylor & Francis, volume 8, número 3, páginas 351–370, 2021.
- WEST, Darrell M; ALLEN, John R. How artificial intelligence is transforming the world. **Report. April**, volume 24, página 2018, 2018.

WIERINGA, Roel. **Design science methodology for information systems and software engineering**. s.l.: Springer, 2014.

WILES, Jackie. **4 Ways to Attract and Retain Talent That Aren't Just About Comp**. [Sine loco: sine nomine], 2022. Disponível em: <https://www.gartner.com/en/articles/4-ways-to-attract-and-retain-talent-that-aren-t-just-about-comp>. Acesso em: 13 de março de 2023.

XUECHENG, Wei; IQBAL, Qaisar; SAINA, Bai. Factors Affecting Employee's Retention: Integration of Situational Leadership With Social Exchange Theory. **Frontiers in Psychology**, Frontiers Media SA, volume 13, 2022.

YADAV, Sandeep; JAIN, Aman; SINGH, Deepti. Early Prediction of Employee Attrition using Data Mining Techniques. In: 2018 IEEE 8th International Advance Computing Conference (IACC). s.l.: IEEE, 2018. Páginas 349–354. DOI: [10.1109/IADCC.2018.8692137](https://doi.org/10.1109/IADCC.2018.8692137).

YAHIA, Nesrine Ben; HLEL, Jihen; COLOMO-PALACIOS, Ricardo. From big data to deep data to support people analytics for employee attrition prediction. **Ieee Access**, Ieee, volume 9, páginas 60447–60458, 2021.

YUAN, Jia. Research on employee turnover prediction based on machine learning algorithms. In: IEEE. 2021 4th international conference on artificial intelligence and big data (icaibd). s.l.: IEEE, 2021. Páginas 114–120.

YUNMENG, Zhang; CHENGYI, Zhang. The Application of the Decision Tree Algorithm Based on K-means in Employee Turnover Prediction. In: 1. S.I. s.l.: Scopus, 2019. volume 1325. DOI: [10.1088/1742-6596/1325/1/012123](https://doi.org/10.1088/1742-6596/1325/1/012123). Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85075825026&doi=10.1088%2f1742-6596%2f1325%2f1%2f012123&partnerID=40&md5=3a3f322c0f5baadce93ea7a9a3a54852>>.

ZHANG, Shichao. Challenges in KNN classification. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, volume 34, número 10, páginas 4663–4675, 2021.

ZHANG, Shichao; LI, Jiaye. KNN classification with one-step computation. **IEEE Transactions on Knowledge and Data Engineering**, Ieee, volume 35, número 3, páginas 2711–2723, 2021.

ZHAO, Yue; HRYNIEWICKI, Maciej K.; CHENG, Francesca; FU, Boyang; ZHU, Xiaoyu. Employee turnover prediction with machine learning: A reliable approach. **Advances in Intelligent Systems and Computing**, Scopus, s.l., volume 869, páginas 737–758, 2018. DOI: [10.1007/978-3-030-01057-7_56](https://doi.org/10.1007/978-3-030-01057-7_56). Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060456909&doi=10.1007%2f978-3-030-01057-7_56&partnerID=40&md5=3914441d9acb14a7dd38a68064e3e020>.

ZHU, Qianwen; SHANG, Jiaxing; CAI, Xinjun; JIANG, Linli; LIU, Feiyi; QIANG, Baohua. CoxRF: Employee Turnover Prediction Based on Survival Analysis. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. s.l.: s.n., 2019. Páginas 1123–1130. DOI: [10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00212](https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00212).

ŽIVKOVIĆ, Ana; ŠEBALJ, Dario; FRANJKOVIĆ, Jelena. Prediction of the Employee Turnover Intention Using Decision Trees. In: 26TH International Conference on Enterprise Information Systems. s.l.: s.n., 2024. Páginas 325–336.

APÊNDICE A – Tabela de aspectos relevantes para migração ou retenção de talentos

Tabela 16: Aspectos relevantes para migração ou retenção de talentos verificados no MSL

Aspectos relevantes para retenção ou migração de talentos
Promoção de Carreira (Plano de Carreira, Oportunidades de Carreira, Nível de Orientação de Carreira, Esperança de Carreira, Salário/Remuneração);
Gestão de Recompensas (Regalias e comissões, Bônus, Abonos);
Treinamento & Desenvolvimento (Oportunidade de Desenvolvimento Profissional, Oportunidade de Desenvolvimento Pessoal);
Estilo de Gestão (Estilo de Liderança, Orientação para o Colaborador, Orientação para o Trabalho);
Desafio insuficiente (suporte organizacional, variedade, inovação, experimentação, oportunidade de aprimoramento de habilidades);
Condições de Serviço (Termos e Condições, Regras e regulamentos rígidos, Sentimento de Insegurança, Desinteresse, Desvinculação do trabalho, Desvinculação da organização, Desvinculação entre colegas);
Flexibilidade nos Horários de Trabalho (Horário rígido, Longo horário de trabalho, Excesso de trabalho, Falta de descanso);
Condição de trabalho (Falta de atividades de apoio entre colegas, Questão de saúde física, Uso de produtos químicos perigosos, Trabalho de longa duração, Falta de descanso);
Localização (viagem longa para chegar ao trabalho, sem facilidade de transporte da organização);
Ambiente de trabalho;
Satisfação no trabalho;

Continua na próxima página

Tabela 16 – *Continuação da página anterior*

Aspectos relevantes para retenção ou migração de talentos
Interdependência de tarefa digital;
Percepção de suporte organizacional;
Percepção de alternativas de oportunidade de trabalho;
Compromisso organizacional;
Trabalho com metodologia ágil;
Políticas de emprego;
Equilíbrio entre vida profissional e pessoal;
Inovação no local de trabalho;
Cultura da empresa;
Tempo de empresa;
Dados Demográficos (Gênero, etnia, cidade, índice de desenvolvimento da cidade, nível educacional, área de formação, nível de experiência, tipo de universidade, tempo de experiência, tamanho da empresa, tipo de empresa, diferença em tempo do último emprego para atual);
Nível de emprego;
Desempenho do 1º ano;
Satisfação com a Carreira;
Green hard Talent Management;
Green soft Talent Management.

APÊNDICE B – Relação de estudos primários selecionados na RSL

Tabela 17: Estudos relevantes selecionados

ID	Título	Citação
E _{RSL} 1	<i>A data-driven risk measurement model of software developer turnover</i>	(MA et al., 2020)
E _{RSL} 2	<i>A machine learning approach for employee retention prediction</i>	(MARVIN; JACKSON; ALAM, 2021)
E _{RSL} 3	<i>A Machine Learning Study on the Model Performance of Human Resources Predictive Algorithms</i>	(SHI, 2022)
E _{RSL} 4	<i>A machine learning-based analytical framework for employee turnover prediction</i>	(WANG; ZHI, 2021)
E _{RSL} 5	<i>A Supervised Machine Learning Method for Predicting the Employees Turnover Decisions</i>	(ARROMRIT et al., 2023)
E _{RSL} 6	<i>A Talent Management Tool using Propensity to Leave Analytics</i>	(RAMAMURTHY et al., 2015)
E _{RSL} 7	<i>An extensive analytical approach on human resources using random forest algorithm</i>	(PAPINENI et al., 2021)
E _{RSL} 8	<i>Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making</i>	(MARÍN DÍAZ; GALÁN HERNÁNDEZ; GALDÓN SALVADOR, 2023)
E _{RSL} 9	<i>Analyzing Employee Attrition Using Machine Learning: the New AI Approach</i>	(KRISHNA; SIDHARTH, 2022a)
E _{RSL} 10	<i>Analyzing Employee Turnover Based on Job Skills</i>	(WANG; CHEN, 2018)
E _{RSL} 11	<i>Applying machine learning algorithms to determine and predict the reasons and models for employee turnover</i>	(SHANKAR; VYAS; TEWARI, 2024)
E _{RSL} 12	<i>Artificial Intelligence-Enabled Knowledge Management Using a Multidimensional Analytical Framework of Visualizations</i>	(BHUPATHI; PRABU; GOH, 2023)
E _{RSL} 13	<i>Comparative Study of Predictive Models to Estimate Employee Attrition</i>	(SEELAM et al., 2022)

Continua na próxima página

Tabela 17 – Continuação da página anterior

ID	Título	Citação
<i>E_{RSL} 14</i>	<i>Comparing Logistic Regression and Tree Models on HR Data</i>	(PATIL et al., 2023)
<i>E_{RSL} 15</i>	<i>CoxRF: Employee Turnover Prediction Based on Survival Analysis</i>	(ZHU et al., 2019)
<i>E_{RSL} 16</i>	<i>Early Prediction of Employee Attrition using Data Mining Techniques</i>	(YADAV; JAIN; SINGH, 2018)
<i>E_{RSL} 17</i>	<i>Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover</i>	(POURKHODABAKHSH; MAMOUDAN; BOZORGIAMIRI, 2023)
<i>E_{RSL} 18</i>	<i>Employee Attrition: Prediction, Analysis Of Contributory Factors And Recommendations For Employee Retention</i>	(MITRAVINDA; SHETTY, 2022)
<i>E_{RSL} 19</i>	<i>Employee turnover in multinational corporations: a supervised machinelearning approach</i>	(VEGLIO; ROMANELLO; PEDERSEN, 2025)
<i>E_{RSL} 20</i>	<i>Employee Turnover Prediction by Machine Learning Techniques</i>	(SARDAR; CHOURASIYA; VIJYALAKSHMI, 2023)
<i>E_{RSL} 21</i>	<i>Employee turnover prediction with machine learning: A reliable approach</i>	(ZHAO et al., 2018)
<i>E_{RSL} 22</i>	<i>Evaluation of Machine Learning Models for Employee Churn Prediction</i>	(SISODIA; VISHWAKARMA; PUJAHARI, 2017)
<i>E_{RSL} 23</i>	<i>Exploiting Connections among Personality, Job Position, and Work Behavior: Evidence from Joint Bayesian Learning</i>	(SHEN et al., 2023)
<i>E_{RSL} 24</i>	<i>Exploiting linkedin to predict employee resignation likelihood</i>	(JESUS; JÚNIOR; BRANDÃO, 2018)
<i>E_{RSL} 25</i>	<i>Exploiting the Contagious Effect for Employee Turnover Prediction</i>	(TENG et al., 2019)
<i>E_{RSL} 26</i>	<i>Exploratory Data Analysis and Classification of Employee Retention based on Logistic Regression Model</i>	(BONGALE; DHARRAO; UROLAGIN, 2023)
<i>E_{RSL} 27</i>	<i>From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction</i>	(YAHIA; HLEL; COLOMO-PALACIOS, 2021)
<i>E_{RSL} 28</i>	<i>HR Analytics: Employee Attrition Analysis using Random Forest</i>	(KRISHNA; SIDHARTH, 2022b)
<i>E_{RSL} 29</i>	<i>HRPA: Human Resource Prediction Analytics</i>	(PANDEY; MEHTA, 2022)
<i>E_{RSL} 30</i>	<i>Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-CriteriaDecision-Making Approach</i>	(SRIVASTAVA; EACHEMPATI, 2021)
<i>E_{RSL} 31</i>	<i>Job satisfaction and turnover decision of employees in the Internet sector in the US</i>	(CHANG et al., 2023)

Continua na próxima página

Tabela 17 – Continuação da página anterior

ID	Título	Citação
<i>E_{RSL} 32</i>	<i>Machine Learning in HR Analytics: A Comparative Study on the Predictive Accuracy of Attrition Models</i>	(BASHA et al., 2024)
<i>E_{RSL} 33</i>	<i>Machine Learning Predictive Models for preventing Employee Turnover costs</i>	(COSTA; CARVALHO, 2022)
<i>E_{RSL} 34</i>	<i>Modeling the Impact of Person-Organization Fit on Talent Management WithStructure-Aware Attentive Neural Networks</i>	(SUN et al., 2021)
<i>E_{RSL} 35</i>	<i>Potential of Artificial Intelligence in Boosting Employee Retention in the Human Resource Industry</i>	(PAIGUDE et al., 2023)
<i>E_{RSL} 36</i>	<i>Predicting Employee Attrition through Machine Learning</i>	(RAJESWARI et al., 2022)
<i>E_{RSL} 37</i>	<i>Predicting Employee Attrition Using Machine Learning Techniques</i>	(FALLUCCHI et al., 2020)
<i>E_{RSL} 38</i>	<i>Predicting Employee Attrition using Supervised Learning Classification Models</i>	(HABOUS; OUBENAALLA et al., 2021)
<i>E_{RSL} 39</i>	<i>Predicting employee turnover intention in IT & ITeS industry using machine learning algorithms</i>	(MONISAA THARANI; VIVEK RAJ, 2020)
<i>E_{RSL} 40</i>	<i>Predicting Employees' Turnover in IT Industry using Classification Method with Feature Selection</i>	(MOHAMAD et al., 2021)
<i>E_{RSL} 41</i>	<i>Prediction of the Employee Turnover Intention Using Decision Trees</i>	(ŽIVKOVIĆ; ŠEBALJ; FRANJKOVIĆ, 2024)
<i>E_{RSL} 42</i>	<i>Predictive Analytics for Employee Attrition: A Comparative Study of Machine Learning Algorithms</i>	(KRISHNA; DWIVEDI; MURTI, 2023)
<i>E_{RSL} 43</i>	<i>Predictive model of employee attrition based on stacking ensemble learning</i>	(CHUNG et al., 2023)
<i>E_{RSL} 44</i>	<i>Research on Employee Turnover Prediction Based on Machine Learning Algorithms</i>	(YUAN, 2021)
<i>E_{RSL} 45</i>	<i>Retention Is All You Need</i>	(MOHIUDDIN et al., 2023)
<i>E_{RSL} 46</i>	<i>RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis</i>	(JIN et al., 2020)
<i>E_{RSL} 47</i>	<i>Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches</i>	(PORKODI; SRIHARI; VIJAYAKUMAR, 2022)
<i>E_{RSL} 48</i>	<i>Supervised Machine Learning Approach for Employee Attrition Analysis</i>	(RATTAN et al., 2023)
<i>E_{RSL} 49</i>	<i>Talent management by predicting employee attrition using enhanced weighted forest optimization algorithm with improved random forest classifier</i>	(PORKODI, S.; SRIHARI, S.; VIJAYAKUMAR, N., 2022)

Continua na próxima página

Tabela 17 – Continuação da página anterior

ID	Título	Citação
$E_{RSL} 50$	<i>The Application of the Decision Tree Algorithm Based on K-means in Employee Turnover Prediction</i>	(YUNMENG; CHENGYI, 2019)

APÊNDICE C – Formulário de avaliação de qualidade da RSL

Tabela 18: Formulário de avaliação de qualidade da RSL

#	Questão	Avaliação
1.	Os objetivos da pesquisa são apresentados de forma clara e coerente, com uma justificativa adequada para o uso de IA na retenção de talentos?	() Sim () Parcialmente () Não
2.	O estudo utiliza uma metodologia empírica robusta (por exemplo, experimentos, estudos de caso ou análise de dados reais) para demonstrar a eficácia da IA na retenção de talentos?	() Sim () Parcialmente () Não
3.	A metodologia é claramente descrita e apropriada para o contexto de retenção de talentos?	() Sim () Parcialmente () Não
4.	O estudo considera a replicabilidade dos resultados (ou a possibilidade de generalização para outros contextos)?	() Sim () Parcialmente () Não
5.	A coleta de dados é bem detalhada, com informações sobre a origem, tamanho da amostra e processos de coleta?	() Sim () Parcialmente () Não
6.	O estudo propõe soluções ou recomendações aplicáveis que podem ser implementadas no contexto real?	() Sim () Parcialmente () Não
7.	Os resultados são claramente apresentados, com dados quantitativos e/ou qualitativos suficientes para sustentar as conclusões?	() Sim () Parcialmente () Não
8.	O artigo apresenta métricas claras de sucesso ou avaliação da previsão de turnover ou da retenção de talentos?	() Sim () Parcialmente () Não
9.	O estudo aborda de forma clara as limitações e possíveis vieses relacionados ao uso de IA ou da retenção de talentos?	() Sim () Parcialmente () Não

Continua na próxima página

Tabela 18 – *Continuação da página anterior*

#	Questão	Avaliação
10.	O estudo especifica claramente como a IA foi utilizada (ex.: aprendizado de máquina, processamento de linguagem natural)?	() Sim () Parcialmente () Não
11.	O estudo explica claramente todas as etapas, desde a coleta até a análise dos dados, garantindo que não haja lacunas no processo?	() Sim () Parcialmente () Não
12.	A coleta e a análise de dados estão em conformidade com os princípios éticos? Há uma declaração de que seguiu princípios éticos?	() Sim () Parcialmente () Não
13.	O estudo descreve claramente o contexto da aplicação (por exemplo, setor de TI, tamanho da empresa) e sua relevância para a previsão de turnover ou retenção de talentos?	() Sim () Parcialmente () Não