UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS ARARÁNGUA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

Etson dos Santos Rodrigues

Método preditivo de desligamentos pré-breakeven no setor tecnológico via aprendizado de máquina

Araranguá
2026

Etson dos Santos Rodrigues

Método preditivo de desligamentos pré-breakeven no setor tecnológico via aprendizado de máquina

Dissertação submetida ao Programa de Pós Graduação em Tecnologias da Informação e Comunicação da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Mestre em Tecnologias da Informação e Comunicação.
Orientador: Prof. Roderval Marcelino, Dr.
Coorientador: Prof. Antonio Carlos Sobieranski, Dr.

Araranguá
2026

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Rodrigues, Etson dos Santos
Método preditivo de desligamentos pré-breakeven no setor tecnológico via aprendizado de máquina / Etson dos Santos Rodrigues ; orientador, Roderval Marcelino, coorientador, Antonio Carlos Sobieranski, 2026.
131 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Campus Araranguá, Programa de Pós-Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2026.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. People Analytics. 3. Machine Learning. 4. Turnover. 5. Breakeven. I. Marcelino, Roderval. II. Sobieranski, Antonio Carlos. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Tecnologias da Informação e Comunicação. IV. Título.

Etson dos Santos Rodrigues

Método preditivo de desligamentos pré-breakeven no setor tecnológico via aprendizado de máquina

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 26 de fevereiro de 2026, pela banca examinadora composta pelos seguintes membros:

Prof. Vilson Gruber, Dr.
Universidade Federal de Santa Catarina

Prof. Alexandre Leopoldo Gonçalves, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Tecnologias da Informação e Comunicação

ICP-Edu
Documento assinado digitalmente
Cristian Cechinel
Data: 17/03/2026 11:15:46-0300
CPF: ***.628.449-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Cristian Cechinel, Dr.
Coordenador do Programa

ICP-Edu
Documento assinado digitalmente
Roderval Marcelino
Data: 17/03/2026 11:42:16-0300
CPF: ***.867.489-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Roderval Marcelino, Dr.
Orientador

Araranguá, 2026

Aos meus pais, à minha esposa e à minha filha. Amo vocês.

AGRADECIMENTOS

Aos meus pais, Benedito Rodrigues e Zoleide dos Santos, minha eterna gratidão pela base sólida que construíram e pela segurança que sempre me forneceram ao longo de toda a minha vida. Vocês foram o alicerce fundamental para que eu pudesse trilhar este caminho com confiança.

À minha esposa, Vanessa Goulart, e à minha amada filha, Celinne Rodrigues, pelo amor incondicional. Agradeço imensamente pela paciência e pela compreensão diante das minhas ausências e das muitas horas de convívio familiar abdicadas em prol da conclusão desta etapa. Vocês são a minha maior motivação.

Ao meu orientador, Prof. Dr. Roderval Marcelino, expresso meu sincero agradecimento pelo suporte acadêmico, pela paciência durante todo o processo de orientação e, acima de tudo, por ter aceitado o desafio de guiar esta pesquisa. Seus ensinamentos foram cruciais para o desenvolvimento deste trabalho.

A esta universidade, ao seu corpo docente, à direção e à administração, agradeço pela excelência no ensino, pela infraestrutura disponibilizada e por oportunizarem a realização deste trabalho.

“Sem dados, você é apenas mais uma pessoa com uma opinião.”
(frase amplamente atribuída a W. Edwards Deming, autoria não comprovada)

RESUMO

A presente dissertação investiga o fenômeno da rotatividade voluntária (*turnover*) em organizações de tecnologia, propondo uma abordagem inovadora de *people analytics* orientada pela sustentabilidade financeira. O objetivo central consistiu no desenvolvimento e validação de um modelo preditivo focado na identificação de desligamentos que ocorrem antes do ponto de equilíbrio econômico (*breakeven*) do colaborador, transcendendo as métricas generalistas de retenção. A método adotada integrou um inventário diversificado de atributos validado em ampla revisão bibliográfica e variáveis financeiras e contratuais inéditas, estruturando um *dataset* sintético enriquecido e aderente ao contexto de empresas brasileiras de base tecnológica. O estudo empírico comparou dois cenários de modelagem, um focado no risco *pré-breakeven* e outro generalista, submetendo-os a algoritmos de aprendizado de máquina supervisionado, sendo eles *logistic regression, random forest, lightGBM, k-neighbors, support vector classification e multi-layer perceptron* com otimização de hiperparâmetros voltada à maximização da métrica *recall*. Os resultados evidenciaram a supremacia das arquiteturas de *ensemble learning*, notadamente o *random forest*, que obteve um *recall* médio de 0,8353 no cenário focado no risco *pré-breakeven*, superando a abordagem generalista e demonstrando a incapacidade de modelos lineares em capturar a não linearidade do fenômeno. Conclui-se que a integração de dados financeiros à modelagem de comportamento humano refina a capacidade preditiva, permitindo que as organizações priorizem estratégias de retenção onde o passivo de investimento é crítico, otimizando a alocação de recursos corporativos.

Palavras-chave: *People Analytics. Machine Learning. Turnover. Breakeven.*

ABSTRACT

This dissertation investigates the phenomenon of voluntary turnover in technology organizations, proposing a novel people analytics approach driven by financial sustainability. The core objective consisted of developing and validating a predictive model aimed at identifying departures occurring prior to the employee's economic breakeven point, transcending generalist retention metrics. The adopted method integrated a diverse inventory of attributes validated through extensive literature review with novel financial and contractual variables, structuring an enriched synthetic dataset tailored to the context of Brazilian technology-based companies. The empirical study compared two modeling scenarios, one focused on pre-breakeven risk and a generalist one, subjecting them to supervised machine learning algorithms, namely Logistic Regression, Random Forest, LightGBM, K-Neighbors, Support Vector Classification, and Multi-Layer Perceptron, with hyperparameter optimization aimed at maximizing the recall metric. Results evidenced the superiority of ensemble learning architectures, notably Random Forest, which achieved an average recall of 0.8353 in the pre-breakeven risk scenario, outperforming the generalist approach and demonstrating the inability of linear models to capture the phenomenon's non-linearity. It is concluded that integrating financial data into human behavior modeling enhances predictive capacity, allowing organizations to prioritize retention strategies where the investment liability is critical, thereby optimizing corporate resource allocation.

Keywords: People Analytics; Machine Learning; Turnover; Breakeven.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AED	Análise Exploratória de Dados
ALE	*Accumulated Local Effects*
ALM	*Application Lifecycle Management*
AUC	*Area Under the Curve*
B/E	*Breakeven*
BI	*Business Intelligence*
CEO	*Chief Executive Officer*
CHRO	*Chief Human Resources Officer*
CPF	Cadastro de Pessoas Físicas
CTGAN	*Conditional Generative Adversarial Networks*
DOI	*Digital Object Identifier*
ERP	*Enterprise Resource Planning*
FN	*False Negatives*
FP	*False Positives*
GUID	*Globally Unique Identifier*
HCM	*Human Capital Management*
HR	*Human Resources*
HRIS	*Human Resource Information System*
IA	Inteligência Artificial
IEEE	*Institute of Electrical and Electronics Engineers*
KNN	*K-Nearest Neighbors*
KPI	*Key Performance Indicator*
LGBM	*Light Gradient-Boosting Machine*
LGPD	Lei Geral de Proteção de Dados
LLM	Large Language Models
MI	*Management Information*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*
PDI	Plano de Desenvolvimento Individual
PIB	Produto Interno Bruto
PME	Pequenas e Médias Empresas
PPGTIC	Programa de Pós-Graduação em Tecnologias da Informação e Comunicação
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
QA	*Quality Assurance*
RF	*Random Forest*
RFE	*Recursive Feature Elimination*
RGPD	Regulamento Geral sobre a Proteção de Dados
RH	Recursos Humanos
RNA	Redes Neurais Artificiais
ROC	*Receiver Operating Characteristic*
RSL	Revisão Sistemática da Literatura
SHAP	*SHapley Additive exPlanations*
SIRH	Sistema de Informação de Recursos Humanos
SMOTE	*Synthetic Minority Over-sampling Technique*
STEM	*Science, Technology, Engineering and Mathematics*
SVC	*Support Vector Classification*
SVM	*Support Vector Machine*

SVR	*Support Vector Regression*
TI	Tecnologia da Informação
TN	*True Negatives*
TP	*True Positives*

SUMÁRIO

1 INTRODUÇÃO

A crescente disponibilidade de dados nas organizações tem impulsionado a adoção de técnicas analíticas avançadas para a extração de conhecimento acionável e o direcionamento de decisões estratégicas. Entretanto, a aplicação de tais técnicas analíticas a conjuntos de dados organizacionais, especificamente no setor tecnológico, requer métodos cuidadosamente adaptados às nuances desse ambiente intrinsecamente dinâmico.

Este dinamismo é acentuado por um hiato estrutural entre a demanda por competências digitais e a oferta de profissionais qualificados no mercado brasileiro. Segundo o Google for Startups (2022), foi estimado que, entre 2021 e 2025, a demanda projetada atingiria 800 mil novos postos de trabalho, enquanto o sistema educacional graduaria apenas 53 mil indivíduos anualmente, resultando em um déficit acumulado superior a 530 mil especialistas no período. Tal assimetria impõe riscos severos ao crescimento econômico, posicionando o Brasil como o terceiro país entre os membros do G20 com maior perda de oportunidade de aumento do produto interno bruto (PIB) derivada da carência de habilidades tecnológicas (GOOGLE FOR STARTUPS, 2022). Atualmente, cerca de 80% dos empregadores brasileiros reportam dificuldades severas no preenchimento de vagas, refletindo um cenário onde a escassez global de talentos atingiu seu maior nível em 17 anos (MANPOWERGROUP, 2023).

No âmbito operacional, o custo da rotatividade é intensificado pela extensão do ciclo de vida de integração do colaborador técnico, como profissionais de tecnologia. Lawson e Hendrick (2024) revelam que o processo de contratação e *onboarding* de um profissional técnico estende-se, em média, por 10,2 meses até que a produtividade plena seja alcançada. Durante este ciclo de investimento, a vulnerabilidade organizacional é acentuada: aproximadamente 38% das novas contratações enfrentam o desligamento nos primeiros seis meses de contrato (LAWSON; HENDRICK, 2024). Sob a perspectiva financeira, tais saídas precoces representam um dreno de recursos críticos, ocorrendo antes que o colaborador alcance o seu ponto de equilíbrio econômico (*breakeven*), tornando a predição de desligamentos prematuros uma prioridade estratégica.

Para compreender essas dinâmicas, as estruturas teóricas sobre a rotatividade voluntária evoluíram historicamente de modelos simples para modelos multidimensionais. Nesse contexto, o trabalho de Mitchell et al. (2001) introduziu o conceito de *job embeddedness* (integração ao trabalho), postulando que a retenção é impulsionada por uma teia composta por vínculos organizacionais, *fit* (ajuste) e pelo sacrifício percebido associado à saída. Embora essa estrutura evidencie que fatores não afetivos frequentemente superam a satisfação no trabalho nas decisões de permanência, ela avalia primariamente os custos psicológicos e materiais do desligamento sob a ótica do colaborador. Nota-se, contudo, uma lacuna crítica na modelagem do sacrifício de investimento sob a perspectiva recíproca da organização, especificamente na distinção entre os desligamentos que ocorrem após a amortização dos custos de contratação e aqueles que se materializam como prejuízo financeiro líquido, aqui definidos como *turnover pré-breakeven*.

Neste contexto, surge o *people analytics*, que pode ser definido como: a aplicação sistemática de modelagem preditiva usando estatística inferencial em dados existentes de recursos humanos (RH) referentes a pessoas para orientar as tomadas de decisões sobre possíveis fatores causais que são a chave para indicadores de desempenho de RH. Isto nos permite testar modelos estatísticos com auxílio de algoritmos de *machine learning* (ML) e predizer quais situações podem estimular o alto desempenho, ou o que pode causar a saída de um funcionário da organização (EDWARDS; EDWARDS, 2016, p. 19, tradução nossa).

A relevância estratégica dessa abordagem é ratificada por Ferrar e Green (2021), que argumentam que a análise de pessoas deve estar intrinsecamente atrelada à entrega de valor comercial. Para os autores, o *people analytics* deve servir tanto aos gestores, para decisões baseadas em fatos, quanto aos investidores e à sociedade, superando a visão setorizada do RH. Ademais, Guenole, Ferrar e Feinzig (2017) destacam que a inteligência de dados pode personalizar a experiência do colaborador, recomendando desde cursos modulares e trilhas de carreira até feedbacks em tempo real, promovendo um ecossistema de desenvolvimento contínuo.

1.1 PROBLEMA DE PESQUISA

Embora o *people analytics* apresente um potencial disruptivo para a gestão estratégica de talentos, sua implementação eficaz defronta-se com desafios substanciais. A mera acumulação volumosa de dados não se traduz, intrinsecamente, em *insights* precisos ou ações efetivas para a retenção de colaboradores. Nesse sentido, Yahia, Hlel e Palacios (2021) advertem sobre a necessidade de uma transição metodológica de um paradigma focado puramente em *big data* (volume) para uma abordagem de *deep data*, enfatizando que a priorização da qualidade e da relevância profunda das informações é um requisito fundamental para a formulação de predições assertivas sobre a rotatividade.

No setor de tecnologia, caracterizado por alta dinamicidade e escassez crônica de profissionais qualificados, a rotatividade voluntária (*turnover*) impõe ônus severos às organizações, transcendendo as perdas operacionais para atingir diretamente a sustentabilidade econômica. Conforme delineado por Isson e Harriott (2016), o custo associado à perda de um colaborador pode variar entre 1,5 e 2,0 vezes o seu salário anual. Essa vulnerabilidade financeira é acentuada no período inicial de contratação, dinâmica explicada pelo conceito de *daily breakeven* (ponto de equilíbrio diário). O autor ilustra que, durante os meses iniciais de recrutamento, *onboarding* e treinamento, o colaborador representa essencialmente um custo, atingindo apenas tardiamente o ponto em que o seu desempenho entrega uma contribuição líquida positiva (ISSON; HARRIOTT, 2016).

Apesar da gravidade desse sacrifício de investimento, observa-se uma limitação nas métricas tradicionais de recursos humanos, as quais tendem a tratar o turnover como um fenômeno homogêneo. Estudos recentes corroboram que o desligamento de profissionais é regido por interações não lineares e altamente complexas (AVRAHAMI et al., 2022). Contudo, a modelagem generalista falha ao não realizar a distinção entre a saída de um profissional veterano, que já amortizou o seu custo de contratação, e a de um colaborador recém-contratado. A incapacidade de prever e tratar isoladamente a evasão que ocorre antes do alcance do ponto de equilíbrio econômico do cargo resulta em prejuízos financeiros diretos que os modelos atuais não conseguem mitigar.

Diante da necessidade de alinhar a ciência de dados em recursos humanos à proteção do capital organizacional, e buscando superar a referida lacuna científica e

prática, este estudo é norteado pelo seguinte problema de pesquisa: Como um método preditivo fundamentado em *people analytics* pode ser estruturado para identificar o risco de desligamento prematuro (*pré-breakeven*) de colaboradores no setor tecnológico, visando mitigar prejuízos de investimento organizacional?

1.2 OBJETIVOS

Desenvolver um método preditivo baseado em aprendizado de máquina (*machine learning*) capaz de identificar o risco de desligamento prematuro (*pré-breakeven*) de colaboradores no setor tecnológico, visando mitigar prejuízos de investimento organizacional.

Para dar suporte à consecução deste objetivo geral, foram estabelecidos os seguintes objetivos específicos:

a) Identificar e consolidar um inventário de atributos preditivos (comportamentais, contratuais, de mercado e financeiros) determinantes para a retenção de talentos no setor de tecnologia;
b) Construir um *dataset* estruturado, incorporando variáveis financeiras inéditas e a variável-alvo "desligamento *pré-breakeven*", para refletir o cenário de retenção no mercado tecnológico brasileiro;
c) Propor cenários de avaliação para contrastar a eficácia de um método direcionado estritamente ao risco financeiro (*pré-breakeven*) em relação à abordagem preditiva generalista de *turnover*;
d) Estruturar um pipeline algorítmico de classificação otimizado para a métrica de *recall*, entregando uma arquitetura preditiva capaz de minimizar estrategicamente a incidência de falsos negativos.

1.3 JUSTIFICATIVA

A relevância desta investigação fundamenta-se na necessidade de aprimorar as métricas de gestão de talentos no setor tecnológico, um ambiente caracterizado por alta volatilidade de capital humano e elevados custos de aquisição e capacitação técnica. A justificativa central reside na insuficiência das abordagens tradicionais de *turnover*, as quais tendem a tratar a rotatividade como um fenômeno homogêneo,

equiparando, erroneamente, o desligamento de um colaborador experiente ao de um recém-contratado que ainda não retornou o investimento inicial.

Sob a ótica econômico-financeira, a pesquisa justifica-se pela introdução do conceito de desligamento *pré-breakeven* como variável-alvo. Diferentemente da rotatividade operacional padrão, a evasão de colaboradores antes do ponto de equilíbrio econômico representa um prejuízo direto de investimento, impactando a sustentabilidade financeira da organização. Portanto, o desenvolvimento de um modelo preditivo capaz de identificar especificamente esse risco preenche uma lacuna na literatura de *people analytics*, promovendo a integração entre a gestão comportamental e a eficiência financeira.

Do ponto de vista metodológico e científico, o estudo se justifica pela complexidade não linear inerente aos fatores que motivam a evasão no setor de tecnologia, realizando a aplicação e a comparação de arquiteturas avançadas de *machine learning* apoiando-se em métricas baseadas em *recall*, visando mitigar o erro estratégico dos falsos negativos em cenários de alto custo de substituição.

Por fim, a pesquisa oferece uma contribuição aplicada ao estruturar um método robusto de atributos enriquecidos e técnicas de anonimização compatíveis com a LGPD. Isso instrumentaliza as organizações para transitar de uma gestão de RH reativa para uma abordagem preditiva e financeiramente orientada, otimizando a alocação de recursos de retenção onde a exposição ao risco financeiro é mais crítica.

1.4 INTERDISCIPLINARIDADE E ADERÊNCIA AO PPGTIC

A presente investigação vincula-se estritamente à linha de pesquisa com ênfase em tecnologia computacional do Programa de Pós-Graduação em Tecnologias da Informação e Comunicação (PPGTIC). A área de *people analytics* propõe a aplicação de conceitos e técnicas avançadas provenientes deste domínio, tais como arquiteturas de machine learning e análise de dados complexos, para o aprimoramento de processos decisórios estratégicos na gestão de capital humano.

Cabe ressaltar que este trabalho caracteriza-se como uma pesquisa pioneira no âmbito do PPGTIC, visto que não foram identificados, nos registros do programa, estudos precedentes dedicados especificamente à modelagem preditiva de desligamentos de trabalhadores ou construções algorítmicas análogas voltadas à mitigação da rotatividade laboral sob a ótica da sustentabilidade financeira. A

abordagem aqui proposta visa transformar a gestão de recursos humanos em uma unidade de inteligência estratégica, integrando a eficiência computacional à lucratividade organizacional. Tal convergência entre a ciência da computação e a gestão de pessoas é fundamental para o desenvolvimento de vantagens competitivas sustentáveis e para a otimização da alocação de recursos onde a exposição ao risco financeiro é mais crítica.

1.5 ESTRUTURA DO TRABALHO

Para viabilizar a compreensão do percurso metodológico e a concatenação lógica das ideias que norteiam esta pesquisa, a presente dissertação está organizada em seis capítulos principais, além das referências e apêndice, estruturados da seguinte forma:

O Capítulo 1 Introdução, apresenta a contextualização do estudo, destacando o cenário do mercado de tecnologia e os desafios da rotatividade de profissionais. Neste capítulo, são definidos o problema de pesquisa, o objetivo geral e os objetivos específicos, bem como a justificativa e a aderência interdisciplinar do estudo ao Programa de Pós-Graduação em Tecnologias da Informação e Comunicação (PPGTIC).

O Capítulo 2 Revisão de Literatura, estabelece o pilar teórico e conceitual que subsidia a pesquisa. Aborda os fundamentos da gestão de talentos, a evolução conceitual do *people analytics* e os princípios de *data science* aplicados a recursos humanos. Adicionalmente, explora os indicadores de performance e os principais frameworks de modelagem analítica na área.

O Capítulo 3 Estado da Arte, expõe os resultados de uma revisão sistematizada da literatura. O capítulo detalha os procedimentos de busca, realiza uma análise estatística das publicações globais e apresenta uma matriz de síntese dos trabalhos correlatos. A partir desta análise, identifica-se a lacuna científica relacionada à contextualização financeira do *turnover*, justificando o ineditismo da variável de *breakeven* adotada neste trabalho.

O Capítulo 4 Proposta Metodológica, descreve minuciosamente o delineamento empírico da pesquisa. Detalha a formulação das hipóteses de negócio, as estratégias de modelagem (Cenário A e Cenário B) e a definição da variável-alvo pautada no ponto de equilíbrio econômico (*breakeven*). O capítulo também abrange a

origem e o enriquecimento do *dataset* sintético, a análise exploratória dos dados (AED), as etapas de pré-processamento (incluindo anonimização em conformidade com a LGPD e balanceamento com SMOTETomek) e os parâmetros de treinamento dos algoritmos de *machine learning*.

O Capítulo 5 Resultado e Avaliações, dedica-se à interpretação dos achados empíricos. Apresenta o desempenho comparativo dos modelos preditivos por meio de matrizes de confusão, curvas *precision-recall*, curvas ROC e validação cruzada. O capítulo culmina na discussão da eficácia estratégica do algoritmo com melhor desempenho na identificação de desligamentos precoces e analisa a importância das variáveis (*feature importance*) para a predição.

O Capítulo 6 Conclusão, sintetiza as evidências encontradas, respondendo à pergunta de pesquisa e confirmando as hipóteses levantadas. Discute as implicações teóricas e as contribuições práticas da inserção de dados financeiros no *people analytics*, além de elencar as limitações do estudo e propor direcionamentos para trabalhos futuros, como a adoção de agentes de inteligência artificial.

Por fim, o documento inclui o Apêndice A, que detalha uma Metodologia de *People Analytics* de 4 Fases, fornecendo um guia processual e estruturado (comparativo ao framework Scrum) para a implementação de projetos baseados em dados no setor de recursos humanos.

2 REVISÃO DE LITERATURA

A presente seção estabelece o pilar teórico e conceitual que subsidia a modelagem preditiva e a análise empírica desenvolvidas nesta dissertação, dedicando-se à revisão sistemática da literatura pertinente ao campo de *people analytics*. O objetivo primordial é delinear o paradigma analítico necessário para a análise de desligamentos no setor tecnológico.

A revisão inicia-se com uma explanação sobre a Gestão de Talentos e o *people analytics*, conceituando este último como a aplicação sistemática de modelagem preditiva, utilizando estatística inferencial em dados existentes de Recursos Humanos (RH) para orientar decisões sobre fatores causais que impactam indicadores de desempenho.

Em um segundo momento, o foco se desloca para o *data science* aplicado à gestão de talentos, que representa o conjunto de princípios e técnicas para a extração de conhecimento a partir de dados. Esta seção examinará a distinção entre análises descritivas, preditivas e prescritivas, e detalhará os algoritmos para obtenção do modelo preditivo, enfatizando o papel do *machine learning* (aprendizado de máquina) supervisionado (como classificação e regressão) e não supervisionado (como *clustering*). Algoritmos específicos, como árvores de decisão, redes neurais e métodos ensemble, serão discutidos como ferramentas essenciais para lidar com a complexidade e o volume dos dados de RH.

Por fim, a revisão abordará os indicadores de performance de *people analytics* e os modelos de *people analytics*, apresentando *frameworks* estruturais como a Metodologia *Value Chain*, o Modelo *Focus-Impact-Value* e *o Modelo Eight Step for Purposeful Analytics*. Tais modelos fornecem os métodos para conectar dados brutos a *insights* estratégicos, desde a definição de questões de negócio e construção de hipóteses até a implementação e avaliação de resultados. A elucidação desses fundamentos metodológicos é necessária para justificar o delineamento do estudo proposto na Seção 4.

2.1 A GESTÃO DE TALENTOS E O PEOPLE ANALYTICS

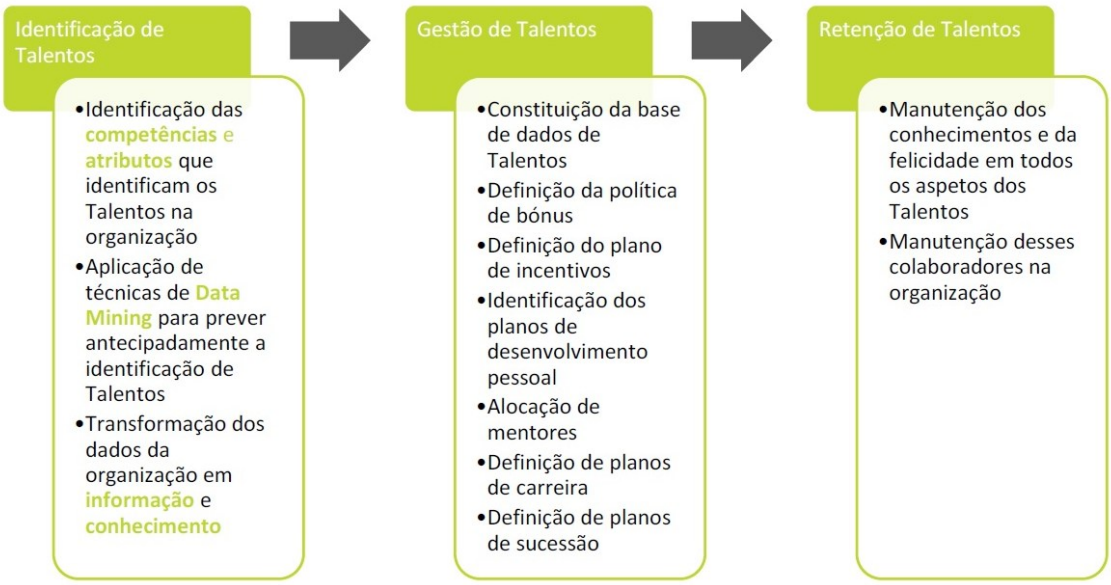
Conforme Mendes (2021), os gestores enfrentam a complexa e promissora tarefa de reestruturar processos e introduzir inovações disruptivas na gestão de

recursos humanos. É essencial que se aproveitem plenamente os avanços tecnológicos em benefício do ser humano, de modo a maximizar as contribuições individuais e assim promover o fortalecimento e o desenvolvimento sustentado da organização, bem como o progresso da comunidade e da sociedade em sua totalidade.

Sadath (2013 apud CANAIS, 2016) ressalta que conhecer os colaboradores e os talentos da organização é responsabilidade fundamental da função de Recursos Humanos. Esta, por meio de uma abordagem estruturada, identifica e aprimora os talentos, evitando a saída dos colaboradores, garantindo, desse modo, a atualização permanente de conhecimentos necessários e mantendo-os satisfeitos em todos os aspectos.

A Figura 1 apresenta de forma adaptada diversos fatores que, atuando de forma estruturada, contribuem para uma eficaz gestão na área da retenção de talentos (DELOITTE, 2015; PRICEWATERHOUSECOOPERS, 2015; MITCHELL; HOLTOM; LEE, 2001; ARISS, 2014 apud CANAIS, 2016):

Figura 1 - Gestão de Talentos



Fonte: Canais (2016, p. 6).

De acordo com Mendes (2021), na gestão de talentos, a integração de inteligência artificial, *big data* e *business intelligence (BI)* permitem cruzar dados de desempenho, projetos, formação e perfil comportamental dos colaboradores. Esse processo gera um *nine box* para mapeamento de talentos e sucessores em tempo

real, além de automatizar o plano de desenvolvimento individual (PDI), facilitando o acompanhamento pelos gestores e RH.

Embora o avanço de tecnologias, como a inteligência artificial e a *big data*, melhorem a gestão de talentos ao otimizar processos e facilitar o desenvolvimento individual, elas também podem expor desafios persistentes, como o viés discriminatório.

Segundo Mendes (2021), relatos de profissionais em redes sociais como o LinkedIn indicam um viés discriminatório nos processos seletivos do mercado de trabalho. Mulheres frequentemente enfrentam questionamentos sobre estado civil, idade e maternidade. Durante as entrevistas, é comum perguntarem sobre seus relacionamentos, planos de gravidez e a possibilidade de faltar ao trabalho devido a doença dos filhos. Essa prática reflete a construção social que marginaliza a maternidade, tratando-a como um impedimento para o desempenho esperado no ambiente profissional.

De acordo com Mendes (2021), aproximadamente 62% dos recém-graduados universitários que se identificam como lésbicas, gays ou bissexuais ocultam sua orientação sexual ao ingressar no mercado de trabalho. A situação das pessoas transexuais é ainda mais grave, enfrentando exclusão quase total dos processos seletivos e acesso extremamente limitado ao mercado de trabalho.

Conforme descrito por Havard Business Review (2018), as pessoas são iludidas pelo que o psicólogo David Armor, de Yale, descreve como a "ilusão de objetividade", que consiste na falsa crença de que estamos isentos de inclinações que prontamente identificamos nos outros. Estas inclinações inconscientes podem contradizer nossas convicções conscientes, como a crença de que a etnia não influencia decisões de contratação ou de que estamos imunes a conflitos de interesse. A pesquisa psicológica frequentemente revela essas tendências não intencionais, indicando que até mesmo pessoas bem-intencionadas podem permitir que pensamentos e sentimentos inconscientes influenciem decisões que parecem ser objetivas.

Compreender inclinações inconscientes é crucial para melhorar a gestão de talentos. Ações como recomendações personalizadas e feedback em tempo real podem aprimorar habilidades e reduzir preconceitos, promovendo um ambiente de trabalho mais justo e alinhado com os objetivos organizacionais.

Muitos trabalhadores gostariam de receber recomendações para aprimorar sua experiência no trabalho. Exemplos de sugestões para colaboradores, segundo Guenole, Ferrar e Feinzig (2017, p. 41, tradução nossa):

a) Recomendação de cursos modulares para aprimorar as habilidades dos funcionários;
b) Informações sobre benefícios relevantes à medida que um trabalhador entra em novas fases da vida (por exemplo, um novo filho, casamento ou compra de casa);
c) Mudanças internas de cargo e carreira que melhor atendam às habilidades e conhecimentos do trabalhador;
d) Oportunidades de contribuir para projetos em toda a empresa com base na experiência e no conhecimento de um indivíduo;
e) Fornecimento de *feedback* de desempenho em tempo real por meio de feedback social entre gerentes e funcionários e entre pares.

2.1.1 A Adoção de Analytics

Muitas organizações já estão percebendo os benefícios da análise. Um relatório da PWC de 2014, escrito pela The Economist Intelligence Unit, concluiu que 89% dos executivos de grandes empresas inquiridos já utilizavam *big data* para tomar decisões ou planeavam começar a fazê-lo nos próximos três anos. Especificamente em RH, em seu relatório CHRO (*Chief Human Resources Officer*) de 2016, a IBM descobriu que o número de diretores de recursos humanos que usam análises preditivas para tomar decisões mais informadas sobre a força de trabalho em atividades de RH aumentou aproximadamente 40% em dois anos. A evidência da tendência é clara, como relata o Global Human Capital Trends 2017 da Deloitte: "*People analytics*, uma disciplina que começou como um pequeno grupo técnico que analisava o envolvimento e a retenção, tornou-se agora *mainstream*" (PWC, 2014 *apud* GUENOLE; FERRAR; FEINZIG, 2017, p. 37, tradução nossa).

O Google utiliza o *people analytics* para apoiar as decisões em todas as etapas do ciclo de um funcionário dentro da organização. No início dos anos 2000, o candidato poderia ser submetido a uma dúzia de entrevistas. Porém, aplicando a devida análise, o Google concluiu que após quatro entrevistas a chance de o conhecimento acumulado pelos entrevistadores levar à decisão correta era de 86%,

sendo que as entrevistas extras aumentavam somente em 1% essa probabilidade. Atualmente, o recrutamento das equipes não técnicas e técnicas é feito com quatro e cinco entrevistas, respectivamente (REVISTA HSM MANAGEMENT, 2016).

Na IBM, a ex-CEO (*Chief Executive Officer*) Ginni Rometty afirmou que a empresa economizou “quase US$ 300 milhões em custos de retenção em seu programa de desgaste preditivo” (Rosenbaum, 2019). Utilizando análise de pessoas, identificou as pessoas com maior risco de saída e prescreveu ações antecipadamente para ajudar os gestores a tomar as decisões corretas (FERRAR; GREEN, p.73, tradução nossa).

A empresa global de dados Nielsen, descobriu que para cada 1% de desgaste de funcionários, poderia evitar US$ 5 milhões em custos de negócios.2 A análise também revelou outros *insights* (como a mobilidade interna como o principal impulsionador da retenção), mas foi a ligação direta entre uma visão de pessoas e uma métrica financeira que permitiu aos executivos de negócios se conectarem e ficarem entusiasmados com a análise de pessoas. A Nielsen combinou estatísticas com uma forte comunicação com as partes interessadas e uma narrativa clara sobre o valor do negócio e os benefícios para os funcionários (FERRAR; GREEN, p.92, tradução nossa).

Outro exemplo inovador de como a análise está quantificando o valor vem da Unilever. Seu diretor de recursos humanos destacou que para cada US$ 1,00 que a empresa investe no bem-estar dos funcionários, ela proporciona um retorno de US$ 2,50 (GREEN, 2019 apud FERRAR; GREEN, p.92, tradução nossa).

Nos últimos 40 anos, mais ou menos, os Recursos Humanos entregaram programas estruturados, desenvolveram políticas e implementaram as melhores práticas para permitir que os executivos e gerentes gerenciassem pessoas em um padrão cíclico, por exemplo, através de avaliações de desempenho anuais, programas específicos para aumento de salário e ciclos de planejamento de sucessão. No entanto, a demanda tem mudado e novas necessidades de informações estão surgindo (GUENOLE; FERRAR; FEINZIG, 2017, p. 41, tradução nossa):

No Quadro 1, é exibida uma comparação em períodos tradicionais, atuais e projeção do futuro em relação às solicitações dos gestores de Recursos Humanos para temas cotidianos da área:

Quadro 1 - Exemplos de RH tradicionais, atuais e futuras solicitações de gestores

Tema	Tradicional	Atual	Futuro
Recrutamento	Eu preciso recrutar alguém. Como eu faço isso?	Eu preciso de alguém novo. Como eu sei de onde vêm as melhores pessoas e quem será o mais adequado para minha equipe?	Você pode me recomendar as pessoas mais adequadas antes de precisarmos delas, tendo assim um banco pronto de candidatos adequados?
Aprendendo e desenvolvendo	Quais cursos estão disponíveis para pessoas de vendas?	Eu quero algo para ajudar o Joe com uma reunião com cliente que ele tem amanhã—um pequeno vídeo seria ótimo.	Você pode notificar Joe sobre suas necessidades de aprendizado agora e sobre suas necessidades no próximo ano? Por favor, recomende as ações de desenvolvimento que pessoas de perfil semelhantes tomaram e envie-lhe artigos que ele possa ler.
Compensação	Gabrielle pediu demissão e eu gostaria de oferecer-lhe um aumento de salário para ver se podemos mantê-la.	Eu gostaria de ser notificado quando as pessoas correm o risco de sair.	Eu gostaria de receber informações de salário e desempenho referente a todo o meu pessoal para me ajudar a manter-me a par das condições do mercado em cada localidade. Eu também gostaria de recomendações sobre maneiras de reter minhas pessoas-chave.
Saúde e bem estar	Quais são os benefícios desta empresa?	Podemos ter benefícios adaptados aos funcionários por localização, nível, idade e experiência?	Podemos enviar notificações às pessoas (antecipadamente) sobre os benefícios de saúde e bem-estar que atendem tanto às suas necessidades de estilo de vida quanto às nossas necessidades na empresa?
Liderança	Quais são os planos de sucessão da minha equipe?	Quem são os melhores líderes da empresa? Quem melhor se adapta às nossas necessidades futuras e aos nossos valores para garantir o nosso sucesso?	Quais comportamentos, habilidades e atributos se ajustam melhor às necessidades de liderança atuais e futuras de nossa empresa? Como cultivamos nosso pessoal atual, recrutamos novas pessoas e gerenciamos a sucessão para permitir um futuro estável à medida que nosso mercado evolui?

Fonte: Adaptado pelo autor (GUENOLE; FERRAR; FEINZIG, 2017, p. 42).

A análise das solicitações dos gestores de RH ao longo do tempo, como mostrado no Quadro 1, destaca a evolução nas demandas e expectativas na gestão de talentos. Com o ambiente organizacional se tornando mais complexo, a análise preditiva de RH se torna crucial.

Definimos análise preditiva de RH como a aplicação sistemática de modelagem preditiva usando estatísticas inferenciais para dados existentes relacionados ao pessoal de RH, a fim de informar julgamentos sobre possíveis fatores causais que

impulsionam os principais indicadores de desempenho relacionados a RH (EDWARDS; EDWARDS, 2016, p. 19, tradução nossa).

O *people analytics* começa com uma questão ou objetivo de negócios de gerenciamento de talentos e, em seguida, integra fontes de dados distintas para criar previsões para o futuro, que podem então ser usadas para delinear as ações das empresas com resultados mensuráveis. (ISSON; HARRIOTT, 2016, p.72, tradução nossa).

Ferrar e Green (2021) destacam os seguintes pontos em relação a doção de análise de pessoas:

a) RH não tem a ver com RH, mas sim com o negócio. A análise de pessoas não se trata de medir as atividades de RH encontradas em *scorecards* ou painéis, ou mesmo de informações ou informações fascinantes, mas de ajudar a entregar resultados (FERRAR; GREEN, 2021, p.37, tradução nossa);
b) A análise de pessoas começa definindo os resultados desejados para as partes interessadas. Qualquer organização tem alguma versão de um *balanced scorecard* com resultados em cinco áreas: resultados para funcionários, estratégia, clientes, resultados financeiros e comunitários. Ser claro sobre esses resultados desejados oferece uma proclamação clara do que é mais importante (variáveis dependentes em termos analíticos) (FERRAR; GREEN, 2021, p.37, tradução nossa);
c) A análise de pessoas requer uma compreensão dos caminhos ou iniciativas para produzir os resultados desejados pelas partes interessadas. No meu próprio trabalho sobre capital humano, identificamos quatro domínios (talento, liderança, organização e departamento de RH) onde as iniciativas podem ser concebidas e entregues (variáveis independentes em termos analíticos) (FERRAR; GREEN, 2021, p.37, tradução nossa).

Bersin (2012 apud ISSON; HARRIOTT) criou um modelo de maturidade analítica de RH para explicar os diferentes níveis de adoção de análise de RH. Ele define esses estágios com base em quatro níveis de maturidade:

a) O nível 1 é denominado relatório operacional reativo e reflete empresas onde a análise de RH se concentra principalmente em relatórios

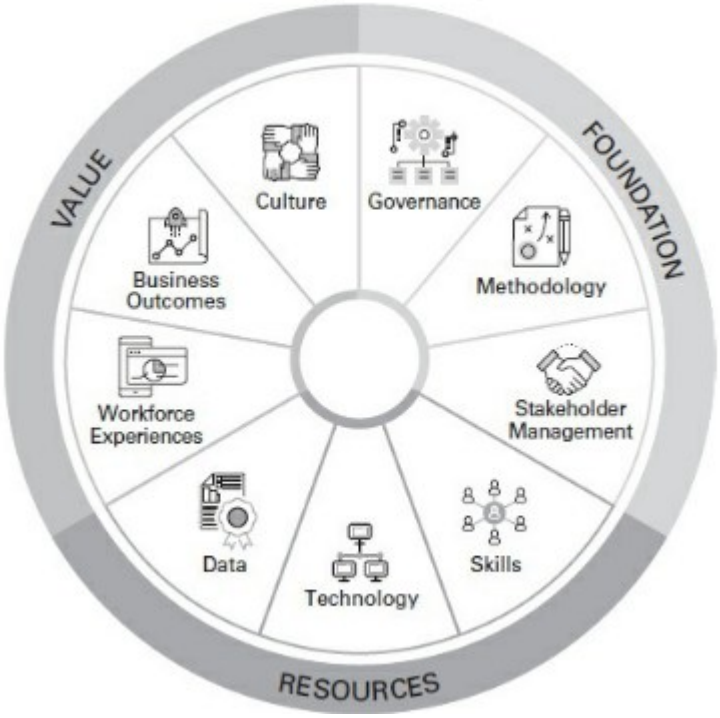
operacionais *ad hoc*. Este nível reage às demandas do negócio, é caracterizado pelo isolamento de dados e é difícil de analisar. (BERSIN, 2012 apud ISSON; HARRIOTT, 2016, p.80, tradução nossa);

b) O nível 2 é denominado relatórios proativos avançados e reflete empresas onde a análise de RH se concentra em relatórios operacionais para avaliar a tomada de decisões multidimensionais. (BERSIN, 2012 apud ISSON; HARRIOTT, 2016, p.80, tradução nossa);
c) O nível 3 é denominado análise estratégica e reflete empresas onde a análise de RH se concentra no desenvolvimento de análises estatísticas de modelos de pessoas e análises de dimensões para compreender a causa e fornecer *insights* acionáveis. (BERSIN, 2012 apud ISSON; HARRIOTT, 2016, p.80, tradução nossa);
d) O nível 4 é denominado análise preditiva e reflete o desenvolvimento de cenários de modelos preditivos para planejamento, análise e mitigação de riscos e integração com planejamento estratégico (BERSIN, 2012 apud ISSON; HARRIOTT, 2016, p.80, tradução nossa).

As organizações mais bem sucedidas em análise de pessoas procuram a excelência em cada uma das nove dimensões e abordam-nas de uma forma que seja adequada (e na ordem apropriada) para a sua empresa. Descobrimos que as organizações que proporcionaram o maior impacto concentram-se nas três categorias simultaneamente. O modelo das Nove Dimensões não é sequencial. Não precisa ser “feito” em uma ordem específica (FERRAR; GREEN, 2021, p.44, tradução nossa).

A Figura 2 exibe o modelo das Nove Dimensões e suas três categorias: Fundação, Recursos e Valor:

Figura 2 - Nine Dimensions for Excellence in People Analytics



Fonte: Ferrar e Green (2021, p. 44)

Fundação: A análise de pessoas precisa de uma base sólida com os elementos certos implementados para permitir o sucesso no futuro, antes que o trabalho se torne muito complexo. Isto está enraizado numa governança forte, metodologias claras e uma gestão eficaz das partes interessadas (FERRAR; GREEN, 2021, p.47, tradução nossa).

Recursos: A análise de pessoas deve ter impacto para ser credível. Isto requer equilibrar os recursos certos, incluindo conhecimentos da própria equipa, tecnologias apropriadas e dados robustos e extensos (FERRAR; GREEN, 2021, p.53, tradução nossa).

Valor: A análise de pessoas tem a responsabilidade de agregar valor à organização e à sua força de trabalho. Isto é obtido através do fornecimento de experiências aprimoradas à força de trabalho, criando impacto por meio de resultados de negócios e desenvolvendo uma cultura baseada em dados para análise (FERRAR; GREEN, 2021, p.59, tradução nossa).

2.1.2 O Impacto do Turnover nas organizações

A rotatividade de funcionários geralmente se refere a todos os que saem de uma organização, incluindo aqueles que se demitem, são demitidos, se aposentam ou saem da organização por qualquer outro motivo (EDWARDS; EDWARDS, 2016, p.299, tradução nossa).

O custo da rotatividade de funcionários pode ser substancial e foi projetado em 93-200 por cento do salário de cada pessoa que sai, dependendo de sua habilidade, nível de responsabilidade e quão difícil é substituí-los (GRIFFETH; HOM, 2001 apud EDWARDS; EDWARDS, 2016, p.299, tradução nossa).

Quando isto é acumulado entre todos os que abandonam uma organização, o custo é substancial e é, portanto, uma área chave onde uma análise sólida das causas e o diagnóstico de problemas podem dar um contributo extremamente valioso para os resultados financeiros da organização. Definir um valor para o custo da rotatividade de funcionários é valioso porque pode ajudar o business case a justificar programas e intervenções para resolver o problema. Os custos da rotatividade voluntária variam dependendo da organização e do tipo de função (EDWARDS; EDWARDS, 2016, p.299, tradução nossa).

A Figura 3 fornece mais detalhes de alguns destes custos em custos de separação, custos de substituição e custos de formação (GRIFFETH; HOM, 2001 apud EDWARDS; EDWARDS, 2016, p.299, tradução nossa):

Figura 3 - Cost of employee turnover

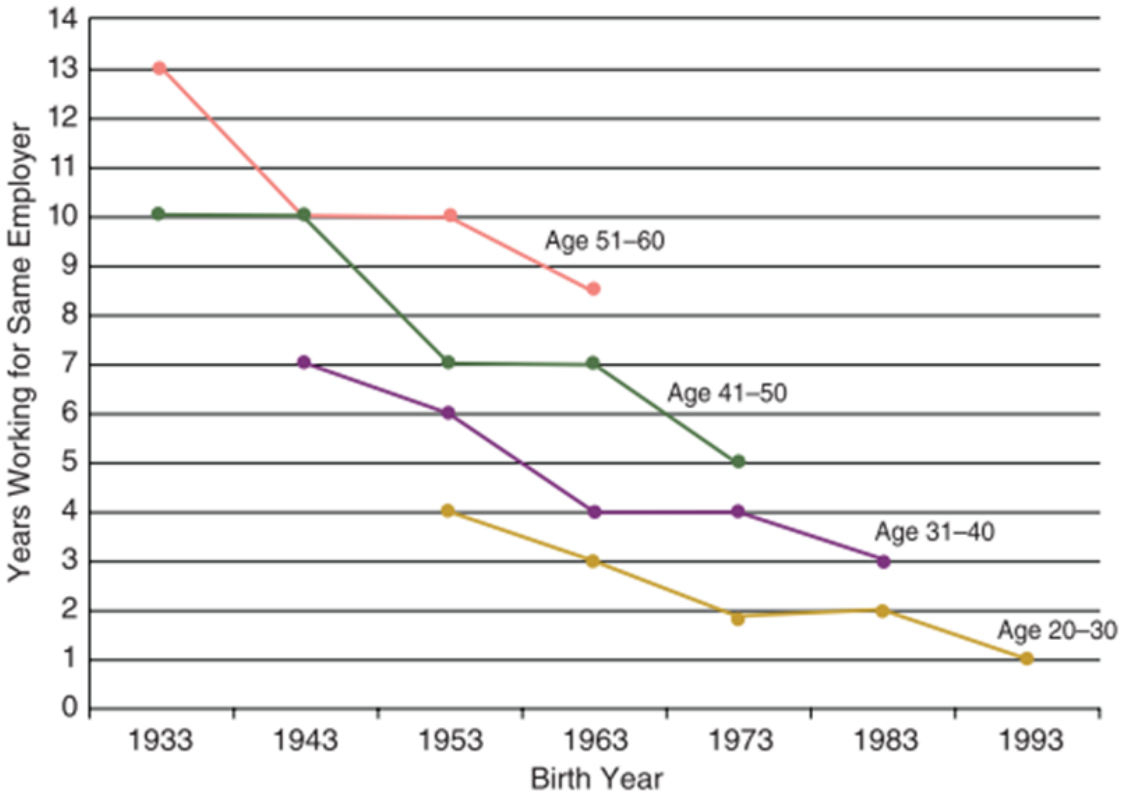
Separation Costs	Replacement Costs	Training Costs
• Exit interviews • Administrative costs • Separation benefits such as unused vacation time paid out • Lost revenues due to vacancy • Productivity declines • Overtime costs for extra labour during job vacancy • Hiring temporary personnel during job vacancy • Client reassignment costs (if a customer-facing role)	• Job advertisements • Recruitment • Administrative processing • Entrance interviews • Applicant selection • Assessment centres/testing • Travelling and relocation expenses • Pre-employment screening • Medical examination (if applicable)	• Formal orientation • Formal job training • Offsite training • On-the-job training • Productivity inefficiency

Fonte: Edwards e Edwards (2016, p.300)

Em algumas funções, pode levar até 12 meses para que o novo ingressante seja atualizado, construa redes internas, ganhe a confiança do cliente, etc., ao nível de competência e produtividade experimentado pelo antecessor, e as organizações estão buscando métodos de integração eficientes para reduzir esse tempo de produtividade (WATKINS, 2013 apud EDWARDS; EDWARDS, 2016, p.300, tradução nossa).

Como se pode ver na Figura 4, ao olharmos para as pessoas entre os 20 e os 30 anos, podemos ver que a permanência média no emprego era de quatro anos entre os nascidos em 1953 (*baby boomers*), quando tinham entre 20 e 30 anos de idade. No entanto, para as pessoas entre os 20 e os 30 anos nascidas em 1993 (geração *millennials*), a permanência média no emprego é de apenas um ano (ISSON; HARRIOTT, 2016, p.40, tradução nossa):

Figura 4 - Median Job Tenure by Age and Birth Cohort



Fonte: Isson e Harriott (2016, p.40)

2.1.2.1 Rotatividade voluntária

Os três principais tipos de rotatividade voluntária de funcionários são rotatividade de aposentadoria, rotatividade de promoções e rotatividade de mudança

de empregador. Para obter insights sobre a rotatividade voluntária, os dados devem ser perfilados com base nestas categorias (ISSON; HARRIOTT, 2016, p.140, tradução nossa):

a) Compreender o motivo da saída;
b) Destino;
c) Correlação entre faturamento e canal de aquisição;
d) Rotatividade por série e desempenho;
e) Rotatividade por dados demográficos para cada função de trabalho.

Com base em modelos estatísticos, deve-se identificar os funcionários existentes que correm o risco de sair, bem como o motivo provável da saída e a data potencial.

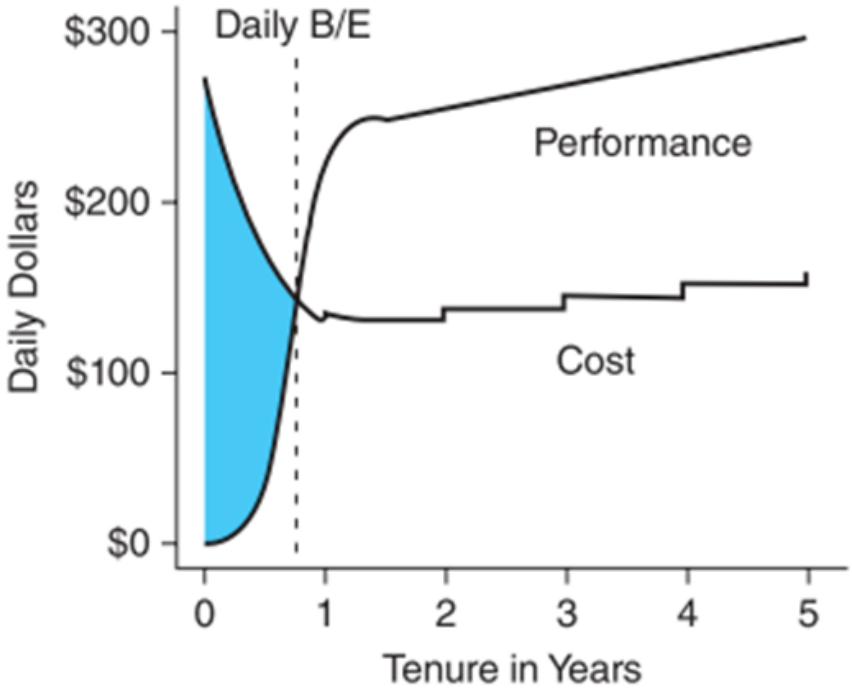
2.1.2.2 Rotatividade Involuntária

A rotatividade involuntária também deve ser analisada por dados demográficos, desempenho da empresa e fatores externos, como recuperação e desaceleração econômica, questões de desempenho e canal e custo de aquisição (ISSON; HARRIOTT, 2016, p.140, tradução nossa).

2.1.2.3 Daily Breakeven

A Figura 5 mostra um gráfico estilizado de custo/desempenho para um funcionário ao longo de cinco anos de mandato. No início, os custos de recrutamento, integração e formação são elevados, enquanto o desempenho é zero. Cada dia que o funcionário chega, a empresa perde dinheiro. Aos nove meses, os limites se cruzam e pela primeira vez, o funcionário entrega uma contribuição líquida positiva. Isso é chamado de “ponto de equilíbrio diário” ou breakeven (B/E)” (ISSON; HARRIOTT, 2016, p.289, tradução nossa).

Figura 5 - Cost and Performance for One Employee



Fonte: Isson e Harriott (2016, p.289)

Para além deste ponto, mantém-se um padrão de longo prazo de contribuição positiva líquida. Estas curvas cruzadas são por vezes chamadas de “tesouras quantitativas”, a zona sombreada à esquerda é uma realidade inevitável dos negócios. Para diminuir os custos gerais devido à rotatividade de funcionários, algo precisa mudar nessas curvas. No entanto, o *breakeven* diário não é a história completa (ISSON; HARRIOTT, 2016, p.289, tradução nossa).

Perder um funcionário significa, no mínimo, uma transferência de conhecimento e experiência para, talvez, um concorrente, bem como uma perda real de produtividade e uma mudança, mesmo que momentânea, no moral do pessoal. Além dos custos bem conhecidos de procurar substitutos (anúncios em mídias sociais, *headhunters*, etc.), encontrar um candidato adequado (entrevistas e períodos de experiência), treinamento (curvas de aprendizado e custos de treinamento) e, finalmente, contratação (custos de contratação, pacotes de realocação, bônus de inscrição, etc.), a rotatividade pode rapidamente se tornar um grande custo comercial para uma empresa. De acordo com Josh Bersin, diretor da Bersin, “o custo de perder um funcionário é de 1,5 a 2,0 vezes o salário anual da pessoa (ISSON; HARRIOTT, 2016, p.308, tradução nossa).

2.1.3 Fontes de Dados de Gestão de Pessoas

Dados úteis relacionados a RH são compostos de muitos tipos diferentes de informações e podem incluir o seguinte (EDWARDS; EDWARDS, 2016, p.26, tradução nossa):

a) Habilidades e qualificações;
b) Medidas de competências específicas;
c) Treinamento frequentado;
d) Níveis de envolvimento dos funcionários;
e) Dados de satisfação do cliente;
f) Registros de avaliação de desempenho;
g) Dados de pagamento, bônus e remuneração.

Existem seis categorias principais de dados essenciais para a análise do planejamento da força de trabalho (ISSON; HARRIOTT, 2016, p.133, tradução nossa):

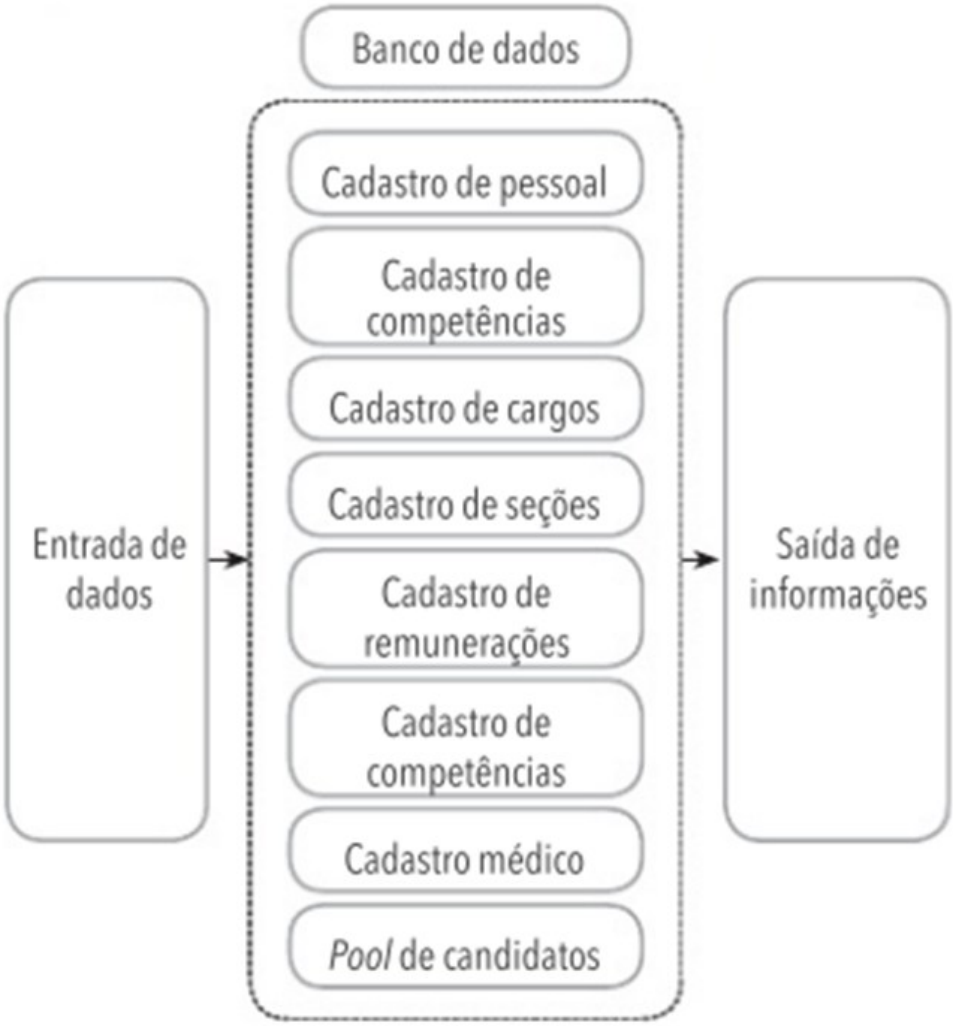
a. Dados de talentos: Inclui dados de *enterprise resource planning (*ERP*)* de RH, dados de recrutamento, dados de rotatividade e retenção, dados de engajamento, dados de treinamento, dados de desempenho, dados de preferências e atitudes, dados de satisfação de funcionários e dados de planejamento de sucessão e aposentadoria;
b. Dados de mercado: Referem-se a informações como dados de mídia social, dados de inteligência competitiva (obter uma compreensão do que sua concorrência está fazendo no que diz respeito ao gerenciamento de talentos) e dados de voz do candidato (o que os candidatos estão dizendo sobre sua organização);
c. Dados de negócios: Inclui métricas de indicadores-chave de desempenho das empresas, como vendas, receitas, clientes, receitas por funcionário, tamanho médio dos pedidos, taxa de retenção, novos negócios, contagem e taxa de recuperação, dados de folha de pagamento, remuneração orçamentária e dados de benefícios, e financiar dados de ERP;
d. Dados econômicos e industriais: Referem-se a dados de referência da indústria e dados macroeconómicos, tais como o produto interno bruto e o índice de preços no consumidor;

e. Dados de estatísticas trabalhistas: Inclui dados sobre a força de trabalho, dados sobre emprego, vagas de emprego, taxas de desemprego, salários, taxa de desemprego, demissões e demissões, crescimento populacional e previsão;
f. Dados de graduação universitária: Refere-se à graduação por disciplina com foco em cargos difíceis de preencher, como aqueles nas áreas *science, technology, engineering and mathematics* (STEM).

As plataformas tecnológicas do Sistema de Informação de Recursos Humanos (SIRH) armazenam informações sobre os funcionários. Muitas vezes existem versões operacionais para as diferentes subfunções de RH (por exemplo, folha de pagamento, recrutamento e aprendizagem), e alguns provedores oferecem um SIRH integrado com módulos que cobrem múltiplas subfunções de RH. Esses sistemas servem a dois propósitos principais. Primeiro, eles formalizam muitas das políticas e práticas de RH da organização em um fluxo de trabalho que reflete a visão da organização sobre as melhores práticas de RH. Em segundo lugar, capturam todos os dados gerados sobre os funcionários (e, em alguns casos, pelos funcionários) no desempenho das responsabilidades da função de RH (GUENOLE; FERRAR; FEINZIG, 2017, p.200, tradução nossa).

A Figura 6 ilustra um banco de dados de gestão de pessoas, destacando a estrutura e os componentes essenciais para armazenamento e análise de dados dentro de um *human resource information system* (HRIS):

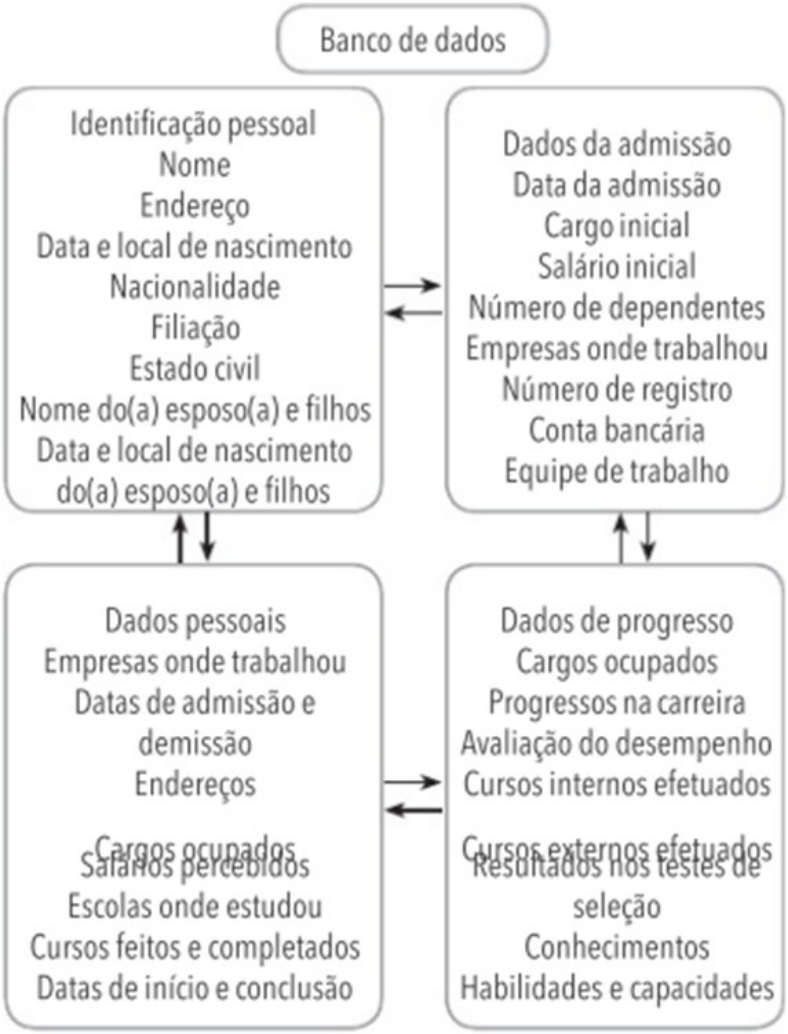
Figura 6 - Exemplo de um banco de dados de gestão de pessoas



Fonte: Chiavenato (2014, p.452)

A Figura 7 ilustra a complexidade e sensibilidade dos dados geridos de recursos humanos, incluindo informações pessoais e profissionais dos funcionários. Esses dados exigem medidas rigorosas de segurança para proteger informações sensíveis e evitar riscos como o roubo de identidade:

Figura 7 - Insumos básicos para o banco de dados de gestão de pessoas



Fonte: Chiavenato (2014, p.453)

Os conjuntos de dados de RH geralmente contêm informações confidenciais, incluindo detalhes pessoais sobre o histórico profissional dos funcionários, avaliações de desempenho, remuneração e códigos de identificação nacional, como números de seguro social ou de seguro nacional. Alguns destes elementos de dados são considerados informações pessoais sensíveis nos Estados Unidos e têm designações semelhantes em outros países. Estas são informações que, se comprometidas ou divulgadas, podem resultar em danos substanciais, como roubo de identidade (GUENOLE; FERRAR; FEINZIG, 2017, p.146, tradução nossa).

Dada a natureza pessoal e o risco potencial associado ao tratamento de dados de RH, são fortemente recomendados cuidados e precauções adicionais. Trabalhe com seu departamento jurídico para obter orientação, aconselhamento e aconselhamento. Certifique-se de compreender e aderir às políticas e práticas da sua

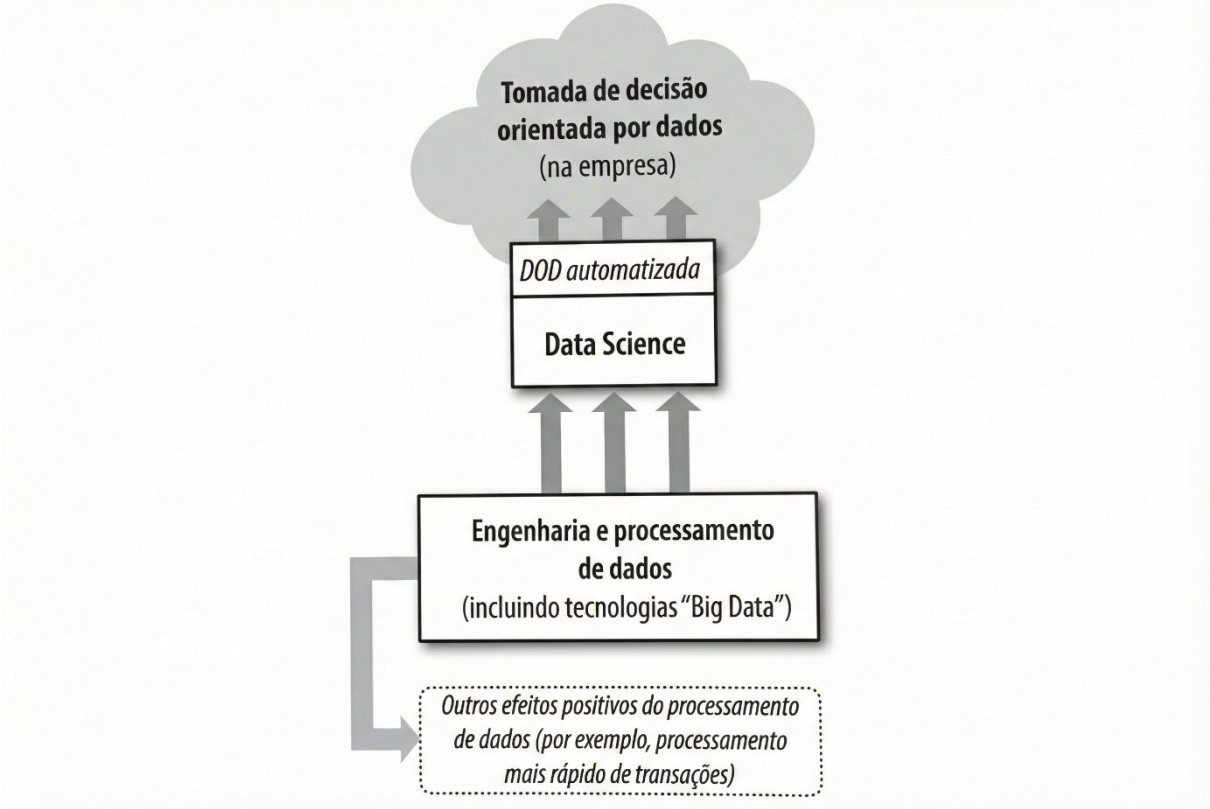
organização e esteja atento às regulamentações específicas do país em relação aos dados dos funcionários (GUENOLE; FERRAR; FEINZIG, 2017, p.147, tradução nossa).

2.2 DATA SCIENCE APLICADO A GESTÃO DE TALENTOS

Conforme Provost e Fawcett (2016), em um patamar avançado, ciência de dados representa um conjunto de princípios fundamentais direcionados à extração de conhecimento a partir de dados. Já a mineração de dados refere-se ao processo de extração de conhecimento, utilizando tecnologias que aplicam esses princípios.

Conforme ilustrado na Figura 8, *data science* envolve princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados:

Figura 8 - Data science no contexto dos diversos processos relacionados a dados na organização



Fonte: Provost e Fawcett (2016, p.5)

Para podermos aproveitar o potencial da análise preditiva de RH, todos nós dependemos dos dados atuais e históricos disponíveis. A análise preditiva de RH depende totalmente de bons dados; não podemos procurar padrões nos dados quando os dados disponíveis são limitados e incompletos. Assim, o sucesso da

análise de RH depende completamente da disponibilidade de boas informações relacionadas às pessoas (EDWARDS; EDWARDS, 2016, p.22, tradução nossa).

Algumas das técnicas analíticas que utilizamos visam explorar as relações entre muitos tipos diferentes de dados (ou variáveis), a fim de identificar “preditores” de alguns resultados importantes de RH (como o desempenho dos funcionários ou a rotatividade de pessoal). Este tipo de análise é utilizado para identificar tendências e relações entre múltiplos fatores na esperança de obter informações que sugiram as possíveis causas de variação do fenômeno que esperamos prever (EDWARDS; EDWARDS, 2016, p.24, tradução nossa).

Os relatórios descritivos de RH normalmente produzidos pelas equipes de *management information* (MI) referem-se à ilustração de um conjunto de dados, um ‘instantâneo’ do que está a ocorrer naquele momento, apresentando o estado atual da situação. Conjuntos de dados abrangem várias partes de uma organização; por exemplo: dados de recrutamento, como tempo de contratação, custos de contratação por funcionário, produtividade do recrutador; dados demográficos da força de trabalho, como idade, sexo, tempo de serviço; ou dados de número de funcionários, como porcentagem de rotatividade e custos por funcionário. Para a grande maioria das funções de RH, este é o limite da capacidade analítica de RH (EDWARDS; EDWARDS, 2016, p.110, tradução nossa).

A análise vem em três níveis: descritivo, preditivo e prescritivo. O descritivo fala do que aconteceu até o presente (ISSON; HARRIOTT, 2016, p.11).

A preditiva revela o que deve ser feito para atingir os objetivos futuros. Prescritivo diz como fazer isso. Quando um paciente diz ao médico que tem congestão nasal, isso é descritivo. O médico aplica seu conhecimento para determinar que tipo de tratamento irá aliviar ou curar a doença. Isso é preditivo. O papel que o paciente leva à farmácia explica como será administrado o tratamento. Essa é a receita (ISSON; HARRIOTT, 2016, p.11).

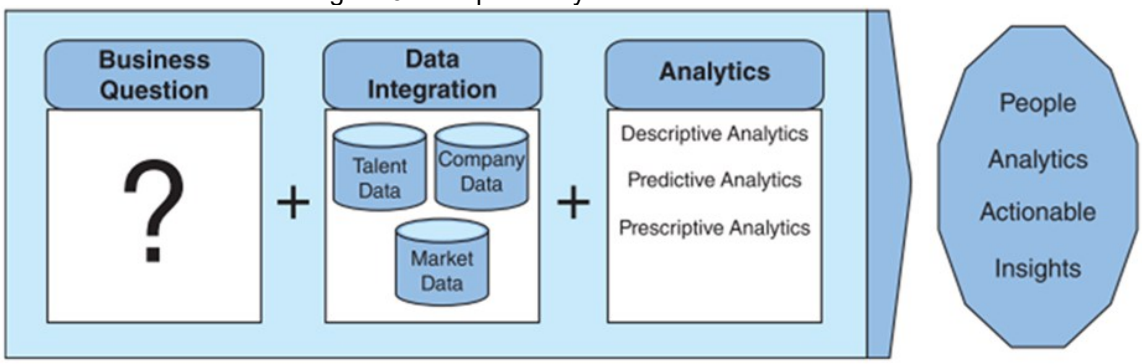
A distinção entre análises descritivas, preditivas e prescritivas mostra como os dados são usados para decisões, desde identificar o estado atual até recomendar ações. No ambiente organizacional, a *big data* amplia essa capacidade de análise, referindo-se a grandes volumes de dados que superam os métodos tradicionais de análise (ISSON; HARRIOTT, 2016, p.11).

O termo *big data* é frequentemente utilizado para descrever grandes conjuntos de dados que são tão grandes que os métodos tradicionais de armazenamento, análise e partilha de dados podem ser inadequados (EDWARDS; EDWARDS, 2016, p.105, tradução nossa).

Big data pode ser definida como ativos de informações de alto volume, alta velocidade e/ou alta variedade que exigem novas formas de processamento para permitir uma melhor tomada de decisão, descoberta de insights e otimização de processos (GARTNER, 2012 apud EDWARDS; EDWARDS, 2016, p.105, tradução nossa).

De acordo com Isson e Harriot (2016, p74), todo o processo de criação de insights acionáveis com *people analytics* está resumido na Figura 9:

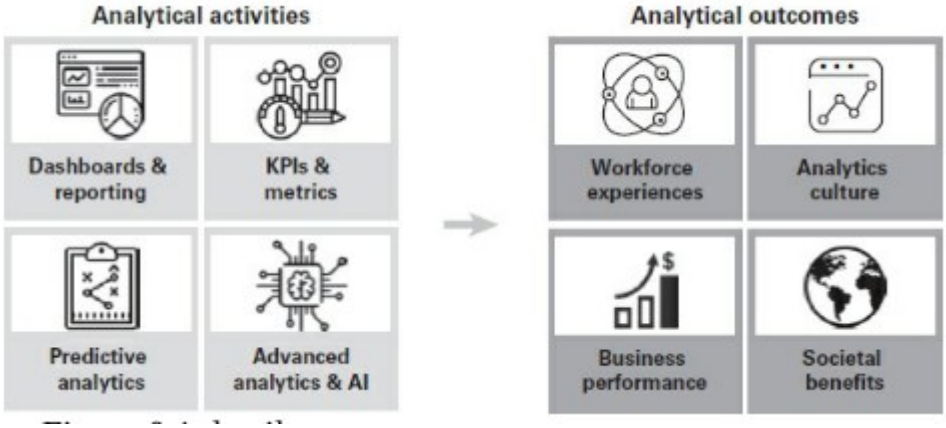
Figura 9 - People Analytics Virtuous Process



Fonte: Isson e Harriot (2016, p.74)

Como é exibido na Figura 10, a definição de análise de pessoas consiste em um conjunto de atividades (FERRAR; GREEN, 2021, p.76, tradução nossa):

Figura 10 - People analytics consists of multiple activities and outcomes

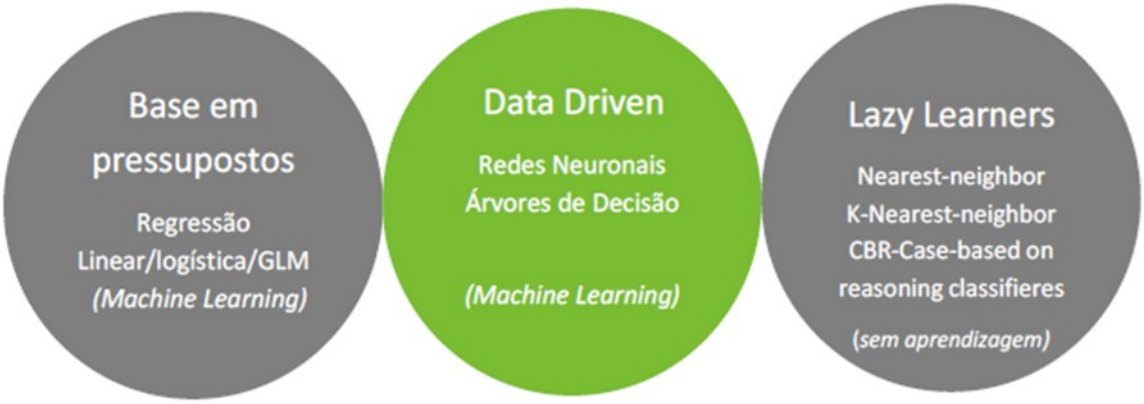


Fonte: Ferra e Green (2021, p. 76)

a) Painéis e relatórios: compartilhando informações e insights de dados pessoais em relatórios e painéis, muitas vezes de maneira formal, padronizada e repetível;
b) Principais *key performance indicator*s (KPIs) e métricas: medem as métricas mais importantes para a empresa e aquelas de valor para executivos de alto escalão, conselhos e investidores;
c) Análise preditiva: utilizando técnicas estatísticas e outras técnicas de análise matemática para prever, prever e planejar cenários para o futuro;
d) Análise avançada e inteligência artificial (IA): aplicação de técnicas e tecnologias muito avançadas para fornecer insights e recomendações, utilizando, por exemplo, aprendizagem automática, IA, aprendizagem profunda e computação cognitiva.

A Figura 11 ilustra três abordagens distintas no contexto de *data mining*:

Figura 11 - Exemplos de abordagens a dados em Data Mining



Fonte: Canais (2016, p. 10)

De acordo com Berry e Linoff (2004, apud CANAIS, 2016) As técnicas de mineração de dados são amplamente aplicáveis em diversos campos devido à sua eficácia no desenvolvimento de modelos descritivos e preditivos, conforme demonstrado por numerosos estudos. As atividades principais em mineração de dados incluem:

Modelos Descritivos:

a) Clustering;

b) Regras de associação;
c) Análise de links;
d) Visualização.

Modelos Preditivos:

a) Classificação (para variáveis categóricas);
b) Regressão (para variáveis contínuas).

2.2.1 Algoritmos para obtenção do modelo preditivo

Segundo Lee, (2019), a inteligência artificial (IA) aplicada aos negócios explora bancos de dados para identificar correlações ocultas que frequentemente passam despercebidas aos seres humanos. Utilizando decisões e resultados históricos de uma organização, a IA treina algoritmos com dados rotulados que podem superar até mesmo os profissionais mais experientes. Isso ocorre porque os humanos tendem a fazer previsões baseadas em preditores fortes, ou seja, em um número limitado de dados altamente correlacionados com um resultado específico, geralmente em uma relação clara de causa e efeito.

De acordo com Russell (2021), a preocupação com decisões automatizadas decorre do potencial viés algorítmico, onde algoritmos de aprendizado de máquina podem gerar decisões enviesadas em áreas como empréstimos, moradia, emprego, seguros, liberdade condicional, sentenças judiciais e admissões universitárias. O uso explícito de critérios como raça nessas decisões tem sido ilegal por décadas em muitos países e é proibido pelo artigo 9 do Regulamento Geral sobre a Proteção de Dados (RGPD) em diversos contextos.

Provost e Fawcett (2016) defendem que na ciência de dados, a aprendizagem de máquina é geralmente associada à aplicação de algoritmos de indução a conjuntos de dados e é frequentemente considerada sinônima da etapa de modelagem no processo de mineração de dados. Este campo de estudo científico se dedica ao desenvolvimento de algoritmos de indução e outros que possuem a capacidade de aprendizado.

De acordo com Géron (2019), o aprendizado de máquina é eficaz para problemas que exigem ajustes detalhados ou regras extensas, muitas vezes superando abordagens tradicionais. É útil para resolver problemas complexos sem

soluções claras tradicionais, adaptar-se a novos dados e lidar com grandes volumes de informação. Exemplos de aplicações incluem a classificação de imagens de produtos em linhas de produção usando redes neurais convolucionais, a classificação automática de artigos de notícias e comentários ofensivos através do processamento de linguagem natural, a detecção de fraudes com cartões de crédito por meio da detecção de anomalias e a segmentação de clientes com base em suas compras para estratégias de marketing personalizadas.

De acordo com Rouse (2009 apud L'HEUREUX; GROLINGER; CAPRETZ, 2017), sobre as categorias de machine learning, as duas principais categorias de tarefas de aprendizado são: aprendizado supervisionado quando ambas as entradas e suas saídas desejadas (rótulos) são conhecidas, onde o sistema aprende a mapear entradas para prever as saídas e aprendizado não supervisionado quando as saídas desejadas não são conhecidas e o próprio sistema descobre a estrutura dos dados.

Classificação e regressão são exemplos de aprendizado supervisionado: na classificação, as saídas assumem valores discretos, enquanto na regressão as saídas são contínuas. Exemplos de algoritmos de classificação são o *k-nearest neighbour*, regressão logística, e *support vector machine* (SVM), enquanto exemplos de regressão incluem *support vector regression* (SVR), regressão linear, e regressão polinomial (ROUSE, 2009 apud L'HEUREUX; GROLINGER; CAPRETZ, 2017).

Segundo Provost e Fawcett (2016), a aprendizagem supervisionada envolve o desenvolvimento de um modelo que estabelece uma relação entre variáveis independentes (atributos) e uma variável dependente específica (variável alvo). Este modelo prevê o valor da variável alvo com base nos atributos, geralmente utilizando uma função probabilística.

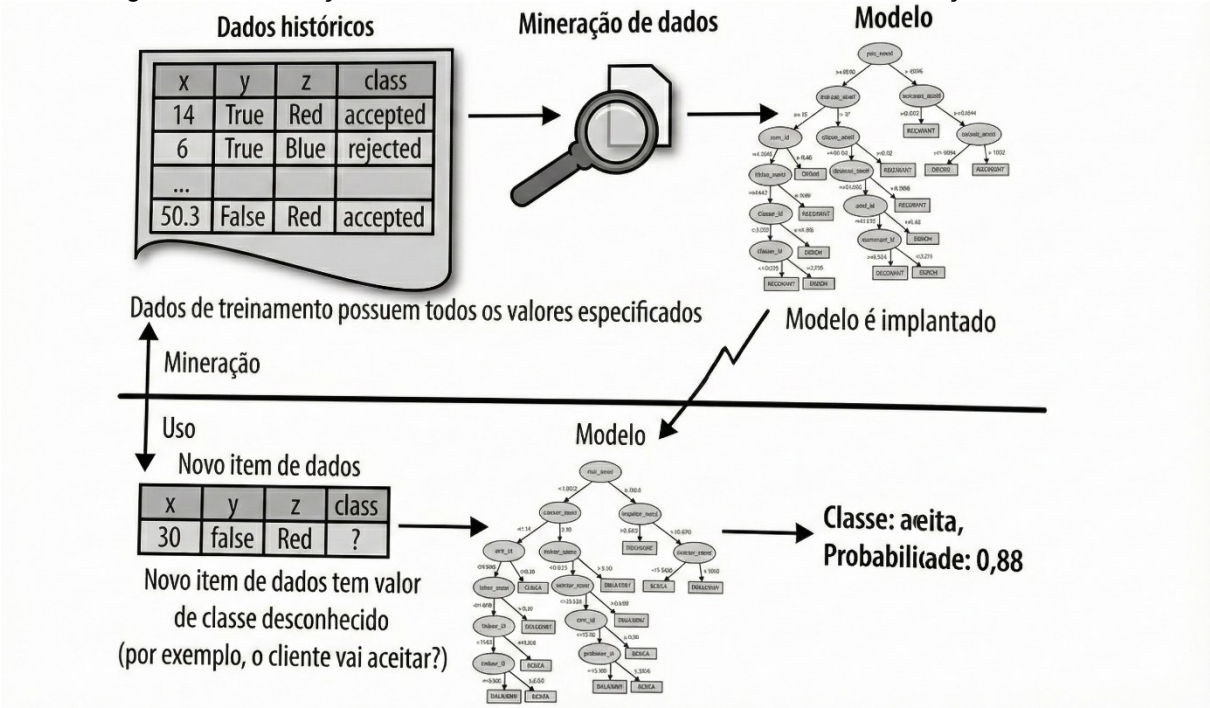
Alguns algoritmos, como redes neurais, podem ser usados para classificação e regressão. O aprendizado não supervisionado inclui clusterização que agrupam objetos com base em critérios de similaridade estabelecidos*; k-means*, por exemplo. A análise preditiva baseia-se no aprendizado de máquina para desenvolver modelos construídos com dados passados, na tentativa de prever o futuro; numerosos algoritmos, incluindo SVR, redes neurais e *naïve bayes*, podem ser usados para esse fim (ROUSE, 2009 apud L'HEUREUX; GROLINGER; CAPRETZ, 2017).

De acordo com Provost e Fawcett (2016), um modelo é uma abstração da realidade desenvolvida para um propósito específico. Sua simplificação se baseia em

pressupostos sobre os elementos considerados relevantes ou irrelevantes para essa finalidade, bem como nas limitações de informações e na viabilidade de manipulação dos dados.

Através da Figura 12, Provost e Fawcett (2016) ilustram o uso dos resultados de mineração de dados:

Figura 12 - Mineração de dados versus o uso dos resultados de mineração de dados



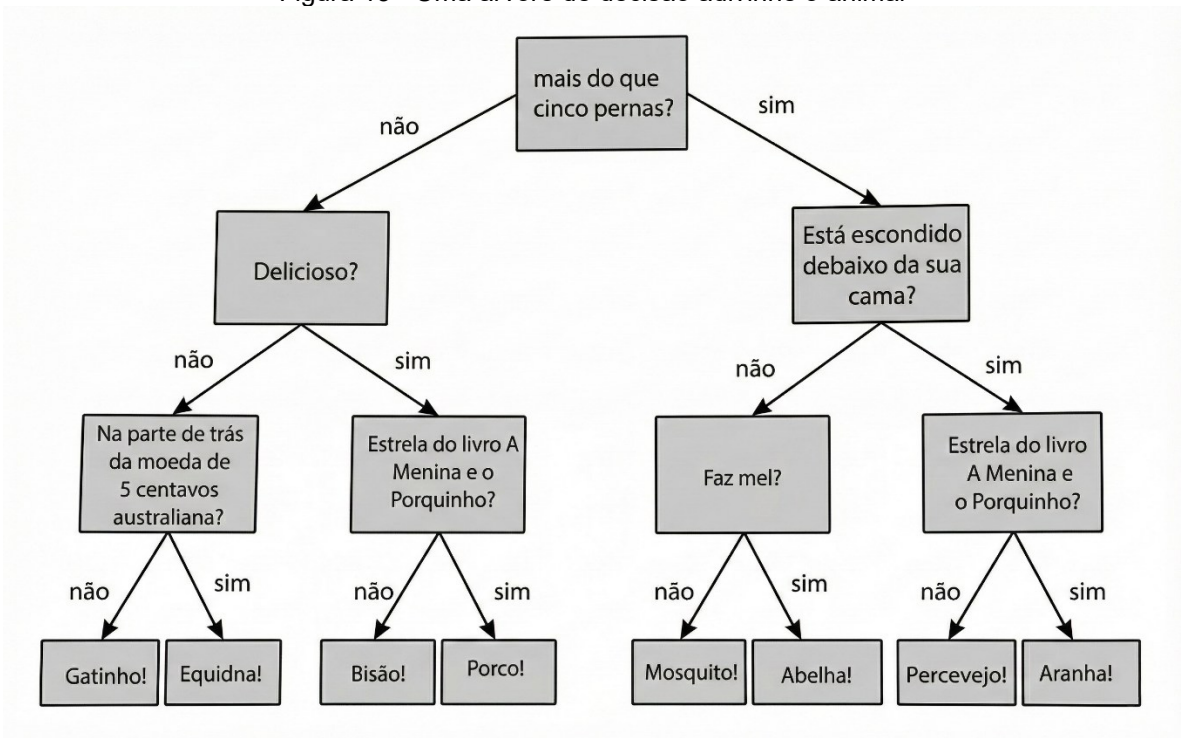
Fonte: Provost e Fawcett (2016, p. 26)

2.2.1.1 Árvores de decisão

Segundo Grus (2016), as árvores de decisão são amplamente recomendadas devido à sua simplicidade de entendimento e interpretação, com um processo de previsão totalmente transparente. Diferentemente de outros modelos, elas conseguem manipular facilmente uma combinação de atributos numéricos (como número de pernas) e categóricos (como delicioso/não delicioso), além de serem capazes de classificar dados mesmo com atributos ausentes.

Através da Figura 13, Grus (2016) exemplifica como uma arvore de decisão define o resultado:

Figura 13 - Uma árvore de decisão adivinhe o animal



Fonte: Grus (2016, p. 202)

De acordo com Provost e Fawcett (2016), a árvore de decisão realiza a segmentação dos dados, onde cada ponto de dados segue exclusivamente um caminho na árvore, culminando em uma única folha. Assim, cada folha representa um segmento específico, definido pelos atributos e valores ao longo do caminho percorrido.

Segundo Provost e Fawcett (2016), os nós internos de uma árvore de decisão são denominados "nós de decisão", pois utilizam os valores dos atributos para determinar o caminho a seguir na árvore.

2.2.1.2 Redes Neurais

De acordo com Lee (2019), esta metodologia emula a arquitetura cerebral ao criar camadas de neurônios artificiais que recebem e transmitem informações, replicando a estrutura das redes neurais biológicas.

De acordo com Provost e Fawcett (2016), uma rede neural pode ser concebida como uma hierarquia de modelos, onde as camadas inferiores processam as características iniciais e cada camada subsequente aplica modelos simples, como regressões logísticas, aos resultados das camadas anteriores. Este processo permite

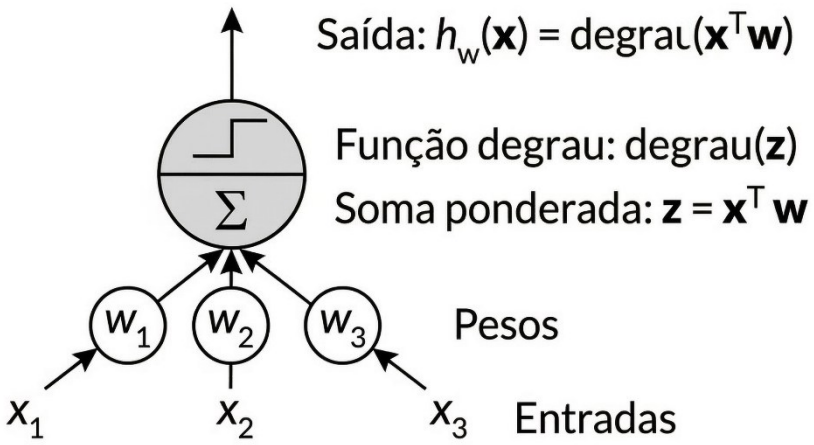
que a rede aprenda a partir de modelos básicos iniciais e, posteriormente, combine os resultados para formar decisões mais complexas, funcionando como uma série de "especialistas" cujas opiniões são integradas.

Provost e Fawcett (2016) argumentam que uma rede neural artificial é um modelo preditivo inspirado no funcionamento cerebral. Assim como o cérebro consiste em neurônios interconectados, cada neurônio na rede neural processa as saídas dos neurônios precedentes e realiza um cálculo, disparando se esse cálculo ultrapassar um determinado limiar.

De acordo com Provost e Fawcett (2016), redes neurais artificiais são compostas por neurônios artificiais que realizam cálculos sobre suas entradas. Utilizadas amplamente em *deep learning*, essas redes são eficazes em tarefas como reconhecimento de caligrafia e detecção facial. No entanto, a natureza de "caixa-preta" dessas redes dificulta a compreensão dos processos internos e redes neurais de grande escala podem apresentar desafios significativos durante o treinamento.

Conforme Géron (2019), a Figura 14 define um neurônio artificial que calcula uma soma ponderada de suas entradas e aplica uma função de degrau:

Figura 14 - Neurônio artificial que calcula uma soma ponderada de suas entradas e aplica uma função de degrau



Fonte: Géron (2019, p. 427)

2.3 INDICADORES DE PERFORMANCE DE PEOPLE ANALYTICS

Dados úteis relacionados a RH são compostos por muitos tipos diferentes de informações e podem incluir o seguinte (EDWARDS; EDWARDS, 2016, p. 22, tradução nossa):

a. Competências e qualificações;
b. Medidas de competências específicas;
c. Treinamentos participados;
d. Níveis de engajamento dos funcionários;
e. Dados de satisfação do cliente;
f. Registros de avaliações de desempenho;
g. Dados de salários, bônus e remuneração.

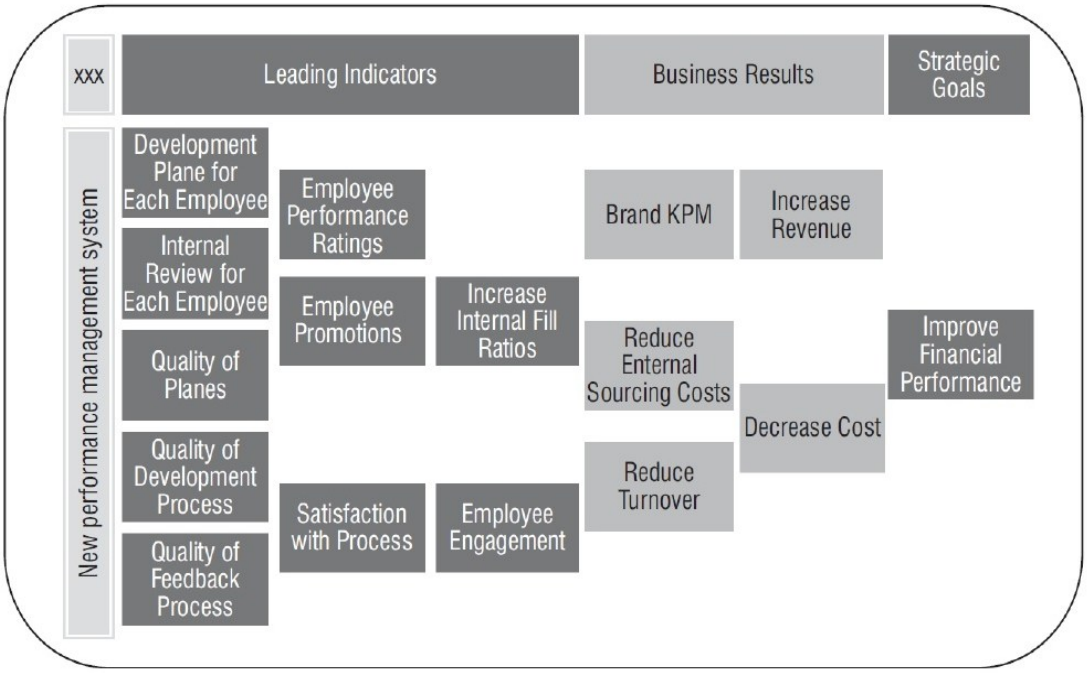
Jantan, Hamdan, & Othman (2010 apud CANAIS, 2016), realizaram alguns estudos recorrendo a técnicas de *data mining* e abordando a temática da retenção de talentos e apresentam 3 componentes de fatores relacionados com a performance do académico, conforme exibido na Figura 15:

Figura 15 - Componentes de fatores relacionados com performance do académico

Fonte: Canais (2016, p. 20)

Pease, Byerly e Fitz-Enz (2013 apud CANAIS, 2016) sugerem os indicadores de performance para potenciar o uso do people analytics, conforme exposto na Figura 16:

Figura 16 - Mapa de indicadores de performance para People Analytics



Fonte: Pease, Byerly e Fitz-Enz (2013 apud CANAIS, 2016, p. 8).

A Figura 17 apresenta um conjunto abrangente de variáveis utilizadas em modelos preditivos de rotatividade de talentos, destacando a complexidade multifacetada do fenômeno. As variáveis são categorizadas em quatro dimensões principais: dados sociodemográficos e geodemográficos, motivação e equilíbrio entre vida pessoal e profissional, desenvolvimento pessoal e profissional, e atratividade do mercado de talentos (ISSON; HARRIOTT, 2016, p. 321):

Figura 17 - Variáveis para um Modelo Preditivo de Rotatividade de Talentos

Sociodemographic and Geodemographic Data	Motivation: Work–Life Balance	Development: Personal and Professional	Talent Market Attractiveness
Gender Age group Age at hire Minority Job level and category Education Experience on current occupation Location Distance to work Distance to commute Parking Commute type (public transportation, own car, or other) Length of service Salary Skills Time to position	Relationship with manager Clarity of goals Confidence in management Fit with company culture and value Type of projects Variety of projects Project completions Ad hoc versus long-term projects Manager feedback frequency, recency, and style Job satisfaction factors Bonding with teammates and colleagues Recognition (reward, trip, or simple thank-you event) Work schedule Flexible hours (arrival, departure)	Specialization and development of technical skills Development of management skills Professional development offered Training or continuing education fees allowance Conference altendance Speaking at conference Training outside and inside the company Opportunity to develop as a whole person Opportunity to fully use skills at work Perceived company support for career development or opportunities offered Development for roles and occupations Performance review ranking Performance review rewards For public organizations, stock options ownership	Unemployment rate by industry and by occupation Unemployment by education level Job openings Job openings by occupation Job openings by region Online job openings by occupation, region, and industry GDP by industry Cost per hire Quits by industry and occupation categories Salary benchmark by roles, occupations, experience, and location Supply-and-demand ratio by occupation, by education level, by region, and by industry Talent shortage indicator Competition's offer to attract similar talent Talent open data Online presence Social media presence (Linkedin, Twitter, Facebook, GitHub, Stack Overflow) Social media behavior Profile change (photo, experience, expertise, education, certification, career) Profile update Online posts Followers/following influencers

Fonte: Isson e Harriot (2016, p. 321)

A Figura 18 destaca fontes de informação essenciais para a análise organizacional, incluindo bancos de dados de RH e pesquisas de satisfação dos colaboradores. Ela também aborda dados de desempenho de vendas e operacionais. Esses dados integrados oferecem insights valiosos, ajudando as organizações a tomar decisões informadas e estratégicas (EDWARDS; EDWARDS, 2016):

Figura 18 - Exemplos de fontes de informação

Information Source	Description	Example
HR database (such as SAP or Oracle)	Information describing the employee across a wide range of people-management activities, including employee personal details, performance, diversity data, promotion information.	Age, gender, education level, role, salary, performance ratings, disabilities, tenure, sickness absence, leaver information, country location, team leader, etc.
Employee attitude survey data (often stored in survey programmes and exported to Excel)	Survey results from annual employee attitude surveys. Either managed in-house or with an external survey provider organization. Typically containing employee engagement data. The information will depend purely on the questions asked in the survey; however, the examples to the right will give an idea of what can be included. Further information on employee engagement and survey data can be found in Chapter 5 (employee attitude surveys – engagement and workforce perceptions).	Job-strain level, employee engagement, job satisfaction, person–organization fit and perceptions of justice.
Customer satisfaction survey data (often stored in survey programmes and exported to Excel)	Often run by marketing or sales functions, customer satisfaction survey data often provides extremely useful information about customer preferences and, when linked to employee profile and operational data, can help organizations to identify how changes to employee skills and behaviours, as well as any operational or process changes, can have a direct impact on the customer experience. Further information on employee engagement and survey data can be found in the employee engagement case study in Chapter 5.	Customer ratings on specific services, branches and customer-facing employees, customer loyalty, customer preferences, customer satisfaction, likelihood of further business.
Sales performance data	Information usually owned by the sales function, recording details of sales performance and revenues. This information is key to determining organizational revenues and the success or otherwise of salespeople and teams in meeting targets. When linked with employee data and operational information it can also help determine the conditions under which salespeople and teams are more likely to succeed or, put another way, what individual or team characteristics tend to make for a more positive bottom line.	Average monthly sales, number of new customers introduced, meeting or not meeting target, individual or team daily, weekly or monthly sales.
Operational performance data	Information or data measuring the successful running of the business. Usually referring to efficiency.	Supermarket scan rates, call-centre call durations, average number of calls dropped out, average number of queries reseived on first contact, time taken to onboard a new customer, etc.

Fonte: Edwads e Edwards (2016, p.35)

2.4 MODELOS DE PEOPLE ANALYTICS

Esta seção investiga *frameworks* teóricos e práticos que facilitam a conexão de dados em insights estratégicos, essenciais para a otimização da gestão de talentos. A seção explora abordagens como a Metodologia *Value Chain*, o Modelo *Focus-Impact-Value* e o Modelo *Eight Step for Purposeful Analytics*, que destacam a importância de integrar técnicas analíticas robustas para validar hipóteses, identificar padrões e auxiliar na toma de decisões organizacionais.

2.4.1 Metodologia Value Chain

Uma metodologia usada por muitos é a cadeia de valor analítica. Foi delineado pela equipe de análise de pessoas do Google já em 2011 e está disponível no site *re:Work* do Google (DEKAS, 2011 apud FERRAR; GREEN, p.191, tradução nossa).

A base desta abordagem é afastar-se das opiniões que impulsionam a ação, passando a utilizar dados, métricas e análises para validar primeiro essas opiniões, crenças e hipóteses. Uma vez revelados os *insights*, estes podem ser usados para apoiar a tomada de decisão necessária para agir (FERRAR; GREEN, 2021, p.192, tradução nossa).

A Figura 19, exemplifica as etapas da metodologia *Value Chain* (FERRAR; GREEN, 2021):

Figura 19 - Metodologia Value Chain

Fonte: Ferrar e Green (2021, p. 192)

2.4.2 Modelo Focus-Impact-Value

O trabalho da equipe de *people analytics* requer foco para ter sucesso na criação de impacto e na entrega de valor. O modelo descrito na figura 2.6 tem uma abordagem simples, embora eficaz (FERRAR; GREEN, 2021, p.192, tradução nossa):

a) Em que devo me concentrar? Os profissionais que se concentram nos desafios empresariais ligados à estratégia de pessoal da empresa têm mais sucesso nos seus esforços;
b) Como posso melhorar meu impacto? As organizações que constroem bases sólidas e não se dedicam à resolução de problemas de tecnologia e dados criam mais impacto a longo prazo;
c) Como posso criar mais valor? Líderes que priorizam seu trabalho com o objetivo final em mente agregam mais valor para o negócio e para os próprios funcionários.

A Figura 20, ilustra os pilares do modelo *Focus-Impact-Value* (FERRAR; GREEN, 2021):

Figura 20 - Modelo Focus-Impact-Value

Fonte: Ferrar e Green (2021, p. 195)

2.4.3 Modelo Eight Step for Purposeful Analytics

Esta metodologia possui oito passos que podem ser agrupados em três partes, conforme apresentado na Figura 21.

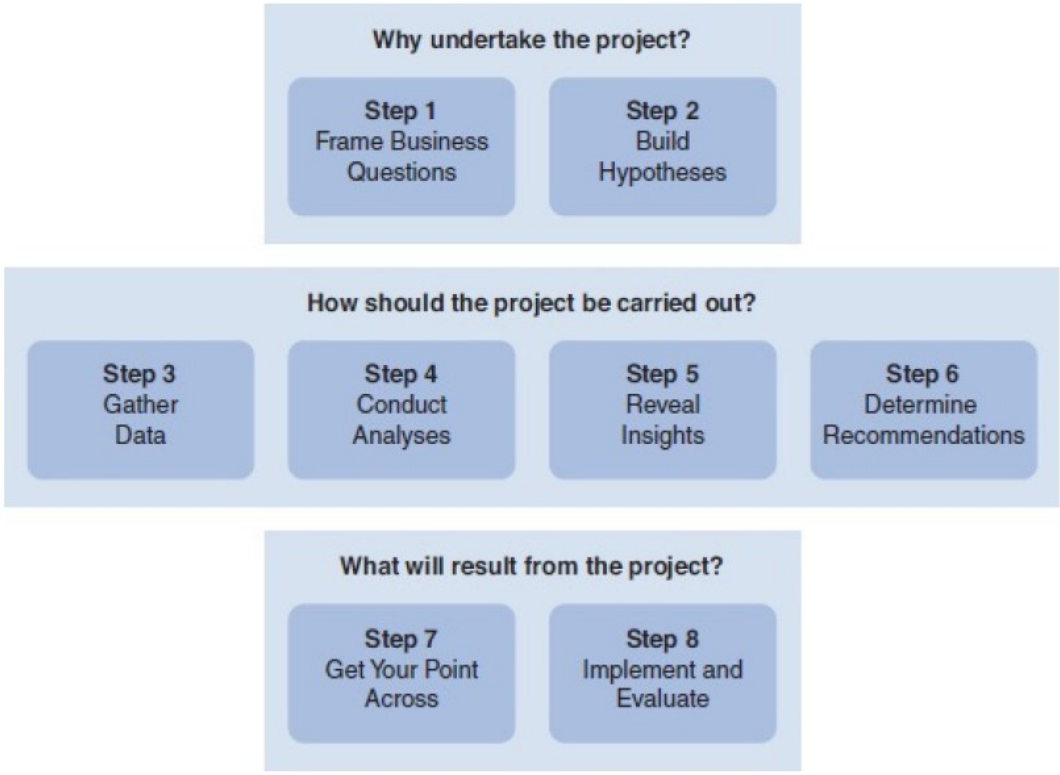
2.4.3.1 Por que realizar o projeto?

Antes de iniciar qualquer projeto de análise, é essencial que seja definido o motivo pelo qual está sendo realizado, visualizando qual o problema de negócio se quer resolver e qual valor que o resultado trará ao negócio. Com isto em mente, é possível definir se realmente vale apena investir força neste projeto ou realocar recursos em outro que trará melhores resultados.

Quando se trata de análise da força de trabalho, as pessoas muitas vezes começam com os dados, mas começando com o fim em mente é uma abordagem

melhor. Em outras palavras, primeiro deve-se definir o que se está tentando mudar e por que (GUENOLE; FERRAR; FEINZIG, 2017, p. 62, tradução nossa).

Figura 21 - Modelo Eight Step for Purposeful Analytics



Fonte: Guenole, Ferrar e Feinzig (2017, p. 62).

Passo 1. Definir questões de negócio: este passo deve vir primeiro, para evitar a realização da análise errada e para dar ao projeto a melhor chance de sucesso. Questões de negócio claramente bem definidas garantem que o trabalho de projeto ou análise realmente são necessários (GUENOLE; FERRAR; FEINZIG, 2017, p. 63, tradução nossa).

Passo 2. Construir hipóteses: formular hipóteses com antecedência ajuda a proteger conclusões com base em relações observadas em seus dados, ao invés de relações subjacentes genuínas. Ao não se definir hipóteses com antecedência à análise, após observar a extração das informações, pode-se começar a formular ideias e conclusões distorcidas que não eram o objetivo inicial do projeto e não agregam valor ao negócio. Hipóteses apropriadas também facilitam a seleção da análise mais adequada ao projeto em questão (GUENOLE; FERRAR; FEINZIG, 2017, p. 63, tradução nossa).

Criar uma hipótese com uma escrita clara e objetiva contribui para testar crenças que gestores têm em relação aos problemas de negócio. A hipótese pode ser uma declaração acompanhada de uma suposição como: se fizermos isso, então isso irá acontecer. Ressaltando sempre a importância de a hipótese ser testável, para que se possa comparar o resultado.

2.4.3.2 Como o projeto deve ser realizado?

Esta etapa do processo é extremamente complexa sob o ponto de vista técnico, por envolver o domínio de tecnologia para estruturar os dados, definindo fontes de informações como bancos de dados de SIRHs, planilhas *excel* e demais fontes de informações, efetuando limpeza e padronização de dados possibilitando extrair conhecimento, devendo sempre utilizar como direcionadores as hipóteses de negócio. Aplica-se também a esta etapa a definição do melhor método estatístico para obter as respostas às hipóteses de negócio criadas em etapas anteriores.

Importante lembrar que as escolhas que se faz nesta etapa influenciam a validade dos resultados do projeto. Esta parte da metodologia provavelmente será a mais complexa, devido ao profundo conhecimento técnico exigido, bem como das habilidades necessárias ao seu desenvolvimento (GUENOLE; FERRAR; FEINZIG, 2017, p. 65, tradução nossa).

Passo 3. Coletar dados: a etapa de coleta de dados requer identificação dos dados mais relevantes para testar as hipóteses e determinar se a qualidade dos dados é suficiente para prosseguir. Devem ser tomadas decisões sobre coletar dados existentes, coletar novos dados ou fazer ambos. Esta etapa também inclui a gestão de quaisquer aspectos legais e éticos relativos aos desafios de privacidade de dados (GUENOLE; FERRAR; FEINZIG, 2017, p. 65, tradução nossa).

Passo 4. Realizar análises: este passo é o que muitas pessoas consideram a parte real da análise. É neste ponto que a metodologia e as estatísticas são aplicadas aos dados para testar as hipóteses e para fornecer a base para as ideias (GUENOLE; FERRAR; FEINZIG, 2017, p. 66, tradução nossa).

Passo 5. Revelar Insights: um dos pedidos mais frequentes na análise da força de trabalho é, segundo os autores pesquisados, “Trazer-me informações, não dados”. Isso porque, os dados são úteis e os resultados analíticos são interessantes, mas a compreensão do contexto e das implicações dos resultados é o que leva aos

insights. Os analistas da força de trabalho devem descobrir insights por dois motivos principais. Em primeiro lugar, os analistas não podem assumir que os patrocinadores de projetos e as partes interessadas são capazes de obter as conclusões mais pertinentes. Os patrocinadores de projetos podem não ser especialistas no tema, de modo que os profissionais de análise da força de trabalho necessitam tornar, esta, uma das suas principais tarefas. O segundo motivo é mais sutil: se os analistas apresentarem apenas dados e análises sem *insights*, os executivos e os patrocinadores de projetos podem tirar as próprias conclusões para melhor se adequarem aos seus preconceitos (GUENOLE; FERRAR; FEINZIG, 2017, p. 66, tradução nossa).

Passo 6. Determinar as recomendações: assim como há necessidade de as organizações obterem dados para gerar informação, os analistas precisam das informações para gerar recomendações. Parte-se, então, da seguinte pergunta: “Se a minha informação é importante o suficiente para destacar, então, o que o negócio deve fazer sobre isso?” Cabe ressaltar que os projetos de análise são todos sobre ajudar o negócio a melhorar seu desempenho, então, embora as informações sejam interessantes, apenas as recomendações ajudarão a melhorar o negócio (GUENOLE; FERRAR; FEINZIG, 2017, p. 67, tradução nossa).

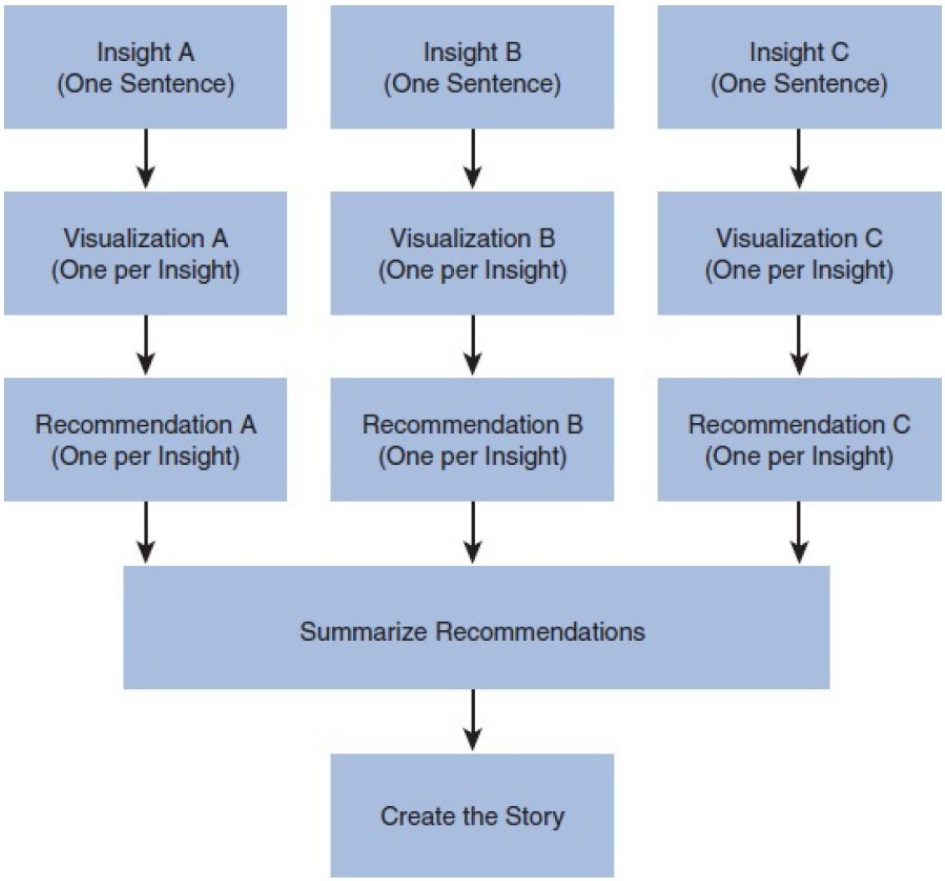
2.4.3.3 Qual será o resultado do projeto?

Projetos de análise bem-sucedidos concluem com decisões. Essas decisões podem gerar mudanças na organização ou podem reforçar o status quo, de qualquer forma, há necessidade de que uma decisão clara seja tomada. A mudança de condução, exige uma comunicação efetiva dos resultados da análise e um processo transparente (GUENOLE; FERRAR; FEINZIG, 2017, p. 68, tradução nossa).

Passo 7. Exponha o seu ponto de vista: todos os projetos de análise têm um momento de verdade. Isso geralmente acontece quando se comunica os resultados do projeto com o patrocinador ou outras partes interessadas para obter seu ponto de vista (GUENOLE; FERRAR; FEINZIG, 2017, p. 68, tradução nossa).

Para Guenole, Ferrar e Feinzig (2017, p. 69): “As histórias devem criar inspirações através de insights, visualizações e recomendações” (tradução nossa), conforme proposto na Figura 22:

Figura 22 - Insights, visualizações e recomendações para a história



Fonte: Guenole, Ferrar e Feinzig (2017, p. 69).

Passo 8. Implementar e avaliar: o passo de implementação e avaliação tem três objetivos discretos. Primeiro, garante que as decisões sejam tomadas como resultado do projeto desenvolvido. Em segundo lugar, ele formula ações para implementação com base nessas decisões. Finalmente, busca avaliar se o projeto retornou valor à organização (GUENOLE; FERRAR; FEINZIG, 2017, p. 70, tradução nossa).

3 ESTADO DA ARTE

Esta seção empreende uma análise aprofundada da literatura científica referente à aplicação de *people analytics*. Mediante uma revisão sistemática, busca-se elucidar o estado da arte neste campo de estudo, abrangendo os algoritmos de aprendizado de máquina empregados, as variáveis preditoras consideradas e os procedimentos metodológicos adotados em contexto global.

3.1 REVISÃO DA LITERATURA

Ao investigar um tema específico, frequentemente encontram-se resultados e abordagens que podem ser divergentes. Uma estratégia eficaz para enfrentar tais inconsistências e embasar a pesquisa é fundamentar-se em produções científicas de alta qualidade e relevância relacionadas ao assunto. Nesse contexto, para garantir o rigor na busca por evidências sem perder o foco aplicado do estudo, adotou-se a abordagem metodológica de revisão sistematizada da literatura.

A revisão sistematizada caracteriza-se como um método de investigação secundária que utiliza a literatura acadêmica como fonte de dados, orientando-se por etapas estruturadas e transparentes para responder a um questionamento de pesquisa previamente definido (MELO et al., 2023). Embora não adote a totalidade dos protocolos exaustivos exigidos por uma Revisão Sistemática da Literatura (RSL) estrita, como o preenchimento de checklists da diretriz PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) ou a execução de meta-análises, a revisão sistematizada incorpora os princípios fundamentais da RSL: a transparência e a reprodutibilidade nos procedimentos de busca, triagem e seleção de artigos (MELO et al., 2023).

Segundo Melo et al. (2023), a estruturação de um protocolo de busca minimiza os vieses do pesquisador, permitindo a identificação de lacunas científicas e o levantamento de metodologias validadas. Dessa forma, a aplicação de estratégias de busca direcionadas e critérios de elegibilidade claros culmina em uma síntese objetiva das evidências científicas disponíveis, conferindo a esta etapa do estudo uma inerente tendência à imparcialidade.

Nas subseções seguintes, serão detalhados os procedimentos metodológicos empregados para a condução desta revisão sistematizada. A subseção 3.2 aborda as bases de dados e as estratégias de busca utilizadas para a execução da revisão. Em seguida, a subseção 3.3 apresenta a análise estatística e bibliométrica das publicações encontradas. Por fim, a subseção 3.4 discute a análise descritiva e integrativa dos trabalhos correlatos incluídos na revisão, evidenciando o estado da arte e o posicionamento desta dissertação frente à literatura atual.

3.2 PROCEDIMENTOS ADOTADOS

A exploração dos dados da pesquisa foi conduzida utilizando quatro bases de dados como principais fontes de informação: Web of Science, Institute of Electrical and Electronics Engineers (IEEE) Xplore, ScienceDirect e Scopus. O acesso a essas plataformas foi realizado por meio do Portal Capes, acessível pelo endereço eletrônico www.periodicos.capes.br, com conexão estabelecida à rede da Universidade Federal de Santa Catarina.

O Web of Science é amplamente reconhecido como uma plataforma essencial para a descoberta científica e acadêmica. Integrando informações de mais de 34 mil periódicos e disponibilizando mais de 217 milhões de registros de metadados, oferece uma base robusta para análises bibliométricas e identificação de tendências científicas, consolidando-se como um recurso indispensável para pesquisas de excelência (WEB OF SCIENCE, 2024).

O Scopus é uma das maiores bases de dados de resumos e citações de literatura científica revisada por pares, abrangendo mais de 28.300 periódicos ativos e 368.000 livros. Essa abrangência permite que pesquisadores identifiquem tendências emergentes, colaborem com especialistas e realizem análises bibliométricas detalhadas, sendo uma ferramenta essencial para a condução de pesquisas acadêmicas de alta qualidade (SCOPUS, 2024).

O ScienceDirect é uma plataforma líder que oferece acesso a uma vasta coleção de publicações científicas e médicas revisadas por pares. Com mais de 18 milhões de conteúdos provenientes de mais de 4.000 periódicos acadêmicos e 30.000 e-books, o ScienceDirect abrange áreas como ciências físicas, engenharia, ciências da vida, ciências da saúde, ciências sociais e humanas (SCIENCEDIRECT, 2024).

O IEEE Xplore é a principal plataforma digital para descoberta e acesso a conteúdo científicos e técnicos publicados pelo Institute of Electrical and Electronics Engineers (IEEE) e seus parceiros editoriais. Com mais de 6 milhões de documentos, incluindo mais de 1,5 milhão de artigos de pesquisa, 4 milhões de artigos de conferências, 14 mil normas técnicas, 66 mil livros e capítulos, e mais de 500 cursos educacionais online, o IEEE Xplore é uma fonte abrangente para engenharias elétrica, ciência da computação e áreas correlatas (IEEE XPLORE, 2024).

Esta pesquisa visa identificar produções científicas que empregam técnicas de aprendizado de máquina aplicadas a dados de recursos humanos, com o objetivo de otimizar a tomada de decisões nessa área, particularmente no que tange à retenção de talentos. A busca sistemática nas quatro bases de dados mencionadas utilizou os seguintes termos em inglês (dada a natureza internacional das bases): *people analytics*, *workforce analytics*, HR *analytics* e *human resource analytics* como palavras-chave para filtrar os estudos relevantes à presente investigação.

3.3 ANÁLISE ESTATÍSTICA

Na investigação das produções científicas pertinentes à temática, inicialmente foi utilizado o termo *people analytics* nas quatro bases de dados, resultando em 240 estudos identificados, distribuídos em 76 na Web of Science, 22 na IEEE Xplore, 13 na Science Direct e 129 na Scopus. Considerando que *people analytics* é uma expressão de uso relativamente recente para o estudo de dados de recursos humanos, a pesquisa foi ampliada para incluir os termos *workforce analytics*, HR *analytics* e *human resource analytics* como palavras-chave adicionais. Com essa ampliação, o total de estudos relacionados ao tema aumentou para 713, sendo 218 na Web of Science, 56 na IEEE Xplore, 35 na Science Direct e 404 na Scopus.

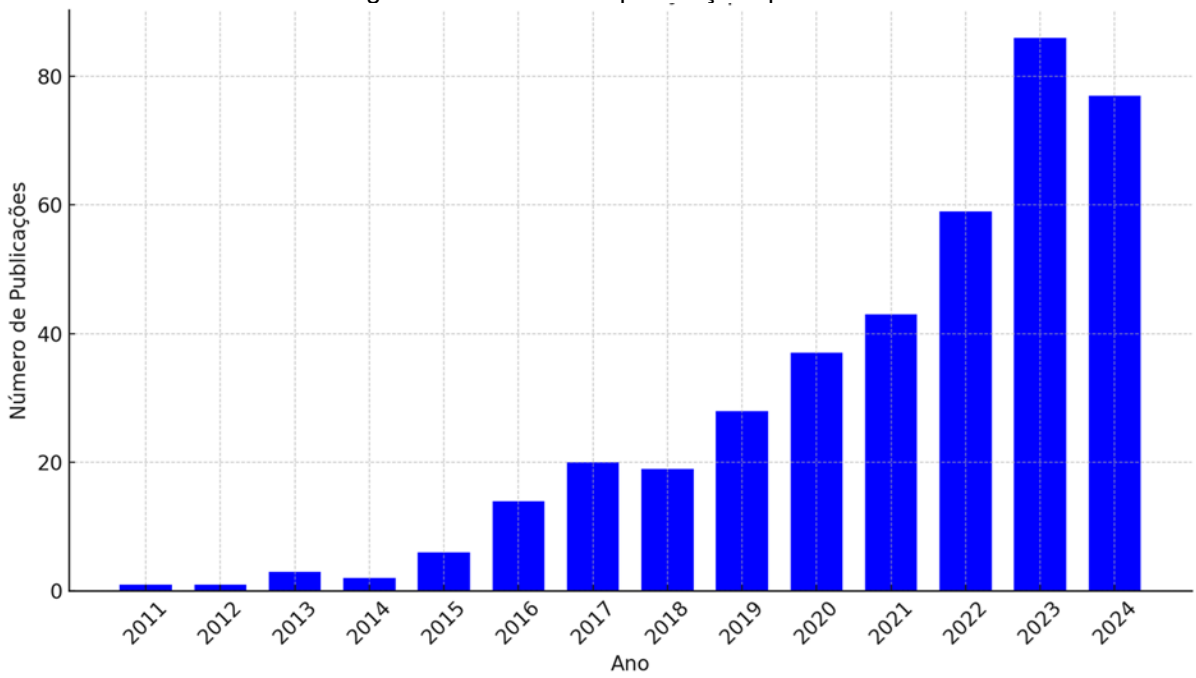
Contudo, tornou-se necessária a depuração dos registros, devido à presença de artigos duplicados em múltiplas bases de dados. Utilizando o *digital object identifier* (DOI) como critério de identificação de duplicatas, e priorizando a manutenção de registros originários de bases diferentes da Web of Science e Scopus (visando à diversificação, dado o volume expressivo de registros nessas duas bases), obteve-se um conjunto final de 396 pesquisas, após a remoção das duplicatas. A distribuição final ficou assim configurada: 21 na Web of Science, 56 na IEEE Xplore, 35 na Science Direct e 284 na Scopus. Contudo, tornou-se necessária a depuração dos registros,

devido à presença de artigos duplicados em múltiplas bases de dados. Utilizando o DOI como critério de identificação de duplicatas, e priorizando a manutenção de registros originários de bases diferentes da Web of Science e Scopus (visando à diversificação, dado o volume expressivo de registros nessas duas bases), obteve-se um conjunto final de 396 pesquisas, após a remoção das duplicatas. A distribuição final ficou assim configurada: 21 na Web of Science, 56 na IEEE Xplore, 35 na Science Direct e 284 na Scopus.

A pesquisa considerou todos os resultados existentes nas bases citadas até a data de 31 de julho de 2024. A seguir, foi realizado a geração de alguns gráficos, a fim de identificar tendências, tipos de publicações e participação de países nos temas abordados.

A Figura 23 demonstra um incremento contínuo nas pesquisas relacionadas ao tema de *people analytics* ao longo dos anos, especialmente após 2020, possivelmente devido a avanços na área de inteligência artificial. É importante destacar que, para o ano de 2024, foram consideradas apenas as pesquisas até 31 de julho, indicando uma tendência de que este ano também apresente um acréscimo significativo em comparação ao ano anterior:

Figura 23 - Número de publicações por ano

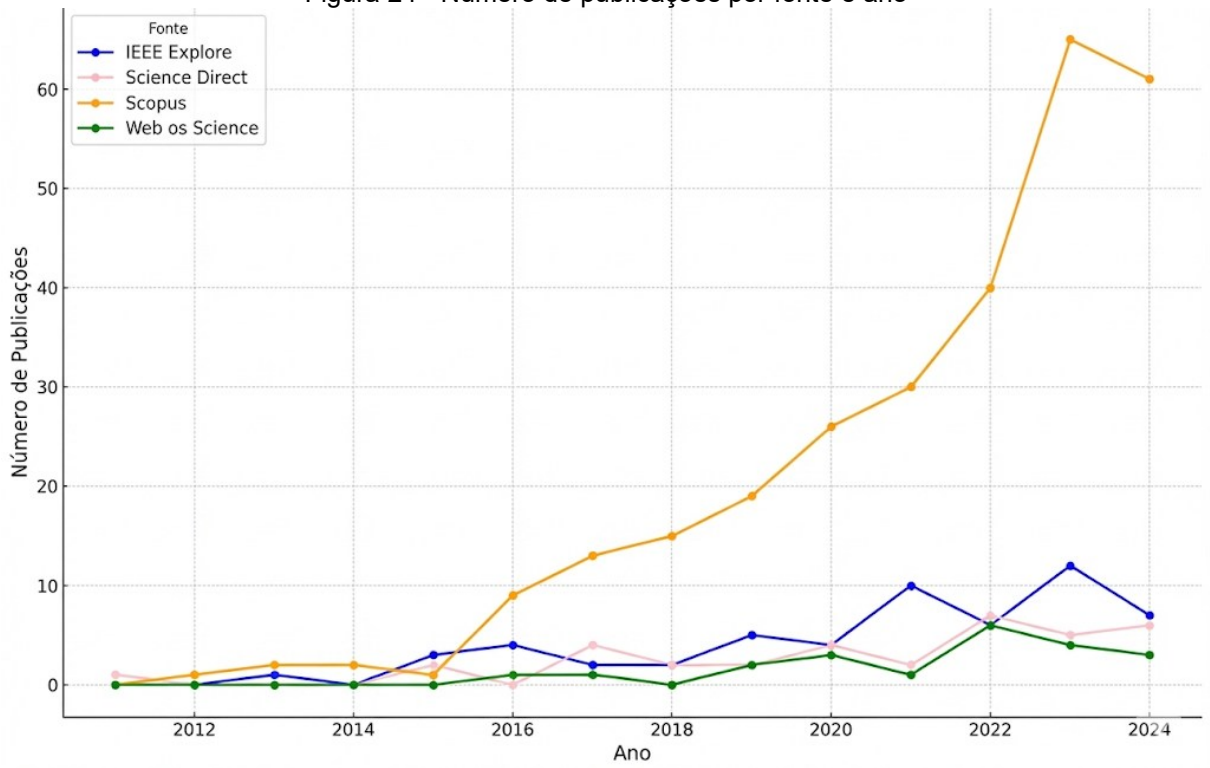


Fonte: Elaborada pelo autor (2026).

Ao comparar a quantidade de artigos publicados e suas fontes, a Figura 24 evidencia a predominância da base de dados Scopus como principal contribuinte. Isso permanece evidente mesmo após o processo de remoção de duplicatas, no qual se buscou preservar registros do mesmo artigo em outras fontes para garantir diversidade.

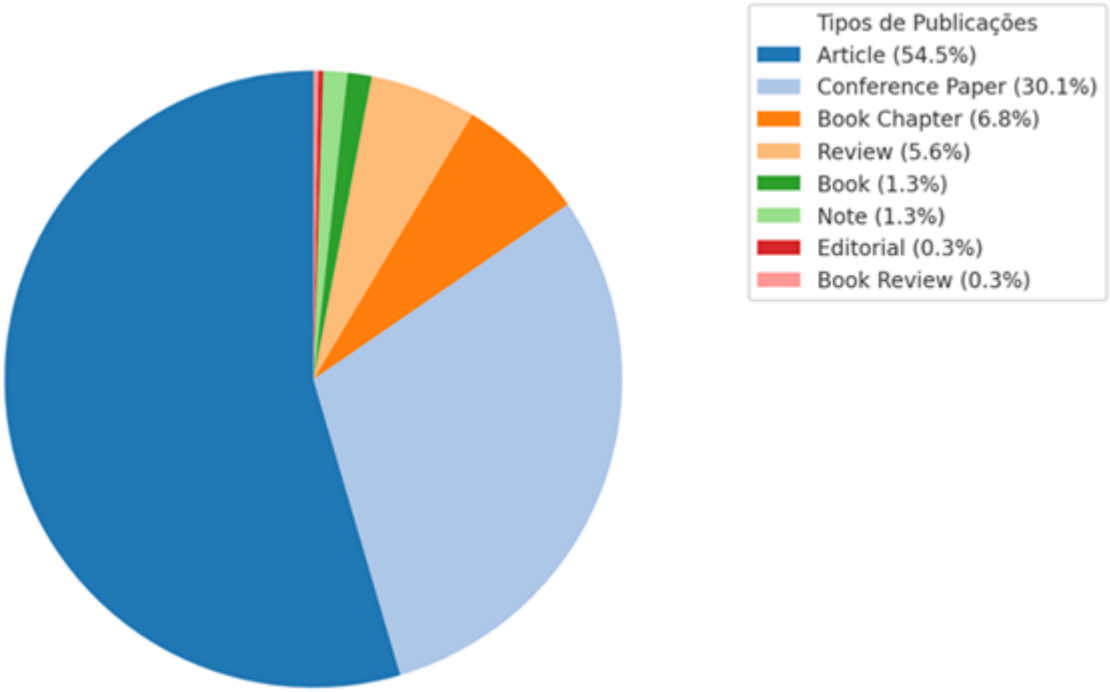
Ao avaliar os tipos de publicações entre as 396 pesquisas analisadas, identificamos que 54,5% correspondem a artigos, enquanto 30,1% são trabalhos acadêmicos apresentados em conferências científicas ou acadêmicas, conforme ilustrado na Figura 25.

Figura 24 - Número de publicações por fonte e ano



Fonte: Elaborada pelo autor (2026).

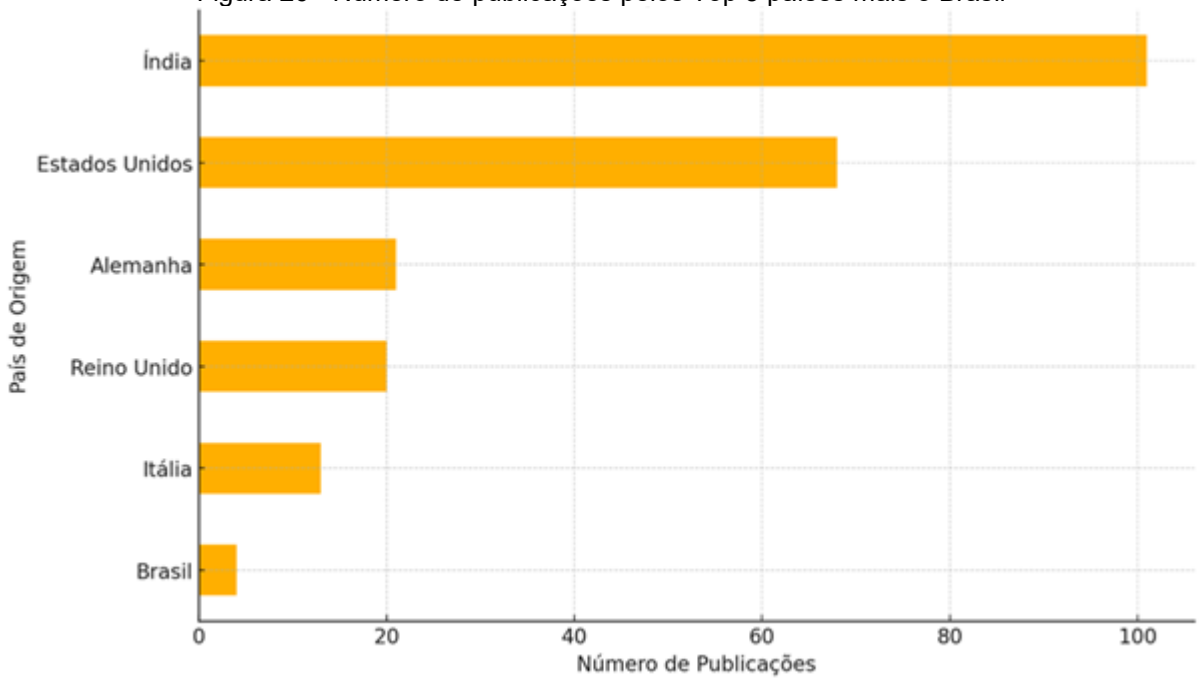
Figura 25 - Percentual de pesquisas por tipo de documentos



Fonte: Elaborada pelo autor (2026).

A Figura 26 também revela a contribuição de cada país para o tema abordado, destacando-se a Índia e os Estados Unidos, com 101 e 68 publicações, respectivamente, enquanto o Brasil apresenta apenas 4 contribuições, sugerindo uma menor exploração acadêmica do tema no contexto brasileiro. É importante salientar que a contribuição dos países foi determinada com base na instituição de origem dos pesquisadores. Além disso, como cada pesquisa pode envolver pesquisadores de diferentes países, ela pode ser contabilizada múltiplas vezes no gráfico:

Figura 26 - Número de publicações pelos Top 5 países mais o Brasil



Fonte: Elaborada pelo autor (2026).

3.4 ANÁLISE DESCRITIVA

A aplicação de algoritmos de *machine learning* à gestão de recursos humanos, especificamente para a predição da rotatividade de funcionários (*turnover*), tem evoluído de estatísticas descritivas simples para modelagens preditivas altamente complexas. A literatura recente concentra-se predominantemente na comparação de algoritmos de classificação para maximizar a precisão e na identificação dos principais impulsionadores do desligamento.

Para estruturar a discussão do estado da arte e afastar-se da análise isolada de publicações, os trabalhos correlatos mais relevantes identificados nesta pesquisa foram classificados e interconectados por meio de dimensões de análise. O Quadro 2 apresenta a matriz de síntese da literatura, cruzando os autores pesquisados com suas respectivas abordagens algorítmicas, tratamentos de dados e lacunas identificadas.

Quadro 2 - Matriz de Síntese dos Trabalhos Correlatos

Autor(es) / Ano	**Foco Temático**	**Abordagem Algorítmica**	**Qualidade de Dados / Explicabilidade**	**Lacuna / Limitação**
Yuan, Kroon e Kramer (2021)	Intenção de rotatividade em pequenas e médias empresas (PMEs).	Bagged tree e modelos de efeitos fixos/aleatórios.	Foco em dados agrupados (estrutura multinível).	Não aborda o impacto financeiro da rotatividade.
Avrahami et al. (2022)	Dinâmica antecedente da rotatividade a longo prazo.	Modelos generalistas de aprendizado de máquina em Big Data.	Uso de grande volume de dados (700 mil colaboradores).	Trata o turnover como evento binário homogêneo, sem diferenciação econômica.
Chung et al. (2023)	Previsão de evasão com base em fatores psicológicos e de ambiente.	Stacking Ensemble (Random Forest + Redes Neurais Artificiais).	Lida com features baseadas em satisfação subjetiva.	Falha em capturar dimensões financeiras e de mercado externo.
Biswas et al. (2023)	Intenção de saída de profissionais de TI.	Gradient Boosting, Random Forest e KNN.	Incorporação de dados externos (redes sociais) via teoria S-O-R.	Foca na intenção de saída e não no prejuízo consolidado.
Chornous e Gura (2020)	Integração de sistemas (HCM, BI) para análise preditiva da força de trabalho.	Modelos preditivos acoplados a plataformas de segurança (ex: IBM SPSS).	Foco na proteção e segurança de dados corporativos sensíveis.	Alta dependência de infraestrutura cara e questões éticas na captura de dados.
Heidemann, Hulter e Tekieli (2024)	O dilema entre desempenho preditivo e transparência no RH.	Algoritmos variados de ML.	Métodos pós-hoc de explicabilidade (SHAP, ALE).	Explicações locais podem ser instáveis para decisões de alto risco.
Shafie et al. (2024)	Rotatividade por agrupamento e perfis de risco.	Redes Neurais Artificiais otimizadas + Clustering.	Aumento de dados (CTGAN) para resolver severo desbalanceamento.	Alta dependência da quantidade e generalização de dados.
Karna et al. (2020)	Efeitos do turnover na estimativa de esforço de projetos ágeis.	Análise empírica baseada em ciclo de vida de aplicativos (ALM).	Avaliação quantitativa de perdas operacionais em equipes.	Não propõe um framework preditivo financeiro para as saídas.
Musanga, Tarambiwa e Zvarevashe (2022)	Otimização da predição de churn de funcionários.	Random Forest, Gradient Boosting com eliminação recursiva (RFE).	Aplicação estrita de seleção de features para acurácia.	Baseia-se apenas em variáveis tradicionais do dataset da IBM.
Yahia, Hlel e Palacios (2021)	Identificação de fatores-chave (ex: viagens) para retenção.	Ensemble learning e Deep Learning.	Transição de Big Data (volume) para Deep Data (qualidade).	Foco em benefícios secundários, carecendo de variáveis contratuais diretas.

Fonte: Elaborada pelo autor (2026).

Nas subseções seguintes, as evidências compiladas na matriz de síntese são discutidas criticamente, sendo agrupadas conforme as principais dimensões de pesquisa encontradas no campo do *people analytics*.

3.4.1 Aprendizado de Máquina (Ensemble Learning) e Desempenho Algorítmico

Existe um consenso crescente na literatura acadêmica quanto à superioridade dos métodos de *ensemble learning* sobre estimadores individuais e modelos lineares clássicos no tratamento de dados de RH, frequentemente caracterizados por relações altamente não-lineares.

Chung et al. (2023) desenvolveram um modelo de empilhamento (*stacking ensemble*) que combinou *Random Forest* (RF) e Redes Neurais Artificiais (RNA), alcançando precisão superior a modelos individuais como Máquinas de Vetores de Suporte (SVM) ou Regressão Logística. De modo complementar, Biswas et al. (2023) demonstraram que arquiteturas de *Gradient Boosting* são altamente eficazes para prever a intenção de saída entre profissionais de Tecnologia da Informação (TI), especialmente ao incorporar fontes de dados externos, como comportamentos em redes sociais.

A robustez específica do algoritmo *Random Forest* é fortemente respaldada por Musanga, Tarambiwa e Zvarevashe (2022), que utilizaram a Eliminação Recursiva de Características (RFE) para avaliar o desempenho de múltiplos algoritmos. Os autores concluíram que o RF oferece o melhor equilíbrio entre precisão e estabilidade. Essa perspectiva também se alinha aos achados de Yuan, Kroon e Kramer (2021), que, ao aplicarem modelos baseados em árvores agregadas (*bagged trees*) a dados agrupados em Pequenas e Médias Empresas (PMEs), destacaram que ensembles baseados em árvores são particularmente resilientes aos ruídos e variações inerentes a dados comportamentais.

3.4.2 Qualidade dos Dados, Desbalanceamento e Explicabilidade

Para além da seleção algorítmica, estudos recentes abordam os desafios metodológicos subjacentes à arquitetura de dados de *people analytics*, como o

desbalanceamento crítico de classes e a natureza frequentemente oculta (caixa-preta) de modelos complexos.

Shafie et al. (2024) propuseram uma abordagem baseada em agrupamento (*clustering*) integrada a Redes Neurais e empregaram Redes Adversárias Generativas Condicionais (CTGAN) para realizar o aumento de dados (*data augmentation*). O trabalho dos autores evidencia que enfrentar o severo desbalanceamento de classes (uma minoria que deixa a empresa frente a uma vasta maioria que permanece) é um requisito fundamental para aprimorar a métrica de *recall*, uma medida crítica, mas muitas vezes negligenciada em favor da acurácia global.

Em relação à relevância das variáveis preditoras, Yahia, Hlel e Palacios (2021) argumentam a favor de uma mudança de paradigma do conceito de *Big Data* para *Deep Data*. Os autores enfatizam que a qualidade e a profundidade dos atributos (como a identificação de frequência de viagens e nível estrutural de satisfação) superam a simples acumulação de volume de dados na obtenção de previsões confiáveis. Contudo, a adoção de modelos analíticos profundos esbarra na resistência da indústria. Para mitigar esse problema, Heidemann, Hulter e Tekieli (2024) aplicaram métodos explicativos *pós-hoc*, como o SHAP (*SHapley Additive exPlanations*), a dados reais de RH, demonstrando que, embora modelos mais complexos gerem melhores predições, camadas robustas de transparência e explicabilidade são indispensáveis para embasar a tomada de decisão estratégica dos gestores.

3.4.3 A Lacuna na Literatura e o Posicionamento Deste Estudo: Contextualização Financeira do Turnover

Embora os estudos detalhados nas subseções anteriores apresentem avanços significativos na identificação de quem possui probabilidade de deixar a organização, observa-se uma escassez notável de pesquisas que integrem o tempo financeiro da rescisão ao alvo da predição.

Os modelos atuais tendem a tratar a rotatividade voluntária como um evento binário e homogêneo. Por exemplo, Avrahami et al. (2022) analisaram uma base massiva de 700 mil funcionários para mapear os antecedentes dinâmicos do *turnover*, mas não distinguiram o impacto econômico de uma saída que ocorre durante a fase de investimento da empresa em comparação a uma saída que ocorre na fase de

retorno. Semelhantemente, Karna et al. (2020) discutiram os efeitos da rotatividade na estimativa de esforço de projetos de TI, tocando nos custos operacionais, mas não chegaram a propor um *framework* preditivo fundamentado no ciclo de vida econômico do colaborador.

O conceito de ponto de equilíbrio do funcionário (*Daily Breakeven*), definido por Isson e Harriott (2016) como o momento em que o valor acumulado gerado pelo colaborador supera o seu custo de aquisição e treinamento, permanece ausente como variável-alvo (*target*) na literatura de *machine learning* aplicada ao RH.

A contribuição desta dissertação (Ineditismo) reside exatamente no preenchimento desta lacuna científica. Diferente das abordagens generalistas do estado da arte, este trabalho propõe um método preditivo que utiliza o *breakeven* econômico como o filtro central de classificação. Em vez de prever a rotatividade geral, o presente estudo desenvolve uma modelagem supervisionada focada estritamente na predição do desligamento *pré-breakeven* distinguindo de forma objetiva o que é a rotatividade natural da rotatividade que materializa perdas financeiras irrecuperáveis para a organização. Ademais, o estudo inova ao fornecer um inventário taxonômico de atributos enriquecidos com variáveis financeiras e contratuais inéditas, adequadas à realidade do mercado tecnológico, elevando o *people analytics* a uma ferramenta direta de proteção de investimentos organizacionais.

4 PROPOSTA METODOLÓGICA

Esta seção dedica-se à apresentação e ao detalhamento da abordagem metodológica desenvolvida para a análise preditiva de turnover, com ênfase estratégica na identificação do ponto de equilíbrio financeiro (*breakeven*) do cargo. A proposta visa transcender a aplicação isolada de algoritmos, estabelecendo um processo estruturado que integra práticas de gestão de dados com técnicas avançadas de *machine learning*.

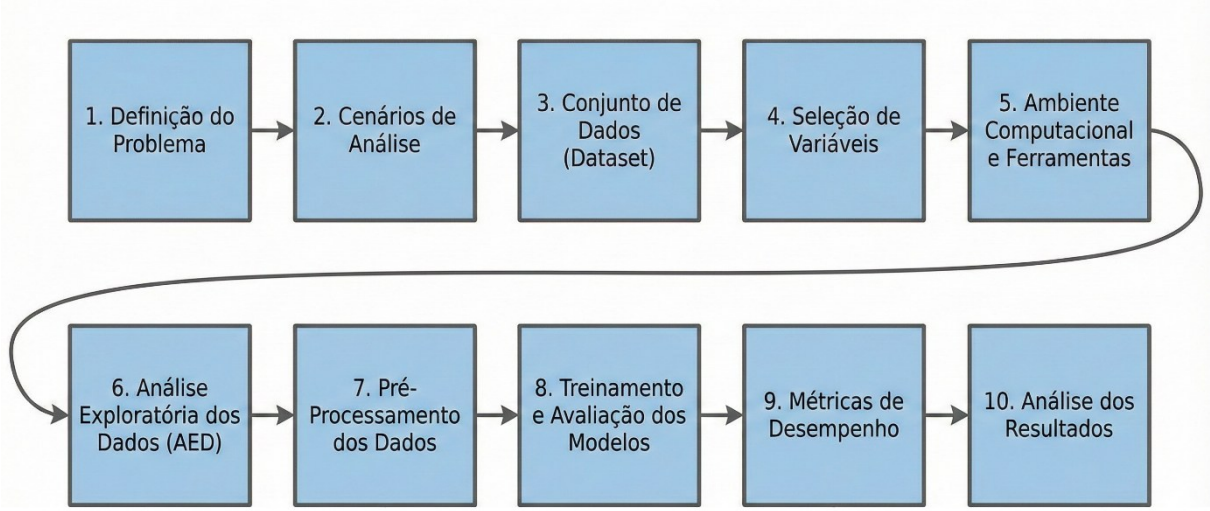
A construção deste modelo preditivo e a seleção dos atributos analisados não foram realizadas de forma arbitrária. A definição das variáveis preditoras fundamentou-se em uma revisão sistemática de publicações científicas, assegurando que os fatores considerados, tais como aspectos demográficos, contratuais e de engajamento, possuam respaldo na literatura acadêmica e relevância comprovada em estudos anteriores sobre retenção de talentos.

Adicionalmente, para garantir a consistência do fluxo analítico, foram adotadas técnicas e atividades da Metodologia de *People Analytics* de 4 Fases, detalhada no Apêndice A deste trabalho. Especificamente, as diretrizes desta metodologia nortearam a definição das hipóteses de negócio (Fase 1 - Planejamento Estratégico e Escopo) e a seleção das técnicas estatísticas e computacionais aplicadas no processo de análise (Fase 2 - Processamento Analítico e Modelagem). Essa adoção permitiu simular um ambiente de tomada de decisão estruturado, alinhando as questões de negócio com a execução técnica.

Para garantir a clareza e a replicabilidade da proposta, a seção está organizada de modo a apresentar sequencialmente o delineamento do estudo, a construção da variável-alvo (*target*) financeira, a descrição do conjunto de dados e as etapas de pré-processamento e modelagem necessárias para a validação dos cenários propostos.

A Figura 27 exemplifica as etapas que foram seguidas no desenvolvimento deste estudo:

Figura 27 - Etapas da proposta metodológica



Fonte: Elaborada pelo autor (2026).

4.1 DELINEAMENTO DO ESTUDO

A presente análise de dados relativa à rotatividade de pessoal (*turnover*) foi aplicada a partir de um conjunto de dados sintético, elaborado pelo autor, de modo que o método analítico proposto não foi implementado em um contexto empresarial real, com representação integral dos distintos papéis funcionais. Entretanto, por meio de uma simulação controlada, são apresentados os resultados correspondentes ao processo analítico.

4.1.1 Definição do Problema e Hipóteses de Pesquisa

A modelagem analítica no campo do *people analytics*, exige uma fundação sólida, frequentemente articulada através da formulação de hipóteses de negócio. Esta prática não apenas direciona a análise, conforme explanado pela Metodologia *Value Chain* (FERRAR; GREEN, 2021) ao sugerir a validação de crenças e opiniões por meio de dados, mas também assegura o alinhamento do projeto ao evitar análises inoportunas descrita pela metodologia *Eight Step Model for Purposeful Analytics* (GUENOLE; FERRAR; FEINZIG, 2017).

Em alinhamento com o processo descrito pela metodologia de *people analytics* de 4 fases (Apêndice A), que define a construção de hipóteses como uma atividade crítica na Fase 1 – Planejamento Estratégico e Escopo, este estudo foi estruturado

com base na definição clara do problema de pesquisa e no subsequente desenvolvimento de pressupostos a serem testados.

O objetivo central do presente estudo é a predição da probabilidade de desligamento de colaboradores em organizações do setor tecnológico, especificamente daqueles que não perfizeram o ponto de equilíbrio econômico (*breakeven*) do cargo.

Para subsidiar esta investigação preditiva, foram delineadas as seguintes hipóteses de pesquisa, de autoria do proponente, visando orientar a análise exploratória de dados (AED) e a subsequente modelagem:

a) H1. Funcionários que possuem satisfação com o trabalho baixa possuem maior tendência de pedir demissão.
b) H2. Funcionários vinculados a modalidade de trabalho presencial possuem maior tendência de pedir demissão.
c) H3. Quando a demanda e oferta no mercado de trabalho está alta ou muito alta para o cargo do funcionário, eles possuem maior tendência de pedir demissão.
d) H4. Quando o salário de mercado for maior que o recebido atualmente pelo funcionário, eles possuem maior tendência de pedir demissão.

As hipóteses H1, H2 e H3 derivam de variáveis preditoras que buscam capturar especificidades regionais e contextuais do setor de tecnologia da informação, enquanto a H4 fundamenta-se em um pressuposto de senso comum amplamente reconhecido na área de gestão de pessoas. A validação empírica dessas premissas será subsequentemente detalhada na subseção 4.5.4 (Validação de Hipóteses e Análise Bivariada).

4.1.2 Estratégia de Modelagem e Cenários de Análise

A análise de dados foi estruturada com o intuito de modelar e prever distintos cenários relativos ao ponto de equilíbrio financeiro (*breakeven*) associado ao cargo. Para tanto, foram mobilizados diversos algoritmos de aprendizado de máquina, de modo a gerar estimativas preditivas comparativas e, por conseguinte, determinar o contexto cuja predição revelou melhor desempenho.

Por meio da Figura 28 é possível visualizar os cenários de análise abordados neste estudo:

Figura 28 - Cenários de análise

Cenário	Descrição
A	Modelo treinado com todos funcionários (independente do tempo de breakeven do cargo) e previsão da demissão somente dos funcionários que não atingiram o breakeven do cargo.
B	Modelo treinado com todos funcionários e previsão da demissão de todos os funcionários (sem qualquer impacto com o tempo de breakeven do cargo)

Fonte: Elaborada pelo autor (2026).

Neste estudo, o foco reside no cenário A, cuja análises das variáveis e a validação das hipóteses apresentadas nas seções subsequentes foram conduzidas considerando o total de colaboradores do conjunto de dados e adotando como variável dependente o desligamento anterior ao alcance do ponto de equilíbrio financeiro (*breakeven*). Na seção 5 (Análise dos Resultados) é apresentado um comparativo entre os distintos cenários.

4.1.2.1 Cenário A: Modelo Focado no Breakeven

Esta abordagem treina o modelo sobre o total de funcionários, mas define uma variável-alvo altamente específica: a demissão ocorrida exclusivamente antes que o ponto de equilíbrio financeiro (*breakeven*) do cargo seja atingido.

Ao utilizar a população total, esta abordagem maximiza o poder estatístico e permite que o modelo aprenda a partir do conjunto mais amplo de não demitidos (funcionários que passaram do ponto de *breakeven*, sejam eles ativos ou demitidos tardiamente).

O desafio desta abordagem é eminentemente técnico e reside no severo desbalanceamento de classes. O evento de interesse demissão precoce será, presumivelmente, um evento raro em comparação com a sua contrapartida e exige técnicas de reamostragem para balanceamento de dados.

4.1.2.2 Cenário B: Modelo Generalista

Esta abordagem representa o modelo padrão de *turnover*, treinado sobre o total de funcionários para prever qualquer evento de demissão, independentemente do tempo de serviço ou impacto financeiro.

Este cenário beneficia-se da maior amostra possível e de um evento de interesse (demissão geral), mitigando problemas de desbalanceamento extremo.

Em contrapartida, o modelo não foca no problema de negócio específico. Ele falha ao agregar dois fenômenos distintos: a evasão que gera prejuízo (pré-*breakeven*) e a evasão que representa um custo operacional padrão (pós-*breakeven*).

4.1.3 Origem, Adaptação e Enriquecimento do Conjunto de Dados (Dataset)

A base de dados que sustenta este estudo foi extraída da plataforma Kaggle, na coleção denominada IBM HR Analytics Employee Attrition & Performance (SUBHASH, 2017). Trata-se de um conjunto sintético, desenvolvido pela equipe de cientistas de dados da IBM, com o propósito de reconstruir cenários analíticos relativos à rotatividade de colaboradores (*attrition*) e às dinâmicas de desempenho no contexto organizacional, de modo a viabilizar a experimentação de modelos preditivos.

O repositório de dados em análise é uma matriz de dados estruturada, compreendendo 1.470 observações, onde cada uma representa um funcionário, e 35 variáveis (ou atributos) que descrevem diferentes facetas da experiência profissional e pessoal de cada indivíduo.

A variável de desfecho primária é a *attrition*, um atributo categórico binário que assume os valores *yes* ou *no*, indicando se o funcionário se desligou da organização. Esta variável é o alvo central para modelos de classificação e análises de inferência causal.

O *dataset* compreende um conjunto de variáveis quantitativas e qualitativas que permitem a análise sob diferentes perspectivas. Entre as variáveis presentes, destacam-se dados demográficos e pessoais, indicadores de desempenho e engajamento e aspectos organizacionais e de ambiente de trabalho.

No entanto, para atender aos objetivos específicos desta pesquisa e refletir o cenário de tecnologia da informação brasileiro, o conjunto original foi submetido a

um rigoroso processo de adaptação e enriquecimento, resultando em um *dataset* customizado.

O Quadro 3 apresenta o comparativo entre os atributos originais da IBM e a estrutura final utilizada neste trabalho:

Quadro 3 - Atributos de referência vs atributos dataset IBM

Atributos de Referência	Atributos Dataset IBM
Idade	Age
Genero	Gender
EstadoCivil	MaritalStatus
PortadorNecessidadesEspeciais	*
PossuiCriancaMenorQue4anos	*
Cargo	JobRole
NivelCargo	JobLevel
Departmento	Department
AnosEmpresa	YearsAtCompany
AnosCargoAtual	YearsInCurrentRole
AnosAtualGestor	YearsWithCurrManager
TipoProjeto	*
AnosDesdeUltimaPromocao	YearsSinceLastPromotion
Salario	MonthlyIncome
RecebeuParticipacaoLucrosResultadosUltimos3anos	*
PercepcaoValorBeneficios	*
GrauEscolaridade	Education
AreaFormacao	EducationField
TreinamentoUltimoAno	TrainingTimesLastYear
AvaliacaoDesempenho	PerformanceRating
Absenteismo	*
SatisfacaoGestor	*
SatisfacaoRelacionamentoColegas	RelationshipSatisfaction
SatisfaçãoTrabalho	JobSatisfaction
NívelEnvolvimentoTrabalho	JobInvolvement
EquilíbrioEntreVidaProfissionalPessoal	WorkLifeBalance
OportunidadeUsarPlenamenteHabilidadesTrabalho	*
DistanciaCasa	DistanceFromHome
TipoTransporte	*
ModalidadeTrabalho	*
HorarioFlexivel	*
HoraExtra	OverTime
ViagemNegocios	BusinessTravel
DemandaOfertaVagasMercadoParaCargo	*
DiferencaSalarialComMercadoParaMesmoCargoNivelExperiência	*
Demitido	Attrition
DemissaoAntesAtingirBreakEvenAcumulado	*

Fonte: Elaborada pelo autor (2026).

Conforme evidenciado no Quadro 3, alguns atributos utilizados na estrutura final deste trabalho não constam no conjunto de dados da IBM e foram assinalados por meio de asteriscos.

Em razão da adoção do *dataset* da IBM como base analítica deste estudo, revelou-se imprescindível o desenvolvimento de uma versão customizada, dotada de diversas adequações para atender aos requisitos do estudo. Entre as principais alterações realizadas, destacam-se:

a) A tradução integral dos atributos e de seus valores categóricos, originalmente apresentados em inglês, para o português;
b) A adequação dos cargos e das áreas de formação, que, no *dataset* original, caracterizam uma organização do setor de saúde, de modo a refletirem o contexto de uma empresa de Tecnologia da Informação;
c) A inclusão de dados sintéticos para cada colaborador, relativos aos atributos listados no Quadro 2, os quais não constavam no conjunto de dados primário.

Em síntese, o novo repositório foi concebido para representar um cadastro de colaboradores do setor de tecnologia da informação, composto por registros sintéticos e integrando a totalidade dos atributos definidos nos Quadros 3,4 e 5 (Atributos de Colaboradores), os quais estão minuciosamente descritos na subseção 4.3 (Seleção e Descrição das Variáveis).

4.2 Definição da Variável-Alvo: Ponto de Equilíbrio (Breakeven)

O método proposto introduz como diferencial analítico a variável do ponto de equilíbrio financeiro acumulado, transcendendo a métrica tradicional de turnover baseada apenas na contagem de desligamentos. O conceito baseia-se na premissa econômica de que a contratação de um colaborador representa um investimento inicial (custos de recrutamento, *onboarding*, treinamento e curva de aprendizado) que deve ser recuperado através da produtividade gerada ao longo do tempo.

Neste contexto, o breakeven é definido como o marco temporal onde o valor agregado acumulado entregue pelo colaborador iguala e supera o custo total acumulado investido pela organização. Formalmente, seja V(t) o valor acumulado e C(t) o custo acumulado no tempo t. O Ponto de Equilíbrio Econômico (t_pec) é definido

como o primeiro momento em que a curva de retorno cruza a curva de custos, conforme a Equação 1:

$$t\_pec = \min\{t \mid V(t) \geq C(t)\} \quad (1)$$

Com base nesta definição matemática, constrói-se a variável-alvo Yi para classificar o desligamento. O objetivo da modelagem é identificar especificamente os casos de prejuízo financeiro, diferenciando-os da rotatividade natural ou positiva. A variável assume valor 1 se o desligamento ocorrer antes do tempo de recuperação do investimento, e 0 caso contrário, conforme a Equação 2:

$$Yi = \begin{cases} 1, & \text{se o colaborador saiu da empresa E tempo de casa} < \text{t\_pec} \\ 0, & \text{caso contrário (ativo ou saiu com lucro)} \end{cases} \quad (2)$$

No contexto deste estudo, essa definição matemática corresponde ao atributo DemissaoAntesAtingirBreakevenAcumulado listado no Quadro 5. Desligamentos onde Yi = 1 representam não apenas uma perda operacional, mas um prejuízo financeiro direto.

Para definição do colaborador com a situação de desligamento anterior a recuperação do valor investido, a lógica foi estruturada a partir do cruzamento de dados cadastrais do colaborador e informações de tempo de retorno de investimento por cargo, conforme estabelecido nos Quadros detalhados na subseção 4.3:

a) Dados cadastrais do trabalhador: Os Quadro 3, 4 e 5 (Atributos do Colaborador) fornecem os dados cadastrais do colaborador, disponibilizando função (Ex: engenheiro de *machine learning*, analista de quality assurance (QA)), tempo de casa atual e situação atual do vínculo (ativo ou demitido).
b) Tempo de retorno por cargo: O Quadro 7 (Breakeven do Cargo) fornece a base temporal para o t_pec. Este atributo define o período médio necessário para que um colaborador, em determinada função, atinja o ponto de equilíbrio.

4.3 SELEÇÃO E DESCRIÇÃO DAS VARIÁVEIS

Esta subseção dedica-se a detalhar o conjunto de variáveis utilizadas nesta pesquisa, abrangendo informações diversas sobre os colaboradores. A construção deste conjunto de dados foi fundamentada em uma extensa revisão da literatura acadêmica, complementada por variáveis inéditas, propostas para enriquecer a análise.

Os Quadros 4, 5 e 6 organizam os atributos em categorias temáticas: dados demográficos, carreira e contratuais, benefícios e remuneração, educação e desenvolvimento, produtividade e desempenho, sentimento e engajamento, e condições de trabalho e desligamento. Para cada atributo, são especificados seu nome, o tipo de dado (qualitativo ou quantitativo), uma descrição sucinta e as referências bibliográficas que embasaram sua inclusão no estudo. As variáveis inéditas, desenvolvidas para refletir a realidade do mercado brasileiro e as particularidades da área de tecnologia da informação (TI), estão explicitamente indicadas nos Quadros 4,5 e 6, sob a designação de proveniência Fonte: Do Autor.

Quadro 4 - Atributos de Colaboradores - Parte 1

Categoria	Nome do atributo	Tipo de dado	Descrição	Referência
Atributos Demográficos	Idade	Inteiro	Idade do funcionário em anos	Heidemann et al (2024), Edwads e Edwards (2016), Avrahami et al. (2022), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
	Genero	Texto	Gênero do funcionário (Feminino ; Masculino)	Heidemann et al (2024), Isson e Harriot (2016), Avrahami et al. (2022), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Biswas et al. (2023), Subhash (2017)
	EstadoCivil	Texto	Estado civil do funcionário (Solteiro; Casado; Divorciado)	Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
	PortadorNecessidadesEspeciais	Binário	Indica se o funcionário é portador de necessidades especiais (Sim; Não)	Edwads e Edwards (2016)
	PossuiCriancaMenorQue4anos	Binário	Indica se o funcionário possui filhos menores de 4 anos (Sim; Não)	Heidemann et al (2024)
Contratuais e de Carreira	Cargo	Texto	Função do funcionário (Executivo(a) de Contas; Engenheiro(a) de Machine Learning; Analista de QA; Diretor(a) de Operações; Arquiteto(a) de Soluções; Gerente de Projetos; SDR (Sales Development Representative); Diretor(a) de Tecnologia; Analista de RH)	Edwads e Edwards (2016), Avrahami et al. (2022), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Subhash (2017)
	NivelCargo	Inteiro	Nível de cargo do funcionário (1 = Júnior; 2 = Pleno; 3 = Sênior; 4 = Gerente; 5 = Diretor)	Isson e Harriot (2016), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Subhash (2017)
	Departmento	Texto	A qual departamento o funcionário pertence (Vendas; Pesquisa e Desenvolvimento (P&D); Recursos Humanos)	Subhash (2017)
	AnosEmpresa	Inteiro	Total de anos que o funcionário trabalhou na empresa	Heidemann et al (2024), Isson e Harriot (2016), Edwads e Edwards (2016), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
	AnosCargoAtual	Inteiro	Total de anos que o funcionário trabalhou em sua função atual	Isson e Harriot (2016), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Subhash (2017)
	AnosAtualGestor	Inteiro	Total de anos que o funcionário trabalhou sob seu gerente atual	Chung et al. (2023), Subhash (2017)
	TipoProjeto	Texto	Tipo de projeto em que o funcionário está envolvido (Inovação; Sustentação; Outros)	Isson e Harriot (2016)
	AnosDesdeUltimaPromocao	Inteiro	Total de anos desde que o funcionário teve sua última promoção na empresa	Heidemann et al (2024), Chung et al. (2023), Yahia, Hlel e Palacios (2021), Shafie et al. (2024), Biswas et al. (2023), Subhash (2017)

Fonte: Elaborada pelo autor (2026).

Quadro 5 - Atributos de Colaboradores - Parte 2

Categoria	Nome do atributo	Tipo dado	Descrição	Referência
Remuneração e Benefícios	Salario	Decimal	Renda mensal do funcionário em Real	Heidemann et al (2024), Isson e Harriot (2016), Yahia, Hlel e Palacios (2021), Shafie et al. (2024), Biswas et al. (2023) , Subhash (2017)
	RecebeuParticipacaoLucrosResultadosUltimos3anos	Binário	Indica se o funcionário recebeu participação nos lucros/resultados nos últimos 3 anos (Sim; Não)	Do Autor
	PercepcaoValorBeneficios	Inteiro	Percepção de valor pelo colaborador dos benefícios pagos pela empresa (Vale alimentação, Plano de Saúde, Vale academia, dentre outros) (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)	Do Autor
Educação e Desenvolvimento	GrauEscolaridade	Inteiro	Nível de escolaridade do funcionário (1 = Ensino médio ou abaixo; 2 = Ensino superior; 3 = Especialização; 4 = Mestrado; 5 = Doutorado)	Heidemann et al (2024), Isson e Harriot (2016), Edwads e Edwards (2016), Yahia, Hlel e Palacios (2021), Subhash (2017)
	AreaFormacao	Texto	Área de formação do funcionário (Ciência da Computação; Bioinformática; Marketing; Curso Técnico; Recursos Humanos; Outros)	Avrahami et al. (2022), Chung et al. (2023), Subhash (2017)
	TreinamentoUltimoAno	Inteiro	Total de vezes que o funcionário teve uma sessão de treinamento no último ano	Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
Desempenho e Produtividade	AvaliacaoDesempenho	Inteiro	Classificação de desempenho do funcionário (1 = Baixo; 2 = Bom; 3 = Excelente; 4 = Extraordinário)	Isson e Harriot (2016), Edwads e Edwards (2016), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
	Absenteismo	Inteiro	Indica o percentual de ausência do funcionário no trabalho em relação aos dias trabalhados	Edwads e Edwards (2016)
Engajamento e Sentimento	SatisfacaoGestor	Inteiro	Quão satisfeito o funcionário se sente com o relacionamento com seu gerente (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)	Isson e Harriot (2016)
	SatisfacaoRelacionamentoColegas	Inteiro	Quão satisfeito o funcionário se sente com o relacionamento com seus colegas (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)	Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
	SatisfaçãoTrabalho	Inteiro	Quão satisfeito o funcionário se sente com seu trabalho (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)	Isson e Harriot (2016), Edwads e Edwards (2016), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Shafie et al. (2024), Subhash (2017)
	NivelEnvolvimentoTrabalho	Inteiro	Quão envolvido o funcionário se sente com seu trabalho (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)	Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Biswas et al. (2023), Subhash (2017)
	EquilibrioEntreVidaProfissionalPessoal	Inteiro	Como o funcionário se sente em relação ao equilíbrio entre vida profissional e pessoal (1 = Ruim; 2 = Bom; 3 = Melhor; 4 = Ótimo)	Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
	OportunidadeUsarPlenamenteHabilidadesTrabalho	Inteiro	Avaliação do funcionário sobre a oportunidade de usar plenamente suas habilidades no trabalho (1 = Pouco; 2 = Razoavelmente; 3 = Consideravelmente; 4 = Completamente)	Isson e Harriot (2016)

Fonte: Elaborada pelo autor (2026).

Quadro 6 - Atributos de Colaboradores - Parte 3

Categoria	Nome do atributo	Tipo dado	Descrição	Referência
Condições de Trabalho e Deslocamento	DistanciaCasa	Inteiro	A que distância o funcionário mora do trabalho em quilômetros	Isson e Harriot (2016), Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Subhash (2017)
	TipoTransporte	Texto	Meio de transporte utilizado pelo funcionário para ir ao trabalho (Carro Próprio; Transporte Público; Outros)	Isson e Harriot (2016)
	ModalidadeTrabalho	Texto	Modalidade de trabalho do funcionário (Presencial; Remoto)	Do Autor
	HorarioFlexivel	Binário	Indica se o funcionário tem horário flexível (Sim; Não)	Isson e Harriot (2016)
	HoraExtra	Binário	Se o funcionário fizer horas extras (Sim ; Não)	Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Subhash (2017)
	ViagemNegocios	Texto	Com que frequência o funcionário viaja a negócios (Viaja_Raramente; Viaja_Frequentemente; Não-Viaja)	Chung et al. (2023), Musanga, Tarambiwa e Zvarevashe (2022), Yahia, Hlel e Palacios (2021), Subhash (2017)
Indicadores Externos de Mercado	DemandaOfertaVagasMercadoParaCargo*	Inteiro	Avaliação da demanda e oferta de vagas no mercado para o cargo do funcionário (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)	Do Autor
	DiferencaSalarialComMercadoParaMesmoCargoNivelExperiência*	Decimal	Diferença salarial eentre o valor do salário de mercado para o mesmo cargo e nível de experiência do funcionário e o atualmente recebido pelo funcionário (em Real)	Do Autor
Dados de Desligamento	Demitido	Binário	Indica se o funcionário saiu da empresa de forma voluntária (Sim ; Não)	
	DemissaoAntesAtingirBreakEvenAcumulado*	Binário	Indica se o funcionário pediu demissão voluntária antes de atingir o ponto de equilíbrio (break-even) acumulado (Sim; Não)	Do Autor

Fonte: Elaborada pelo autor (2026).

O Quadro 7 permite a análise de cargos, seus níveis hierárquicos, faixas salariais e a correspondente demanda no mercado de trabalho. O objetivo destas informações é avaliar o impacto na demissão do funcionário, conforme salário, oferta e demanda do cargo no mercado de trabalho:

Quadro 7 - Análise de Cargo no Mercado de Trabalho

Nome do atributo	Tipo de dado	Descrição
Cargo	Texto	Função do funcionário (Executivo(a) de Contas; Engenheiro(a) de Machine Learning; Analista de QA; Diretor(a) de Operações; Arquiteto(a) de Soluções; Gerente de Projetos; SDR (Sales Development Representative); Diretor(a) de Tecnologia; Analista de RH)
NivelCargo	Inteiro	Nível de cargo do funcionário (1 = Júnior; 2 = Pleno; 3 = Sênior; 4 = Gerente; 5 = Diretor)
Salario	Decimal	Estimativa do salário mensal de mercado para o cargo e nível de cargo (em Real)
DemandaOfertaVagasMercadoParaCargo	Binário	Avaliação da demanda e oferta de vagas no mercado para o cargo (1 = Baixo; 2 = Médio; 3 = Alto; 4 = Muito Alto)

Fonte: Elaborada pelo autor (2026).

O Quadro 8 apresenta uma estrutura de dados focada na análise do retorno sobre o investimento baseado no cargo dos funcionários. Esta estrutura de dados permite avaliar a eficiência financeira e o tempo de retorno associado a cada cargo na organização:

Quadro 8 - Breakeven do Cargo

Nome do atributo	Tipo de dado	Descrição
Cargo	Texto	Função do funcionário (Executivo(a) de Contas; Engenheiro(a) de Machine Learning; Analista de QA; Diretor(a) de Operações; Arquiteto(a) de Soluções; Gerente de Projetos; SDR (Sales Development Representative); Diretor(a) de Tecnologia; Analista de RH)
MesesBreakEven	Inteiro	Quantidade de meses para atingir o ponto de equilíbrio (break-even) de investimento no funcionário
MesesBreakEvenAcumulado	Inteiro	Quantidade acumulada de meses em que a produtividade ou receita acumulada gerada pelo funcionário ao longo de todo o seu tempo na empresa ultrapassa o custo acumulado total

Fonte: Elaborada pelo autor (2026).

4.3.1 Justificativa das Variáveis Propostas pelo Autor

A customização do conjunto de dados sintético original exigiu a incorporação de atributos inéditos, projetados para refletir com maior fidelidade as dinâmicas de retenção e as particularidades contemporâneas do mercado de tecnologia. Com o intuito de conferir rigor metodológico à pesquisa, a modelagem dessas novas variáveis foi embasada em literaturas consolidadas de áreas correlatas. A seguir, apresenta-se a fundamentação científica que sustenta a adoção de cada um desses atributos.

A variável “Recebeu Participação nos Lucros/Resultados nos últimos 3 anos” foi incluída por representar a presença de incentivos variáveis atrelados ao desempenho organizacional, os quais são amplamente discutidos na literatura de remuneração estratégica. Milkovich, Newman e Gerhart (2014) destacam que planos de participação nos lucros (*profit sharing*) e esquemas similares de remuneração variável podem reforçar o alinhamento entre objetivos organizacionais e interesses dos empregados, contribuindo para maior comprometimento e retenção, em comparação a estruturas compostas apenas por salário fixo. O manual da WorldatWork sobre recompensas também apresenta a participação nos resultados como componente relevante de programas de *total rewards*, apontando sua utilização por empresas que buscam engajar e reter talentos em ambientes competitivos (WORLDATWORK, 2007). Dessa forma, o histórico de recebimento de PLR (Participação nos Lucros/Resultados) nos últimos anos funciona, neste estudo, como um indicador da exposição do colaborador a incentivos coletivos de desempenho, cuja literatura associa à maior lealdade organizacional e menor propensão ao desligamento voluntário.

A variável “Percepção de Valor dos Benefícios” foi concebida com base no conceito de *total rewards*, que considera que a proposta de valor ao empregado envolve não apenas o salário direto, mas também benefícios, desenvolvimento, bem-estar e reconhecimento (WORLDATWORK, 2007). Obras de gestão de recursos humanos apontam que a satisfação com benefícios, especialmente quando estes são percebidos como relevantes e bem comunicados, constitui um elemento central da satisfação global com o pacote de recompensas e influencia decisões de permanência na organização (DESSLER, 2003). Relatórios profissionais de entidades e consultorias em RH indicam, ainda, que organizações com pacotes de benefícios percebidos como robustos e bem comunicados registram menores taxas de

rotatividade voluntária em relação a empresas cujos benefícios são considerados pouco atrativos pelos colaboradores (MPLOYER ADVISOR, 2025). Assim, a percepção subjetiva do valor dos benefícios é incorporada ao modelo como uma forma de mensurar a avaliação global que o colaborador faz do componente não salarial da recompensa, que a literatura de recompensas vincula à satisfação e à retenção.

A variável “Modalidade de Trabalho” (presencial, híbrida ou remota) reflete os diferentes arranjos de trabalho que se consolidaram, sobretudo após a disseminação do teletrabalho, e que passaram a ocupar posição central nas estratégias de gestão de pessoas. Dessler (2003) discute que práticas de flexibilização, como *home office* e horários flexíveis, compõem importantes ferramentas de atração e retenção, particularmente em mercados de trabalho com escassez de profissionais qualificados. Entidades profissionais, como a Society for Human Resource Management (SHRM), relatam que políticas de trabalho remoto ou híbrido, quando alinhadas às preferências dos colaboradores, tendem a elevar engajamento, sentimento de autonomia e equilíbrio entre vida pessoal e profissional, fatores que, em conjunto, contribuem para reduzir a intenção de desligamento (SHRM, 2022). Dessa forma, a modalidade de trabalho é utilizada como *proxy* da flexibilidade oferecida pela organização e da adequação entre o formato de trabalho e as preferências individuais.

A variável “Demanda/Oferta de Vagas no Mercado para o Cargo” foi introduzida como medida do contexto externo de oportunidades de emprego disponíveis ao colaborador. Ehrenberg e Smith (2012) destacam que a probabilidade de mobilidade laboral aumenta em ambientes caracterizados por baixas taxas de desemprego e elevado número de vagas abertas, pois os trabalhadores tendem a explorar alternativas para melhorar salário e condições de trabalho. O relatório OECD (*Organisation for Economic Co-operation and Development*) *Employment Outlook* 2023 mostra que, em muitos países da OCDE, a combinação de desemprego baixo e elevado número de vagas resultou em mercados de trabalho com escassez de mão de obra, com níveis historicamente altos de oportunidades por trabalhador (OECD, 2023). Documentos correlatos observam que, em períodos de maior escassez de mão de obra, intensificam-se as taxas de desligamentos voluntários, na medida em que empregados se sentem mais confiantes para trocar de empregador em busca de melhores condições (OECD, 2023). Assim, ao incluir uma medida de demanda/oferta de vagas específicas para o cargo e nível de experiência, este estudo procura capturar

o grau de alternativas externas disponíveis ao indivíduo, variável que a literatura de economia do trabalho indica ter influência direta sobre a decisão de *turnover* voluntário.

Por fim, a variável “Diferença Salarial com o Mercado para Mesmo Cargo/Nível de Experiência” baseia-se na teoria da equidade e na literatura de remuneração por mercado, que destacam a importância das comparações salariais externas na formação da satisfação e nos comportamentos de permanência. A teoria da equidade de Adams, amplamente discutida em manuais de comportamento organizacional, como o de Robbins, argumenta que indivíduos comparam suas recompensas (como salário) com as de outros em posições equivalentes, e percepções de injustiça distributiva tendem a gerar insatisfação e, potencialmente, comportamentos de saída da organização (ROBBINS, 2009). A literatura de compensação orientada ao mercado, sintetizada em obras como a de Milkovich, Newman e Gerhart, enfatiza que a competitividade salarial em relação ao mercado é requisito básico para atração e retenção de talentos, recomendando o uso de pesquisas salariais externas para orientar faixas internas (MILKOVICH; NEWMAN; GERHART, 2014). Dessa forma, a diferença salarial em relação ao mercado é utilizada aqui como indicador objetivo de competitividade da remuneração individual, cuja relevância para comportamento de permanência é amplamente reconhecida na literatura de compensação e de comportamento organizacional.

4.4 AMBIENTE COMPUTACIONAL E FERRAMENTAS

A análise de dados conduzida pelo autor, procedeu-se mediante a utilização da linguagem Python e de um conjunto de bibliotecas analíticas apropriadas, aplicado a um conjunto de dados sintético customizado para abranger os atributos discriminados nos Quadros 3,4 e 5 (Atributos de Colaboradores).

O desenvolvimento completo do projeto está disponibilizado em repositório público no GitHub (https://github.com/etson-rodrigues/people-analytics-ufsc-project), sob licença MIT License.

No desenvolvimento técnico da análise de dados, empregou-se uma diversidade abrangente de bibliotecas em Python, com ênfase em manipulação de dados, rotinas de pré-processamento, implementação e avaliação de modelos de

aprendizado de máquina, bem como para outras funções de suporte, conforme detalhado no Quadro 9:

Quadro 9 - Bibliotecas em Python

Biblioteca	Função
inflection	Converte strings entre diferentes convenções de nomenclatura (camelCase, snake_case).
unidecode	Realiza a transliteração de texto Unicode para uma representação aproximada em ASCII, removendo acentos e diacríticos.
uuid	Gera identificadores únicos universais (UUIDs) para garantir a unicidade de objetos ou registros.
pandas	Biblioteca essencial para a manipulação e análise de dados, provendo estruturas de dados de alta performance como o DataFrame.
numpy	Pacote fundamental para a computação científica, oferecendo suporte a arrays multidimensionais e uma vasta gama de funções matemáticas.
collections	Módulo nativo do Python que implementa tipos de dados especializados. A classe Counter é utilizada para a contagem de elementos, como na análise da distribuição de classes.
matplotlib	Biblioteca para a criação de visualizações de dados estáticas, animadas e interativas de forma programática.
seaborn	Interface de alto nível, baseada no matplotlib, para a criação de gráficos estatísticos informativos e visualmente atrativos.
scipy	Ecossistema de software de código aberto para matemática, ciência e engenharia. Utilizada para testes estatísticos (chi2_contingency) e para gerar distribuições de probabilidade (randint, loguniform) na otimização de hiperparâmetros.
scikit-learn (sklearn)	Framework abrangente para aprendizado de máquina. Os módulos utilizados no projeto englobam: • Pré-processamento: Normalização de características (MinMaxScaler). • Seleção e Validação de Modelos: Divisão de dados (train_test_split), validação cruzada estratificada (StratifiedKFold) e busca de hiperparâmetros (RandomizedSearchCV). • Modelos: Implementação de algoritmos como Regressão Logística (LogisticRegression), Random Forest (RandomForestClassifier), K Neighbors (KNeighborsClassifier), SVM e Multilayer Perceptron (MLPClassifier). • Métricas: Avaliação de performance dos modelos (accuracy_score, roc_auc_score, f1_score, etc.).
imbalanced-learn (imblearn)	Biblioteca especializada no tratamento de conjuntos de dados desbalanceados, fornecendo técnicas de reamostragem como o algoritmo híbrido SMOTETomek.
lightgbm	Framework de Gradient Boosting que utiliza algoritmos baseados em árvores, notável pela sua alta velocidade de treinamento e eficiência no uso de memória.

Fonte: Elaborada pelo autor (2026).

4.5 ANÁLISE EXPLORATÓRIA DOS DADOS (AED)

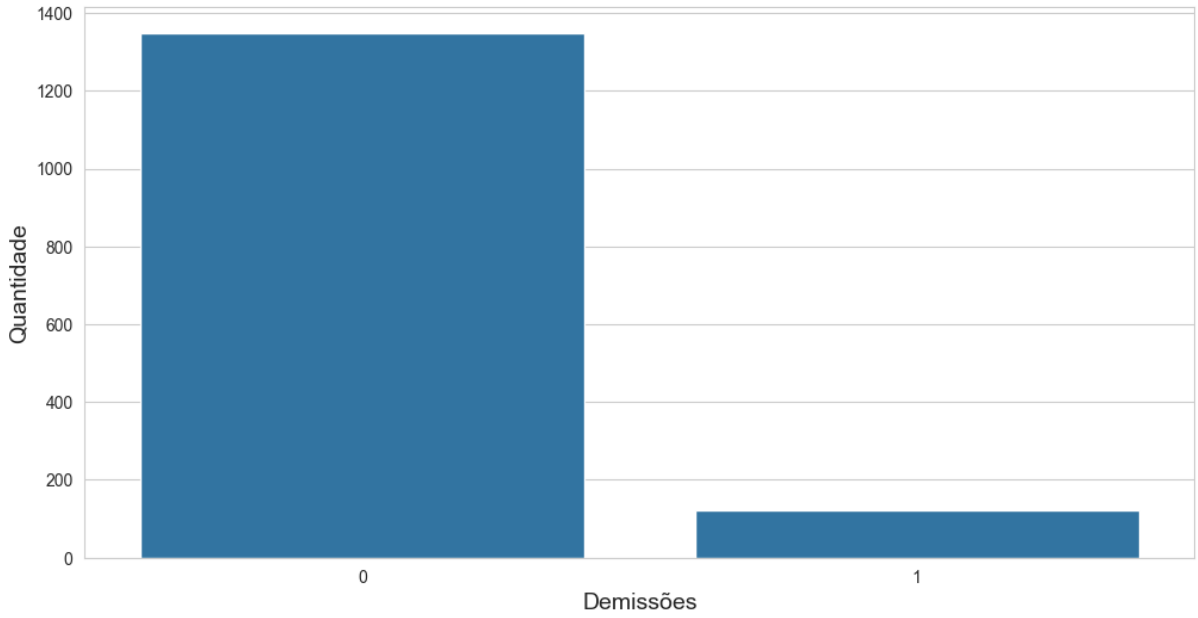
A fase de análise exploratória dos dados (AED) constitui um estágio metodológico crucial em qualquer empreendimento analítico, atuando como o alicerce para a modelagem preditiva subsequente. Conforme estabelecido por Bruce e Bruce (2019), a AED transcende a mera manipulação superficial de conjuntos de dados, buscando, primordialmente, a obtenção de uma compreensão intuitiva e perspicaz do fenômeno em estudo por meio de técnicas de sumarização e visualização.

A articulação dos procedimentos de AED nesta subseção estrutura-se em uma progressão lógica, iniciando-se pelo exame da variável de interesse e, sequencialmente, avançando para a análise das variáveis preditoras e a validação das premissas teóricas.

4.5.1 Distribuição e Balanceamento da Variável Alvo

A Figura 29 exibe a distribuição da variável-alvo demissão_antes_atingir_break_even_acumulado no *dataset* utilizado no estudo. Observa-se que 1349 registros correspondem à classe não (colaboradores que permaneceram até o ponto de equilíbrio financeiro) e apenas 121 registros à classe sim (colaboradores desligados antes do *breakeven*):

Figura 29 - Total de demissões



Fonte: Elaborada pelo autor (2026).

Esse desnível evidencia um forte desbalanceamento de classes, condição essa que foi tratada na etapa de modelagem para evitar vieses na previsão de desligamentos precoces.

4.5.2 Análise Exploratória de Variáveis Numéricas

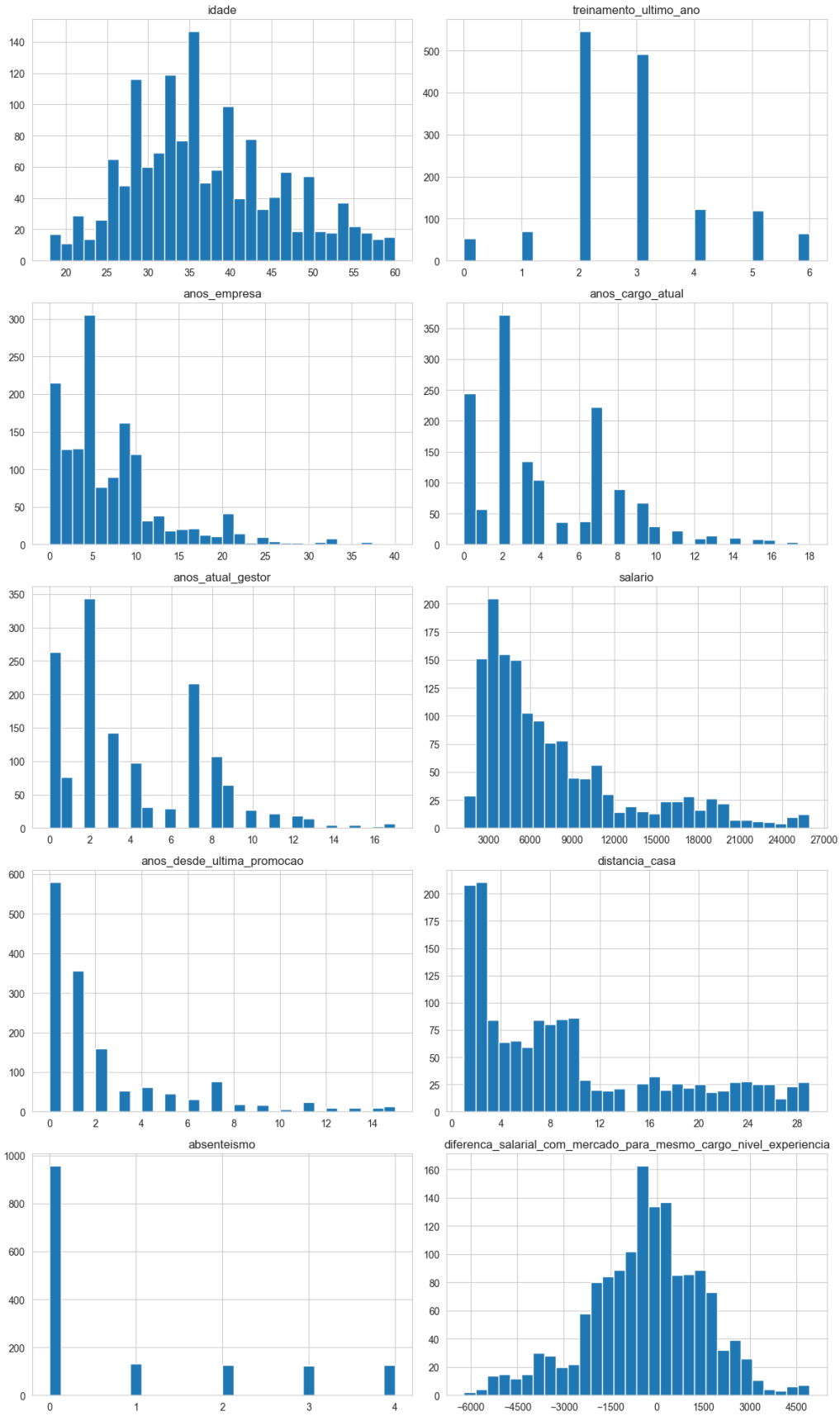
A Figura 30 apresenta um painel de histogramas que descreve as distribuições de frequência dos atributos numéricos da amostra.

As distribuições para anos_empresa, anos_cargo_atual e anos_atual_gestor exibem uma forte assimetria à direita. Isso indica que a grande maioria dos colaboradores possui um tempo de casa e de função relativamente baixo

(concentração nos primeiros anos), com uma longa extensão de funcionários mais antigos.

O histograma da variável diferenca_salarial_com_mercado_para_mesmo_cargo_nivel_experiencia apresenta uma distribuição aproximadamente normal, com a notável característica de ser centrada em um valor negativo (próximo de -1.800 Reais). Esta é uma evidência quantitativa de que a política de remuneração da empresa é, na média, de pagar abaixo da média de mercado.

Figura 30 - Distribuições de frequência dos atributos numéricos



Fonte: Elaborada pelo autor (2026).

4.5.3 Análise Exploratória de Variáveis Categóricas

A Figura 31 exibe um painel de distribuições de frequência univariadas para um conjunto de atributos categóricos.

A visualização permite traçar um perfil modal da organização, identificando que os colaboradores são majoritariamente compostos por profissionais com especialização, em regime de trabalho remoto, e em cargos de nível pleno e júnior. Notavelmente, os níveis de satisfação com o trabalho, percepção de valor dos benefícios e demanda de mercado para os cargos são fortemente assimétricos, com a grande maioria dos colaboradores concentrada nas categorias alto e muito alto.

Figura 31 - Distribuições de frequência univariadas para um conjunto de atributos categóricos

Fonte: Elaborada pelo autor (2026).

4.5.4 Validação de Hipóteses e Análise Bivariada

Nesta subseção, é estabelecida a confrontação empírica das proposições teóricas do estudo com os dados observados, alinhando-se à fase de análise exploratória de dados (AED).

A análise bivariada constitui um pilar fundamental da AED (análise exploratória dos dados), dedicando-se à investigação da associação entre pares de variáveis. Conforme Bruce e Bruce (2019), a AED visa a obtenção de uma compreensão intuitiva e perspicaz dos dados por meio de sumarização e visualização. Neste contexto, a análise bivariada emprega métodos estatísticos e gráficos para quantificar a relação entre os preditores e a variável-alvo, fornecendo o discernimento necessário para estruturar a modelagem preditiva subsequente.

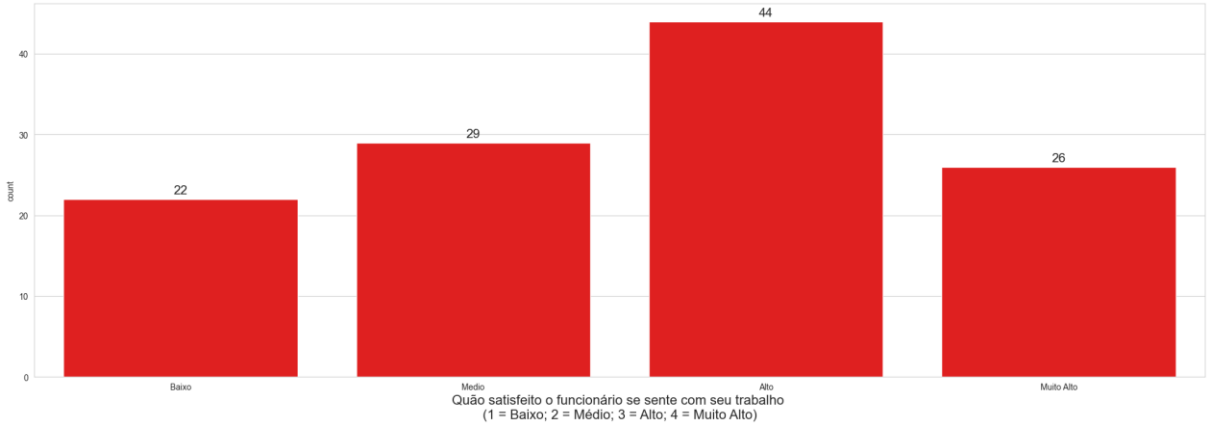
O procedimento detalhado nesta subseção é ilustrado pela Figura 32, que conecta a variável dependente (rotatividade de funcionários) às hipóteses previamente formuladas na subseção 4.1.1, as quais serão submetidas ao teste empírico por meio da análise bivariada:

Figura 32 - Hipóteses formuladas

Fonte: Elaborada pelo autor (2026).

H1. Funcionários que possuem satisfação com o trabalho baixa possuem maior tendência de pedir demissão: A hipótese é falsa, segundo a Figura 33. Observa-se que a maior ocorrência de desligamentos ocorre precisamente entre os colaboradores que reportam um nível alto de satisfação, com um total de 44 colaboradores. Este volume supera não apenas o grupo muito alto (26 colaboradores), mas também o grupo médio (29 colaboradores) e baixo (22 colaboradores). De fato, o número de demissões entre os altamente satisfeitos é o dobro do número de demissões entre aqueles que se declaram pouco satisfeitos.

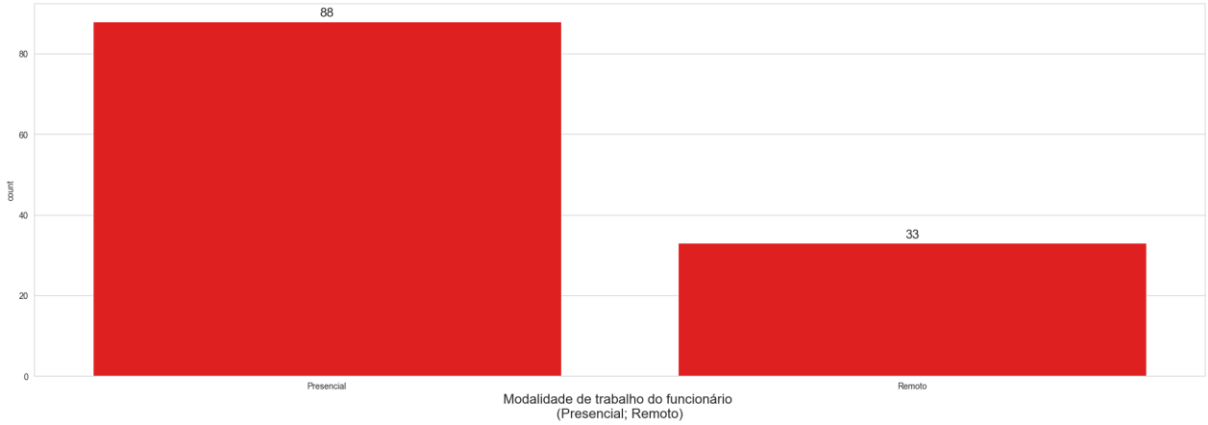
Figura 33 - Quantidade de demissões pelo campo satisfação no trabalho



Fonte: Elaborada pelo autor (2026).

H2. Funcionários vinculados a modalidade de trabalho presencial possuem maior tendência de pedir demissão: A hipótese se mostrou verdadeira. A Figura 34 indica uma disparidade quantitativa notável: o contingente de colaboradores em regime presencial que se desligaram da organização (88 colaboradores) é aproximadamente 2,67 vezes superior ao número de desligamentos no regime remoto (33 colaboradores):

Figura 34 - Quantidade de demissões pelo campo modalidade de trabalho

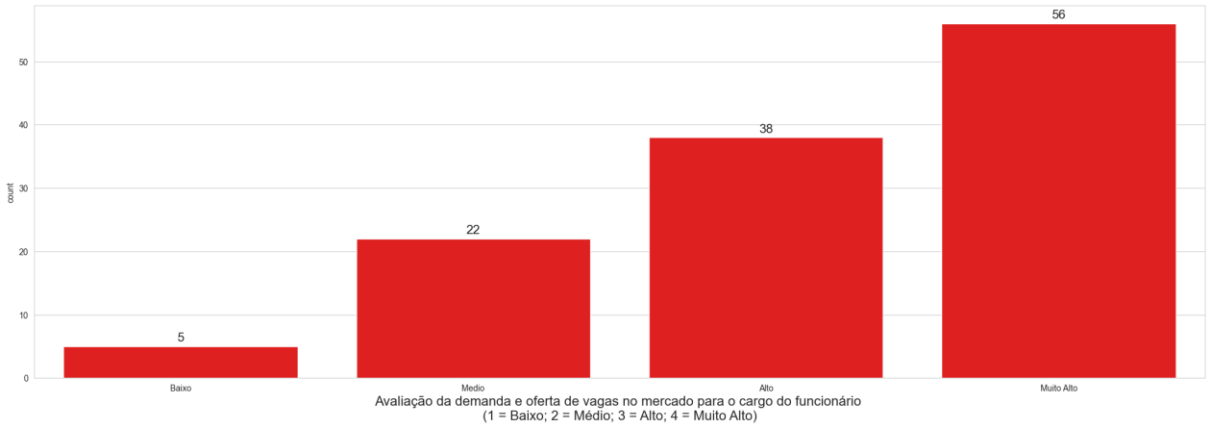


Fonte: Elaborada pelo autor (2026).

H3. Quando a demanda e oferta no mercado de trabalho está alta ou muito alta para o cargo do funcionário, os mesmos possuem maior tendência de pedir demissão: A hipótese se mostrou verdadeira. A Figura 35 expõe uma relação robusta entre as variáveis. Conforme a percepção da demanda de mercado para o cargo aumenta, passando de baixo (5 Colaboradores) para médio (22 Colaboradores), alto

(38 Colaboradores) e muito alto (56 Colaboradores), o volume absoluto de demissões cresce de forma consistente e acentuada:

Figura 35 - Quantidade de demissões pelo campo demanda e oferta de vagas no mercado para o cargo do funcionário

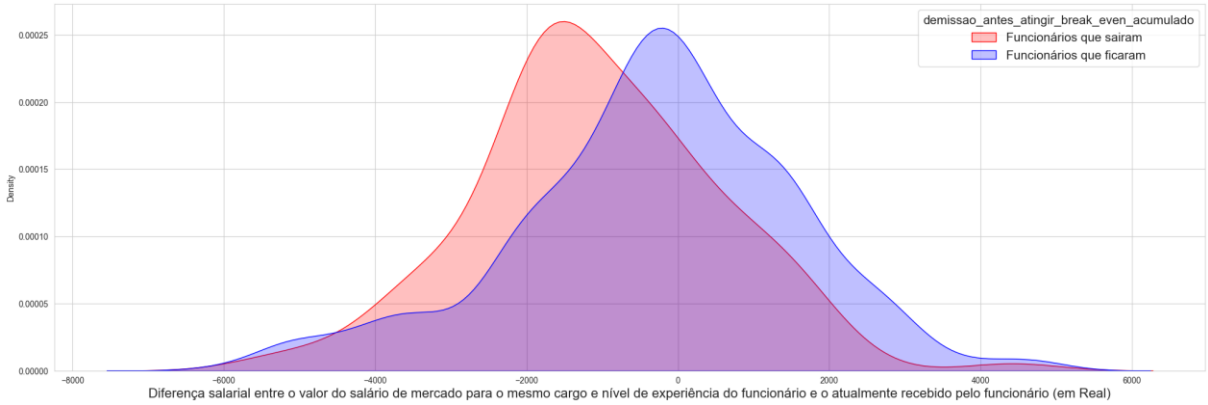


Fonte: Elaborada pelo autor (2026).

H4. Quando o salário de mercado for maior que o recebido atualmente pelo funcionário, os mesmos possuem maior tendência de pedir demissão: Essa hipótese se mostrou verídica. A Figura 36 apresenta duas curvas de densidade de probabilidade, que modelam a distribuição da defasagem salarial para dois grupos distintos: colaboradores que se demitiram (funcionários que saiam, em vermelho) e colaboradores que permaneceram na empresa (funcionários que ficaram, em azul). A variável no eixo horizontal representa a diferença entre o salário do colaborador e o salário de mercado para a função (salário atual - salário de mercado).

O item mais interessante é que o pico da distribuição dos colaboradores que se demitiram (curva vermelha) está localizado em território negativo, aproximadamente em R$ -1.800. Isso indica que o grupo com a maior densidade de demissões era composto por indivíduos que recebiam, em média, um salário abaixo do benchmark de mercado para suas funções.

Figura 36 - Probabilidade de demissão por diferença salarial de mercado



Fonte: Elaborada pelo autor (2026).

4.5.4.1 Síntese das Hipóteses

O Quadro 10 apresenta os resultados das hipóteses criada pelo autor e sua veracidade perante o conjunto de dados (conforme subseção 4.1.1).

Quadro 10 - Resultados das hipóteses

ID	Hipóteses	Conclusão
H1	Funcionários vinculados a modalidade de trabalho presencial possuem maior tendência de pedir demissão	VERDADEIRO
H2	Funcionários que possuem satisfação baixa possuem maior tendência de pedir demissão	FALSO
H3	Quando a demanda e oferta no mercado de trabalho está alta ou muito alta para o cargo do funcionário, os mesmos possuem maior tendência de pedir demissão	VERDADEIRO
H4	Quando o salário de mercado for maior que o recebido atualmente pelo funcionário, os mesmos possuem maior tendência de pedir demissão	VERDADEIRO

Fonte: Elaborada pelo autor (2026).

4.6 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento compreende um conjunto de atividades estruturadas destinadas a converter dados brutos em um formato adequado para a aplicação de algoritmos de machine learning (GÉRON, 2019). Essas atividades são cruciais, pois a maioria dos algoritmos de aprendizado de máquina não opera eficientemente com características faltantes, ruídos ou atributos não escalonados (GÉRON, 2019).

De acordo com VULPEN (2019), a fase de pré-processamento é amplamente reconhecida pelo axioma "entra lixo, sai lixo" (*garbage in, garbage out*), apoiando o posicionamento de Bruce e Bruce (2019) que afirmam que a acurácia dos resultados é diretamente proporcional à qualidade dos dados de entrada.

Em face da natureza sensível e da heterogeneidade do conjunto de dados de *people analytics*, esta subseção detalha os procedimentos adotados, que incluem a anonimização de dados sensíveis, a transformação de atributos para escalas compatíveis, a gestão de dados categóricos e a implementação de técnicas de amostragem para reequilíbrio de classes.

4.6.1 Aspectos Éticos e Anonimização

Durante a extração de dados de colaboradores provenientes de distintos sistemas de gestão, é comum a captura de atributos sensíveis, tais como o cadastro de pessoas físicas (CPF) e o nome, conforme definido pelas disposições da lei geral de proteção de dados (LGPD).

A lei geral de proteção de dados Pessoais (LGPD), formalizada pela lei nº 13.709, de 14 de agosto de 2018, constitui o quadro normativo brasileiro que dispõe sobre o tratamento de dados pessoais, abrangendo tanto os meios digitais quanto os físicos, realizado por qualquer pessoa natural ou jurídica, pública ou privada (BRASIL, Lei nº 13.709, de 2018).

Em consonância com o disposto legal, informações como nome e cadastro de pessoas físicas (CPF) são definidas como dado pessoal, por se relacionarem a um indivíduo natural identificado ou identificável. Dada a natureza sensível dessas informações, o rigor da LGPD impõe que todas as operações de tratamento que englobam desde a coleta até a eliminação, sejam conduzidas sob os princípios de segurança e prevenção, exigindo a adoção de medidas técnicas e administrativas aptas a proteger os dados de acessos não autorizados e de situações ilícitas. Deste modo, o manuseio de atributos que permitem a identificação direta do titular demanda uma cautela incondicional e a implementação de processos de anonimização como prática fundamental para a mitigação de riscos.

Para ilustrar o processo de tratamento dessas informações neste estudo, a planilha inicial recebeu dados fictícios de matrícula, nome e CPF, sobre os quais se aplicou uma estratégia de anonimização, conforme visualizado na Figura 37:

Figura 37 - Amostra com dados fictícios de matrícula, nome e CPF

	A	B	C	D	E	F	G
1	Matricula	Nome	CPF	Idade	Genero	EstadoCivil	PortadorNecessidadesEspeciais
2	1	Vitoria Miranda Vieira	119.339.571-23	58	Feminino	Casado	Não
3	2	Davi Rocha Monteiro	109.411.188-00	58	Masculino	Solteiro	Não
4	3	Thiago Pereira Azevedo	501.211.355-19	55	Masculino	Solteiro	Não
5	4	Sergio Vieira Ramos	363.782.456-65	55	Masculino	Casado	Não
6	5	Vitor Cardoso Martins	115.612.858-70	52	Masculino	Casado	Não
7	6	Bernardo Costa Moraes	492.686.896-26	51	Masculino	Divorciado	Não
8	7	Carlos Souza Melo	952.566.382-54	52	Masculino	Solteiro	Não
9	8	Daniel Peixoto Gomes	368.130.431-96	52	Masculino	Casado	Sim
10	9	Miguel Peixoto Miranda	213.387.950-15	55	Masculino	Divorciado	Não

Fonte: Elaborada pelo autor (2026).

No presente estudo, procedeu-se à substituição das colunas de matrícula, nome e CPF por identificadores globais únicos ou *globally unique identifiers* (GUIDs) gerados aleatoriamente, os quais atuaram como chaves primárias de cada registro ao longo de toda a análise. Paralelamente, elaborou-se uma planilha auxiliar de correlação, relacionando cada GUID aos respectivos dados originais de matrícula, nome e CPF.

Como resultado, a base de dados disponibilizada para as etapas subsequentes passou a conter apenas o identificador anonimizado, assegurando a completa dissociação dos elementos sensíveis, conforme visualizado na Figura 38:

Figura 38 - Amostra com identificador anonimizado

	A	B	C	D	E
1	id	idade	genero	estado_civil	portador_necessidades_especiais
2	d875de1f-9819-4a3b-a836-2577ce769d69	58	Feminino	Casado	Nao
3	13f5281e-38b8-42a7-9b50-157f32698125	58	Masculino	Solteiro	Nao
4	d4003cc0-22b6-4534-8eff-f62411ff1c9b	55	Masculino	Solteiro	Nao
5	d95adbe6-6f00-433c-a91d-5863223be1bc	55	Masculino	Casado	Nao
6	2fd6050d-4b4b-4b0d-afe6-6e6413b8215d	52	Masculino	Casado	Nao
7	b0ac0a21-dc89-4aeb-b21e-454b3047efe7	51	Masculino	Divorciado	Nao
8	d4995dfd-ec54-4291-ab46-4601e6fa27cf	52	Masculino	Solteiro	Nao
9	9321206c-c388-4740-bb9c-6fc90c765fd9	52	Masculino	Casado	Sim
10	27e44a1e-ac91-440c-995a-d9a685313ebf	55	Masculino	Divorciado	Nao

Fonte: Elaborada pelo autor (2026).

Por fim, enfatiza-se que o processo de anonimização deve ser restrito ao número mínimo de agentes, preferencialmente um único profissional ou, idealmente, automatizado já na extração, de modo a impedir o contato de especialistas técnicos com dados sensíveis e, assim, minimizar os riscos de exposição.

4.6.2 Transformação e Normalização de Atributos

No presente estudo, aplicou-se técnicas de pré-processamento de dados no *dataset*, cujo objetivo central é a conversão dos dados brutos em uma matriz de características otimizada para a modelagem. Iniciou-se com a estratificação das variáveis em subconjuntos numéricos, binários e categóricos, o que permitiu a aplicação de técnicas de transformação customizadas.

O pipeline de transformação incluiu a normalização dos atributos numéricos por meio do *MinMaxScaler*, ajustando-os para que se encaixem em um intervalo específico entre 0 e 1, para mitigar vieses de escala em algoritmos sensíveis a magnitude. Para as variáveis categóricas, foi empregada uma estratégia adaptada ao contexto, sendo os atributos de baixa cardinalidade submetidos a uma codificação binária, ao passo que, para os de alta cardinalidade, optou-se pela codificação por frequência (*frequency encoding*). Esta técnica foi selecionada para evitar a "maldição da dimensionalidade", um problema comum em métodos como *one-hot encoding*.

4.6.3 Balanceamento de Classes (*SMOTETomek*)

Para mitigar o desbalanceamento inerente ao conjunto de treinamento, foi aplicada uma técnica híbrida de reamostragem, o *synthetic minority over-sampling technique tomek links* (SMOTETomek). Este método combina a sobreamostragem da classe minoritária através da criação de exemplos sintéticos SMOTE com a subamostragem por meio da remoção de *tomek links* (pares de observações de classes opostas que são vizinhos mais próximos). O objetivo foi criar um conjunto de dados de treinamento balanceado, permitindo que os algoritmos de aprendizado não desenvolvessem um viés preditivo em favor da classe majoritária.

4.7 TREINAMENTO E AVALIAÇÃO DOS MODELOS

Na etapa de modelagem, foi realizado a validação de modelos preditivos, inseridos em um paradigma de classificação binária. O objetivo primário consistiu na predição da variável dependente demissao_antes_atingir_break_even_acumulado, que operacionaliza o risco de rescisão voluntária de colaboradores antes que estes alcancem seu ponto de equilíbrio financeiro para a organização.

4.7.1 Particionamento dos Dados

O conjunto de dados pré-processado foi segmentado em conjuntos de treinamento (80%) e teste (20%). A técnica de divisão estratificada foi fundamental, assegurando que a proporção original de classes da variável-alvo, notadamente desbalanceada, fosse preservada em ambas as amostras. Esta abordagem é crucial para evitar viés na avaliação do modelo em dados não vistos.

4.7.2 Seleção dos Algoritmos de Classificação

Na etapa de modelagem preditiva, a seleção do conjunto de algoritmos para a tarefa de classificação binária exige uma abordagem criteriosa que contemple a diversidade metodológica e a eficácia na generalização. O objetivo primordial é a identificação da arquitetura que otimiza a predição da variável dependente, mitigando os riscos de sobreajuste e subajuste inerentes à complexidade dos dados. Para o presente estudo, foi conduzida uma análise comparativa de desempenho entre seis algoritmos de classificação distintos, cada qual representando paradigmas analíticos relevantes no campo da ciência de dados:

a) Regressão logística (*logistic regression*): Este modelo, fundamentalmente linear, é amplamente empregado em tarefas de classificação binária, estimando a probabilidade de uma instância pertencer a uma classe específica (GÉRON, 2019). Opera mediante a aplicação de uma função logística (sigmoide) para mapear a combinação linear das características de entrada a uma probabilidade no intervalo, sendo particularmente valorizado por sua interpretabilidade (GÉRON, 2019).
b) Floresta aleatória (*random forest*): É um paradigma preditivo classificado como um método de ensemble learning (aprendizado por agrupamento), reconhecido por sua eficácia e popularidade em tarefas de classificação e regressão (BRUCE; BRUCE, 2019). Uma extensão crucial deste método reside na amostragem aleatória de variáveis preditoras em cada ponto de divisão das árvores, além da amostragem dos registros, o que confere robustez ao modelo e mitiga o risco de sobreajuste (BRUCE; BRUCE, 2019).

c) *Light gradient-boosting machine* (LGBM): Este algoritmo se insere no contexto mais amplo do *boosting*, uma técnica geral de ensemble que visa aprimorar a precisão do modelo por meio do ajuste sequencial de preditores (BRUCE; BRUCE, 2019). Ao contrário do *bagging*, o *boosting* ajusta uma série de modelos sucessivos, atribuindo um peso maior aos registros que apresentaram erros substanciais nas rodadas anteriores, com o objetivo de minimizar o erro residual (BRUCE; BRUCE, 2019). O *boosting* é frequentemente modelado como uma otimização de uma função de custo, sendo o *boosting* gradiente estocástico a forma mais geral e amplamente utilizada (BRUCE; BRUCE, 2019).

d) *K-neighbors* (KNN): Trata-se de uma técnica de classificação e previsão de natureza intuitiva e simples, que se diferencia dos métodos estatísticos clássicos por ser orientada por dados e não impor uma estrutura linear predefinida (BRUCE; BRUCE, 2019). No contexto da classificação, a metodologia envolve identificar os K registros mais similares (vizinhos) e atribuir ao novo registro a classe majoritária entre esses vizinhos (BRUCE; BRUCE, 2019). A similaridade é determinada por uma métrica de distância, sendo a Distância Euclidiana a escolha mais comum (BRUCE; BRUCE, 2019).

e) *Support vector classification* (SVC): Modelo versátil que visa encontrar o hiperplano ótimo de separação com a maior margem possível entre as classes, sendo essa margem definida pelos vetores de suporte (GÉRON, 2019). Em domínios de dados não lineares, a SVC utiliza o truque do kernel para mapear as instâncias a um espaço de dimensão superior, onde a separação linear se torna factível (GÉRON, 2019).

f) *Multi-layer perceptron* (MLP): Consiste em uma arquitetura de rede neural artificial (*feedforward*) composta por múltiplas camadas de neurônios, incluindo camadas ocultas (GÉRON, 2019). O treinamento do MLP é realizado por meio do algoritmo de retropropagação, que ajusta iterativamente os pesos de conexão utilizando o Gradiente Descendente (GÉRON, 2019), capacitando-o a modelar relações não lineares de alta complexidade (GÉRON, 2019).

4.7.3 Ajuste de Hiperparâmetros (Hyperparameter Tuning)

Nesta fase do estudo, foi aplicado otimização e avaliação dos modelos de aprendizado de máquina, utilizando os seis algoritmos de classificação: *logistic regression*; *random forest*; *light gradient-boosting machine*; *K neighbors*; *support vector classification*; e *multi-layer perceptron*.

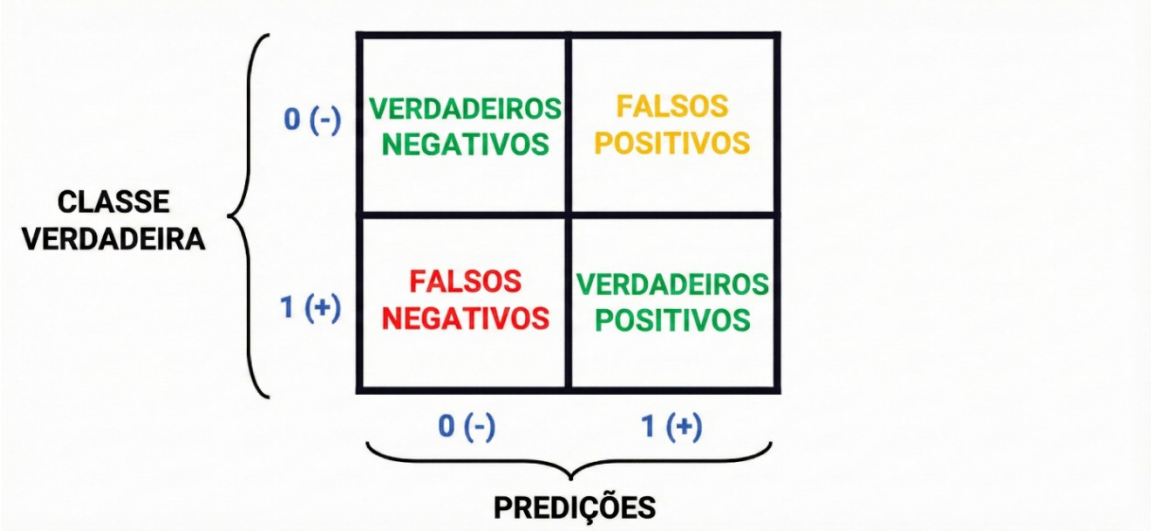
A primeira etapa consistiu na sintonia fina dos hiperparâmetros do classificador, onde foi optado pela técnica *RandomizedSearchCV*. Este método realiza uma busca estocástica em um espaço de hiperparâmetros predefinido, amostrando um número fixo de combinações (foram utilizadas 50 interações no presente estudo). A otimização foi orientada pela maximização da métrica de *recall* para a classe positiva (funcionários que saem), justificada pelo alto custo para o negócio de um falso negativo (não prever a saída de um funcionário em risco).

4.8 MÉTRICAS DE DESEMPENHO

No desenvolvimento do modelo de aprendizado de máquina do presente estudo, houve um enfoque voltado para a otimização de suas métricas de desempenho, objetivando a minimização dos falsos negativos e a maximização dos verdadeiros positivos.

A Figura 39 ilustra uma matriz de confusão, instrumento crucial na avaliação do desempenho de modelos classificatórios em algoritmos de aprendizado de máquina:

Figura 39 - Matriz de confusão



Fonte: Elaborada pelo autor (2026).

No contexto de turnover de funcionários, as classes possuem definições específicas. A **classe verdadeira** (*true class*) refere-se ao que realmente aconteceu, ou seja, se o funcionário de fato saiu ou ficou. Já as **predições** (*predictions*) representam o que o modelo de Machine Learning previu.

Em relação à classificação binária, considera-se **1 (+)** como a classe positiva, indicando que "o funcionário vai sair" (*turnover*), enquanto o **0 (-)** representa a classe negativa, significando que "o funcionário vai ficar".

Analisando cada um dos quadrantes, considerando o contexto de turnover de funcionários:

a) **Verdadeiros positivos** (*true positives* (TP)):
   - O que aconteceu: O funcionário realmente saiu da empresa (classe verdadeira = 1);
   - O que o modelo previu: O modelo previu corretamente que o funcionário iria sair (predição = 1);

b) **Verdadeiros negativos** (*true negatives* (TN)):
   - O que aconteceu: O funcionário realmente ficou na empresa (classe verdadeira = 0);
   - O que o modelo previu: O modelo previu corretamente que o funcionário iria ficar (predição = 0);

c) **Falsos positivos** (*false positives* (FP)):
   - O que aconteceu: O funcionário na verdade ficou na empresa (classe verdadeira = 0);
   - O que o modelo previu: O modelo previu erradamente que o funcionário iria sair (predição = 1);

d) **Falsos negativos** (*false negatives* (FN)):
   - O que aconteceu: O funcionário realmente saiu da empresa (classe verdadeira = 1);
   - O que o modelo previu: O modelo previu erradamente que o funcionário iria ficar (predição = 0).

Ao invés de focar na acurácia geral, o presente estudou focou no *recall* como métrica. O *recall*, também referida como sensibilidade ou taxa de verdadeiros positivos, quantifica a proporção de instâncias que são corretamente identificadas

como pertencentes à classe positiva pelo classificador, em relação a todas as instâncias que são verdadeiramente positivas (BRUCE; BRUCE, 2019).

Matematicamente, o *recall* é representado pela fórmula (TP / (TP + FN)), no contexto específico da análise, traduz-se como a proporção de (funcionários que saíram e o modelo previu que sairiam) / (total de funcionários que realmente saíram).

Embora possa induzir a alguns alarmes falsos (falsos positivos) e, consequentemente, a iniciativas de retenção direcionadas a colaboradores que, a priori, não apresentariam intenção de desligamento, a prioridade estratégica foi evitar a ocorrência de falsos negativos, ou seja, impedir que um colaborador valioso efetue a saída de forma imprevista.

5 RESULTADO E AVALIAÇÕES

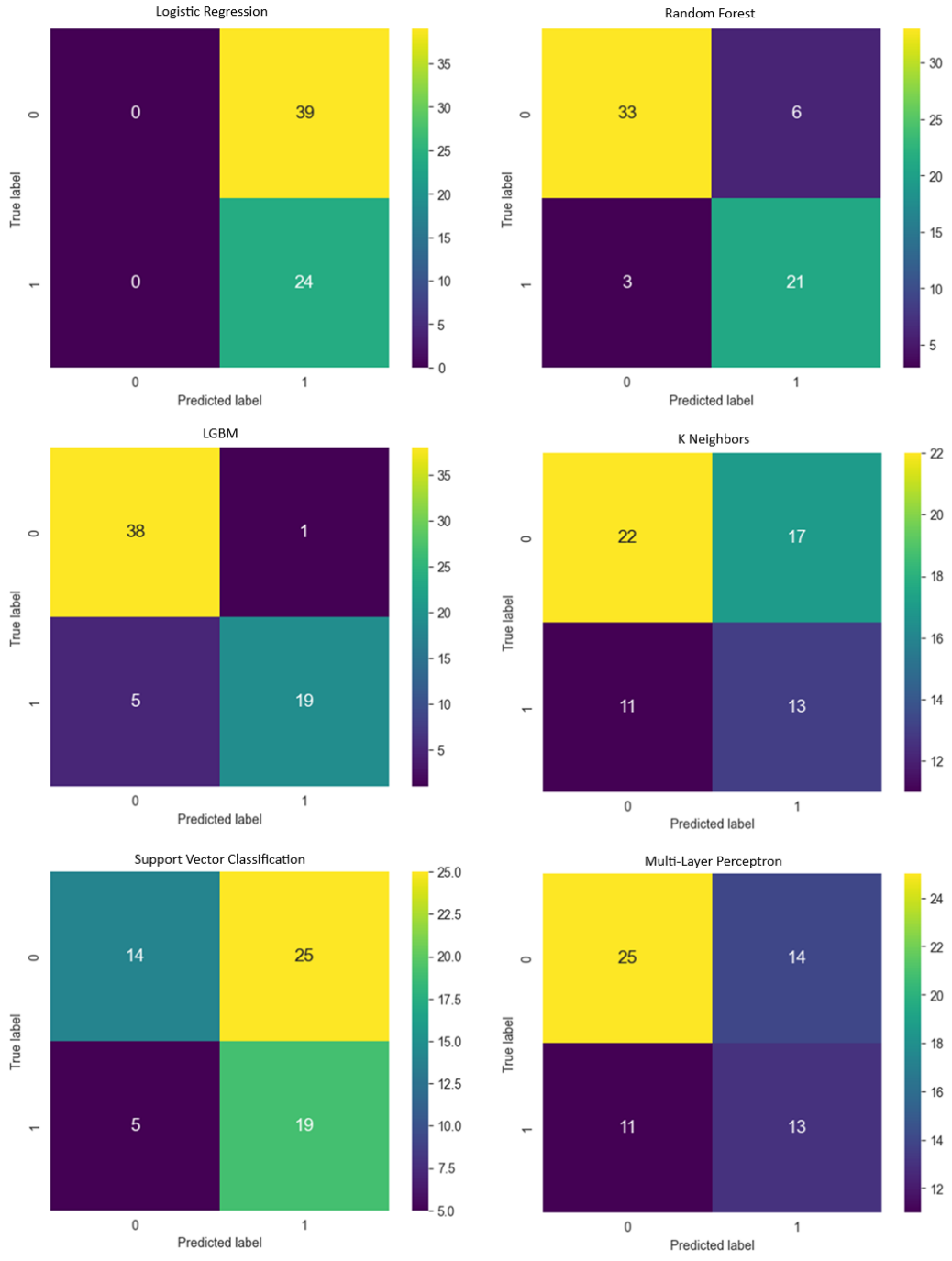
Esta seção dedica-se à análise e à interpretação dos resultados provenientes do processo de modelagem preditiva, em estrito alinhamento com a proposta metodológica delineada na seção 4. A avaliação registrada é fundamental para validar empiricamente a premissa central do estudo, que avalia a maior eficácia da predição de desligamentos quando a variável-alvo é ligada ao risco de rescisão antes que o colaborador atinja o ponto de equilíbrio econômico (*breakeven*) do cargo (cenário A), em contraste com uma abordagem generalista do *turnover* (cenário B).

Para fundamentar essa validação, o desempenho dos algoritmos é avaliado sob múltiplas dimensões analíticas, seguindo uma lógica de aprofundamento progressivo. Inicialmente, apresenta-se a análise de erros via matrizes de confusão, fundamentais para sintetizar os *trade-offs* operacionais e a incidência de falsos negativos. Na sequência, as curvas *precision-recall* são empregadas para mensurar a sensibilidade dos modelos, priorizando a mitigação de erros críticos em um contexto de desbalanceamento de classes. Subsequentemente, as curvas *receiver operating characteristic* (ROC) são analisadas para aferir a capacidade discriminatória global das arquiteturas testadas. A robustez e a capacidade de generalização dos resultados são asseguradas por meio de técnicas de validação cruzada; por fim, examina-se a *feature importance* do modelo de melhor desempenho, visando elucidar a lógica interna do algoritmo e os preditores determinantes para o fenômeno do desligamento precoce.

5.1 ANÁLISE DE ERROS: MATRIZES DE CONFUSÃO

A Figura 40 ilustra as matrizes de confusão obtidas para o cenário A, permitindo a compreensão dos resultados preditivos:

Figura 40 - Matriz de confusão do cenário A



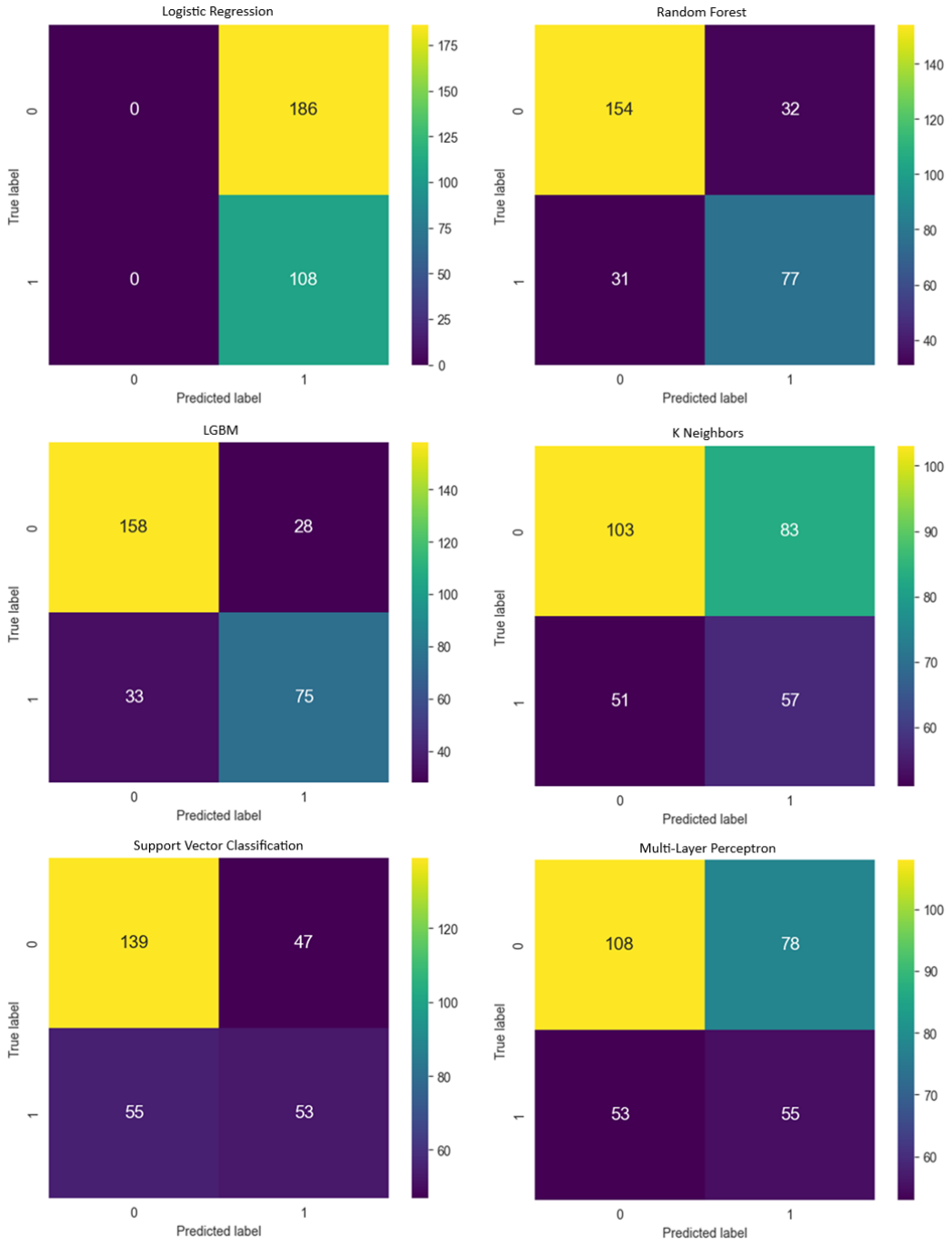
Fonte: Elaborada pelo autor (2026).

A análise quantitativa das matrizes de confusão para o cenário A, operando em um limiar de decisão otimizado, sintetiza os *trade-offs* operacionais e expõe a inadequação de modelos como regressão logística, KNN e MLP. O melhor desempenho está entre os ensembles de árvores, onde o LightGBM adotou um perfil de alta precisão (P=0.95), minimizando falsos positivos (FP=1) ao custo de um bom *recall* (R=0.79, FN=5). Em contrapartida, o *random forest* demonstrou sua

superioridade estratégica ao alinhar-se com o objetivo de maximizar o *recall*, alcançando o menor número de falsos negativos (FN=3), o erro de maior custo para o negócio, e consequentemente, o maior *recall* (R=0.875) entre os modelos viáveis.

A Figura 41 exibe as matrizes de confusão obtidas para o cenário B, permitindo a interpretação dos resultados preditivos:

Figura 41 - Matriz de confusão do cenário B



Fonte: Elaborada pelo autor (2026).

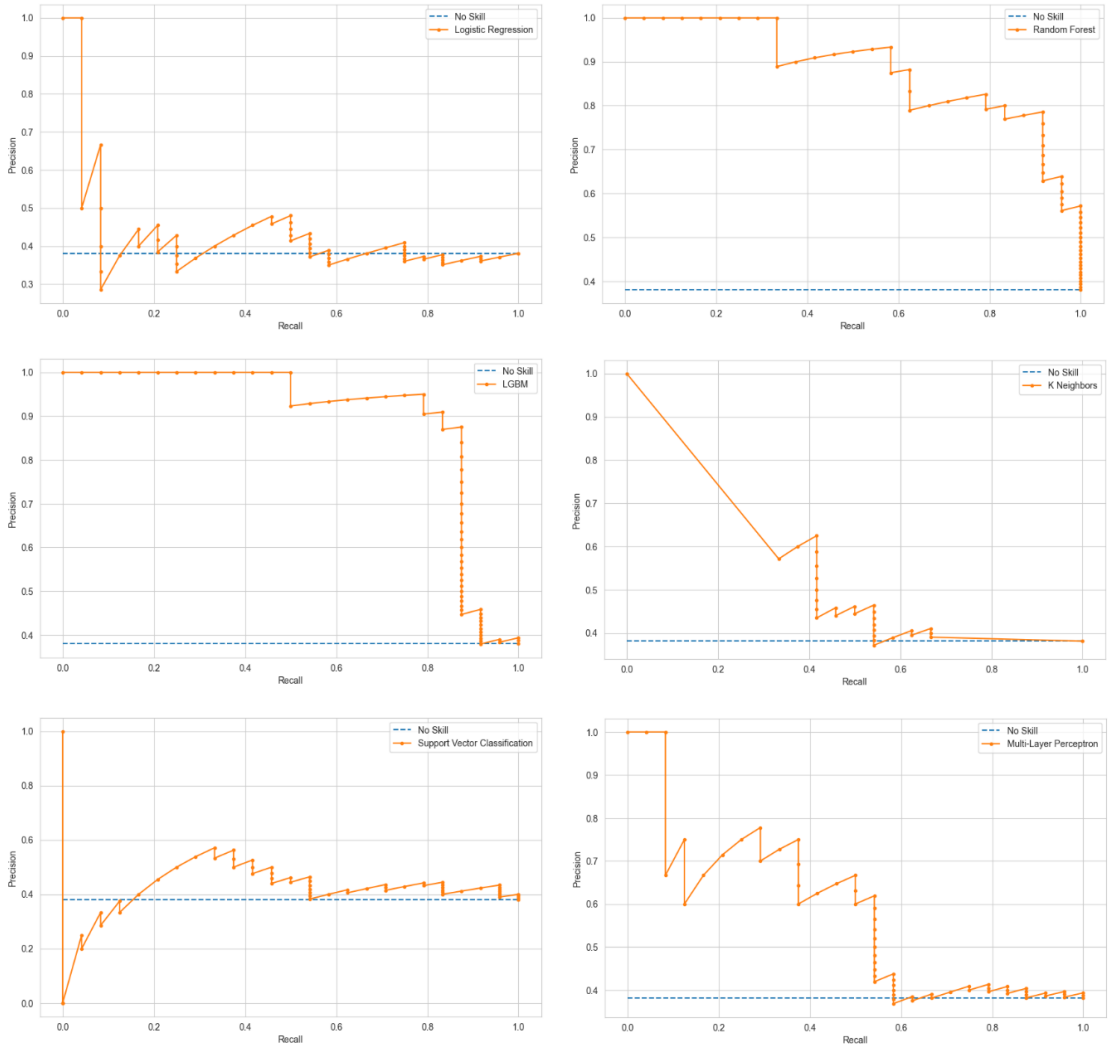
A análise quantitativa das matrizes de confusão para o cenário B, expõe a falha generalizada dos modelos não-*ensemble*: a regressão logística se mostrou um classificador com desempenho pífio em precisão (P=0.367), e os modelos KNN, SVC e MLP se mostraram inadequados, falhando em identificar aproximadamente metade da classe positiva (FN = 51/55). O desempenho viável se restringiu aos *ensembles* de árvores, onde o *random forest* (FN=31, R=0.713) demonstrou uma ligeira superioridade na maximização do recall sobre o LightGBM (FN=33, R=0.694). No entanto, o maior *recall* alcançado neste cenário é claramente inferior ao obtido no cenário A.

5.2 AVALIAÇÃO VIA CURVAS PRECISION-RECALL

A Figura 42 delineia as curvas *precision-recall* referentes ao cenário A, permitindo a análise comparativa do desempenho preditivo das arquiteturas algorítmicas avaliadas.

A avaliação dos seis algoritmos no cenário A, revelou que os métodos de ensemble baseados em árvores são categoricamente superiores para a tarefa de maximizar o *recall*. Enquanto o modelo *logistic regression*, *k neighbors* e *support vector classification* falharam em discriminar além da linha de base (P = 0.38), e o MLP exibiu um desempenho instável, tanto o LGBM quanto o *random forest* (RF) demonstraram capacidade preditiva substancial. Contudo, o LGBM, apesar da alta precisão em *recall* moderada (P = 0.6), sofreu um colapso de performance em P > 0.8. O *random forest*, em contrapartida, manteve-se com ótima performance, apresentando uma curva de degradação suave e mantendo alta precisão (P > 0.8) mesmo em regimes de altíssimo *recall* (R > 0.9).

Figura 42 - Curva precision-recall do cenário A

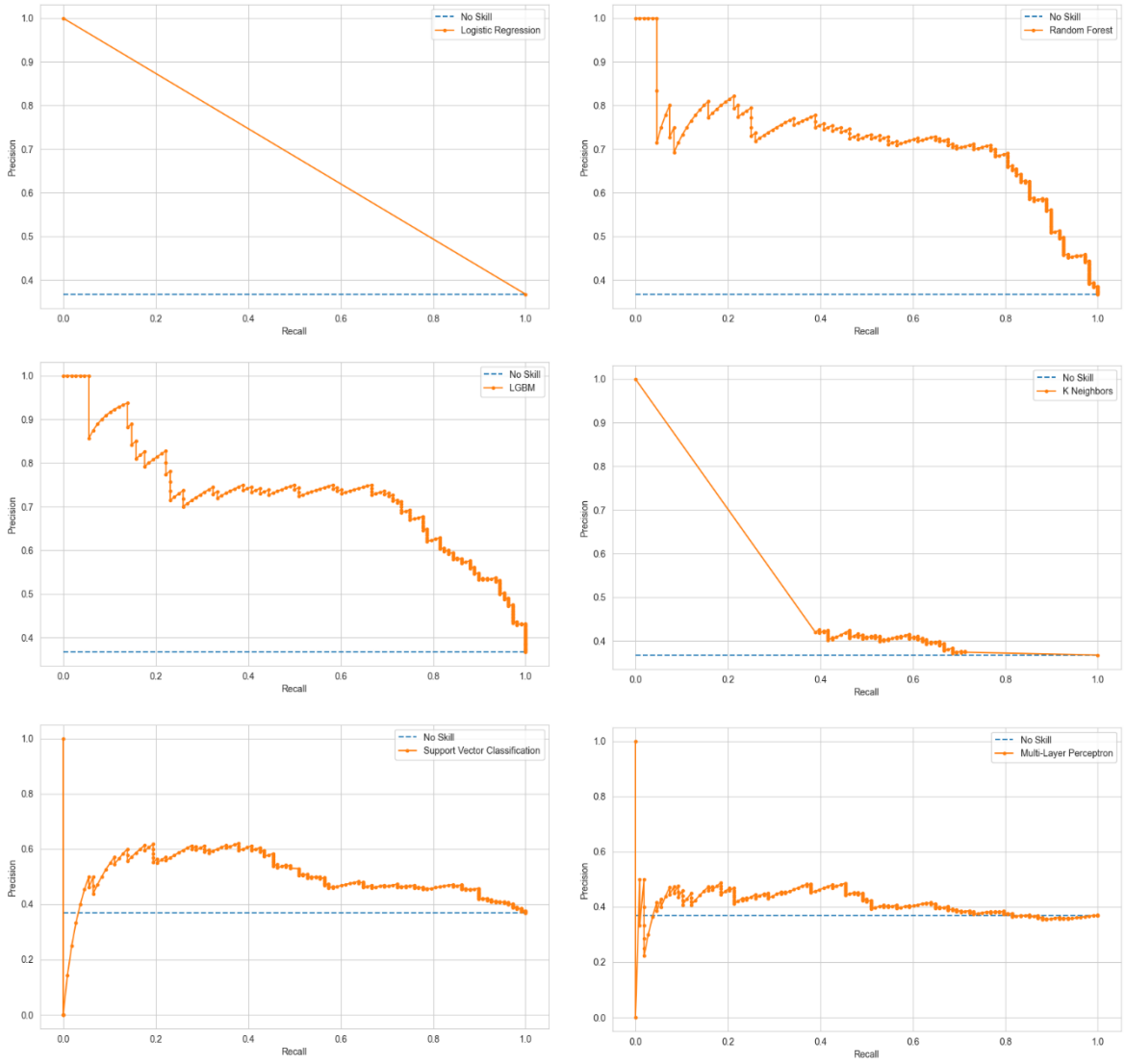


Fonte: Elaborada pelo autor (2026).

A Figura 43 exibe as curvas *precision-recall* referentes ao cenário B, possibilitando a comparação do desempenho preditivo dos algoritmos avaliados.

Em análise ao cenário B, os resultados demonstram a inadequação de arquiteturas como KNN, SVC e MLP, cujas performances foram instáveis e majoritariamente inferiores à linha de base (P = 0.38). Embora a regressão logística tenha fornecido um desempenho linear estável, os ensembles baseados em árvores (*random forest* e LightGBM) confirmaram sua dominância. Contudo, uma análise comparativa revela que o desempenho obtido neste cenário generalista foi inferior ao alcançado pelo modelo focado em determinar colaboradores demitidos antes do *breakeven* do cargo do cenário A.

Figura 43 - Curva precision-recall do cenário B

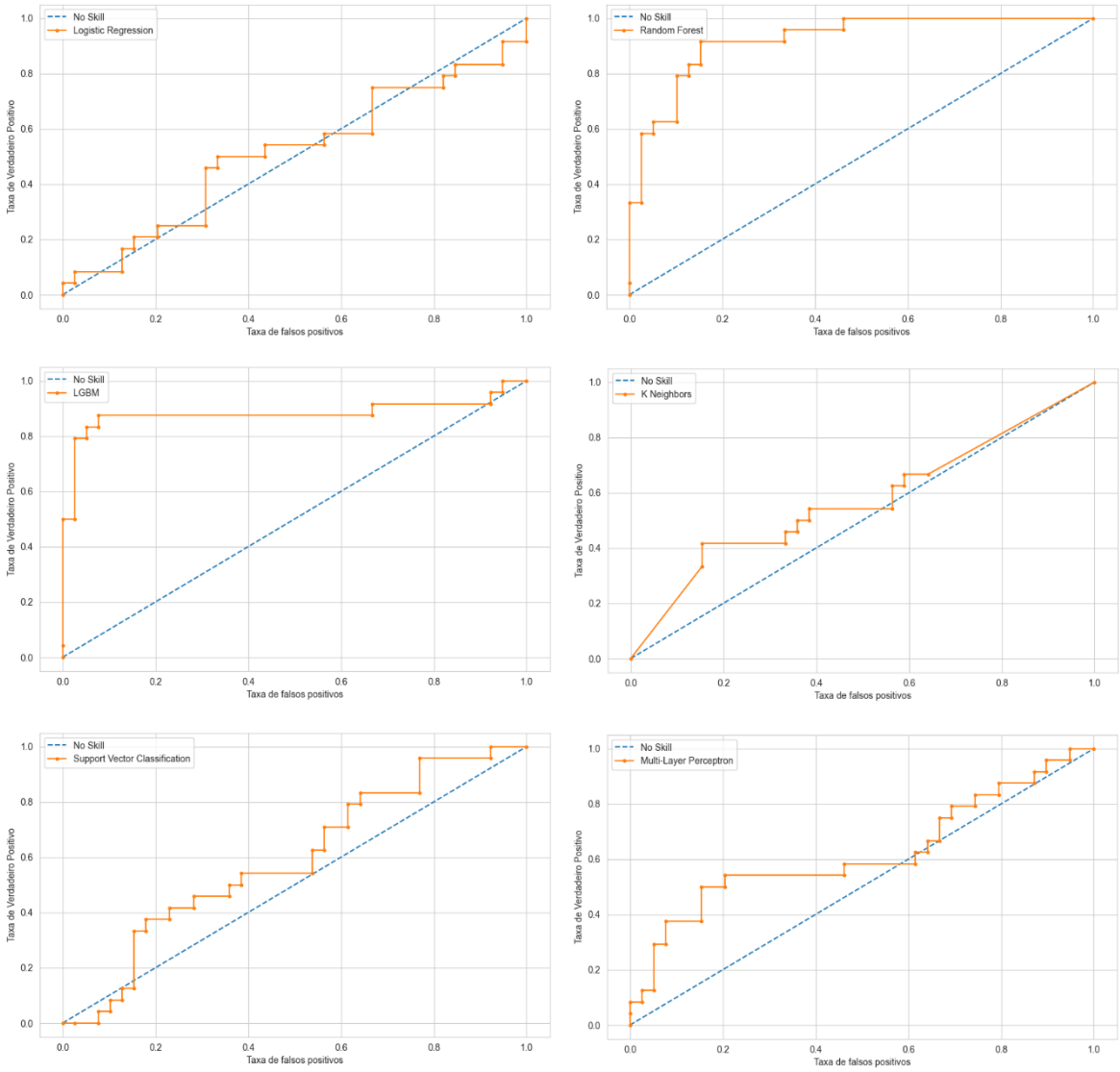


Fonte: Elaborada pelo autor (2026).

5.3 CAPACIDADE DISCRIMINATÓRIA (CURVAS ROC)

A Figura 44 apresenta as curvas *receiver operating characteristic* (ROC) referentes ao cenário A, exibindo o desempenho dos modelos avaliados:

Figura 44 - Curva receiver operating characteristic do cenário A

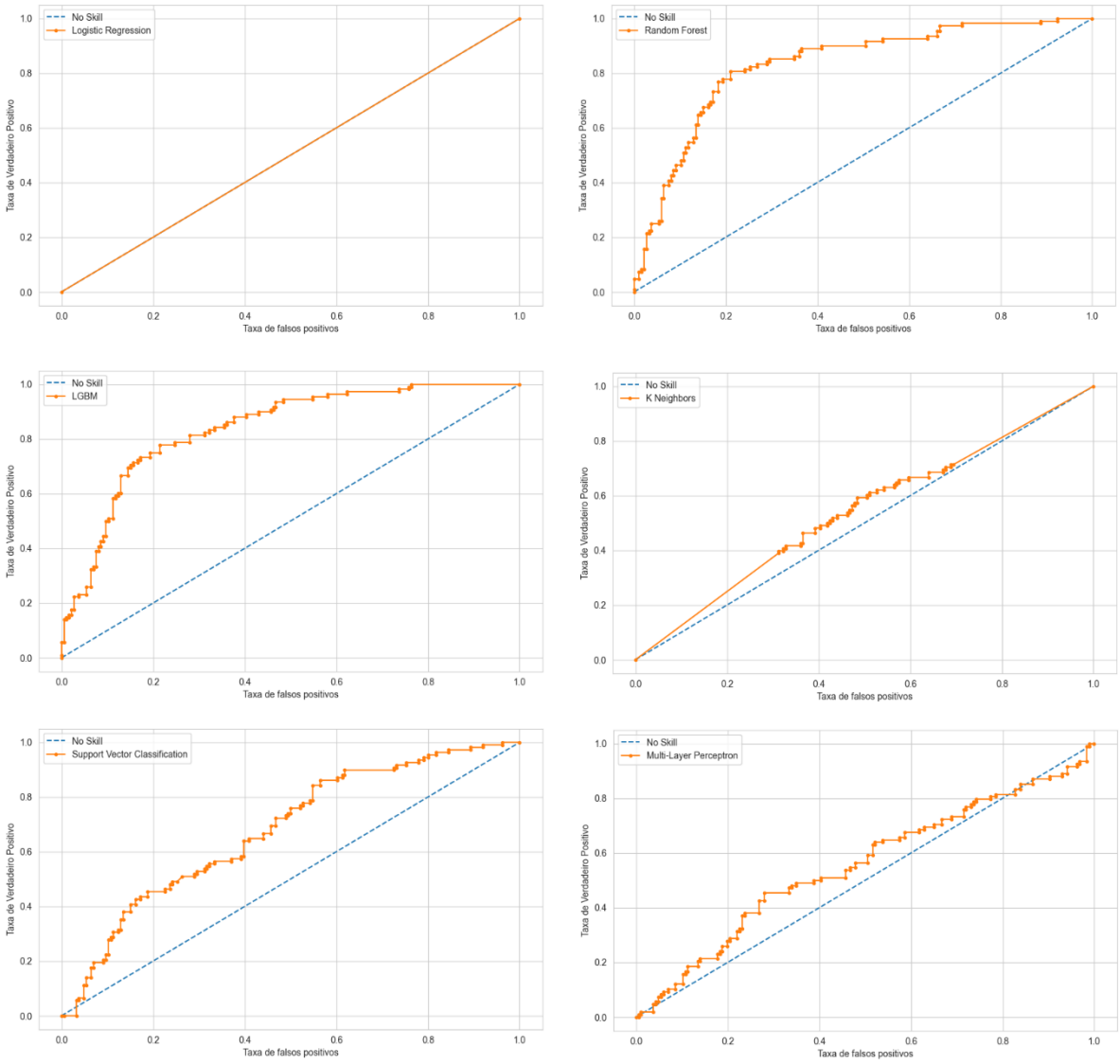


Fonte: Elaborada pelo autor (2026).

A análise da curva *receiver operating characteristic* (ROC) para o cenário A, corrobora a hierarquia de desempenho dos modelos em termos de capacidade discriminatória. Os algoritmos *logistic regression*, *k-neighbors*, SVC e MLP exibem curvas ROC sobrepostas à linha de base, confirmando sua incapacidade de separar as classes positiva e negativa. Em contraste, os *ensembles* baseados em árvores demonstram poder preditivo, com o *random forest* emergindo como o modelo superior, apresentando uma AUC-ROC visivelmente maior, reforçando sua seleção como modelo ótimo já estabelecida pela análise da curva PR.

A Figura 45 demonstra as curvas *receiver operating characteristic* (ROC) referentes ao cenário B, apresentando o desempenho dos modelos analisados:

Figura 45 - Curva receiver operating characteristic do cenário B



Fonte: Elaborada pelo autor (2026).

A análise da curva ROC para o cenário B, revelou uma falha categórica da maioria dos modelos testados. A regressão logística, com uma AUC-ROC contornando a libra base, demonstra uma total ausência de poder discriminatório, provando que os determinantes do *turnover* geral não são linearmente separáveis com este modelo. Similarmente, os modelos KNN, SVC e MLP falham em superar significativamente o desempenho aleatório. Em acentuado contraste, os *ensembles* baseados em árvores (*random forest* e LGBM) exibem curvas AUC-ROCs elevadas, estabelecendo-se como os únicos modelos estatisticamente viáveis para esta tarefa generalista.

5.4 GENERALIZAÇÃO DO MODELO (VALIDAÇÃO CRUZADA)

Para obter uma estimativa fidedigna e imparcial da capacidade de generalização do modelo no mundo real, foi implementada um método de validação final. Construiu-se um pipeline computacional (ImbPipeline) que integra duas operações sequenciais:

a) Reamostragem com SMOTETomek: Técnica híbrida que primeiro aplica o *synthetic minority over-sampling technique* (SMOTE) para gerar instâncias sintéticas da classe minoritária e, em seguida, utiliza o *tomek links* para remover exemplos sobrepostos entre as classes, refinando o espaço de decisão.
b) Classificador: O modelo com os hiperparâmetros já otimizados.

Este pipeline foi então avaliado por meio de uma validação cruzada estratificada de 10 folds (StratifiedKFold).

O processo de validação cruzada foi aplicado aos cenários A e B, contemplando os seis algoritmos em análise. Os resultados decorrentes desta execução encontram-se detalhados nas Tabelas 1 e 2:

Tabela 1 - Validação cruzada do cenário A

Algoritmo	**Precisão Média (DP)**	**Recall Médio (DP)**	**ROC AUC Médio (DP)**
Logistic Regression	0.2013 (+/- 0.0333)	1.0000 (+/- 0.0000)	0.9214 (+/- 0.0325)
Random Forest	0.6418 (+/- 0.1023)	0.8353 (+/- 0.2210)	0.9776 (+/- 0.0200)
LGBM	0.7598 (+/- 0.2202)	0.7526 (+/- 0.2660)	0.9778 (+/- 0.0195)
K Neighbors	0.2100 (+/- 0.0854)	0.4538 (+/- 0.1760)	0.7126 (+/- 0.0972)
SVC	0.1479 (+/- 0.0201)	1.0000 (+/- 0.0000)	0.9033 (+/- 0.0657)
Multi-Layer Perceptron	0.4782 (+/- 0.2608)	0.5385 (+/- 0.2571)	0.9211 (+/- 0.0551)

Fonte: Elaborada pelo autor (2026).

Conforme exposto pelo Tabela 1, os algoritmos *logistic regression* e o SVC, apesar de um *recall* médio perfeito (1.0), se demonstraram como classificadores irrelevantes, como evidenciado por suas precisões médias baixíssimas (0.2013 e 0.1479, respectivamente), enquanto KNN e MLP exibiram performance abaixo do aceitável. A escolha viável se restringiu aos *ensembles*, onde o LightGBM, embora apresentando médias balanceadas (R=0.7526), demonstrou uma instabilidade de performance proibitiva (DP de recall +/- 0.2660). O *random forest*, em contrapartida,

consolidou-se como um ótimo modelo, alcançando o *recall* médio mais elevado (0.8353) entre os classificadores.

Tabela 2 - Validação cruzada do cenário B

Algoritmo	Precisão Média (DP)	Recall Médio (DP)	ROC AUC Médio (DP)
Logistic Regression	0.2571 (+/- 0.3367)	0.7000 (+/- 0.9165)	0.5000 (+/- 0.0000)
Random Forest	0.7312 (+/- 0.1364)	0.6907 (+/- 0.1386)	0.8378 (+/- 0.1014)
LGBM	0.7250 (+/- 0.1566)	0.6500 (+/- 0.1211)	0.8360 (+/- 0.0988)
K Neighbors	0.3924 (+/- 0.0866)	0.4778 (+/- 0.1404)	0.5401 (+/- 0.1008)
SVC	0.5375 (+/- 0.1754)	0.5278 (+/- 0.2129)	0.6743 (+/- 0.1368)
Multi-Layer Perceptron	0.3726 (+/- 0.0518)	0.5870 (+/- 0.3838)	0.5368 (+/- 0.0619)

Fonte: Elaborada pelo autor (2026).

De acordo com a Tabela 2, o algoritmo *logistic regression*, com uma ROC AUC média de 0.5000 (+/- 0.0000), provou estatisticamente a natureza não-linear do fenômeno, enquanto KNN e MLP falharam em apresentar capacidade discriminatória (AUC +/- 0.54). Apenas os ensembles de árvores se mostraram viáveis, com o *random forest* emergindo como o modelo superior, alcançando o *recall* médio mais alto (0.6907) e a maior AUC-ROC média (0.8378). Contudo, o melhor desempenho para o cenário B é claramente inferior ao alcançado pela mesma arquitetura no cenário A (0.8353).

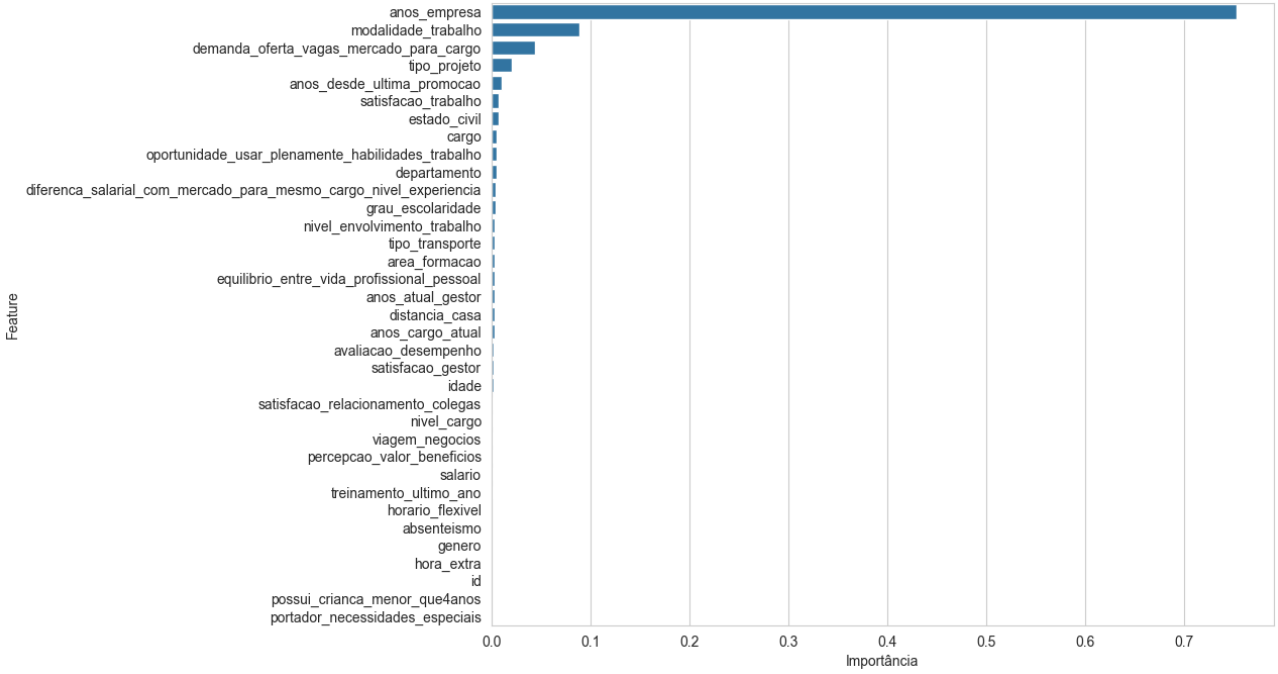
5.5 FEATURE IMPORTANCE DO MODELO COM MELHOR DESEMPENHO

Para além da mera aferição de desempenho (como *f1-score* ou *recall*), esta subseção investiga a interpretabilidade do modelo com melhor desempenho, buscando compreender a razão deste algoritmo obter sucesso. Conforme evidenciado nas seções 5.1 a 5.4, métodos de ensemble baseados em árvores são categoricamente superiores, com destaque para o modelo *random forest para o cenário A.*

Através da análise de *feature importance* (importância de variáveis), objetivou-se compreender a lógica interna que o classificador utilizando *random forest* desenvolveu durante o treinamento.

A Figura 46 exibe gráficos de barras horizontais ilustrando a contribuição relativa de cada variável preditora para o modelo:

Figura 46 - Feature importance do modelo random forest para o cenário A



Fonte: Elaborada pelo autor (2026).

A análise sobre a importância das features indicou que o modelo evidencia um ponto: os fatores de tempo de serviço, como anos_empresa, eram os mais relevantes. No entanto, uma análise mais a fundo sugere que essa forte influência do tempo de serviço é um viés metodológico, vindo da própria forma de treinamento do cenário A. Ao aprender com todos os dados, os modelos pegaram uma "regra simples": quem tem pouco tempo de empresa, sai. Porém, essa regra não é válida para o grupo-alvo (*pré-breakeven*), onde todos, por definição, têm pouco tempo de casa (variância nula). A conclusão é que os reais impulsionadores do *turnover* precoce residem nos fatores secundários, como diferenca_salarial e modalidade_trabalho, que foram ofuscados pelo fator de tempo de serviço.

5.6 DISCUSSÃO DOS RESULTADOS

A análise dos resultados empíricos obtidos neste estudo corrobora a hipótese central da pesquisa, demonstrando que a modelagem preditiva orientada o desligamento *pré-breakeven* (cenário A) apresenta uma eficácia superior às abordagens generalistas de predição de *turnover* (cenário B).

A supremacia das arquiteturas de *ensemble learning*, notadamente o algoritmo *random forest*, sobre modelos lineares como a regressão logística, alinha-se aos

estudos de Avrahami et al. (2022) e Chung et al. (2023). Tais resultados evidenciam que o fenômeno da rotatividade voluntária no setor tecnológico é regido por interações não lineares e complexas entre as variáveis, as quais modelos estatísticos tradicionais falham em capturar, resultando em ausência de capacidade discriminatória (AUC próxima de 0.50 no cenário B).

Sob a ótica da métrica de *recall*, a priorização do cenário A mostrou-se estrategicamente acertada. Ao alcançar um *recall* médio de 0,8353 na validação cruzada, o modelo proposto mitiga eficazmente o risco dos falsos negativos (FN). No contexto de *people analytics*, conforme preconizado por Isson e Harriott (2016), o custo de perder um talento não identificado supera o custo operacional de intervir preventivamente em um falso positivo. Portanto, a capacidade do modelo em identificar a vasta maioria dos colaboradores em risco de saída precoce valida sua aplicabilidade como ferramenta de suporte à decisão.

Adicionalmente, a análise de *feature importance* revelou uma dinâmica interessante: embora houvesse um viés metodológico esperado em relação às variáveis temporais (devido à natureza da variável-alvo *pré-breakeven*), o modelo atribuiu relevância secundária significativa a fatores como diferenca_salarial e modalidade_trabalho, confirmando hipóteses levantadas no *dataset* com dados sintéticos.

Em suma, os resultados transcendem a mera validação algorítmica. Eles atestam que a integração de dados financeiros (conceito de *breakeven*) à modelagem de dados de recursos humanos refina a capacidade preditiva, permitindo que a organização atue cirurgicamente na retenção de colaboradores que ainda representam um passivo de investimento, otimizando assim a sustentabilidade econômica da gestão de talentos.

6 CONCLUSÃO

A presente dissertação dedicou-se à investigação e ao desenvolvimento de um modelo computacional preditivo voltado à mitigação do *turnover* voluntário em organizações de tecnologia, introduzindo uma abordagem inovadora centrada no conceito de ponto de equilíbrio econômico (*breakeven*) do colaborador. A pesquisa foi motivada pela necessidade crítica de alinhar as práticas de *people analytics* a questões de sustentabilidade financeira organizacional, transcendendo as métricas tradicionais de rotatividade que tratam de maneira isolada perdas operacionais e prejuízos de investimento.

No que tange ao objetivo geral, o estudo obteve êxito ao estruturar e validar um modelo de classificação binária capaz de inferir, com elevada sensibilidade, a probabilidade de desligamento de colaboradores antes do retorno sobre o investimento inicial de contratação e treinamento. A estratégia metodológica de segregar a análise em dois cenários distintos, cenário A (focado no *breakeven*) e cenário B (generalista), provou-se determinante. Os resultados empíricos demonstraram que a delimitação do problema de negócio (cenário A) não apenas refinou a precisão dos algoritmos, mas também potencializou a métrica de *recall*, essencial para a identificação preventiva de talentos em risco.

Uma contribuição metodológica significativa deste trabalho reside na estruturação e disponibilização de um inventário organizado de amplos atributos de colaboradores, consolidado nos Quadros 4, 5 e 6. Esta compilação não apenas sintetiza variáveis validadas em ampla revisão bibliográfica, mas também introduz variáveis inéditas de autoria própria, desenhadas para capturar as nuances de empresas brasileiras de base tecnológica. Este artefato serve, portanto, como um material referencial robusto para outras organizações que almejem implementar iniciativas de *people analytics*, oferecendo um ponto de partida validado para a engenharia de recursos (*feature engineering*).

Do ponto de vista técnico, a pesquisa evidenciou a superioridade das arquiteturas de *ensemble learning* sobre modelos lineares clássicos na modelagem de fenômenos de recursos humanos. A performance robusta do algoritmo *random forest*, em detrimento da regressão logística, evidenciou a natureza não linear e complexa das variáveis que regem a retenção de talentos no setor tecnológico.

Adicionalmente, a análise de *feature importance* e a validação das hipóteses permitiram desmistificar a premissa de que a satisfação subjetiva é o único preditor de saída, revelando que fatores estruturais como a modalidade de trabalho e a competitividade salarial exercem influência preponderante na decisão de desligamento precoce, obviamente ponderando que essas conclusões foram obtidas a partir de um *dataset* com dados sintéticos e que cada organização deverá possuir seus próprios motivadores para desligamento precoce de colaboradores.

As implicações práticas deste estudo para a área de gestão de pessoas e tecnologia da informação são relevantes. O método proposto oferece um caminho para que organizações transitem de uma gestão reativa para uma abordagem preditiva e financeiramente orientada. Ao integrar dados financeiros (custo e retorno) à modelagem de comportamento humano, o estudo fornece um método para priorizar ações de retenção onde elas são economicamente mais críticas, otimizando a alocação de recursos corporativos.

Sob a perspectiva do impacto organizacional, a adoção do método proposto transcende a mera modernização tecnológica do departamento de Recursos Humanos, consolidando o *people analytics* como um mecanismo direto de proteção ao capital corporativo. Em um mercado de tecnologia marcado pela severa escassez de profissionais qualificados e por custos de substituição que podem atingir de 1,5 a 2,0 vezes o salário anual do indivíduo (ISSON; HARRIOTT, 2016), a predição assertiva de saídas precoces ataca frontalmente a vulnerabilidade financeira das empresas. O impacto materializa-se na capacidade da organização de abandonar políticas de retenção pulverizadas e de alto custo, passando a alocar orçamentos e esforços de engajamento de forma cirúrgica naqueles talentos que ainda configuram um passivo de investimento. Dessa forma, ao intervir antes que o custo irrecuperável do recrutamento e *onboarding* se converta em prejuízo líquido, a gestão de pessoas é elevada ao patamar de unidade de inteligência estratégica, atuando ativamente para a sustentação da eficiência econômica e da vantagem competitiva da corporação.

Entretanto, é imperativo reconhecer as limitações inerentes ao delineamento da pesquisa. A utilização de um conjunto de dados sintético, embora adaptado para refletir a realidade do mercado de tecnologia, impõe cautela na generalização direta dos resultados para ambientes de produção específicos sem prévia recalibragem.

Como recomendações para trabalhos futuros, sugere-se a aplicação do método proposto em dados reais de corporações de tecnologia, permitindo a validação externa do modelo e a análise de variáveis não estruturadas, como processamento de linguagem natural em entrevistas de desligamento e avaliações de clima.

Outra recomendação, seria a evolução deste modelo preditivo para um ecossistema de agentes de inteligência artificial (agentes de IA) autônomos. Sugere-se o desenvolvimento de agentes que não apenas predigam o risco de saída, mas que atuem proativamente na retenção de talentos, interagindo de forma personalizada com colaboradores e gestores para sugerir intervenções, planos de carreira ou ajustes de benefícios em tempo real. A integração desses agentes com *large language models* (LLMs) poderia inaugurar uma nova fronteira na gestão de talentos, onde a IA deixa de ser uma ferramenta de consulta para se tornar um parceiro ativo na estratégia organizacional.

REFERÊNCIAS

AVRAHAMI, Dan et al. A human resources analytics and machine-learning examination of turnover: implications for theory and practice. **International Journal of Manpower,** v. 43, n. 6, p. 1405-1424, 2022. ISSN 0143-7720. DOI: 10.1108/IJM-12-2020-0548. Disponível em: https://doi.org/10.1108/IJM-12-2020-0548. Acesso em: 22 out. 2024.

BISWAS, Ashish Kumar et al. An ensemble learning model for predicting the intention to quit among employees using classification algorithms. **Decision Analytics Journal,** v. 9, 2023. ISSN 2772-6622. DOI: 10.1016/j.dajour.2023.100335. Disponível em: https://doi.org/10.1016/j.dajour.2023.100335. Acesso em: 22 out. 2024.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Dispõe sobre a Proteção de Dados Pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Brasília, DF: Presidência da República. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 09 dez. 2025.

BRUCE, Peter; BRUCE, Andrew. **Estatística Prática para Cientistas de Dados**: 50 Conceitos Essenciais. Rio de Janeiro: Alta Books, 2019.

CANAIS, Maria do Rosário Ligeiro Pinto. **People analytics aplicado à retenção de talentos nas organizações**. 2016. 72f. Dissertação (Mestrado em Gestão da Informação) - Instituto Superior de Estatística e Gestão de Informação Universidade Nova de Lisboa. Lisboa, 2016. Disponível em: https://run.unl.pt/bitstream/10362/19251/1/TGI0059.pdf. Acesso em: 22 out. 2023.

CHIAVENATO, Idalberto. **Gestão de pessoas:** o novo papel dos Recursos Humanos nas organizações. 4. ed. Barueri: Manole, 2014.

CHORNOUS, Galyna O.; GURA, Viktoriya, L. Integration of Information Systems for Predictive Workforce Analytics: Models, Synergy, Security of Entrepreneurship. **European Journal of Sustainable Development,** v. 9, n. 1, p. 83-98, 2020. ISSN: 2239-5938. DOI: 10.14207/ejsd.2020.v9n1p83. Disponível em: https://doi.org/10.14207/ejsd.2020.v9n1p83. Acesso em: 22 out. 2024.

CHUNG, Doohee et al. Predictive model of employee attrition based on stacking ensemble learning. **Expert Systems with Applications,** v. 215, 2023. ISSN 0957-4174. DOI: 10.1016/j.eswa.2022.119364. Disponível em: https://doi.org/10.1016/j.eswa.2022.119364. Acesso em: 22 out. 2024.

DESSLER, G. **Administração de recursos humanos**. 2. ed. São Paulo: Pearson Prentice Hall, 2003.

EDWARDS, Martin; EDWARDS, Kirsten. **Predictive HR analytics**: mastering the HR metric. Londres: Kogan Page, 2016.

EHRENBERG, R. G.; SMITH, R. S. **Modern labor economics**: theory and public policy. 12. ed. New York: Routledge, 2012.

FERRAR, Jonathan; GREEN, David. **Excellence In People Analytics**: how to use workforce data to create business value. Londres: Kogan Page, 2021.
GÉRON, Aurélien. **Mãos A Obra:** aprendizado De máquina com Scikit-Learn, Keras & TensorFlow. 2ª Edição. Rio de Janeiro: Alta Books, 2019.

GOOGLE FOR STARTUPS. **Panorama de talentos em tecnologia**: a escassez de profissionais de tecnologia no Brasil e seu consequente impacto no ecossistema de startups. São Paulo: Google, 2022. Disponível em: https://services.google.com/fh/files/events/relatorio_gfs_panorama_de_talentos_em_tecnologia.pdf. Acesso em: 14 jan. 2026.

GRUS, Joel. **Data science do zero:** primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016.

GUENOLE, Nigel; FERRAR, Jonathan; FEINZIG, Sheri. **The power of people**: how successful organizations use workforce analytics to improve business performance. Nova Jersey: Pearson FT Press, 2017.

HARVARD BUSINESS REVIEW. **Gerenciando pessoas:** Os melhores artigos da Harvard Business Review sobre como liderar equipes. Rio de Janeiro: GMT Editores, 2018.

HEIDEMANN, Ansgar; HULTER, Svenja M.; TEKIELI, Michael. Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency. **The International Journal of Human Resource Management,** v. 35, n. 14, p. 2343-2366, 2024. ISSN 0958-5192. DOI: 10.1080/09585192.2024.2335515. Disponível em: https://doi.org/10.1080/09585192.2024.2335515. Acesso em: 22 out. 2024.

IEEE XPLORE. **About IEEE Xplore Digital Library**. 2024. Disponível em: https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-ieee-xplore. Acesso em: 1 dez. 2024.

ISSON, Jean Paul; HARRIOTT, Jesse S. **People analytics in the era of big data:** changing the way you attract, acquire, develop, and retain talent. Nova Jersey: John Wiley & Sons, 2016.

LAWSON, Adrienn; HENDRICK, Stephen. **2024 State of Tech Talent Report**: survey-based insights into the current state of technical talent acquisition, retention, and management globally. São Francisco: The Linux Foundation, 2024. Disponível em: https://training.linuxfoundation.org/wp-content/uploads/2024/04/2024-Tech-Talent-Report-FINAL.pdf. Acesso em: 14 jan. 2026.

LEE, Kai-Fu. **Inteligência artificial**: como os robôs estão mudando o mundo, a forma como amamos, nos relacionamos, trabalhamos e vivemos. Rio de Janeiro: Globo Livros, 2019.

L'HEUREUX, Alexandra; GROLINGER, Katarina; CAPRETZ, Miriam. Machine learning with big data: challenges and approaches. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 2017, Boston. **Proceedings** [...]. Boston: IEEE, 2017. p. 3110-3115. Disponível em: https://ieeexplore.ieee.org/document/7906512. Acesso em: 22 out. 2023.

KARNA, Hrvoje et al. The Effects of Turnover on Expert Effort Estimation. **Journal of Information and Organizational Sciences,** v. 44, n. 1, p. 51-81, 2020. ISSN 1846-3312. DOI: 10.31341/jios.44.1.3. Disponível em: https://doi.org/10.31341/jios.44.1.3. Acesso em: 22 out. 2024.

MANPOWERGROUP. **Escassez de Talentos 2023**: Brasil. São Paulo: ManpowerGroup, 2023. Infográfico. Disponível em: https://7793757.fs1.hubspotusercontent-na1.net/hubfs/7793757/Brasil_BR_TS_Infografico.pdf?__hstc=13131435.00217e4d5b055d77dc2a067c37d224ac.1768413842828.1768413842828.1768413842828.1&__hssc=13131435.1.1768413842828&__hsfp=1484a697f1d42ce304a6da4845cdd27b&hsCtaTracking=4416c9f1-c8ba-4f10-92b0-97a44a46768d%7C934ff08d-20cd-45d8-8506-4ba16939b4fc. Acesso em: 14 jan. 2026.

MENDES, Francisco. **Gestão do RH 4.0**: Digital, humano e disruptivo. São Paulo: Literare Books, 2021.

MELO, Bernardo et al. **Revisão Sistemática de Literatura (RSL):** um guia da teoria à prática. Barreiros, PE: Ed. dos Autores, 2023. Disponível em: https://repositorio.ifpe.edu.br/xmlui/handle/123456789/997. Acesso em: 10 dez. 2025.

MILKOVICH, G. T.; NEWMAN, J. M.; GERHART, B. **Compensation**. 11. ed. New York: McGraw-Hill, 2014.

MITCHELL, Terence R. et al. Why people stay: using job embeddedness to predict voluntary turnover. **Academy of Management Journal**, Briarcliff Manor, v. 44, n. 6, p. 1102–1121, 2001.

MPLOYER ADVISOR. **How benefits correlate to attraction and retention**. 2025. Disponível em: https://mployeradvisor.com/blog/how-benefits-correlate-to-attraction-and-retention. Acesso em: 6 mar. 2026.

MUSANGA, Vengai; TARAMBIWA, Edmore; ZVAREVASHE, Kudakwashe. A Supervised Machine Learning Model to Optimize Human Resources Analytics for Employee Churn Prediction. In: ZIMBABWE CONFERENCE OF INFORMATION AND COMMUNICATION TECHNOLOGIES (ZCICT), 1., 2022, Harare. **Proceedings** [...]. Harare: IEEE, 2022. p. 1-6. Disponível em: https://doi.org/10.1109/ZCICT55726.2022.10045987. Acesso em: 22 out. 2024.

OECD. **OECD employment outlook 2023**. Paris: OECD Publishing, 2023. Disponível em: https://www.oecd.org/en/publications/oecd-employment-outlook-2023_08785bba-en.html. Acesso em: 6 mar. 2026.

PROVOST, Foster; FAWCETT, Tom. **Data science para negócios**. Rio de Janeiro: Alta Books, 2016.

REVISTA HSM MANAGEMENT. **People analytics, a fronteira tech na gestão de RH.** 2016. Disponível em: https://hsmmanagement.com.br/people-analytics-a-fronteira-tech-na-gestao-de-rh/. Acesso em: 22 mai. 2024.

ROBBINS, S. P. **Comportamento organizacional**. 11. ed. São Paulo: Pearson Prentice Hall, 2009.

RUSSELL, Stuart. **Inteligência artificial a nosso favor**: como manter o controle sobre a tecnologia. São Paulo: Editora Schwarcz, 2021.

SCIENCEDIRECT. **ScienceDirect:** Elsevier's premier platform of peer-reviewed scholarly literature. 2024. Disponível em: https://www.elsevier.com/products/sciencedirect. Acesso em: 1 dez. 2024.

SCOPUS. **Scopus:** comprehensive, multidisciplinary, trusted abstract and citation database. 2024. Disponível em: https://www.elsevier.com/solutions/scopus. Acesso em: 1 dez. 2024.

SHAFIE, Mohanmad Reza et al. A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation. **Decision Analytics Journal,** v. 11, 2024. ISSN 2772-6622. DOI: 10.1016/j.dajour.2024.100461. Disponível em: https://doi.org/10.1016/j.dajour.2024.100461. Acesso em: 22 out. 2024.

SHRM. **2022 employee benefits survey**. Alexandria, VA: Society for Human Resource Management, 2022.

SUBHASH, Pavan. **IBM HR Analytics Employee Attrition & Performance**. v. 1, 2017. Disponível em: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data. Acesso em: 31 jul. 2025.

VULPEN, Erik van. **The Basic Principles of People Analytics**. 2ª Edição. [S.l.]: AIHR, 2019. ISBN 9781097268757. Disponível em: https://www.aihr.com/resources/The_Basic_principles_of_People_Analytics.pdf. Acesso em: 8 dez. 2025.

WORLDATWORK. **The WorldatWork handbook of compensation, benefits & total rewards**: a comprehensive guide for HR professionals. Hoboken: John Wiley & Sons, 2007.

WEB OF SCIENCE. **Web of Science Trust the difference**. 2024. Disponível em: https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/. Acesso em: 1 dez. 2024.

YAHIA, Nesrine Bem; HLEL, Jihen; PALACIOS, Ricardo Colomo. From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. **IEEE Access,** v. 9, p. 60447-60458, 2021. DOI: 10.1109/ACCESS.2021.3074559. ISSN 2169-3536. Disponível em: https://doi.org/10.1109/ACCESS.2021.3074559. Acesso em: 22 out. 2024.

YUAN, Shuai; KROON, Brigitte; KRAMER, Astrid. Building prediction models with grouped data: A case study on the prediction of turnover intention. **Human Resource Management Journal**, v. 34, n. 1, p. 20-38, 2021. ISSN 1748-8583. DOI: 10.1111/1748-8583.12396. Disponível em: https://doi.org/10.1111/1748-8583.12396. Acesso em: 22 out. 2024.

APÊNDICE A – METODOLOGIA DE PEOPLE ANALYTICS DE 4 FASES

Esta metodologia possui como objetivo, implementar um processo estruturado e simplificado para implantação de *people analytics* em uma organização, utilizando uma abordagem orientada por dados para gerar *insights* acionáveis e suportar a tomada de decisão estratégica.

A.1 DEFINIÇÃO DE PAPÉIS E RESPONSABILIDADES

Para o sucesso do projeto, é fundamental a clara definição dos papéis envolvidos, conforme definido na Figura A. 1:

Figura A. 1 - Papéis e responsabilidades

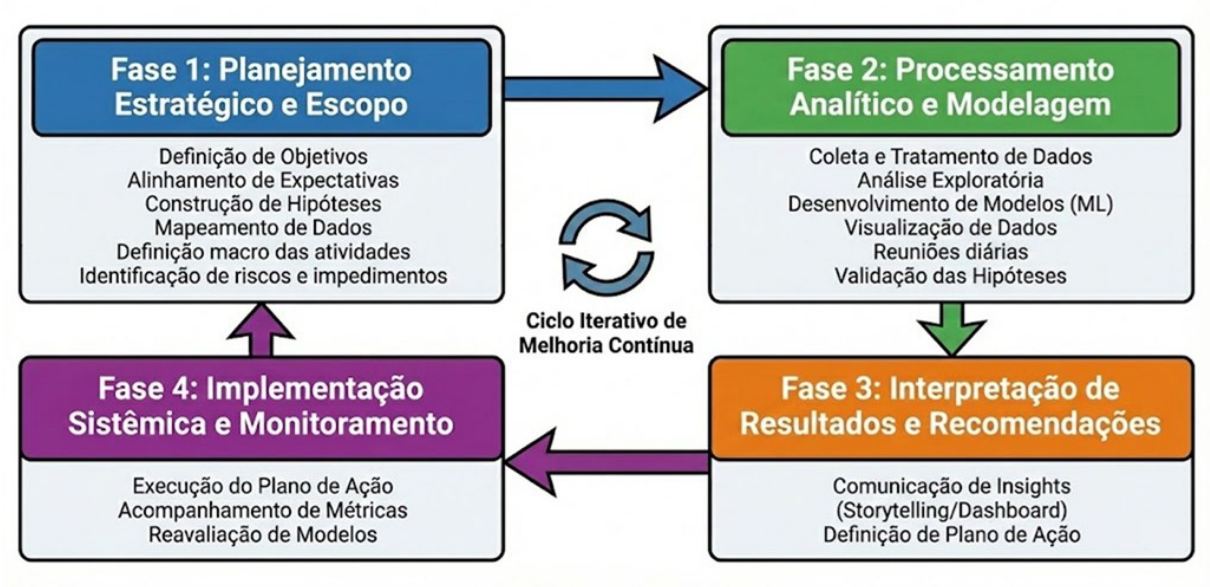
Papel	Responsabilidades Chave	Perfil
Stakeholders	- Definir os objetivos de negócio (ex: reduzir o turnover em X%). - Fornecer hipóteses iniciais baseadas em sua percepção do negócio. - Aprovar as recomendações e alocar recursos para a implementação das ações.	Diretores, Gerentes e Líderes de Setor com interesse direto na retenção de talentos.
Especialista de Negócio	- Atuar como ponte entre Stakeholders e a equipe técnica. - Traduzir os problemas de negócio em questões analíticas. - Validar e refinar hipóteses. - Interpretar os resultados da análise, revelando insights para o negócio. - Propor recomendações e planos de ação baseados nos dados.	Profissional de Recursos Humanos (Business Partner) com profundo conhecimento dos processos e desafios da gestão de pessoas na organização.
Especialista(s) Técnico(s)	- Extrair, tratar e consolidar dados de diversas fontes (SQL, APIs, planilhas). - Realizar análises exploratórias e estatísticas. - Desenvolver e validar modelos de Machine Learning (usando Python, R). - Criar dashboards e visualizações interativas (Power BI, Tableau).	Profissionais com habilidades técnicas em Engenharia de Dados, Ciência de Dados, Análise de BI. A quantidade de profissionais dependerá da complexidade e do porte da empresa.

Fonte: Elaborada pelo autor (2026).

A.2 CICLO DE VIDA DA METODOLOGIA: ABORDAGEM EM QUATRO FASES

O processo é dividido em quatro fases cíclicas, garantindo agilidade e alinhamento contínuo, conforme ilustrado pela Figura A. 2:

Figura A. 2 - Ciclo de vida da metodologia



Fonte: Elaborada pelo autor (2026).

A.2.1 Fase 1: Planejamento Estratégico e Escopo

O objetivo desta fase é garantir que o projeto esteja alinhado com as metas estratégicas da empresa e que os requisitos técnicos sejam claramente compreendidos.

A.2.1.1 Atividade 1: Reunião de Kick-off Estratégico

A primeira etapa consiste na **Reunião de *Kick-off* Estratégico**, que conta com a presença dos *stakeholders* e do especialista de negócio. A pauta do encontro divide-se em dois momentos. Inicialmente, realiza-se o **alinhamento de expectativas**, onde os *stakeholders* comunicam o que esperam do projeto e como o sucesso será medido (por exemplo: "esperamos identificar os três principais motivos de saída para agir sobre eles").

Em seguida, procede-se à **construção de hipóteses**. Com a facilitação do especialista de negócio, o grupo levanta as suposições iniciais focando no "o quê" e

no "porquê". Algumas hipóteses comuns incluem: a percepção de que colaboradores com baixa avaliação de desempenho recente são mais propensos a sair; a influência da distância entre a residência e o trabalho; ou a ideia de que a falta de um plano de carreira claro aumenta a chance de turnover.

O **resultado esperado** é uma relação simples contendo o objetivo do projeto, as métricas de sucesso e a lista de hipóteses a serem investigadas. Recomenda-se o registro desses itens em softwares de gestão como JIRA ou Trello.

A.2.1.2 Atividade 2: Reunião de Alinhamento Técnico

A etapa seguinte consiste na **Reunião de Alinhamento Técnico**, envolvendo o especialista de negócio e os especialistas técnicos. O encontro inicia-se com a **tradução da necessidade**, momento em que o especialista de negócio apresenta as metas e hipóteses previamente definidas pelos *stakeholders*.

Na sequência, a equipe técnica realiza o **mapeamento de dados**, identificando onde encontrar as informações necessárias para testar as hipóteses. As fontes podem incluir, por exemplo, o sistema de RH (para dados cadastrais), o sistema de avaliação de desempenho e planilhas de pesquisa de clima.

Além disso, é feita a **definição macro das atividades**, desenhando-se um plano de alto nível que abrange as etapas de coleta, tratamento, análise e visualização dos dados. Por fim, ocorre a **identificação de riscos e impedimentos**, um ponto crítico para antecipar atrasos, como a falta de acesso a bancos de dados, a não padronização de informações ou a necessidade de licenças específicas para ferramentas.

O **resultado esperado** é a elaboração de um plano técnico inicial, com as fontes de dados mapeadas, as tarefas técnicas cadastradas e uma lista de impedimentos encaminhada para resolução pelo especialista de negócio.

A.2.2 Fase 2: Processamento Analítico e Modelagem

Nesta fase, a equipe técnica dedica-se à transformação de dados brutos em inteligência, operando em um ciclo ágil gerenciado por reuniões diárias. As atividades centrais, executadas pelo especialista técnico, iniciam-se com a **coleta e preparação**

dos dados, etapa que envolve extrair, limpar, padronizar e enriquecer a base (criando novas variáveis, como "tempo de casa" ou "tempo desde a última promoção").

Na sequência, realiza-se a **análise exploratória e modelagem**, visando validar ou refutar as hipóteses iniciais por meio de estatística. É neste momento que ocorre o treinamento de algoritmos de *machine learning* para identificar padrões complexos e prever a probabilidade de saída do colaborador (*turnover*). O ciclo encerra-se com a **visualização de dados**, construindo dashboards interativos (utilizando ferramentas como Power BI) para explorar os fatores de risco.

A gestão do processo ocorre por meio de **reuniões diárias** de 15 minutos entre os especialistas. O especialista técnico relata o progresso e aponta impedimentos (como dúvidas de negócio ou problemas de acesso), enquanto o especialista de negócio atua na remoção desses obstáculos. Para garantir a transparência, recomenda-se o uso de quadros de atividades baseados em metodologias ágeis (como *Scrum* ou *Kanban*) e a estimativa do esforço das tarefas utilizando técnicas como a sequência de Fibonacci.

Através da Figura A. 3, é possível visualizar um exemplo de registro de atividades:

Figura A. 3 - Quadro de Atividades

Fonte: Elaborada pelo autor (2026).

A.2.3 Fase 3: Interpretação de Resultados e Recomendações

O objetivo desta fase é traduzir os achados técnicos em uma narrativa clara e convincente que motive os *stakeholders* à ação. A atividade central é a **Reunião de Comunicação e Visualização de Dados**, que conta com a apresentação do especialista de negócio e o suporte técnico do especialista técnico.

A reunião estrutura-se inicialmente na **apresentação da análise (via *storytelling*)**, onde o especialista de negócio utiliza *dashboards* para ilustrar os resultados, mantendo a narrativa focada no impacto para o negócio e não apenas na técnica. Demonstra-se, por exemplo, como modelos preditivos podem identificar que a ausência de um plano de carreira é um fator de risco crítico em determinados setores.

Na sequência, procede-se à **revelação de *insights* e determinação de recomendações**. Com base nos dados, propõem-se ações concretas, como a criação de *workshops* com líderes para estruturar planos de desenvolvimento individual (PDI), ao constatar que a percepção de falta de desenvolvimento supera a questão salarial como causa de saída.

O **resultado esperado** é a consolidação de um plano de ação aprovado pelos *stakeholders*, contendo a definição clara de responsáveis e prazos.

A.2.4 Fase 4: Implementação Sistêmica e Monitoramento

A análise de dados só gera valor efetivo quando as ações propostas são executadas e seus resultados mensurados. Nesta fase final, ocorre primeiramente a implementação, onde os *stakeholders* e suas equipes executam o plano de ação aprovado, contando com o suporte consultivo do especialista de negócio.

Paralelamente, estabelece-se a rotina de avaliar e monitorar. A equipe técnica atualiza periodicamente o *dashboard* para acompanhar a taxa de *turnover* e verificar se as ações estão surtindo o efeito desejado. Além disso, o próprio modelo preditivo passa por reavaliações constantes, sendo retreinado com novos dados sempre que necessário para manter sua acurácia.

Dessa forma, o ciclo recomeça de maneira iterativa, gerando novos *insights* e refinando continuamente as estratégias de retenção da empresa.

A.3 ANÁLISE COMPARATIVA: METODOLOGIA PROPOSTA VERSUS FRAMEWORK SCRUM

A metodologia proposta, embora não seja uma implementação "pura" do Scrum, incorpora muitos de seus princípios e práticas para promover agilidade e alinhamento.

O Quadro A. 1 detalha essa relação, destacando as semelhanças conceituais e as diferenças práticas:

Quadro A. 1 - Framework SCRUM vs Metodologia de análise de dados

Componente	Framework Scrum (Puro)	Metodologia de Análise de Dados (Adaptação Proposta)	Análise Comparativa (Semelhanças e Diferenças)
Papéis	**Product Owner (PO):** Maximiza o valor do produto, gerenciando e priorizando o Product Backlog. **Scrum Master (SM):** Garante o processo, remove impedimentos e serve ao time. **Time de Desenvolvimento:** Constrói o incremento do produto.	**Especialista de Negócio:** Define prioridades (como PO), traduz necessidades e remove impedimentos (como SM). **Especialista(s) Técnico(s):** Extrai dados, desenvolve modelos e cria visualizações (como o Time de Desenvolvimento).	**Semelhança:** Clara separação entre o "o quê" (negócio) e o "como" (técnico). **Diferença Chave:** Fusão dos papéis de PO e SM. O Especialista de Negócio acumula a responsabilidade de priorizar o trabalho e facilitar o processo. Isso otimiza a comunicação, mas elimina a figura de um guardião do processo puramente neutro (o SM).
Eventos	**Sprint Planning:** Reunião para planejar o trabalho do Sprint. **Sprint:** Um ciclo com tempo fixo (time-box) de 2 a 4 semanas para criar um incremento. **Daily Scrum:** Reunião diária de 15 min para sincronização e identificação de impedimentos. **Sprint Review:** Apresentação do incremento aos stakeholders para obter feedback. **Sprint Retrospective:** Reunião para inspecionar e adaptar o processo de trabalho.	**Reunião de Alinhamento Técnico (Fase 1):** Funciona como um planejamento, onde as hipóteses são convertidas em um plano de trabalho técnico. **Ciclo de Execução e Análise (Fase 2):** Não possui um time-box rígido; dura o necessário para investigar um conjunto de hipóteses. **Reuniões Diárias:** Equivalente direto da Daily Scrum, com o mesmo propósito. **Reunião de Comunicação e Visualização:** Análoga à Sprint Review, onde os insights e o dashboard são apresentados aos Stakeholders.	**Semelhança:** Adota uma cadência de planejamento, sincronização diária e revisão com as partes interessadas. **Diferença Chave:** Ausência de Sprints com time-box fixo e da Sprint Retrospective. O fluxo de trabalho é mais contínuo (semelhante ao Kanban) do que iterativo com tempo fixo. A falta de uma Retrospectiva formal significa que a melhoria do processo pode ocorrer de forma ad-hoc, em vez de ser um evento institucionalizado.
Artefatos e Práticas	**Product Backlog:** Lista ordenada de tudo o que é necessário no produto. **Sprint Backlog:** Itens do Product Backlog selecionados para o Sprint, mais o plano para entregá-los. **Incremento:** A soma de todos os itens do Backlog concluídos durante um Sprint.	**Lista de Hipóteses:** Gerada na Fase 1, funciona como um "Backlog de Investigação". **Quadro de Atividades (Quadro Scrum):** Durante a Fase 2, as atividades técnicas são gerenciadas em um quadro. O esforço pode ser estimado (ex: Fibonacci) para prever a capacidade. **Insights, Modelo Preditivo e Dashboard:** O resultado tangível do ciclo de análise; o "incremento" de conhecimento e valor gerado.	**Semelhança:** Utiliza artefatos para dar transparência ao trabalho (lista de prioridades, plano de tarefas, entregável de valor). A adoção de um quadro e estimativas aumenta o alinhamento com práticas ágeis. **Diferença Chave:** O "backlog" é composto por hipóteses a serem testadas, não por funcionalidades a serem construídas. O "incremento" é um ativo de conhecimento (um modelo, um insight), que é intrinsecamente mais exploratório e incerto do que uma feature de software.

Fonte: Elaborada pelo autor (2026).