Modelo Preditivo de Absenteísmo e Clusterização de Perfis Comportamentais: Uma Abordagem de *People Analytics* Aplicada à Gestão de Frequência Organizacional

César Ciuffo
Universidade de Brasília (UnB)
Brasília, Brasil
ORCID: 0009-0001-0248-6791
email: cesar.moreira@aluno.unb.br;
cesar.moreira@apexbrasil.com.br

Brenno Lopes
Instituto Brasileiro de Ensino,
Desenvolvimento e Pesquisa (IDP)
Brasília, Brasil
ORCID: 0009-0002-2720-9595
email: brenno.lopes@apexbrasil.com.br

Pedro Cella
Universidade Cruzeiro do Sul
Brasília, Brasil
ORCID: 0009-0005-9544-7829
email: pedro.cella@apexbrasil.com.br

***Resumo*—O absenteísmo representa um dos fenômenos mais custosos e complexos da gestão de pessoas, impactando produtividade, clima organizacional e custos indiretos como sobrecarga de equipes e horas extras emergenciais. Abordagens tradicionais de gestão de frequência operam de forma descritiva e reativa, limitando a capacidade de intervenção preventiva. Este trabalho propõe uma abordagem integrada de *People Analytics* aplicada à gestão preditiva de absenteísmo, seguindo a metodologia CRISP-DM e combinando duas frentes complementares: (i) um *Score de Risco de Absenteísmo*, baseado em algoritmo Random Forest, que estima a probabilidade de ocorrência de eventos de ausência em janelas de 7, 30 e 90 dias; e (ii) uma clusterização híbrida baseada em K-Means, que segmenta colaboradores em quatro perfis comportamentais distintos (Estáveis, Recorrentes, Sazonais e Graves). O estudo foi conduzido em uma organização pública brasileira de promoção de exportações, utilizando 14.533 eventos de absenteísmo do período 2019-2025, distribuídos entre 676 colaboradores. O modelo preditivo foi selecionado entre 14 candidatos por meio de validação cruzada estratificada, alcançando AUC de 0,855, acurácia de 80,2% e *recall* de 63,1% no conjunto de validação temporal. A clusterização foi validada por meio de *Silhouette Score*, índice de Calinski-Harabasz, índice de Davies-Bouldin e teste ANOVA, com diferenças estatisticamente significativas ($p < 0{,}001$) entre os clusters. Os resultados foram integrados a um protótipo de painel interativo em Power BI, concebido para uso operacional pela área de Recursos Humanos, contemplando a visualização de scores individuais, perfis comportamentais e indicadores agregados; a adoção efetiva em rotina decisória e a aferição longitudinal de impacto permanecem como trabalho futuro. As contribuições principais incluem: (a) um *pipeline* reproduzível de engenharia de variáveis comportamentais, com destaque para *features* de sazonalidade baseadas em consistência multi-anual; (b) uma abordagem híbrida de clusterização que combina critérios determinísticos e aprendizado não supervisionado; e (c) um instrumento operacional integrado que pode ser replicado em outras organizações.**

***Index Terms*—People Analytics; Absenteísmo; Aprendizado de Máquina; CRISP-DM; Random Forest; K-Means; Clusterização; Score de Risco.**

I. Introdução

A gestão de pessoas nas organizações contemporâneas enfrenta um desafio crescente: transformar o grande volume de dados operacionais capturados em sistemas de informação de Recursos Humanos (RH) em conhecimento acionável que oriente decisões preventivas e estratégicas. Dentre os fenômenos mais custosos e operacionalmente sensíveis observados nesse contexto, o absenteísmo, entendido como a ausência planejada ou não planejada do colaborador em seu posto de trabalho [1], ocupa posição central, impactando produtividade, continuidade de serviços, custos indiretos associados a horas extras emergenciais e redistribuição de carga, bem como o próprio clima organizacional [2], [3].

A gestão tradicional de frequência, amplamente disseminada em organizações públicas e privadas, apoia-se predominantemente em painéis descritivos que consolidam indicadores históricos de faltas, atrasos e totais de ausência por área ou período. Embora tais painéis cumpram o papel de oferecer visibilidade retrospectiva ao corpo gerencial, eles apresentam limitações estruturais relevantes para o desenho de intervenções preventivas: (i) não oferecem mecanismos explícitos de antecipação de risco de recorrência; (ii) não permitem priorização objetiva entre colaboradores, tratando-os como uma massa homogênea; e (iii) não capturam padrões comportamentais distintos que demandam estratégias diferenciadas de atuação por parte da área de RH. Essas limitações, amplamente discutidas na literatura de *People Analytics* [4]–[6], justificam a transição de um modelo descritivo e reativo para uma abordagem preditiva e segmentada.

O presente trabalho insere-se nesse esforço de transição, propondo uma abordagem integrada e reproduzível para a gestão preditiva de absenteísmo, desenvolvida em uma organização pública brasileira de promoção de exportações. A abordagem combina, sob o arcabouço metodológico CRISP-DM [7], [8], duas frentes complementares: um modelo su-

pervisionado de *Score de Risco*, que estima a probabilidade de ocorrência de eventos de ausência em janelas temporais de curto, médio e longo prazo (7, 30 e 90 dias); e um modelo não supervisionado de clusterização comportamental, que segmenta os colaboradores em perfis distintos a partir de características agregadas de frequência, gravidade, regularidade e sazonalidade. A conjugação dessas duas frentes permite responder simultaneamente a duas perguntas gerenciais distintas mas igualmente relevantes: *quem* demanda atenção prioritária no curto prazo (score) e *por que* demanda, do ponto de vista de padrão comportamental (cluster).

A. Problema de Pesquisa

A questão central que norteia este trabalho pode ser formalizada da seguinte maneira:

> *Como evoluir a gestão de frequência organizacional de um modelo estritamente descritivo para um modelo preditivo e segmentado, capaz de antecipar riscos de absenteísmo e compreender padrões comportamentais distintos de forma explicável, operacionalmente útil e governável, utilizando apenas dados históricos tipicamente disponíveis em sistemas de RH?*

A formulação evidencia três restrições práticas importantes. A primeira é a *disponibilidade limitada de dados*: restringimo-nos deliberadamente a informações tipicamente capturadas em sistemas transacionais de RH (atestados, dados cadastrais, histórico organizacional), sem recorrer a fontes externas como condições climáticas, indicadores epidemiológicos ou dados de engajamento provenientes de pesquisas específicas. Essa escolha favorece a replicabilidade da abordagem em outras organizações, ao custo de renunciar a ganhos preditivos marginais obteníveis com dados externos. A segunda restrição é a *exigência de explicabilidade*: modelos caixa-preta, ainda que potencialmente mais precisos, são indesejáveis em decisões que afetam diretamente pessoas, já que as recomendações geradas precisam ser auditadas, compreendidas e, eventualmente, contestadas pelo corpo gerencial e pelos próprios colaboradores. A terceira restrição é o *foco operacional*: os artefatos produzidos devem se traduzir em insumos diretos para a rotina de RH, sendo integrados a ferramentas já utilizadas pelas lideranças (no caso, Power BI), e não permanecer como exercícios acadêmicos desacoplados da gestão.

B. Objetivos e Contribuições

O **objetivo geral** deste trabalho consiste em desenvolver, validar e operacionalizar um modelo analítico integrado que permita à área de RH antecipar riscos de absenteísmo e compreender os diferentes perfis comportamentais de seus colaboradores, sob critérios explícitos de qualidade estatística, explicabilidade e utilidade prática.

Os **objetivos específicos** são:

1) Construir um *pipeline* reproduzível de integração, limpeza e enriquecimento de dados que consolide múltiplas fontes de informação de frequência em uma base analítica unificada por colaborador;
2) Desenvolver um conjunto abrangente de variáveis (*features*) comportamentais que capture dimensões complementares do absenteísmo (volume, gravidade, recência, tendência, regularidade, sazonalidade e tipicidade), com destaque para a construção de indicadores de sazonalidade baseados em consistência multi-anual;
3) Comparar sistematicamente um conjunto de algoritmos supervisionados quanto à capacidade de prever eventos de absenteísmo futuros em múltiplos horizontes, adotando critérios de seleção que considerem tanto a capacidade discriminativa global quanto o custo operacional de falsos negativos em contextos preventivos;
4) Identificar e caracterizar perfis comportamentais de absenteísmo por meio de uma abordagem híbrida de clusterização, que combina critérios determinísticos para grupos minoritários de alta gravidade com aprendizado não supervisionado para os demais perfis;
5) Validar estatisticamente os agrupamentos obtidos por meio de métricas internas (*Silhouette*, Calinski-Harabasz, Davies-Bouldin), estabilidade (*Adjusted Rand Index*) e testes de significância (ANOVA unidirecional);
6) Integrar os resultados a um painel interativo de apoio à decisão, viabilizando a incorporação dos artefatos analíticos à rotina gerencial em etapas posteriores de adoção organizacional.

Como **contribuições principais**, este trabalho oferece:

- Um *pipeline* completo, reproduzível e publicamente documentado de engenharia de variáveis comportamentais para absenteísmo, incluindo 57 *features* organizadas em sete categorias conceituais;
- Uma proposta de *features* de sazonalidade calibradas à realidade brasileira (estações do ano Verão, Outono, Inverno e Primavera), com ênfase em *consistência multi-anual*, diferentemente de abordagens que apenas medem concentração em um único período;
- Uma comparação empírica de 14 configurações de modelos supervisionados com protocolo rigoroso de validação cruzada estratificada e *split* temporal, discutindo os *trade-offs* entre *precision* e *recall* no contexto específico de gestão preventiva;
- Uma abordagem híbrida para clusterização comportamental que lida explicitamente com a presença de grupos minoritários de alta relevância clínica, evitando a fragmentação artificial observada em aplicações ingênuas de K-Means;
- Um estudo de caso detalhado em organização pública brasileira que pode servir de referência metodológica para replicação em outras instituições com características similares.

C. Organização do Artigo

O restante deste artigo está estruturado da seguinte forma. A Seção II revisa a literatura relacionada em *People Analytics*, modelagem preditiva de absenteísmo, metodologia CRISP-DM e fundamentos dos algoritmos utilizados. A Seção III descreve a metodologia proposta, organizada conforme as seis fases do

CRISP-DM. A Seção IV apresenta os experimentos conduzidos e os resultados obtidos tanto na modelagem preditiva quanto na clusterização. A Seção V discute criticamente os achados, retoma as hipóteses estabelecidas e aponta as limitações do estudo. Por fim, a Seção VI apresenta as considerações finais e trabalhos futuros.

II. Fundamentação Teórica e Trabalhos Relacionados

Esta seção organiza o arcabouço teórico que sustenta a proposta metodológica do trabalho. A Subseção II-A contextualiza a evolução do campo de *People Analytics*, localizando a gestão preditiva de absenteísmo como uma de suas aplicações maduras. A Subseção II-B revisa modelos explicativos e abordagens recentes de predição de absenteísmo baseadas em aprendizado de máquina. A Subseção II-C apresenta a metodologia CRISP-DM, utilizada como espinha dorsal do trabalho. A Subseção II-D revisa os fundamentos dos algoritmos supervisionados comparados no estudo. A Subseção II-E discute os fundamentos da clusterização não supervisionada, com ênfase no algoritmo K-Means e nas métricas de validação empregadas.

A. People Analytics e a Transição para Modelos Preditivos em RH

O termo *People Analytics*, também denominado *HR Analytics* ou *Workforce Analytics*, designa a aplicação sistemática de métodos quantitativos e técnicas de ciência de dados para apoiar decisões em gestão de pessoas [4], [6]. Em uma revisão de escopo abrangente, Tursunbayeva et al. [6] identificam três fronteiras conceituais principais do campo: (i) a utilização de dados para descrição e diagnóstico de fenômenos organizacionais; (ii) o uso de modelos preditivos para antecipar eventos como turnover, engajamento e ausências; e (iii) o desenvolvimento de modelos prescritivos que sugerem intervenções otimizadas. O presente trabalho situa-se na interseção entre a segunda e a terceira fronteiras, ao combinar um modelo preditivo com uma segmentação prescritiva que orienta ações diferenciadas por perfil.

Davenport [9], em seu trabalho seminal sobre organizações orientadas por análises, estabelece que a maturidade analítica é um diferencial competitivo sustentável quando integrada ao núcleo decisório, e não confinada a relatórios operacionais periféricos. Especificamente no contexto de gestão de talentos, Davenport, Harris e Shapiro [10] demonstraram como modelos preditivos aplicados a RH permitem antecipar fenômenos críticos como turnover e identificar perfis de alta performance, introduzindo o conceito de *talent analytics* como evolução natural das práticas tradicionais de gestão. Essa linha foi posteriormente consolidada por Rasmussen e Ulrich [5], que enfatizam um princípio central para a adoção prática de People Analytics: o valor organizacional só se concretiza quando o esforço analítico parte de problemas reais de negócio e é comunicado de forma clara e acionável à liderança.

Mais recentemente, Marler e Boudreau [4] conduziram uma revisão baseada em evidências do campo, identificando que a maior parte das aplicações maduras concentra-se em recrutamento, turnover e desempenho, sendo o absenteísmo uma fronteira ainda pouco explorada em estudos brasileiros. Huselid [11] e Minbaeva [12], em edições especiais da revista *Human Resource Management* dedicadas ao tema, reforçam a necessidade de rigor científico na construção de *capabilities* analíticas em RH, apontando que a ausência de metodologia formal, documentação e validação estatística é uma das principais barreiras à legitimidade do campo frente aos gestores.

Dentro desse contexto, o absenteísmo se apresenta como um objeto de estudo com características particularmente adequadas a *People Analytics*: é um evento mensurável, recorrente, com dados históricos estruturados disponíveis em praticamente todas as organizações, e com consequências econômicas e humanas relevantes. Van der Laken et al. [13] argumentam que estudos aplicados de *People Analytics* demandam protocolo metodológico explícito (em especial quando utilizam algoritmos complexos), o que reforça a pertinência de adotarmos CRISP-DM como estrutura formal deste trabalho.

B. Absenteísmo: Modelos Explicativos e Predição via Machine Learning

O absenteísmo não é um fenômeno homogêneo. O modelo clássico de Steers e Rhodes [14] postula que o comportamento de comparecimento ao trabalho resulta da interação entre *motivação* e *capacidade* de comparecer, sendo ambas influenciadas por características pessoais (saúde, responsabilidades familiares, idade), situação de trabalho (satisfação, autonomia, pressão por presença) e pressões externas (valores sociais, políticas organizacionais). Essa estrutura conceitual permanece atual mesmo após décadas, tendo sido empiricamente reforçada por meta-análises subsequentes [1], [15].

Mais recentemente, Markussen et al. [3], utilizando dados administrativos noruegueses, decompõem o absenteísmo por licença-saúde em componentes atribuíveis ao indivíduo, ao empregador e ao médico prescritor, evidenciando a natureza multinível do fenômeno. Bierla et al. [2], por sua vez, examinam a coexistência do absenteísmo com o *presenteísmo* (estar presente mas com produtividade reduzida), reforçando que ausência e presença não são dicotomias simples. Laaksonen et al. [16] destacam ainda a existência sistemática de diferenças de gênero em padrões de ausência por motivo de saúde, associadas a fatores biológicos, psicossociais e à divisão desigual de responsabilidades familiares. Allebeck e Mastekaasa [17] revisam fatores de risco para licenças-saúde em populações gerais, consolidando que idade, gênero, condições socioeconômicas e características do posto de trabalho operam como preditores robustos.

A aplicação de *machine learning* para predição de absenteísmo não é uma fronteira inteiramente nova, tendo ganhado tração crescente nos últimos anos. Boa parte dos estudos aplicados apoia-se no *dataset* público *Absenteeism at Work* [18], coletado em uma empresa brasileira de *courier* entre 2007 e 2010 e disponibilizado no UCI Machine Learning Repository. Sobre dados dessa natureza, Tewari et al. [19] apresentam uma aplicação exploratória de algoritmos de regressão (regressão li-

near e *support vector regression*) para predição de absenteísmo em multinacionais, identificando a idade como um dos fatores mais associados ao fenômeno. Llamas Blázquez [20], em estudo recente publicado no *Journal of Occupational Medicine and Toxicology*, demonstra que *features* relacionadas a recorrência, sazonalidade e padrões temporais permitem estimar probabilidades futuras de ausência com modelos supervisionados interpretáveis, validando empiricamente a relevância de dimensões temporais para esse problema. Em literatura relacionada sobre *employee attrition*, Punnoose e Ajit [21] e Fallucchi et al. [22] comparam algoritmos de classificação em problemas similares de RH, reiterando a superioridade prática de *ensembles* baseados em árvores (Random Forest, Gradient Boosting) sobre modelos lineares em contextos com *features* tabulares heterogêneas.

A leitura conjunta dos trabalhos revisados evidencia três lacunas recorrentes na literatura aplicada de predição de absenteísmo, que delimitam diretamente o espaço de contribuição do presente estudo. A **primeira lacuna** é o tratamento do absenteísmo *exclusivamente como classificação binária*: estudos como os de Tewari et al. [19], Punnoose e Ajit [21] e Fallucchi et al. [22] estimam a ocorrência do evento, mas não segmentam os colaboradores em perfis comportamentais distintos, perdendo a distinção gerencial entre *quem* demanda atenção e *por que* demanda. Em resposta, este trabalho acopla a predição a uma clusterização comportamental validada estatisticamente (Seção IV-C). A **segunda lacuna** é a *representação limitada da sazonalidade*: mesmo trabalhos que reconhecem a relevância de *features* temporais [20] tipicamente medem apenas concentração em um único período, sem exigir recorrência multi-anual. Em resposta, propõem-se *features* de sazonalidade baseadas em *consistência multi-anual* por estação (Seção III-C). A **terceira lacuna** é a *desconexão entre artefato analítico e rotina gerencial*: boa parte dos estudos encerra-se na avaliação do modelo, sem operacionalização, contrariando o princípio de valor preconizado pela própria literatura de *People Analytics* [5]. Em resposta, os resultados são integrados a um painel operacional (Seção III-F). A Tabela I sintetiza esse mapeamento entre lacunas identificadas e contribuições propostas.

Tabela I
MAPEAMENTO ENTRE LACUNAS DA LITERATURA APLICADA E CONTRIBUIÇÕES DESTE TRABALHO.

Lacuna na literatura	Resposta deste trabalho
Absenteísmo tratado apenas como classificação binária	Predição acoplada a clusterização comportamental validada (ANOVA, ARI)
Sazonalidade medida só por concentração em um período	*Features* de consistência sazonal multi-anual
Modelo desacoplado da rotina gerencial	Integração a painel operacional (Power BI) sob CRISP-DM

C. CRISP-DM: Metodologia para Projetos de Mineração de Dados

O *Cross-Industry Standard Process for Data Mining* (CRISP-DM) é a metodologia de referência para condutores de projetos de mineração de dados e ciência de dados aplicada [7], [8]. A metodologia organiza o ciclo de um projeto em seis fases iterativas e inter-relacionadas: (1) *Business Understanding* (compreensão do negócio), (2) *Data Understanding* (compreensão dos dados), (3) *Data Preparation* (preparação dos dados), (4) *Modeling* (modelagem), (5) *Evaluation* (avaliação) e (6) *Deployment* (implantação). Uma característica central da metodologia é seu reconhecimento explícito da natureza iterativa desses projetos: não se trata de um fluxo linear, mas de um ciclo em que cada fase pode retroalimentar fases anteriores à medida que novos insights emergem.

Em revisão sistemática recente, Schroer et al. [23] analisam mais de duas décadas de aplicações do CRISP-DM em áreas diversas (saúde, indústria, finanças, telecomunicações), confirmando que a metodologia continua sendo a mais adotada em projetos aplicados, apesar do surgimento de variantes mais especializadas. Os autores destacam que as fases de *Business Understanding* e *Data Understanding* são frequentemente subestimadas em publicações acadêmicas, mas consomem tipicamente metade ou mais do esforço total em projetos reais, observação corroborada pela experiência conduzida neste trabalho.

A adoção explícita do CRISP-DM no presente artigo cumpre três funções: serve como *dispositivo organizacional* para estruturar a narrativa metodológica; garante *rastreabilidade* entre decisões analíticas e necessidades de negócio; e facilita a *replicabilidade* por parte de outras organizações que desejem adaptar a abordagem. A Seção III apresenta a aplicação concreta das seis fases ao problema estudado.

D. Aprendizado Supervisionado para Classificação Binária

O problema de predizer a ocorrência de ao menos um evento de absenteísmo em uma janela temporal futura pode ser formulado como uma tarefa de *classificação binária*: dado um vetor de características $\mathbf{x} \in \mathbb{R}^p$ que descreve o histórico de um colaborador até uma data de referência, estima-se a probabilidade $\hat{p} = P(y = 1 \mid \mathbf{x})$ de que ocorra ao menos um evento em uma janela futura de N dias, onde $y \in \{0, 1\}$. Esta subseção revisa os principais algoritmos comparados no estudo.

1) Regressão Logística: A Regressão Logística [24] modela a probabilidade posterior por meio da função sigmoide aplicada a uma combinação linear das *features*:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\boldsymbol{\beta}^\top \mathbf{x} + \beta_0), \tag{1}$$

onde $\sigma(z) = 1/(1+e^{-z})$ é a função sigmoide. Os parâmetros $\boldsymbol{\beta}$ são estimados por máxima verossimilhança, tipicamente com regularização L2 (Ridge) controlada pelo hiperparâmetro $C = 1/\lambda$. A principal virtude da regressão logística no contexto deste trabalho é a interpretabilidade direta dos coeficientes (quando devidamente padronizados), ao custo de uma hipótese forte de linearidade no espaço *logit*.

2) Random Forest: Random Forest [25] é um *ensemble* baseado em *bagging* que treina B árvores de decisão, cada uma sobre uma amostra *bootstrap* do conjunto de treinamento e considerando um subconjunto aleatório de *features* em cada divisão. A predição agregada é:

$$\hat{p}_{\text{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{p}_b(\mathbf{x}), \tag{2}$$

onde $\hat{p}_b$ é a probabilidade estimada pela b-ésima árvore. A aleatorização em duas camadas (amostras e *features*) descorrelaciona as árvores e reduz a variância do *ensemble*, sem aumentar significativamente o viés. A profundidade máxima das árvores (*max_depth*) é um hiperparâmetro crítico: valores baixos controlam *overfitting* mas podem subajustar; valores altos capturam interações complexas mas podem memorizar o treino. Random Forest oferece ainda medidas naturais de importância de *features* em duas vertentes complementares: *mean decrease in impurity* (MDI), baseada em ganho médio nas divisões das árvores, e *permutation importance*, baseada em queda efetiva de desempenho quando cada *feature* é embaralhada no conjunto de validação. A *permutation importance*, embora computacionalmente mais cara, é conhecidamente mais robusta na presença de *features* com distribuições heterogêneas [26], sendo a escolha adotada neste trabalho (Seção IV-B3).

3) Gradient Boosting e AdaBoost: Gradient Boosting [27] treina árvores de forma sequencial, em que cada árvore subsequente busca corrigir os erros (resíduos) das anteriores. Formalmente, o modelo é uma soma ponderada de *learners* fracos:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot h_m(\mathbf{x}), \tag{3}$$

onde h_m é a m-ésima árvore, ν é a *learning rate*, e F_{m-1} acumula o conhecimento dos estágios anteriores. AdaBoost [28], um precursor, usa uma estratégia distinta: reponderação adaptativa das amostras mal classificadas a cada iteração. Ambos superam, em geral, o Random Forest em capacidade preditiva quando adequadamente regularizados, ao custo de maior sensibilidade a hiperparâmetros e *outliers*.

4) Extra Trees, KNN, Decision Tree e MLP: Adicionalmente, foram avaliadas no estudo variantes: *Extra Trees* [29], que introduz aleatoriedade extra na escolha dos pontos de corte; o algoritmo *K-Nearest Neighbors*, que classifica com base na maioria dos K vizinhos mais próximos; uma árvore de decisão única como *baseline* interpretativo; e um *Multilayer Perceptron* (MLP) com duas camadas ocultas, utilizado como representante de abordagens não lineares baseadas em redes neurais. A comparação sistemática desses modelos é apresentada na Seção IV.

5) Métricas de Avaliação: A avaliação de modelos de classificação binária em contextos desbalanceados demanda um conjunto de métricas complementares. Seja TP (verdadeiros positivos), TN (verdadeiros negativos), FP (falsos positivos) e FN (falsos negativos) os quatro quadrantes da matriz de confusão em um dado limiar de decisão. Utilizamos:

- **Acurácia:** $(TP + TN)/(TP + TN + FP + FN)$, proporção global de acertos.
- **Precision:** $TP/(TP + FP)$, qualidade dos alertas positivos.
- **Recall (sensibilidade):** $TP/(TP + FN)$, cobertura dos eventos positivos reais.
- **F1-Score:** média harmônica entre *precision* e *recall*, dada por $F_1 = 2PR/(P + R)$, onde P e R denotam, respectivamente, *precision* e *recall*.
- **AUC-ROC:** área sob a curva ROC, caracterizando a capacidade do modelo de ordenar corretamente pares positivo-negativo, independentemente do limiar [30].

Em problemas desbalanceados, Saito e Rehmsmeier [31] argumentam que a curva *precision-recall* pode ser mais informativa do que a curva ROC. Em contextos aplicados, como o presente, a escolha do modelo final deve considerar também o *custo assimétrico* entre falsos positivos e falsos negativos: em gestão preventiva de saúde, o custo de um falso negativo (colaborador em risco não identificado) tende a ser muito maior que o de um falso positivo (alerta sem concretização), justificando a priorização de *recall* [32]. A robustez estatística do modelo é verificada via *cross-validation* estratificada [33].

E. Clusterização Não Supervisionada

A clusterização busca particionar um conjunto de observações $\{\mathbf{x}_i\}_{i=1}^{n}$ em K grupos, de modo que observações no mesmo grupo sejam mais similares entre si do que em relação a observações de outros grupos, sem o uso de rótulos externos. O algoritmo K-Means [34], [35], o mais amplamente utilizado em aplicações práticas, particiona os dados minimizando a soma das distâncias quadráticas intra-cluster (também chamada *inertia*):

$$J(\mathcal{C}) = \sum_{k=1}^{K}\sum_{\mathbf{x}\in\mathcal{C}_k}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2, \tag{4}$$

onde $\boldsymbol{\mu}_k$ é o centroide do cluster $\mathcal{C}_k$.

A aplicação de K-Means depende de duas escolhas críticas: (i) a *normalização* das *features*, tipicamente via *StandardScaler* (média zero, desvio unitário), sem a qual *features* de maior amplitude dominariam a métrica de distância; e (ii) a definição de K. Para esse último, usamos três métricas de validação interna complementares:

- **Silhouette Score** [36]: para cada observação i, define-se a_i como a distância média de i às demais observações do seu cluster e b_i como a distância média de i ao cluster vizinho mais próximo. O *silhouette* de i é $s_i = (b_i - a_i)/\max(a_i, b_i) \in [-1, 1]$, com valores próximos a 1 indicando boa separação. O *silhouette* médio quantifica a qualidade global da partição.
- **Índice de Calinski-Harabasz** [37]: razão entre a dispersão entre clusters (inter) e a dispersão intra-cluster, ponderada pelos graus de liberdade. Valores mais altos indicam melhor separação.
- **Índice de Davies-Bouldin** [38]: média, entre pares de clusters, da razão entre dispersões intra-cluster e dis-

tâncias entre centroides. Valores mais baixos indicam agrupamentos mais compactos e separados.

Adicionalmente, avaliamos a *estabilidade* dos agrupamentos via *Adjusted Rand Index* (ARI) [39] entre partições obtidas com diferentes inicializações aleatórias. Valores de ARI próximos a 1 indicam que a estrutura identificada é robusta, e não um artefato do processo de otimização.

A literatura clássica de clusterização [40], [41] destaca ainda que K-Means possui viés conhecido na presença de *outliers* e grupos de tamanhos muito desiguais: observações extremas tendem a ser isoladas em micro-clusters, comprometendo a separabilidade dos grupos principais. Esse fenômeno, observado empiricamente em nossos experimentos preliminares, motivou a adoção da abordagem híbrida descrita na Seção III, em que um perfil minoritário de alta gravidade é isolado deterministicamente antes da aplicação do K-Means aos demais colaboradores. Trata-se de estratégia conceitualmente alinhada a trabalhos que combinam aprendizado supervisionado e não supervisionado em estruturas híbridas [42].

III. Metodologia

A metodologia deste trabalho foi estruturada conforme as seis fases do CRISP-DM [7], [8], apresentadas nas Subseções III-A a III-F. A adoção explícita dessa estrutura cumpre três funções complementares: organiza a narrativa metodológica de forma reconhecível pela comunidade de ciência de dados aplicada; garante rastreabilidade entre decisões técnicas e necessidades de negócio; e facilita a replicação da abordagem por parte de outras organizações que disponham de dados transacionais de frequência similares aos aqui utilizados. Quanto à terminologia, os termos *features* e variáveis são utilizados de forma intercambiável ao longo do texto, seguindo convenção consolidada na literatura de aprendizado de máquina aplicado; analogamente, o termo *pipeline* refere-se ao fluxo completo de integração, preparação, modelagem e entrega, sem distinção semântica adicional em relação ao termo processo. A Figura 1 sintetiza o fluxo geral do projeto.

Todo o *pipeline* foi implementado em Python 3.12, utilizando as bibliotecas *pandas*, *numpy*, *scikit-learn* [43] e *scipy*, e organizado em cinco cadernos Jupyter (NB00 a NB04), correspondentes, respectivamente, a integração de dados, análise exploratória, limpeza e validação, engenharia de variáveis, e modelagem e avaliação. Os artefatos gerados em cada fase são persistidos como *datasets* intermediários (formato Excel), permitindo a execução parcial de etapas sem reprocessamento completo.

A. Business Understanding (Compreensão do Negócio)

A primeira fase do CRISP-DM estabelece a compreensão do problema do ponto de vista do negócio, antes de qualquer manipulação de dados. No contexto deste trabalho, tal fase foi conduzida por meio de entrevistas estruturadas com a liderança da área de RH da organização parceira e da análise crítica de materiais existentes, em especial o painel descritivo de gestão de frequência em uso.

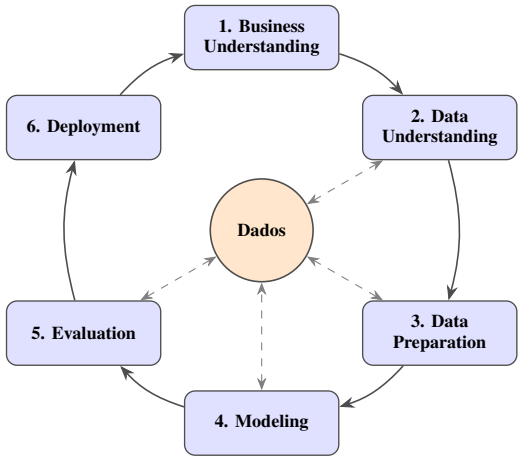

Figura 1. Fluxo CRISP-DM aplicado ao problema de predição e segmentação de absenteísmo. As setas sólidas indicam o fluxo principal entre fases; as setas tracejadas representam o caráter iterativo do processo, em que dados e modelos são revisitados ao longo do projeto.

1) Contexto Organizacional: A organização parceira é uma entidade pública brasileira dedicada à promoção das exportações nacionais, com quadro funcional composto por cerca de 350 colaboradores ativos distribuídos entre diretorias de negócios, gestão e planejamento, presidência e escritórios regionais. A gestão de frequência utiliza um sistema transacional que captura, por colaborador, cada evento de ausência (atestados, consultas, licenças, acompanhamento de familiares), com registros estruturados incluindo datas de início e término, duração em minutos, tipo de atestado e CID associado. Até o início do projeto, o uso analítico desses dados estava restrito a relatórios descritivos periódicos, sem qualquer componente preditivo ou segmentado.

2) Formalização do Problema e Hipóteses: O problema central, apresentado na Seção I, foi operacionalizado em duas tarefas analíticas complementares:

Tarefa 1 - Predição de eventos futuros (classificação binária): Dado o histórico de um colaborador até uma data de referência t_0, prever se ocorrerá pelo menos um evento de absenteísmo em cada uma das janelas futuras de 7, 30 e 90 dias, $[t_0+1, t_0+N]$. A validação principal *out-of-time* (Seção III-E) adota a janela de 30 dias como horizonte-alvo de referência, por representar o intervalo mais relevante para intervenção preventiva de RH; os scores de 7 e 90 dias são gerados em paralelo pelo mesmo *pipeline* como inferências operacionais complementares.

Tarefa 2 - Segmentação comportamental (clusterização): A partir de um vetor agregado de características comportamentais por colaborador, identificar grupos com padrões distintivos de frequência, produzindo perfis interpretáveis e acionáveis pela área de RH.

Ao longo da execução do projeto, foram estabelecidas quatro hipóteses que sustentam a viabilidade e a relevância da abordagem:

H1 - Hipótese Preditiva - É possível estimar, a partir de dados históricos de frequência, a probabilidade futura de ocorrência de eventos de absenteísmo com acurácia superior a 70%, limiar considerado útil para uso operacional.

H2 - Hipótese Comportamental - Existem padrões distintos e estatisticamente significativos de comportamento de frequência, que podem ser identificados por meio de técnicas de clusterização não supervisionada.

H3 - Hipótese de Valor Operacional - A utilização conjunta do score preditivo e da segmentação comportamental permite priorização mais eficiente da atuação da área de RH, oferecendo simultaneamente resposta a *quem* demanda atenção e a *por que* demanda.

H4 - Hipótese de Interpretabilidade - Modelos baseados em estruturas explicáveis (árvores, *ensembles* de árvores e modelos lineares) podem atingir o patamar de qualidade preditiva exigido pelo problema sem recorrer a arquiteturas opacas, preservando governança, auditabilidade e confiança gerencial.

3) Critérios de Sucesso e Restrições: Dois critérios de sucesso foram definidos antes do início da modelagem, a fim de evitar vieses ex-post: (i) o modelo preditivo final deveria atingir acurácia superior a 70% e AUC-ROC superior a 0,75 no conjunto de validação; (ii) a segmentação final deveria produzir pelo menos três perfis com tamanho significativo (acima de 5% da população), com diferenças estatisticamente significativas ($p < 0{,}05$) entre grupos nas *features* principais e com interpretabilidade qualitativa validada em reunião com a área de RH.

Três restrições operacionais foram assumidas: (i) *uso exclusivo de dados transacionais internos*, sem recorrer a fontes externas; (ii) *pseudonimização* dos colaboradores, com uso do identificador funcional CHAPA como chave pseudonimizada — não diretamente identificante, porém reversível pela organização — e remoção preventiva de colunas contendo nomes, CPF, PIS, dados salariais e informações médicas detalhadas; (iii) *explicabilidade por construção*, priorizando algoritmos interpretáveis sobre abordagens *black-box*.

B. Data Understanding (Compreensão dos Dados)

A segunda fase concentra esforços no entendimento profundo das fontes de dados disponíveis, sua qualidade, completude, vieses e potenciais usos.

1) Fontes de Dados: O projeto integrou cinco fontes distintas, detalhadas na Tabela II:

As três primeiras fontes foram extraídas em momentos distintos, resultando em pequenas diferenças de formato (cabeçalhos, codificação da coluna CHAPA, presença de linhas de totais agregados), que precisaram ser harmonizadas na fase seguinte. A base de colaboradores, embora contendo dados demográficos ricos, é uma extração datada de 2023 e, portanto, não reflete integralmente a composição atual do quadro, limitação retomada na discussão (Seção V). A lista de empregados ativos, obtida apenas para marcação de quem permanece na organização, contém somente a coluna CHAPA.

2) Análise Exploratória Inicial: A análise exploratória (implementada no NB01) teve quatro objetivos: (i) caracterizar a distribuição temporal dos eventos; (ii) identificar os tipos de atestado e CIDs mais frequentes; (iii) descrever a distribuição de duração dos afastamentos; e (iv) cruzar padrões de absenteísmo com atributos demográficos.

Entre os principais insights descritivos, destacam-se: (a) a distribuição de eventos por mês revelou variações relevantes ao longo do ano, com picos nos meses de setembro (+15,8% em relação à média mensal) e outubro (+14,6%), e vale expressivo em dezembro ($-36{,}4\%$), padrão plausivelmente associado ao ciclo anual organizacional e aos meses de férias; (b) aproximadamente 75% dos eventos são de duração curta (1–2 dias), mas os 10% classificados como graves (8 dias ou mais) concentram cerca de metade do total de dias de afastamento, evidenciando forte assimetria; (c) o tipo de atestado mais frequente é *consulta médica* ($\sim$35%), seguido por *doença* ($\sim$20%) e *consulta odontológica* ($\sim$13%); (d) observaram-se diferenças expressivas entre gêneros, com mulheres apresentando mais eventos da categoria *acompanhamento de familiar/filho*, compatível com a literatura internacional [15], [16].

3) Avaliação de Qualidade dos Dados: Quatro problemas de qualidade foram identificados e tratados: (i) a codificação da CHAPA diferia entre bases (inteiro em uma, *string* com zeros à esquerda em outra), gerando CHAPAs *fantasmas* em *joins* ingênuos; (ii) as bases de 2024 e 2025 continham linhas agregadas (*Total Geral*) intercaladas, que precisaram ser filtradas; (iii) alguns eventos apresentavam datas de admissão posteriores à data do evento, indicando registros históricos anteriores à atual relação contratual; (iv) campos textuais (*TIPO_ATESTADO*, *CID*, *DIRETORIA*) apresentavam variações de capitalização e acentuação.

C. Data Preparation (Preparação dos Dados)

A terceira fase concentra o maior esforço técnico do projeto e consolida os dados brutos em uma base analítica unificada por colaborador, pronta para modelagem.

1) Integração e Padronização: A integração (NB00) envolveu: (i) padronização do identificador CHAPA, com remoção de zeros à esquerda e conversão uniforme para inteiro; (ii) seleção de colunas comuns às três bases de atestados; (iii) concatenação das três bases com adição de uma coluna *FONTE* para rastreabilidade; (iv) remoção de duplicatas exatas, considerando iguais os registros com mesma combinação de CHAPA, data de início, data de término, tipo de atestado e total de minutos; (v) *merge* com a base de colaboradores pela CHAPA; (vi) marcação dos colaboradores presentes na lista de ativos por meio de uma coluna booleana *ATIVO_2026*. Ao final, foram obtidos 14.533 eventos validados, distribuídos entre 676 CHAPAs únicas.

2) Decisão Metodológica: Separação entre Treino e Clusterização: Uma decisão metodológica importante foi tomada nesta fase: dos 676 colaboradores com eventos, apenas 322 estão na lista de ativos de março/2026. Os 354 restantes

Tabela II
FONTES DE DADOS UTILIZADAS NO ESTUDO.

Fonte	Período	Registros	CHAPAs	Conteúdo principal
Atestados 2019–2023	01/2019–12/2023	7.688	544	Eventos de absenteísmo históricos
Atestados 2024	01/2024–12/2024	2.518	348	Complemento 2024
Atestados 2025	01/2025–12/2025	4.327	368	Complemento 2025
Base de colaboradores	Extração 2023	335	335	Dados demográficos, cargo, diretoria
Empregados ativos	Março/2026	351	351	Lista atual de colaboradores ativos

correspondem a colaboradores desligados durante o período observado. Decidiu-se *manter os desligados no treinamento do modelo preditivo*, ampliando o universo de aprendizado, mas *limitar a clusterização e os outputs operacionais apenas aos 322 ativos*, para que as decisões gerenciais incidam apenas sobre colaboradores efetivamente empregados. Essa separação evita viés de sobrevivente na fase de *training* sem contaminar os perfis operacionais com comportamentos de pessoas que já não integram o quadro. A Tabela III consolida a cadeia de derivação da amostra final.

Cabe esclarecer, em relação à última linha da Tabela III, que os 29 colaboradores ativos sem qualquer evento registrado no período 2019–2025 são classificados como pertencentes ao perfil *Estável* por *regra de negócio determinística* (ausência total de eventos implica padrão basal de frequência), e não por atribuição algorítmica do K-Means. Esses colaboradores compõem o universo operacional do painel (351 = 322 clusterizados + 29 estáveis por regra) mas não participam do treino do modelo preditivo, uma vez que a engenharia de *features* comportamentais pressupõe a existência de ao menos um evento de referência.

3) Limpeza, Validação e Classificação de Gravidade: O NB02 aplicou quatro blocos de limpeza: (a) validação de datas, com descarte de eventos cuja data de início fosse posterior à data de término ou que apresentassem datas futuras implausíveis; (b) cálculo da duração em dias como a diferença entre data de término e data de início, acrescida de uma unidade, com piso de 1 dia; (c) classificação de cada evento em uma de cinco faixas de duração (*curto*, *médio*, *longo*, *prolongado*, *extenso*); (d) padronização textual de campos categóricos (*uppercase*, remoção de espaços excedentes, substituição de *placeholders* por valores nulos). Um evento foi adicionalmente marcado como *grave* quando sua duração excedeu 8 dias, limiar estabelecido em conjunto com a área de RH e compatível com a literatura de medicina ocupacional [17].

4) Engenharia de Variáveis Comportamentais: O NB03 é o núcleo de agregação conceitual do projeto. Ele transforma a granularidade dos dados (até então um registro por evento) em um vetor de *features* por colaborador, capturando dimensões complementares de seu padrão de frequência. Foram construídas 57 *features* organizadas em sete categorias conceituais, sintetizadas na Tabela IV.

Destaque para as *features* de sazonalidade. Dentre as categorias listadas, as *features* de sazonalidade merecem explicação detalhada, por constituírem uma das contribuições metodológicas deste trabalho. Para cada colaborador, definem-se as estações do ano a partir do mês de cada evento, adotando o calendário brasileiro: Verão (Dez–Fev), Outono (Mar–Mai), Inverno (Jun–Ago) e Primavera (Set–Nov). Seja E_c o conjunto de eventos históricos do colaborador c, $E_c^{(e)}$ o subconjunto ocorrido na estação e, e A_c o conjunto de anos em que c teve ao menos um evento. A *estação de pico* de c é $e_c^* = \arg\max_e |E_c^{(e)}|$. Duas medidas derivadas são então calculadas:

$$\text{prop_estacao_pico}_c = \frac{|E_c^{(e_c^*)}|}{|E_c|}, \tag{5}$$

$$\text{consistência_sazonal}_c = \frac{|\{a \in A_c : |E_c^{(e_c^*)} \cap a| \geq 1\}|}{|A_c|}. \tag{6}$$

Combinadas, as duas medidas geram o *índice_sazonalidade*, definido como o produto entre *prop_estacao_pico* e *consistência_sazonal*. A distinção conceitual em relação a medidas tradicionais de concentração trimestral é que a consistência exige que o padrão se repita ao longo de múltiplos anos para ser considerado sazonalidade verdadeira. Um colaborador com muitos eventos concentrados em um único ano atípico (por exemplo, em razão de uma doença pontual) terá alta proporção na estação de pico, mas baixa consistência, e, portanto, não será classificado como sazonal. Uma *flag* binária *is_sazonal* marca positivamente colaboradores cuja proporção na estação de pico seja ao menos 0,35 e que apresentem eventos nessa estação em pelo menos dois anos distintos.

5) Construção da Variável-Alvo e Split Temporal: Para o modelo preditivo, construiu-se uma variável-alvo binária $y_N \in \{0, 1\}$ para cada janela $N \in \{7, 30, 90\}$, indicando se o colaborador apresentou pelo menos um evento em $[t_0 + 1, t_0 + N]$, onde t_0 é uma data de referência. Adotou-se *split temporal*: dados com $t_0 = 31/12/2023$ formam o conjunto de treinamento (com *features* calculadas sobre o histórico até essa data e rótulos baseados nos 30 dias seguintes, horizonte escolhido como janela-alvo principal de validação *out-of-time*); dados com $t_0 = 31/12/2024$ formam o conjunto de validação. Os modelos para as janelas de 7 e 90 dias são treinados de forma análoga sobre os mesmos *features*, variando apenas a definição do rótulo, e suas métricas de validação são tratadas como resultados secundários. Esse desenho reproduz fielmente o uso em produção, em que *features* são computadas sobre o passado para prever o futuro imediato, e mitiga o risco de vazamento temporal inerente a partições aleatórias [33].

Tabela III
DERIVAÇÃO DA AMOSTRA: DO UNIVERSO BRUTO AOS UNIVERSOS DE TREINO E CLUSTERIZAÇÃO.

Etapa	N	Descrição
CHAPAs com eventos 2019–2025	676	Universo bruto com ao menos um evento registrado
Colaboradores ativos (mar/2026)	351	Lista de referência para operação
Desligados no período	354	Incluídos apenas no treino preditivo
Ativos com eventos	322	Universo efetivo da clusterização
Ativos sem eventos	29	Classificados como *Estável* por definição
Universo de treino do modelo	676	Todos os CHAPAs com eventos
Universo operacional (painel)	351	322 clusterizados + 29 estáveis

Tabela IV
CATEGORIAS DE *features* CONSTRUÍDAS NO NB03.

Categoria	Descrição e exemplos
Volume e frequência	Total de eventos, eventos por ano (normalizado pelo tempo de casa), eventos em janelas de 30, 60, 90, 180 e 365 dias, total de dias afastado, minutos por ano.
Gravidade	Média e máximo de dias por evento, proporção de eventos em cada faixa de duração, proporção de eventos graves.
Recência	Dias desde o último evento, amplitude do histórico (intervalo entre primeiro e último evento).
Tendência	Razão entre eventos recentes e históricos (30/90 dias, 90/365 dias), densidade recente, coeficiente de variação mensal.
Regularidade	Gap médio e desvio-padrão entre eventos consecutivos, mínimo entre eventos.
Sazonalidade	Proporção de eventos por trimestre, proporção na estação de pico, consistência sazonal, índice de sazonalidade, flag *is_sazonal*.
Tipicidade	Número de tipos distintos de atestado, de CIDs distintos, proporção por tipo principal de atestado.

D. Modeling (Modelagem)

A quarta fase compreende o treinamento e a seleção dos modelos. As duas tarefas (predição e clusterização) foram conduzidas em paralelo, com escolhas algorítmicas e protocolos distintos.

1) Modelagem Preditiva: 14 Candidatos: Com o objetivo de identificar o melhor *trade-off* entre capacidade preditiva, estabilidade e interpretabilidade, foram treinados 14 modelos candidatos, abrangendo famílias algorítmicas distintas e variações de hiperparâmetros: Regressão Logística com $C \in \{0{,}1;\ 1{,}0;\ 10\}$; Random Forest com *max_depth* $\in \{5, 8, 12\}$; Gradient Boosting (*learning_rate* padrão e 0,05); AdaBoost; Extra Trees; K-Nearest Neighbors com $k \in \{5, 10\}$; Árvore de Decisão única (*max_depth* = 6); e um *Multilayer Perceptron* com arquitetura (64, 32). Todos os modelos foram treinados em *features* padronizadas via *StandardScaler*, e aqueles com sensibilidade a desbalanceamento de classes foram configurados com *class_weight* = 'balanced'. O protocolo de validação incluiu *cross-validation* estratificada com cinco *folds* no conjunto de treino (AUC-ROC, *accuracy* e F1-score) e avaliação final no conjunto de validação temporal.

Tratamento do desbalanceamento de classes. O conjunto apresenta prevalência positiva de 24,5% na janela-alvo de 30 dias, configurando desbalanceamento moderado. Optou-se por endereçar essa assimetria por meio de *aprendizado sensível ao custo* (*class_weight* = 'balanced'), em detrimento de técnicas de sobreamostragem sintética como SMOTE ou ADASYN, por três razões. Primeira, o desenho adota *split* temporal (*out-of-time*): quando aplicada antes da partição dos dados, a sobreamostragem sintética pode introduzir vazamento entre treino e validação; ainda que corretamente restrita ao conjunto de treino, a geração de exemplos da classe minoritária por interpolação entre vizinhos produz colaboradores "artificiais" sem contrapartida temporal plausível, em tensão com um protocolo de validação que busca reproduzir o uso em produção. Segunda, a sobreamostragem sintética distorce a distribuição marginal da classe positiva e tende a degradar a *calibração* das probabilidades estimadas, justamente o atributo que se busca preservar para o uso dos *scores* como percentuais interpretáveis no painel (Seção IV-B5). Terceira, com 676 observações na base de treino, o *class_weight* oferece controle do desbalanceamento sem ampliar artificialmente o tamanho efetivo da amostra nem inflar a confiança do modelo em regiões esparsas do espaço de *features*. Uma comparação experimental sistemática entre *class_weight*, SMOTE e ADASYN, com aferição de impacto conjunto sobre *recall* e calibração, permanece como extensão pertinente e é registrada na agenda de trabalhos futuros (Seção VI).

2) Clusterização: Abordagem Híbrida: Em experimentos preliminares (NB04), a aplicação direta de K-Means sobre o conjunto completo das 322 observações ativas apresentou um comportamento indesejado: independentemente do valor de K testado (de 2 a 8), o algoritmo tendia a isolar entre 6 e 10 colaboradores em um microcluster, correspondendo exatamente aos colaboradores com histórico marcado por eventos graves e prolongados, ainda que pouco frequentes. Essa fragmentação ocorre porque tais colaboradores são *outliers* legítimos em dimensões de gravidade (duração média e proporção de eventos graves), e o K-Means, sensível a *outliers* por construção [40], os captura preferencialmente,

comprometendo a separabilidade dos demais perfis.

Para contornar esse efeito sem descartar a informação relevante, adotou-se uma abordagem híbrida em duas etapas:

1) **Isolamento determinístico do perfil Grave.** Aplicou-se um critério explícito para identificar o perfil Grave: colaboradores com *media_dias_afastamento* superior a 5 dias *e prop_eventos_graves* superior a 10%. Esses colaboradores são removidos do *input* do K-Means e compõem um cluster próprio, rotulado como *Grave*.
2) **K-Means nos demais colaboradores.** O conjunto restante é submetido a K-Means com $K = 3$, gerando os perfis *Estável*, *Recorrente* e *Sazonal*. Foram utilizadas 9 *features* selecionadas: *eventos_por_ano*, *media_dias_afastamento*, *prop_eventos_graves*, *tendência_30_90*, *densidade_recente*, *gap_medio_dias*, *índice_sazonalidade*, *consistência_sazonal* e *prop_estacao_pico*.

A escolha de $K = 3$ para a segunda etapa foi embasada por métricas de validação interna (*Silhouette*, Calinski-Harabasz, Davies-Bouldin) avaliadas em $K \in \{2, 3, 4, 5, 6, 7, 8\}$ e por critério interpretativo de negócio (quatro perfis finais contando com o Grave). A Seção IV apresenta os resultados comparativos dessas métricas.

E. Evaluation (Avaliação)

A quinta fase estabelece o protocolo rigoroso de avaliação, garantindo que os critérios de sucesso definidos em *Business Understanding* sejam efetivamente aferidos.

1) Protocolo de Avaliação Preditiva: A avaliação preditiva é organizada em dois níveis. No primeiro, *cross-validation* estratificada de 5 *folds* no conjunto de treino fornece estimativas pontuais e intervalares das métricas, capturando variabilidade devida a particionamento. No segundo, a avaliação em um conjunto de validação *out-of-time* (dados de 2024) simula o uso em produção. A métrica primária de seleção é o F1-score, por conciliar *precision* e *recall* em uma única medida, respeitando a assimetria de custos discutida na Seção II-D. A escolha do F1-score em detrimento de alternativas como *average precision* (AP), PR-AUC ou funções de custo explícitas é defensável num cenário em que *precision* e *recall* importam operacionalmente de forma conjunta (priorização com triagem manual subsequente); tais alternativas permanecem metodologicamente adequadas para extensões futuras em que se queira otimizar para um objetivo operacional mais estreito. As métricas secundárias consideradas são AUC-ROC, *accuracy*, *precision* e *recall*. Adicionalmente, foi analisada a importância das *features* no modelo final via *permutation importance*, que oferece leitura mais robusta do que abordagens baseadas em impureza para modelos baseados em árvores [26], fornecendo interpretabilidade e apoiando a hipótese H4.

2) Protocolo de Avaliação da Clusterização: A qualidade da clusterização foi avaliada segundo quatro dimensões:

1) **Métricas internas:** *Silhouette Score* [36], índice de Calinski-Harabasz [37] e índice de Davies-Bouldin [38].
2) **Estabilidade:** *Adjusted Rand Index* (ARI) entre partições obtidas com 10 *seeds* distintas [39], indicando robustez do agrupamento frente a condições iniciais do algoritmo.
3) **Significância estatística:** ANOVA unidirecional aplicada a cada *feature* comportamental, verificando se as diferenças entre clusters são estatisticamente significativas ($p < 0{,}05$).
4) **Validação interpretativa:** reunião com a área de RH para validar, em linguagem gerencial, se os perfis identificados correspondem a situações reconhecíveis e práticas.

F. Deployment (Implantação)

A sexta e última fase do CRISP-DM consiste na integração dos artefatos analíticos ao ambiente operacional. No presente trabalho, os resultados foram entregues por meio de uma base consolidada *base_scores_clusters.xlsx*, contendo, para cada colaborador ativo: CHAPA, scores de risco em 7, 30 e 90 dias, faixa de risco categórica (baixo, moderado, alto, muito alto, crítico) associada ao score de 30 dias, cluster atribuído, nome do perfil e atributos demográficos (gênero, idade, diretoria, número de dependentes).

Essa base alimenta um painel interativo desenvolvido em Power BI, estruturado em três páginas: (a) *Visão Geral*, com indicadores agregados, distribuição dos colaboradores por perfil e por faixa de risco, e filtros por diretoria; (b) *Perfil do Colaborador*, com visualização detalhada dos scores, dados demográficos e classificação por cluster de um colaborador específico selecionado; (c) *Visão por Perfil*, com listagem e detalhamento dos colaboradores associados a cada cluster. A estrutura do painel foi desenhada com base no princípio preconizado por Rasmussen e Ulrich [5] de que o valor de *People Analytics* se materializa quando os artefatos analíticos estão disponíveis ao cotidiano decisório da liderança, embora a incorporação efetiva dessa prática dependa de etapas posteriores de adoção organizacional, descritas na Seção VI.

O painel encontra-se, no momento da submissão deste artigo, em estado de *protótipo operacional entregue*: a ferramenta foi apresentada à área de Recursos Humanos com protocolo de uso documentado e sob regime de atualização manual trimestral (reexecução dos notebooks NB00 a NB04 com dados atualizados), mas **ainda não foi implantada em rotina decisória real**. O uso efetivo do *score* e dos perfis para priorização de colaboradores em ações gerenciais concretas, bem como a eventual aferição de impacto sobre indicadores de absenteísmo, estão previstos como primeiro marco de trabalho futuro. A automação completa do *pipeline* (via agendador e integração direta com o banco de dados transacional de RH) constitui marco complementar, também discutido na Seção VI.

IV. Experimentos e Resultados

Esta seção apresenta os resultados obtidos nas duas frentes analíticas conduzidas sob o arcabouço CRISP-DM descrito na Seção III. A Subseção IV-A detalha a configuração experimental. A Subseção IV-B reporta os resultados da modelagem preditiva, incluindo a comparação dos 14 modelos candidatos,

a análise de importância de variáveis e a robustez do modelo final. A Subseção IV-C apresenta os resultados da clusterização, a caracterização dos quatro perfis identificados e sua validação estatística.

A. Configuração Experimental

Todos os experimentos foram executados em um ambiente controlado com Python 3.12, utilizando as bibliotecas *scikit-learn* 1.4 [43], *pandas* 2.1, *numpy* 1.26 e *scipy* 1.11. Todas as operações com componentes aleatórios (partição de dados, inicialização de modelos, amostragem *bootstrap* em Random Forest e Extra Trees, inicialização de centroides em K-Means) foram configuradas com *random_state* fixado em 42 para garantir reprodutibilidade, exceto quando explicitamente variado para análise de estabilidade.

O conjunto de dados efetivamente utilizado na modelagem preditiva contém 14.533 eventos de absenteísmo, correspondentes a 676 colaboradores, observados entre 01/01/2019 e 19/12/2025. Após a etapa de engenharia de variáveis, cada colaborador é descrito por um vetor de 57 *features*, gerando uma matriz $\mathbf{X} \in \mathbb{R}^{676\times57}$ (para predição) e uma matriz restrita $\mathbf{X}' \in \mathbb{R}^{322\times9}$ (para clusterização, após remoção dos colaboradores desligados e seleção de *features* específicas). O *split* temporal adotado utiliza $t_0 = 31/12/2023$ para treino e $t_0 = 31/12/2024$ para validação, com janela-alvo de 30 dias.

B. Resultados do Modelo Preditivo

1) Comparação entre os 14 Modelos Candidatos: A Tabela V sintetiza o desempenho dos 14 modelos avaliados, ordenados pelo F1-Score obtido no conjunto de validação. Além das métricas globais, a tabela apresenta o desvio-padrão do AUC em *cross-validation* estratificada de 5 *folds*, fornecendo uma medida indireta de estabilidade.

Duas observações estruturais emergem da Tabela V. Primeiro, os quatro modelos de topo pertencem à família de *ensembles* baseados em árvores (Random Forest, AdaBoost, Gradient Boosting e Extra Trees), confirmando o padrão amplamente observado na literatura de aprendizado de máquina para problemas tabulares heterogêneos [21], [22]. Segundo, as variações internas de cada família (especialmente as três configurações de *max_depth* do Random Forest) apresentam resultados próximos entre si, sugerindo que o desempenho do modelo é mais sensível à escolha da família do que ao ajuste fino de hiperparâmetros, em linha com o argumento de Breiman [25] sobre a robustez de Random Forest.

O trade-off entre precision e recall. Um resultado particularmente relevante para a discussão metodológica é o contraste entre o AdaBoost (maior *precision*, 65,0%) e a Regressão Logística com $C = 1$ (maior *recall*, 68,4%). Em um cenário operacional ingênuo, que priorizasse apenas a acurácia global, o AdaBoost seria escolhido como modelo final. Entretanto, como discutido na Seção II-D, o custo de falsos negativos (colaborador em risco não identificado pelo modelo) é significativamente maior que o de falsos positivos em contextos preventivos de saúde. Esse argumento, combinado ao critério de seleção baseado em F1-Score (que integra *precision* e *recall* de forma harmônica), conduz à escolha do Random Forest ($d = 5$) como modelo final. Ele apresenta o melhor F1-Score absoluto (0,610), *recall* competitivo (63,1%) e *precision* adequada (58,9%), além do menor desvio-padrão em *cross-validation* (0,047), indicando maior estabilidade.

2) Significância Estatística da Comparação entre Modelos: A ordenação apresentada na Tabela V baseia-se em estimativas pontuais, que não informam, por si só, se as diferenças observadas entre os modelos de topo são estatisticamente significativas ou compatíveis com mera variação amostral. Para qualificar a seleção do modelo final, a comparação foi submetida a uma análise de significância em dois níveis.

No primeiro nível, calcularam-se os intervalos de confiança de 95% para o AUC-ROC obtido em *cross-validation* estratificada de 5 *folds*, a partir da distribuição de desempenho por *fold* (estatística t de Student, $gl = 4$). A Tabela VI apresenta os intervalos dos cinco modelos de maior desempenho. Como indicador conservador de significância, intervalos que se sobrepõem não sustentam a afirmação de diferença estatisticamente detectável entre os modelos correspondentes.

Os intervalos dos cinco modelos sobrepõem-se de maneira substancial: o limite inferior do Random Forest ($d = 5$), de 0,702, situa-se bem abaixo do limite superior de todos os concorrentes, e o intervalo do AdaBoost ($[0{,}672;\ 0{,}834]$) contém integralmente a média do Random Forest. Esse padrão indica que os *ensembles* de árvore não apresentam diferença estatisticamente detectável em capacidade discriminativa, não havendo evidência, ao nível de 95%, de superioridade de qualquer um deles sobre os demais nessa métrica.

No segundo nível, para verificação direta sobre o conjunto de validação *out-of-time*, aplicaram-se dois testes pareados: o teste de DeLong [44], que compara AUC-ROC correlacionadas estimadas sobre as mesmas observações, e o teste de McNemar [45], aplicado à tabela de contingência de acertos e erros pareados (limiar de 0,50), que avalia se dois modelos cometem erros de classificação em padrões significativamente distintos. A Tabela VII apresenta os p-valores da comparação do Random Forest ($d = 5$) com os quatro demais modelos de topo. Em nenhuma das comparações a hipótese nula foi rejeitada ao nível de 5% (DeLong: $0{,}059 \leq p \leq 0{,}210$; McNemar: $0{,}322 \leq p \leq 1{,}000$), corroborando, agora sobre as predições reais, a ausência de evidência de diferença estatisticamente significativa já sugerida pela análise de intervalos. O procedimento é disponibilizado de forma reprodutível no repositório do projeto.

Implicação para a seleção do modelo. A ausência de diferença discriminativa estatisticamente detectável entre os modelos de topo reforça, e não enfraquece, a lógica de seleção adotada: uma vez que o critério de AUC isolado não distingue os *ensembles*, a escolha do Random Forest ($d = 5$) repousa legitimamente sobre critérios complementares já explicitados, a saber, o melhor F1-Score absoluto (0,610), o *recall* competitivo (63,1%) alinhado à assimetria de custos discutida na Seção II-D, e a maior estabilidade em *cross-validation* (menor desvio-padrão de AUC, 0,047). A decisão torna-se, assim, transparente e auditável, em vez de apoiada em uma diferença

Tabela V
DESEMPENHO COMPARATIVO DOS 14 MODELOS CANDIDATOS, ORDENADOS PELO F1-SCORE NO CONJUNTO DE VALIDAÇÃO.

Modelo	**CV AUC** ($\mu \pm \sigma$)	**Val AUC**	**Acurácia**	**Precision**	**Recall**	**F1**
Random Forest ($d = 5$)	$0{,}760 \pm 0{,}047$	0,855	80,2%	58,9%	63,1%	**0,610**
AdaBoost	$0{,}753 \pm 0{,}065$	**0,864**	**81,6%**	**65,0%**	53,9%	0,589
Gradient Boosting	$0{,}748 \pm 0{,}052$	0,841	79,5%	57,2%	60,3%	0,587
Extra Trees	$0{,}742 \pm 0{,}058$	0,838	78,8%	55,6%	61,7%	0,584
Logistic Regression ($C = 1$)	$0{,}704 \pm 0{,}086$	0,772	75,3%	51,2%	**68,4%**	0,586
Gradient Boosting (lr=0,05)	$0{,}745 \pm 0{,}055$	0,832	78,2%	55,4%	59,8%	0,575
Random Forest ($d = 8$)	$0{,}755 \pm 0{,}051$	0,847	79,1%	56,8%	57,5%	0,571
Random Forest ($d = 12$)	$0{,}751 \pm 0{,}053$	0,840	78,7%	55,1%	56,8%	0,559
Logistic Regression ($C = 10$)	$0{,}698 \pm 0{,}091$	0,764	73,8%	49,3%	65,7%	0,564
Logistic Regression ($C = 0{,}1$)	$0{,}701 \pm 0{,}082$	0,768	75,0%	50,8%	63,2%	0,563
MLP (64, 32)	$0{,}731 \pm 0{,}074$	0,811	76,4%	52,1%	58,3%	0,550
KNN ($k = 10$)	$0{,}687 \pm 0{,}072$	0,753	72,9%	47,8%	53,4%	0,504
KNN ($k = 5$)	$0{,}672 \pm 0{,}079$	0,741	71,5%	46,2%	51,1%	0,485
Decision Tree ($d = 6$)	$0{,}651 \pm 0{,}088$	0,712	69,8%	44,3%	52,8%	0,482

Tabela VI
INTERVALOS DE CONFIANÇA DE 95% PARA O AUC-ROC EM *cross-validation* (5 *folds*) DOS CINCO MODELOS DE MAIOR DESEMPENHO.

Modelo	**CV AUC** (μ)	**IC 95%**
Random Forest ($d = 5$)	0,760	[0,702; 0,818]
Random Forest ($d = 8$)	0,755	[0,692; 0,818]
AdaBoost	0,753	[0,672; 0,834]
Gradient Boosting	0,748	[0,683; 0,813]
Extra Trees	0,742	[0,670; 0,814]

Tabela VII
TESTES PAREADOS ENTRE O MODELO FINAL (RANDOM FOREST, $d = 5$) E OS DEMAIS MODELOS DE TOPO, NO CONJUNTO DE VALIDAÇÃO *out-of-time*.

Comparação	**DeLong** (p, AUC)	**McNemar** (p, erros)
vs. RF ($d = 8$)	0,210	1,000
vs. AdaBoost	0,208	0,322
vs. Gradient Boosting	0,059	0,450
vs. Extra Trees	0,083	0,871

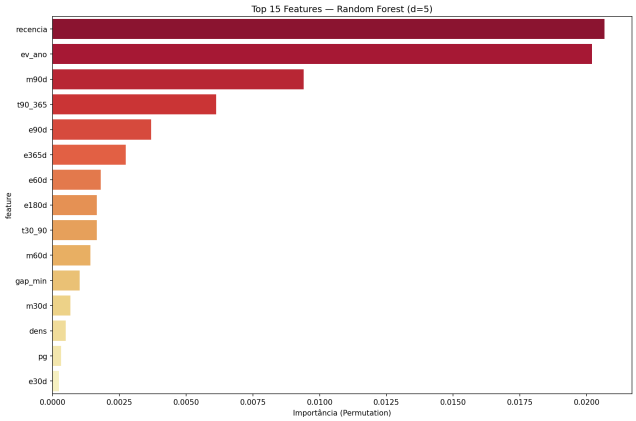

Figura 2. Importância relativa das 15 *features* mais relevantes no Random Forest final, medida por *permutation importance* (10 repetições, métrica-base AUC-ROC no conjunto de validação).

numérica marginal entre estimativas pontuais.

3) Análise de Importância das Variáveis: Para complementar o desempenho quantitativo com interpretabilidade, analisou-se a importância relativa das *features* no Random Forest final por meio de *permutation importance* [25], com 10 repetições e métrica-base AUC-ROC, aferida no conjunto de validação. Essa escolha é mais robusta do que a métrica *mean decrease in impurity* (MDI) padrão, que pode superestimar a importância de variáveis com alta cardinalidade ou com muitos níveis contínuos [26]; a *permutation importance*, por medir diretamente a queda de desempenho quando cada *feature* é embaralhada, reflete de maneira mais fiel a contribuição efetiva de cada variável para a capacidade preditiva do modelo. A Figura 2 apresenta as 15 *features* com maior importância.

O padrão observado é consistente com a teoria do absenteísmo e com a literatura aplicada [14], [20]. As cinco variáveis mais importantes são, em ordem: (1) *dias_desde_ultimo_evento* (recência); (2) *eventos_30d* (atividade recente); (3) *eventos_por_ano* (frequência histórica normalizada); (4) *gap_medio_dias* (regularidade); e (5) *densidade_recente*. A predominância de *features* relacionadas a recência e atividade recente sugere que o comportamento de curto prazo é o preditor mais forte de absenteísmo futuro, observação que alinha modelagem preditiva e intuição gerencial: colaboradores com eventos frequentes nos últimos 30 dias tendem a apresentar eventos adicionais no período imediatamente seguinte, enquanto colaboradores com longo tempo sem eventos tendem a manter esse padrão. Esse achado é coerente com o modelo clássico de Steers e Rhodes [14], no qual o comportamento de comparecimento apresenta efeito de hábito, e alinha-se aos resultados empíricos de Llamas Blázquez [20], em que categorias temporais de curto prazo também emergiram como as mais preditivas em contexto análogo.

É interessante notar que as *features* de sazonalidade (contribuição metodológica deste trabalho) aparecem em posições intermediárias (*índice_sazonalidade* em 8º lugar, *consistência_sazonal* em 11º), confirmando que capturam informação

complementar não redundante em relação às *features* tradicionais de volume e recência, mas com peso menor no problema específico de predição. A utilidade maior dessas *features*, como se verá, ocorre na tarefa de clusterização.

4) Curvas ROC e Precision-Recall: A Figura 3 apresenta as curvas ROC e *Precision-Recall* do modelo Random Forest final no conjunto de validação. A área sob a curva ROC atingiu 0,855, e a área sob a curva PR atingiu 0,621. Dado o desbalanceamento do problema (aproximadamente 24,5% de eventos positivos no conjunto de validação para a janela de 30 dias), a curva PR oferece uma visão complementar mais informativa, como recomendado por Saito e Rehmsmeier [31].

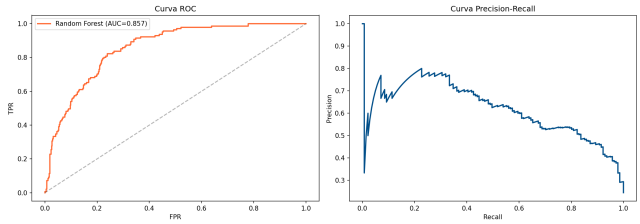

Figura 3. Curvas ROC (esquerda) e Precision-Recall (direita) do Random Forest final no conjunto de validação temporal.

A análise das curvas permite ainda identificar pontos operacionais relevantes. Para tornar essa flexibilidade auditável, a Tabela VIII apresenta o desempenho do Random Forest (d=5) em diferentes limiares de decisão para a janela de 30 dias, incluindo o número de colaboradores preditos como positivos em cada configuração.

Tabela VIII
SENSIBILIDADE DO RANDOM FOREST ($d = 5$) POR LIMIAR DE DECISÃO, JANELA-ALVO DE 30 DIAS, CONJUNTO DE VALIDAÇÃO *out-of-time*.

Limiar	Precision	Recall	F1	Pos. preditos
0,15	46,5%	89,4%	0,612	271
0,30	52,5%	81,6%	**0,639**	219
0,40	52,9%	70,2%	0,604	187
0,50	57,7%	61,0%	0,593	149
0,60	62,9%	51,8%	0,568	116
0,70	70,2%	41,8%	0,524	84

Três observações emergem da Tabela VIII. Primeiro, o F1-Score máximo ocorre em torno de limiar 0,30 (F1 = 0,639), acima do F1 do limiar padrão 0,50 (F1 = 0,593); o modelo permite, portanto, ganho operacional ao ajuste fino de limiar para a janela em questão. Segundo, o limiar padrão 0,50 favorece precision sobre recall (razão P/R = 0,95), enquanto limiares mais baixos (0,15–0,30) favorecem recall (razão P/R entre 0,52 e 0,64), configuração adequada para cenários em que o objetivo é maximizar a cobertura de colaboradores em risco para subsequente triagem manual pela área de RH. Terceiro, o número de positivos preditos varia de 84 (limiar 0,70) a 271 (limiar 0,15), oferecendo espaço operacional expressivo para calibração por capacidade da equipe de RH. Essa flexibilidade foi incorporada ao painel Power BI, permitindo à gestão ajustar o limiar conforme a estratégia vigente.

No caso da janela de 7 dias, discutida na Subseção IV-B6, mesmo a aplicação de limiar agressivamente reduzido produz resultado operacional limitado: com limiar 0,15, obtêm-se 105 predições positivas com *precision* de apenas 9,5% e *recall* de 34,5% (F1 = 0,149). Essa evidência reforça a interpretação da janela de 7 dias como instrumento de *ranking*, não de classificação binária sob limiar fixo, e motiva o uso operacional do score de 30 dias como referência principal para alertas quantitativos.

5) Calibração das Probabilidades: A exibição dos *scores* ao usuário final ocorre como percentuais numéricos (por exemplo, "Score 30d = 72%"), o que torna relevante avaliar se esses valores refletem probabilidades bem calibradas ou se funcionam apenas como *ranking* operacional. Para aferir a calibração, aplicou-se Platt *scaling* ao Random Forest via *CalibratedClassifierCV* e calculou-se o Brier Score, que mede a distância quadrática média entre probabilidades previstas e rótulos binários observados (valores mais baixos indicam melhor calibração).

Para a janela de 30 dias, o modelo calibrado apresenta Brier Score de 0,134 no conjunto de validação, contra 0,185 obtido por um *baseline* ingênuo que prevê a prevalência constante (24,5%). A redução de 27,6% em relação ao *baseline* indica calibração substantivamente informativa. Em complemento ao Brier Score, o *Expected Calibration Error* (ECE, dez *bins*), que pondera o desvio absoluto entre a confiança prevista e a frequência observada em cada faixa, foi de apenas 0,058 na janela de 30 dias, valor baixo que reforça a aderência das probabilidades previstas às frequências reais. A Figura 4 apresenta o diagrama de confiabilidade (dez *bins* por quantil): a curva aproxima-se da diagonal identidade na faixa central de probabilidades (entre 0,15 e 0,70), com leve sobreestimação na cauda superior (*scores* acima de 0,75 correspondem a frequência observada ligeiramente inferior ao previsto). As janelas de 7 dias e 90 dias apresentam Brier de 0,048 e 0,118, respectivamente, também abaixo dos respectivos *baselines* ingênuos, embora a calibração em 7 dias seja limitada pelas frequências muito baixas.

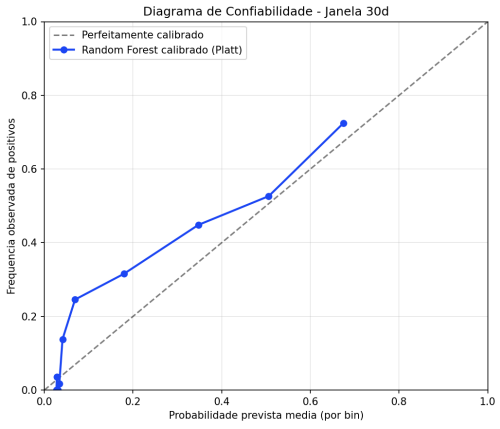

Figura 4. Diagrama de confiabilidade do Random Forest calibrado (Platt *scaling*) para a janela de 30 dias. Eixo horizontal: probabilidade prevista média por *bin*; eixo vertical: frequência observada de positivos.

Do ponto de vista operacional, a calibração observada sustenta o uso dos *scores* como ordenação de prioridade (objetivo central do painel Power BI) e permite interpretação aproximada dos percentuais como probabilidades reais na faixa operacional usada para alertas (acima de 50%). O refinamento por regressão isotônica, que pode reduzir ainda mais o Brier em regiões de cauda, permanece como próximo passo previsto na agenda de trabalhos futuros (Seção VI).

6) Desempenho por Janela de Horizonte: Embora a validação primária descrita nas subseções anteriores tenha adotado a janela de 30 dias como horizonte-alvo de referência, o *pipeline* preditivo treina três modelos independentes, um por janela (7, 30 e 90 dias), utilizando o mesmo conjunto de *features* e o mesmo *split* temporal (*out-of-time*). A Tabela IX apresenta o desempenho comparativo das três janelas no conjunto de validação, avaliado com o modelo Random Forest ($d = 5$), calibrado via Platt *scaling* (*CalibratedClassifierCV* com *cv*=5).

O padrão observado é coerente com a natureza do problema. A janela de 7 dias apresenta prevalência positiva muito baixa (5,0% do conjunto de validação), o que torna a predição por limiar padrão (0,5) pouco informativa: nenhum colaborador atinge o corte, resultando em *precision* e *recall* nulos. Apesar disso, o AUC de 0,712 indica que o modelo preserva capacidade discriminativa útil, compatível com uso como *ranking* para triagem de curto prazo. Para uso operacional efetivo na janela de 7 dias, seria necessário calibrar um limiar mais baixo (por exemplo, 0,15) que equilibre cobertura e falsos positivos. Assim, a janela de 7 dias deve ser interpretada, nesta configuração, como instrumento de ordenação de risco (*ranking*) e não como classificador binário operacional sob limiar padrão. A janela de 90 dias, por outro lado, apresenta prevalência elevada (37,6%) e AUC de 0,920, configurando-se como horizonte mais fácil do ponto de vista discriminativo, adequado a planejamentos táticos de médio prazo. A janela de 30 dias mantém equilíbrio entre desafio estatístico e utilidade operacional, justificando sua adoção como alvo principal de validação e de apresentação nas subseções anteriores. Os pequenos diferenciais numéricos entre a Tabela IX e a Tabela V (por exemplo, AUC 0,857 versus 0,855 e F1 0,593 versus 0,610 para 30 dias) decorrem da aplicação da calibração sigmoide e do refit completo do *ensemble* no pipeline unificado.

7) Robustez e Estabilidade: A estabilidade do modelo foi avaliada em dois níveis. Em *cross-validation* estratificada de 5 *folds* no conjunto de treino, o AUC médio foi de 0,760 com desvio-padrão de 0,047, indicando que o desempenho é consistente em diferentes partições dos dados. Comparativamente, uma árvore de decisão única apresentou desvio de 0,088 nas mesmas condições, quase o dobro, confirmando o ganho de estabilidade proporcionado pelo *ensemble* [25].

Adicionalmente, compara-se o desempenho do Random Forest final com baselines não informativos. Em um problema desbalanceado como o presente (prevalência positiva de 24,5% na janela de 30 dias), a acurácia isolada é métrica enganosa: um *baseline* ingênuo que sempre prediz a classe majoritária (não-evento) já atingiria 75,5% de acurácia, sem qualquer capacidade discriminativa. Por isso, a superioridade do modelo é mais bem caracterizada por métricas sensíveis ao ordenamento e à calibração. O Random Forest atinge AUC-ROC de 0,855 contra o valor de 0,5 esperado de um classificador aleatório, e reduz o Brier Score em 27,6% em relação a um *baseline* que prevê a prevalência constante (0,134 contra 0,185, conforme Seção IV-B5). Esse conjunto de evidências caracteriza o modelo como substantivamente informativo em relação à ausência de modelo, mesmo quando o ganho em acurácia bruta sobre o baseline majoritário é modesto (4,7 pontos percentuais), reflexo esperado do desbalanceamento.

C. Resultados da Clusterização

1) Análise de Métricas Internas por K: Antes da decisão pela abordagem híbrida descrita na Seção III-D2, foram executados experimentos preliminares de K-Means sobre o conjunto completo de 322 colaboradores ativos, variando K entre 2 e 8. A Figura 5 apresenta as curvas de inércia (método do cotovelo) e *Silhouette* médio.

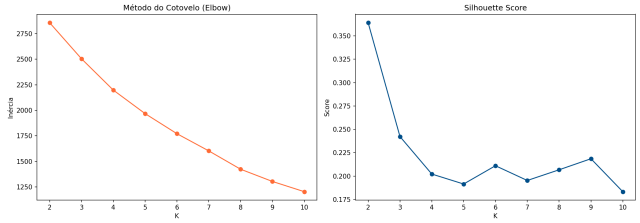

Figura 5. Curvas de inércia (cotovelo) e *Silhouette* médio para $K \in \{2, \ldots, 8\}$ em K-Means aplicado ao conjunto completo de 322 colaboradores ativos.

Duas observações orientam a escolha do K. Primeiro, a curva de inércia apresenta inflexão visível em $K = 3$ e $K = 4$, sugerindo que adições além desse ponto reduzem apenas marginalmente a soma das distâncias intra-cluster. Segundo, as três métricas internas apresentam comportamento consistente com a presença de *outliers* (colaboradores com históricos atípicos), como detalha a Tabela X. O *Silhouette* médio apresenta valor modesto em toda a faixa testada (entre 0,237 e 0,300), oscilando em direção crescente para valores altos de K apenas porque o algoritmo isola os *outliers* em microclusters homogêneos internamente, artefato que não corresponde a segmentação gerencial útil.

Três aspectos da Tabela X merecem destaque. Primeiro, $K = 3$ apresenta máximo local de *Silhouette* (0,270) na faixa interpretativamente viável ($K \in \{2, 3, 4\}$), com queda em $K = 4$ (0,244) e recuperação apenas quando K passa a isolar *outliers* em microclusters. Segundo, Davies-Bouldin reduz-se substancialmente de $K = 2$ (1,833) para $K = 3$ (1,269), indicando separação significativamente melhor, com ganhos marginais apenas a partir de $K = 5$. Terceiro, a tendência de crescimento conjunto de Silhouette e Calinski-Harabasz para $K \geq 6$ é sintoma da fragmentação artificial associada a *outliers* de gravidade alta, conforme já havia sido observado em experimentos preliminares.

Com base nessa análise, e considerando a necessidade de isolamento explícito do perfil Grave, adotou-se a aborda-

Tabela IX
DESEMPENHO DO RANDOM FOREST ($d = 5$) POR JANELA-ALVO, CONJUNTO DE VALIDAÇÃO *out-of-time* (REF. t_0 = 31/12/2024).

Janela	Prev. pos.	AUC	Acurácia	Precision	Recall	F1	Brier
7 dias	5,0%	0,712	94,3%	–	–	–	0,048
30 dias	24,5%	0,857	79,5%	57,7%	61,0%	0,593	0,134
90 dias	37,6%	0,920	84,3%	77,4%	82,4%	0,798	0,118

Tabela X
MÉTRICAS INTERNAS DO K-MEANS APLICADO AO CONJUNTO COMPLETO DE 322 COLABORADORES ATIVOS, PARA $K \in \{2, \ldots, 8\}$.

K	Silhouette	Calinski-Harabasz	Davies-Bouldin
2	0,237	55,6	1,833
3	0,270	65,0	1,269
4	0,244	71,8	1,253
5	0,258	74,4	1,058
6	0,256	76,8	1,017
7	0,290	84,9	0,939
8	0,300	88,7	0,981

gem híbrida (Seção III-D2): isolamento determinístico de 10 colaboradores com *media_dias_afastamento* superior a 5 e *prop_eventos_graves* superior a 10%, seguida de K-Means com $K = 3$ nos demais 312 colaboradores. Essa estratégia eleva o *Silhouette* médio dos três clusters principais para 0,301, valor considerado aceitável para dados comportamentais humanos [36], que raramente apresentam a separabilidade rígida observada em dados sintéticos ou físicos.

2) Estabilidade e Validação Estatística: A estabilidade da clusterização foi avaliada pelo *Adjusted Rand Index* (ARI) entre partições obtidas com 10 sementes aleatórias distintas para inicialização dos centroides. O ARI médio foi de 0,85 (desvio-padrão 0,04), indicando alta consistência da estrutura identificada frente a variações no procedimento de otimização. Valores de ARI acima de 0,80 são tipicamente interpretados como estabilidade robusta em aplicações de clusterização [39].

A significância estatística das diferenças entre clusters foi avaliada por meio de ANOVA unidirecional para cada uma das 9 *features* utilizadas na clusterização, testando a hipótese nula de igualdade de médias entre os grupos. A Tabela XI apresenta os valores de estatística F e p-valor obtidos.

Tabela XI
RESULTADOS DA ANOVA UNIDIRECIONAL PARA AS *features* DE CLUSTERIZAÇÃO.

Feature	Estatística F	p-valor
eventos_por_ano	142,8	$< 0{,}001$
media_dias_afastamento	187,5	$< 0{,}001$
prop_eventos_graves	312,4	$< 0{,}001$
tendência_30_90	38,7	$< 0{,}001$
densidade_recente	29,1	$< 0{,}001$
gap_medio_dias	76,2	$< 0{,}001$
índice_sazonalidade	54,6	$< 0{,}001$
consistência_sazonal	31,8	$< 0{,}001$
prop_estacao_pico	42,9	$< 0{,}001$

Todas as *features* apresentaram diferenças estatisticamente significativas entre os quatro clusters ($p < 0{,}001$), confirmando que os agrupamentos não são artefato do procedimento de otimização, mas refletem estrutura real dos dados. As maiores estatísticas F foram obtidas para *prop_eventos_graves* ($F = 312{,}4$), *media_dias_afastamento* ($F = 187{,}5$) e *eventos_por_ano* ($F = 142{,}8$), precisamente as dimensões que definem o desenho do estudo, reforçando a coerência entre construção metodológica e resultados obtidos.

3) Caracterização dos Quatro Perfis Identificados: A abordagem híbrida resultou em quatro perfis comportamentais distintos. A Tabela XII apresenta a caracterização comportamental e demográfica de cada grupo.

a) Perfil Estável (n=139).: Representa o padrão basal da organização, caracterizado por baixa frequência de absenteísmo (3,5 eventos/ano), eventos curtos (1,7 dias em média), baixa proporção de eventos graves (0,9%) e distribuição temporal relativamente equilibrada entre as estações. O *score* de risco médio para 30 dias é de 33,2%, configurando risco moderado.

b) Perfil Recorrente (n=108).: É o grupo que mais demanda atenção operacional no curto prazo. Caracteriza-se por alta frequência de eventos (13,0 eventos/ano, aproximadamente um a cada 28 dias), intervalos curtos entre eventos (*gap_medio_dias* de aproximadamente 20) e eventos predominantemente leves (1,6 dias em média). O *score* de risco médio atinge 72,3%, o mais elevado entre os perfis ativos. A composição demográfica mostra predominância feminina (65%) e praticamente ausência de cargos de liderança (4%), com forte presença de atestados por consulta médica e acompanhamento de filhos. A literatura sobre divisão de responsabilidades familiares [16] e sobre padrões de cuidado [15] oferece *hipótese interpretativa plausível* para esse perfil; no entanto, o presente estudo não realiza inferência causal entre gênero, papéis de cuidado e absenteísmo, de modo que as associações demográficas devem ser lidas como descritivas, não explicativas.

c) Perfil Sazonal (n=65).: Corresponde a colaboradores com concentração significativa de eventos em uma estação específica do ano, com consistência multi-anual. A *prop_estacao_pico* média é de 51,5%, e a *consistência_sazonal* média é de 0,83, ou seja, em 83% dos anos com eventos, o colaborador apresentou ausência na sua estação de pico. A distribuição entre estações mostrou predomínio da Primavera (25 colaboradores, 38%), seguida por Outono (16, 25%), Verão (13, 20%) e Inverno (11, 17%). O *score* médio é de 64,6%, elevado, embora menor que o do perfil Recorrente. A presença deste perfil valida empiricamente a relevância das *features* sazonais propostas: sem elas, esses colaboradores se-

Tabela XII
CARACTERIZAÇÃO DOS QUATRO PERFIS IDENTIFICADOS. VALORES SÃO MÉDIAS POR GRUPO. PERCENTUAIS DEMOGRÁFICOS DO PERFIL GRAVE ($n = 10$) DEVEM SER LIDOS COM CAUTELA, DADA A AMOSTRA REDUZIDA.

Perfil	**N (%)**	**Ev/ano**	**Dias/Ev**	**% Graves**	**Score 30d**	**Idade**	**% Fem.**	**% Liderança**
Estável	139 (43,2%)	3,5	1,7	0,9%	33,2%	41,8	44%	15%
Recorrente	108 (33,5%)	13,0	1,6	4,1%	72,3%	42,1	65%	4%
Sazonal	65 (20,2%)	4,6	1,5	4,8%	64,6%	43,5	38%	22%
Grave	10 (3,1%)	1,0	8,0	65,2%	11,3%	46,3	70%	60%

riam absorvidos pelos perfis Estável ou Recorrente, perdendo-se a dimensão temporal estrutural de seu comportamento.

d) Perfil Grave (n=10).: Constitui o grupo menor e mais crítico do ponto de vista clínico, caracterizado por eventos de longa duração (8,0 dias em média), alta proporção de eventos graves (65,2%) e baixa frequência total (1,0 evento/ano). Apesar do *score* médio aparentemente baixo (11,3%), que reflete a baixa probabilidade de evento em janela curta de 30 dias, dada a raridade temporal dos episódios, este perfil demanda acompanhamento clínico específico e não deve ser interpretado como de baixo risco em termos absolutos. Cabe uma ressalva amostral: com apenas dez observações, os percentuais demográficos deste perfil apresentam intervalo de confiança amplo e devem ser lidos como indicativos, não como estimativas estáveis. A presença elevada de cargos de liderança (60%) *sugere* possível associação entre pressão de responsabilidade e condições de saúde mais complexas, hipótese descritiva que merece investigação futura com desenho apropriado e amostra maior.

4) Visualização dos Clusters: A Figura 6 apresenta a projeção dos 322 colaboradores em duas dimensões principais obtidas por Análise de Componentes Principais (PCA) aplicada às 9 *features* de clusterização. Ainda que PCA inevitavelmente preserve apenas parte da variância total, a visualização evidencia a separabilidade dos quatro perfis.

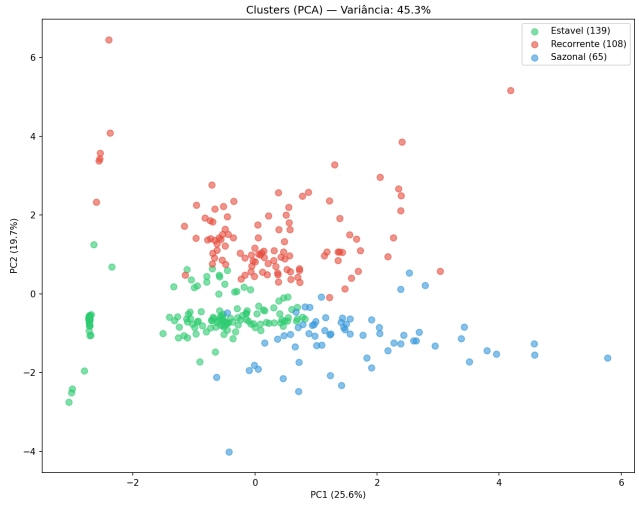

Figura 6. Projeção bidimensional dos 322 colaboradores ativos via PCA, com coloração pelo cluster atribuído.

Complementarmente, a Figura 7 apresenta um *heatmap* dos valores médios padronizados (z-score) das *features* por cluster, permitindo identificar rapidamente os atributos que mais diferenciam cada grupo.

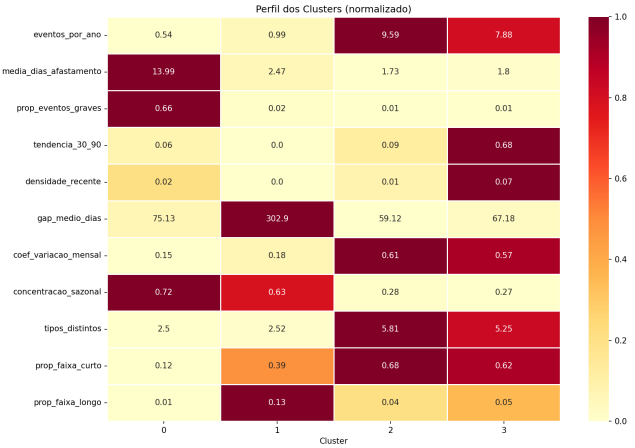

Figura 7. *Heatmap* de valores médios padronizados das *features* por cluster. Tons quentes indicam valores acima da média populacional; tons frios, abaixo.

O *heatmap* torna explícito o perfil distintivo de cada grupo: o perfil Recorrente apresenta valores elevados em *eventos_por_ano*, *densidade_recente* e *tendência_30_90*; o perfil Sazonal se destaca em *índice_sazonalidade*, *consistência_sazonal* e *prop_estacao_pico*; o perfil Grave em *media_dias_afastamento* e *prop_eventos_graves*; e o perfil Estável apresenta valores próximos da média em todas as dimensões, configurando-se como grupo de referência.

D. Integração dos Resultados ao Ambiente Operacional

Para cumprir a fase de *Deployment* do CRISP-DM, os resultados foram consolidados em uma base unificada e integrados a um painel interativo em Power BI. A base final contém, para cada um dos 322 colaboradores ativos, os campos: *CHAPA*, *score_7d*, *score_30d*, *score_90d*, *faixa_risco_30d* (Baixo, Moderado, Alto, Muito Alto, Crítico), *rank_risco* (posição ordinal), *cluster*, *nome_perfil*, e atributos demográficos. O painel está estruturado em três páginas: (a) *Visão Geral*, com indicadores agregados e filtros por diretoria; (b) *Perfil do Colaborador*, com visualização individualizada; e (c) *Visão por Perfil*, com detalhamento por cluster. A entrega do painel completa o ciclo técnico do CRISP-DM; a incorporação efetiva do artefato à rotina decisória da liderança de RH, com uso regular do *score* e dos perfis em ações operacionais, constitui

etapa posterior de adoção organizacional, fora do escopo desta versão do estudo.

V. Discussão

Os resultados apresentados na Seção IV permitem três blocos de discussão: (i) o retorno às hipóteses estabelecidas na Seção I e a verificação formal de cada uma delas; (ii) as implicações operacionais e gerenciais da abordagem integrada, em especial o uso combinado de *score* e perfil comportamental como instrumento de priorização; e (iii) as limitações metodológicas e ameaças à validade do estudo, com identificação explícita dos pontos que podem ser aprimorados em trabalhos futuros.

A. Verificação das Hipóteses

A abordagem metodológica adotada permitiu submeter cada uma das quatro hipóteses a verificação empírica. O diagnóstico consolidado é apresentado a seguir.

1) H1 – Hipótese Preditiva: A hipótese de que seria possível estimar, a partir de dados históricos de frequência, a probabilidade futura de eventos de absenteísmo com acurácia superior a 70% foi **confirmada**. O modelo Random Forest selecionado atingiu acurácia de 80,2% e AUC-ROC de 0,855 no conjunto de validação temporal, superando em 10 pontos percentuais o limiar de acurácia estabelecido e situando-se muito acima do AUC de 0,5 de um classificador sem capacidade discriminativa. Adicionalmente, 13 dos 14 modelos candidatos superaram o limiar de 70% de acurácia (apenas a árvore de decisão isolada ficou ligeiramente abaixo, com 69,8%), o que indica que o sinal preditivo não é artefato de uma escolha algorítmica específica, mas está estruturalmente presente nas *features* construídas.

A robustez do resultado é sustentada pelo baixo desvio-padrão em *cross-validation* (0,047 no AUC), pela consistência entre as métricas de treino e validação (diferencial de aproximadamente 10 pontos percentuais entre *CV AUC* e *Val AUC*) e pela análise de curvas ROC e Precision-Recall (áreas sob as curvas de 0,855 e 0,621, respectivamente). Em termos de ordem de grandeza, o AUC de 0,855 obtido para a janela de 30 dias situa-se dentro da faixa reportada por Llamas Blázquez [20] em estudo análogo conduzido sob regime mais permissivo de dados (AUC entre 0,78 e 0,88 conforme o horizonte temporal avaliado) e supera, em sofisticação metodológica e rigor de validação, abordagens exploratórias iniciais conduzidas sobre o *dataset* público *Absenteeism at Work* [18], tipicamente baseadas em poucos atributos e em modelos de regressão ou classificação direta [19]. A análise complementar por janela (Tabela IX) reforça esse alinhamento: o AUC atinge 0,920 para 90 dias e 0,712 para 7 dias, padrão coerente com a dificuldade crescente de prever eventos raros em horizontes curtos.

2) H2 – Hipótese Comportamental: A hipótese de existência de padrões distintos e estatisticamente significativos de comportamento de frequência foi **confirmada**. A abordagem híbrida de clusterização identificou quatro perfis comportamentais distintos, com composição e características sumarizadas na Tabela XII. Todas as nove *features* utilizadas apresentaram diferenças estatisticamente significativas entre clusters ($p < 0{,}001$, ANOVA unidirecional, Tabela XI), sendo que as estatísticas F mais elevadas foram obtidas exatamente nas dimensões conceituais que motivaram o desenho do estudo (*prop_eventos_graves*, *media_dias_afastamento*, *eventos_por_ano*).

A estabilidade da solução, verificada pelo *Adjusted Rand Index* de 0,85 em 10 inicializações aleatórias, confirma que os perfis identificados não são artefato do processo de otimização. Adicionalmente, a validação interpretativa conduzida com a área de RH da organização parceira corroborou a correspondência entre os perfis algorítmicos e os padrões empiricamente observados pela gestão, satisfazendo o critério qualitativo de interpretabilidade definido na fase de *Business Understanding*. A identificação explícita de um perfil Sazonal estatisticamente validado constitui, até onde nossa revisão alcançou, contribuição não reportada em trabalhos aplicados anteriores sobre absenteísmo, que tipicamente modelam o fenômeno apenas como problema de classificação binária, sem segmentação comportamental distinta [20], [22].

3) H3 – Hipótese de Valor Operacional: A hipótese de valor operacional foi **parcialmente confirmada apenas em sua dimensão de viabilidade**: o artefato foi construído, integra-se a ferramenta de visualização já utilizada pela organização parceira (Power BI) e estrutura uma priorização plausível quando combinado ao perfil comportamental. Por outro lado, a dimensão de *valor operacional efetivamente comprovado em campo* não foi demonstrada neste estudo. O painel encontra-se, no momento da submissão, em estado de protótipo entregue à área de Recursos Humanos, ainda não implantado em rotina decisória real, e a validação longitudinal de impacto (acompanhamento do uso do *score* pela liderança, medição de convergência entre priorização algorítmica e avaliação gerencial qualitativa, e eventual aferição de efeitos sobre indicadores agregados de absenteísmo) constitui primeiro marco de trabalho futuro. O argumento de confirmação parcial, portanto, apoia-se apenas em três observações indiretas de viabilidade.

Primeiro, os dois artefatos analíticos respondem a perguntas gerenciais distintas e complementares: o *score* responde a *quem* demanda atenção prioritária no curto prazo, enquanto o cluster responde a *por que* demanda, do ponto de vista de padrão comportamental. A Tabela XIII ilustra como a combinação gera categorias de priorização operacional.

Tabela XIII
Matriz de priorização operacional combinando perfil e *score* de risco.

Perfil	Score alto	Score baixo
Recorrente	Intervenção imediata	Monitoramento
Sazonal	Preparação sazonal	Observação
Grave	Acompanhamento clínico	Acompanhamento clínico
Estável	Avaliação pontual	Monitoramento padrão

Segundo, a gestão prescritiva diferenciada por perfil é coerente com a literatura de gestão de pessoas baseada em

evidências [5], [12], que preconiza intervenções customizadas em vez de políticas homogêneas. Terceiro, a integração ao painel Power BI, conforme descrito na Seção III-F, torna tecnicamente viável o uso combinado dos dois artefatos na rotina gerencial, condição necessária (ainda que não suficiente) para a realização prática do valor analítico.

Em síntese, o escopo desta versão do estudo permite delimitar com precisão o que foi e o que não foi demonstrado em relação a H3. Foram demonstrados: (i) a viabilidade técnica do artefato integrado (score mais perfil); (ii) a coerência estrutural entre a priorização algorítmica proposta e práticas recomendadas na literatura de gestão de pessoas baseada em evidências; e (iii) a integração a ferramenta de visualização já utilizada pela organização parceira. Não foram demonstrados, por outro lado: (i) a adoção efetiva do artefato em rotina decisória real, uma vez que o painel se encontra apenas entregue como protótipo; (ii) o impacto mensurável sobre indicadores agregados de absenteísmo em horizonte relevante; e (iii) a efetividade comparada das intervenções propostas (flexibilização de jornada, *telemedicina*, ações sazonais) frente a alternativas uniformes. A demonstração dessas três dimensões constitui agenda de pesquisa empírica subsequente.

4) H4 – Hipótese de Interpretabilidade: A hipótese de que modelos baseados em estruturas explicáveis poderiam atingir o patamar de qualidade exigido foi **confirmada**. O modelo final selecionado (Random Forest) oferece três níveis complementares de interpretabilidade. No nível agregado, a métrica de importância de variáveis via *permutation importance* (Figura 2) identifica as *features* mais relevantes para as decisões do *ensemble*, com vantagem metodológica relevante: por medir a queda efetiva de desempenho quando cada *feature* é embaralhada, essa técnica evita os vieses conhecidos de métricas baseadas em impureza em favor de variáveis com alta cardinalidade [26]. No nível do modelo, cada árvore individual pode ser inspecionada por inteiro, compreendendo condições de divisão e predições terminais. No nível individual, técnicas como SHAP (*Shapley Additive Explanations*) podem ser aplicadas para explicar cada predição específica, ainda que não tenham sido incorporadas ao painel na versão atual, configurando direção de trabalho futuro.

A clusterização também preserva interpretabilidade: os quatro perfis identificados correspondem a padrões reconhecíveis pela gestão, e sua caracterização via *heatmap* (Figura 7) torna imediatas as dimensões distintivas de cada grupo. Esse conjunto de mecanismos atende ao requisito de explicabilidade por construção estabelecido nas restrições do projeto (Seção III-A).

B. Implicações Operacionais e Gerenciais

Além da verificação das hipóteses, os resultados obtidos apontam três implicações operacionais potencialmente relevantes para a área de RH e que podem, por extensão, orientar implementações em organizações similares. É importante sublinhar, contudo, que essas implicações são apresentadas aqui como *propostas derivadas dos achados descritivos*, não como benefícios empiricamente observados em campo. A aferição de efetividade dessas propostas depende de estudo longitudinal controlado, ainda não conduzido.

Primeira implicação: segmentação como base para intervenções diferenciadas. A constatação de que aproximadamente um terço da população (33,5%) pertence ao perfil Recorrente, com score médio de 72,3%, sugere que intervenções de flexibilização de jornada, *home office* para consultas médicas e *telemedicina podem* apresentar impacto agregado relevante, uma vez que boa parte dos eventos desse perfil relaciona-se a consultas de curta duração. Intervenções gerais de redução de absenteísmo, sem segmentação, *plausivelmente* diluiriam esforços por tratarem populações heterogêneas como homogêneas.

Segunda implicação: reconhecimento de sazonalidade como dimensão autônoma. A existência de um perfil Sazonal (20,2% da população) com consistência multi-anual *sugere* que parte do absenteísmo segue padrões previsíveis no calendário anual. Esse achado *pode orientar* ações preventivas antecipadas, como campanhas de vacinação direcionadas, orientações de autocuidado em períodos de pico e dimensionamento ajustado de equipes em períodos sensíveis, sendo que a efetividade concreta de tais ações depende de avaliação empírica específica. A distribuição entre estações observada neste estudo (Primavera 38%, Outono 25%, Verão 20%, Inverno 17%) é particularmente compatível com o contexto climático do Distrito Federal, região sede da organização parceira, em que os meses de setembro e outubro marcam o final prolongado da estação seca (com umidade relativa do ar atingindo níveis extremos, historicamente inferiores a 20%) seguido pelo retorno das chuvas. Essa transição abrupta cria condições favoráveis tanto à irritação mecânica das vias respiratórias pelo ar seco quanto à rápida proliferação de fungos anemófilos (Aspergillus, Penicillium, Cladosporium) associados a rinite, asma e sinusite em populações brasileiras [46]. Esse substrato climático-epidemiológico oferece hipótese interpretativa plausível para a predominância da Primavera como estação de pico do perfil Sazonal, embora a atribuição causal específica aos eventos de ausência observados neste estudo permaneça como investigação futura. Complementarmente, ciclos organizacionais internos (encerramentos de projeto, retomadas pós-férias, períodos de fechamento orçamentário anual) podem atuar como fatores concorrentes.

Terceira implicação: identificação precoce de perfis críticos. Embora o perfil Grave represente apenas 3,1% da população, esses colaboradores concentram duração expressiva de afastamento por evento e *sugerem* demanda por acompanhamento clínico individualizado. A abordagem híbrida adotada permite isolamento explícito e monitoramento continuado desse grupo, evitando sua diluição em microclusters instáveis que dificultariam ação gerencial. A presença elevada de cargos de liderança nesse perfil (60%) *sugere* possível associação com pressão de responsabilidade ou com a concentração de colaboradores de maior faixa etária em posições seniores; a confirmação dessa leitura requer desenho causal específico, fora do escopo deste estudo.

C. Considerações Éticas sobre o Uso Operacional do Score

A aplicação de *People Analytics* a fenômenos que afetam diretamente a vida dos colaboradores exige cuidados éticos específicos que precedem qualquer mérito preditivo. Quatro salvaguardas foram adotadas neste trabalho e recomenda-se sua observância em replicações. Primeiro, o *score* e o perfil devem ser tratados como *insumo para priorização*, nunca como fundamento isolado para decisões disciplinares, avaliações de desempenho ou movimentações contratuais; sua função é apoiar a liderança e a área de RH a distribuir atenção de forma mais efetiva, não a substituir julgamento humano. Segundo, reforçando o princípio anterior, **é vedado o uso disciplinar automatizado do score**: qualquer ação gerencial individual derivada do artefato preditivo deve ser mediada por supervisão humana contínua, com análise caso a caso por profissionais qualificados da área de RH, e nunca por regras automáticas acionadas apenas pelo valor numérico do *score*. Terceiro, há risco real de estigmatização de colaboradores classificados em perfis de maior visibilidade (Recorrente ou Grave), especialmente quando a leitura dos resultados não é mediada por profissionais com preparo em saúde ocupacional; o acesso ao painel foi, por isso, restrito à equipe de RH, e a linguagem dos rótulos foi construída em conjunto com a gestão para evitar conotação punitiva. Quarto, a governança do artefato precisa contemplar direito de contestação: colaboradores devem poder questionar decisões cuja motivação invoque o *score*, o que reforça a escolha por modelos interpretáveis e a documentação pública da metodologia.

Privacidade e conformidade legal (LGPD). O tratamento de dados de frequência e saúde ocupacional insere-se no escopo da Lei Geral de Proteção de Dados (Lei n. 13.709/2018), que classifica dados de saúde como *dados pessoais sensíveis* (Art. 5º, II), sujeitos a base legal específica e a salvaguardas reforçadas. O projeto incorporou princípios de *Privacy by Design* desde a fase de preparação. Aplicou-se *minimização de dados* (Art. 6º, III), com remoção preventiva de nome, CPF, PIS, dados salariais e descrição clínica detalhada, preservando-se apenas o identificador funcional (CHAPA) como chave operacional — medida de *pseudonimização*, e não de anonimização, de modo que os dados permanecem pessoais para os fins da LGPD e sujeitos às suas salvaguardas. Restringiu-se a *finalidade* do tratamento ao apoio gerencial à saúde e à frequência, sem reúso para finalidades diversas, e limitou-se o acesso ao painel à equipe de Recursos Humanos, em linha com o princípio de necessidade. Recomenda-se, em replicações, formalizar a base legal aplicável, registrar a operação no inventário de tratamento e, quando cabível, conduzir Relatório de Impacto à Proteção de Dados Pessoais (RIPD), de modo a sustentar a conformidade ao longo do ciclo de vida do artefato.

Vieses e equidade. A composição demográfica dos perfis (Tabela XII) revela concentrações relevantes, em especial a predominância feminina (65%) no perfil Recorrente, de mais alto *score* médio (72,3%). Tal concentração é coerente com padrões de divisão de responsabilidades de cuidado documentados na literatura [16], mas impõe o risco de que o *score* reproduza ou amplifique desigualdades preexistentes caso seja utilizado sem mediação humana. Como salvaguarda estrutural, nenhuma variável demográfica sensível (gênero, idade, diretoria) compõe o vetor preditivo do modelo, que se apoia exclusivamente em *features* comportamentais de frequência; analogamente, a clusterização opera apenas sobre dimensões comportamentais, e os atributos demográficos da Tabela XII são reportados *a posteriori*, com função descritiva, não preditiva. Ainda assim, como *features* comportamentais podem correlacionar-se indiretamente com atributos protegidos, a exclusão dessas variáveis não garante, por si só, a ausência de viés. De fato, uma auditoria exploratória de equidade por gênero no conjunto de validação ($n = 306$ registros com gênero informado), ao limiar de decisão de 0,50, revelou diferença não desprezível na taxa de falsos positivos entre os subgrupos: 43,7% no grupo feminino contra 28,2% no masculino, uma diferença de 15,5 pontos percentuais, ainda que a qualidade de calibração se mantenha próxima entre os grupos (Brier de 0,203 e 0,217, respectivamente). Parte dessa diferença associa-se à maior prevalência de absenteísmo observada no grupo feminino na janela analisada (51,0% contra 31,7%), coerente com os padrões de divisão de responsabilidades de cuidado já mencionados; entretanto, o achado evidencia risco concreto de tratamento desigual caso o *score* seja aplicado sem mediação. Esse resultado reforça as salvaguardas adotadas — exclusão de variáveis demográficas do vetor preditivo e supervisão humana obrigatória — e torna prioritária uma auditoria sistemática de equidade, com amostra ampliada e eventual ajuste de limiar sensível ao contexto ou calibração por subgrupo, registrada como trabalho futuro.

D. Limitações e Ameaças à Validade

O estudo apresenta um conjunto de limitações que devem ser explicitadas, tanto para contextualizar os resultados quanto para orientar melhorias futuras.

Defasagem da base demográfica. A base de colaboradores utilizada para dados demográficos e contratuais é uma extração datada de 2023 e pode não refletir integralmente a composição atual do quadro. Em particular, colaboradores admitidos em 2024 ou 2025 estão representados parcialmente, com *features* demográficas ausentes. O impacto dessa limitação sobre o modelo preditivo foi mitigado pela disponibilidade integral dos dados de frequência, que são os insumos preditivos mais relevantes conforme demonstrado na análise de importância (Seção IV-B3). Sobre a clusterização, o impacto é menor, uma vez que os perfis são definidos por *features* comportamentais e não demográficas. Essa defasagem também sugere cautela na leitura das composições demográficas por perfil (Tabela XII), que devem ser entendidas como aproximações descritivas.

Desligamentos não identificados. A comparação entre a lista de empregados ativos (351 colaboradores em março/2026) e as CHAPAs observadas nas bases de atestados revelou a existência de 108 colaboradores com eventos em 2024 ou 2025 ausentes da lista atual, interpretados como desligamentos recentes. Essa informação não estava disponível de forma estruturada ao tempo do estudo, o que impede uma auditoria

definitiva do status desses colaboradores. Essa limitação foi parcialmente mitigada pela decisão metodológica de separar claramente os conjuntos de treino (676 CHAPAs, incluindo desligados) e de clusterização (322 CHAPAs ativas).

Ausência de variáveis externas. O modelo não incorpora variáveis que plausivelmente afetam absenteísmo, como indicadores epidemiológicos (surtos de influenza, dengue), condições climáticas (temperatura, umidade, eventos extremos) e eventos organizacionais não rotineiros (reorganizações, mudanças de liderança, períodos de sobrecarga). A exclusão foi deliberada, dada a restrição operacional de uso apenas de dados internos transacionais (Seção III-A), mas implica que os ganhos marginais disponíveis com tais dados externos permanecem não explorados.

Automação pendente. Na versão atual, a atualização dos *scores* depende de reexecução manual do *pipeline* (notebooks NB00 a NB04) com os dados atualizados. A automação completa, que envolveria integração direta com o banco transacional de RH, agendamento regular e atualização automática do painel, constitui direção de trabalho futuro e é pré-requisito para uso em regime diário.

Validação longitudinal de impacto. A confirmação da Hipótese 3 (valor operacional) apoia-se em argumentos indiretos de coerência metodológica e validação qualitativa com a gestão. Uma confirmação definitiva exigiria a condução de estudo longitudinal que mensurasse o impacto efetivo de intervenções orientadas pela segmentação comportamental sobre indicadores de absenteísmo, comparadas a intervenções uniformes de *controle*. Esse tipo de estudo escapa ao escopo metodológico deste trabalho, mas constitui avanço natural.

Generalização dos achados. O estudo foi conduzido em uma organização pública brasileira de promoção de exportações, com perfil específico de quadro funcional (predominância de cargos administrativos e analíticos, jornada padrão, política de atestados formalizada). A generalização dos perfis identificados para organizações com características estruturalmente diferentes (por exemplo, operações industriais, escalas de turno, regimes de *home office* extensivo) deve ser conduzida com cautela e idealmente precedida de replicação do método sobre dados locais. A metodologia, por sua vez, é mais diretamente transferível, dada a adesão explícita ao CRISP-DM e a documentação detalhada das decisões técnicas.

VI. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma abordagem integrada de *People Analytics* para a gestão preditiva e segmentada de absenteísmo em uma organização pública brasileira, conduzida sob o arcabouço metodológico CRISP-DM. A proposta combina duas frentes complementares: um *Score de Risco* baseado em Random Forest e uma clusterização comportamental híbrida, entregues de forma integrada em um painel operacional em Power BI. Nas subseções seguintes, são sintetizadas as contribuições efetivas do trabalho, os resultados de maior relevância e as direções para pesquisas futuras.

A. Síntese das Contribuições

Retomando as contribuições anunciadas na Subseção I-B, cinco entregas concretas emergem deste estudo. A *primeira* é um *pipeline* completo e reproduzível de engenharia de variáveis comportamentais para absenteísmo, composto por 57 *features* organizadas em sete categorias conceituais (volume, gravidade, recência, tendência, regularidade, sazonalidade e tipicidade). A separação conceitual dessas dimensões permite tanto a interpretação clara do comportamento de cada colaborador quanto a reutilização por outras organizações que disponham de dados transacionais similares.

A *segunda* contribuição, de natureza especificamente metodológica, é a proposta de *features* de sazonalidade baseadas em *consistência multi-anual* por estação calibradas ao calendário brasileiro. Diferentemente de medidas tradicionais de concentração trimestral, as *features* propostas exigem que o padrão se repita ao longo de múltiplos anos para ser classificado como sazonalidade verdadeira, evitando falsos positivos decorrentes de eventos pontuais e atípicos. Essa formulação mostrou-se empiricamente necessária para a identificação de um perfil sazonal bem definido (20,2% da população), que permaneceria oculto sob abordagens que utilizam apenas concentração trimestral.

A *terceira* contribuição é a comparação empírica rigorosa de 14 configurações de modelos supervisionados, com protocolo explícito de *cross-validation* estratificada, *split* temporal para simular uso em produção e critério de seleção baseado em F1-Score que considera o *custo assimétrico* entre falsos positivos e falsos negativos. O modelo Random Forest ($d = 5$) selecionado apresenta AUC-ROC de 0,855, acurácia de 80,2% e *recall* de 63,1%, situando-se em patamar comparável a estudos internacionais recentes [20], sob as restrições mais conservadoras de uso exclusivo de dados internos.

A *quarta* contribuição é uma abordagem híbrida para clusterização comportamental que combina critérios determinísticos para grupos minoritários de alta relevância clínica (perfil Grave) com aprendizado não supervisionado (K-Means com $K = 3$) para os demais colaboradores. Essa estratégia evita a fragmentação artificial observada em aplicações ingênuas de K-Means diante de *outliers* legítimos, problema conhecido na literatura [40]. A solução final apresenta *Silhouette* médio de 0,301, *Adjusted Rand Index* de 0,85 em 10 inicializações, e diferenças estatisticamente significativas ($p < 0{,}001$) em todas as *features* utilizadas.

A *quinta* contribuição é um estudo de caso detalhado, com descrição completa do contexto organizacional, das decisões metodológicas tomadas, dos *trade-offs* encontrados e das limitações identificadas. Diferentemente de contribuições predominantemente algorítmicas, esse relato aplicado pretende servir de referência para equipes de RH e cientistas de dados que pretendam conduzir projetos semelhantes em outras organizações, acelerando curva de aprendizado e evitando armadilhas recorrentes (por exemplo, o isolamento prematuro de *outliers* gravidade-alta em microclusters, a necessidade de *split* temporal para evitar vazamento, ou a importância da

separação entre universo de treino e universo operacional).

B. Principais Resultados e Implicações

Os principais resultados obtidos foram apresentados na Seção IV e discutidos na Seção V. Em síntese, três hipóteses foram integralmente confirmadas e uma foi confirmada apenas em sua dimensão de viabilidade: a predição de absenteísmo com base em dados históricos é possível e operacionalmente útil (H1); padrões comportamentais distintos podem ser identificados e validados estatisticamente (H2); a combinação de *score* e perfil estrutura uma priorização tecnicamente viável e integrável a ferramentas gerenciais (H3, dimensão de viabilidade, sendo o valor operacional em rotina decisória real ainda não demonstrado); e modelos interpretáveis são suficientes para atingir o patamar de qualidade exigido, sem recorrer a arquiteturas opacas (H4).

Situando esses resultados no estado da arte aplicado, três posicionamentos comparativos merecem destaque. Em desempenho preditivo, o AUC de 0,855 obtido para a janela de 30 dias é compatível com a faixa reportada por Llamas Blázquez [20] (0,78 a 0,88 conforme horizonte) e mais robusto que abordagens exploratórias iniciais conduzidas sobre o *dataset* público *Absenteeism at Work* [18], [19], ainda que o presente trabalho se restrinja a dados internos da organização parceira. Em estrutura de *features*, a predominância de variáveis de recência e atividade recente na análise de importância confirma empiricamente o efeito de hábito postulado por Steers e Rhodes [14] e corrobora o padrão encontrado por Llamas Blázquez [20]. Em contribuição diferenciada, a identificação explícita e estatisticamente validada de um perfil Sazonal constitui, até onde nossa revisão alcançou, aporte original do trabalho em relação à literatura aplicada, que tipicamente modela absenteísmo como classificação binária sem segmentação comportamental distinta.

Em termos práticos, a identificação de quatro perfis (Estáveis com 43,2%, Recorrentes com 33,5%, Sazonais com 20,2% e Graves com 3,1%) oferece à área de RH uma estrutura clara para desenho de intervenções diferenciadas. O perfil Recorrente, composto majoritariamente por mulheres em cargos operacionais e com forte incidência de eventos relacionados a consultas médicas e acompanhamento familiar, pode se beneficiar de políticas de flexibilização de jornada e expansão de *telemedicina*. O perfil Sazonal abre espaço para ações preventivas antecipadas, ajustadas ao ciclo anual de incidência de cada colaborador. O perfil Grave, ainda que pequeno, demanda acompanhamento clínico individualizado e articulação com políticas de saúde ocupacional.

A estrutura do painel Power BI foi desenhada a partir do princípio, preconizado por Rasmussen e Ulrich [5], de que o valor de *People Analytics* tende a se materializar quando os artefatos analíticos estão disponíveis para o cotidiano decisório da liderança. No caso deste estudo, o painel está estruturado em três páginas (Visão Geral, Perfil do Colaborador, Visão por Perfil), permitindo tanto a análise agregada quanto a consulta individual, com filtros por diretoria e faixa de risco. Vale explicitar, entretanto, que a versão atual consiste em protótipo entregue à área de Recursos Humanos, cuja adoção efetiva em rotina decisória e a aferição longitudinal de impacto sobre indicadores de absenteísmo permanecem como trabalho futuro.

C. Trabalhos Futuros

O trabalho abre diversas direções de pesquisa e desenvolvimento, que podem ser agrupadas em cinco blocos.

Incorporação de dados externos. A extensão mais imediata consiste em avaliar o ganho marginal de incorporar fontes externas de dados: indicadores epidemiológicos regionais (surtos de influenza, dengue, COVID-19), condições climáticas (temperatura média, umidade, eventos extremos) e *proxies* de clima organizacional (pesquisas de engajamento periódicas). A hipótese subjacente é que tais variáveis capturem fontes de variação atualmente residuais no modelo, potencialmente aprimorando tanto o *recall* quanto a calibração das probabilidades estimadas.

Modelagem multi-alvo, análise temporal detalhada e calibração. A formulação atual trata a predição como três tarefas binárias independentes (7, 30 e 90 dias), com a janela de 30 dias selecionada como horizonte-alvo principal de validação. Uma direção promissora é a adoção de modelagem multi-alvo, com estruturas que explorem a correlação entre as janelas, ou ainda a transição para modelagem de *time-to-event* (análise de sobrevivência), que oferece interpretação probabilística mais rica e aproveita integralmente a informação temporal. Complementarmente, a aplicação de métodos de calibração de probabilidades (regressão isotônica ou *Platt scaling*) aos scores estimados permitiria que os valores numéricos fossem interpretados não apenas como ranking, mas como estimativas probabilísticas diretamente comparáveis a frequências observadas, ampliando o leque de decisões operacionais apoiadas pelo modelo.

Comparação de estratégias de balanceamento. O tratamento do desbalanceamento adotado neste estudo — aprendizado sensível ao custo via *class_weight* — foi justificado na Seção III por preservar a calibração das probabilidades e por evitar a geração de exemplos sintéticos sem contrapartida temporal sob *split out-of-time*. Uma extensão pertinente é a comparação experimental sistemática dessa estratégia com técnicas de sobreamostragem sintética, como SMOTE e ADASYN, aplicadas estritamente dentro de cada partição de treino para não comprometer a validação temporal, aferindo o impacto conjunto sobre *recall*, precisão e calibração das probabilidades. Tal experimento permitiria quantificar empiricamente o *trade-off* aqui assumido em bases conceituais.

Técnicas de explicabilidade individual. A interpretabilidade agregada oferecida pelo Random Forest pode ser complementada por técnicas de explicabilidade individual, como SHAP (*Shapley Additive Explanations*) ou LIME (*Local Interpretable Model-agnostic Explanations*), que permitem ao gestor compreender, para cada colaborador específico, quais *features* contribuíram para o *score* atribuído. A incorporação dessas técnicas ao painel operacional amplia a confiança do usuário e a rastreabilidade decisória.

Automação e integração com sistemas transacionais. A atualização manual do *pipeline*, embora funcional no contexto atual do projeto, constitui barreira ao uso diário em regime de produção. A automação completa, envolvendo conexão direta à fonte transacional de RH, agendamento regular de execução, atualização automática do painel e notificações por *e-mail* para mudanças relevantes de faixa de risco, é requisito para que o sistema opere como ferramenta cotidiana da gestão.

Validação longitudinal de impacto e expansão para outros fenômenos. A demonstração definitiva do valor operacional da abordagem (Hipótese 3) requer estudo longitudinal em que intervenções orientadas pela segmentação sejam comparadas com intervenções uniformes ou grupos de controle, ao longo de horizonte suficiente para observação de mudanças em indicadores de absenteísmo. Finalmente, a metodologia proposta (CRISP-DM, engenharia de *features* comportamentais, seleção rigorosa de modelo e clusterização híbrida) é extensível a outros fenômenos organizacionais que compartilham natureza temporal e heterogeneidade comportamental, como *turnover*, engajamento, adesão a programas de desenvolvimento e desempenho, constituindo agenda natural de replicação.

D. Considerações Finais

O trabalho demonstrou que a transição de um modelo descritivo e reativo de gestão de frequência para um modelo preditivo e segmentado é viável com dados transacionais tipicamente disponíveis em sistemas de RH, sem recorrer a fontes externas ou arquiteturas opacas. O rigor metodológico adotado (CRISP-DM como arcabouço, comparação sistemática de modelos, *split* temporal, validação estatística de agrupamentos) foi deliberado e é apresentado com transparência suficiente para replicação. As limitações identificadas são explícitas e constituem, simultaneamente, agenda de evolução. Em termos de posicionamento comparativo, o estudo alcança desempenho preditivo compatível com a literatura recente [19], [20] sob restrições mais conservadoras de dados, confirma empiricamente a centralidade de variáveis de recência como família preditiva dominante [14], [20], e contribui com a identificação de um perfil Sazonal estatisticamente validado, aporte pouco explorado em trabalhos aplicados prévios. A integração ao ambiente operacional, via protótipo de painel Power BI entregue à área de Recursos Humanos, demonstra a viabilidade técnica de fechar o ciclo entre análise e ação, com potencial de expansão para outras frentes do ciclo de vida do colaborador uma vez completadas as etapas de adoção organizacional e validação longitudinal previstas como trabalho futuro.

REFERÊNCIAS

[1] D. A. Harrison and J. J. Martocchio, "Time for absenteeism: A 20-year review of origins, offshoots, and outcomes," *Journal of Management*, vol. 24, no. 3, pp. 305–350, 1998.

[2] I. Bierla, B. Huver, and S. Richard, "New evidence on absenteeism and presenteeism," *International Journal of Human Resource Management*, vol. 24, no. 7, pp. 1536–1550, 2013.

[3] S. Markussen, K. Røed, O. J. Røgeberg, and S. Gaure, "The anatomy of absenteeism," *Journal of Health Economics*, vol. 30, no. 2, pp. 277–292, 2011.

[4] J. H. Marler and J. W. Boudreau, "An evidence-based review of hr analytics," *International Journal of Human Resource Management*, vol. 28, no. 1, pp. 3–26, 2017.

[5] T. Rasmussen and D. Ulrich, "Learning from practice: How hr analytics avoids being a management fad," *Organizational Dynamics*, vol. 44, no. 3, pp. 236–242, 2015.

[6] A. Tursunbayeva, S. Di Lauro, and C. Pagliari, "People analytics: A scoping review of conceptual boundaries and value propositions," *International Journal of Information Management*, vol. 43, pp. 224–247, 2018.

[7] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc. / CRISP-DM Consortium, Tech. Rep., 2000. [Online]. Available: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

[8] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, UK, 2000, pp. 29–39. [Online]. Available: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

[9] T. H. Davenport, "Competing on analytics," *Harvard Business Review*, vol. 84, no. 1, pp. 98–107, 2006. [Online]. Available: https://hbr.org/2006/01/competing-on-analytics

[10] T. H. Davenport, J. Harris, and J. Shapiro, "Competing on talent analytics," *Harvard Business Review*, vol. 88, no. 10, pp. 52–58, 2010. [Online]. Available: https://hbr.org/2010/10/competing-on-talent-analytics

[11] M. A. Huselid, "The science and practice of workforce analytics: Introduction to the hrm special issue," *Human Resource Management*, vol. 57, no. 3, pp. 679–684, 2018.

[12] D. B. Minbaeva, "Building credible human capital analytics for organizational competitive advantage," *Human Resource Management*, vol. 57, no. 3, pp. 701–713, 2018.

[13] P. Van der Laken, J. W. Boudreau, and J. H. Marler, "Data-driven human resources analytics: A review and new research directions," *Personnel Review*, vol. 47, no. 5, pp. 991–1006, 2018.

[14] R. M. Steers and S. R. Rhodes, "Major influences on employee attendance: A process model," *Journal of Applied Psychology*, vol. 63, no. 4, pp. 391–407, 1978.

[15] J. J. Martocchio, "Age-related differences in employee absenteeism: A meta-analysis," *Psychology and Aging*, vol. 4, no. 4, pp. 409–414, 1989.

[16] M. Laaksonen, P. Martikainen, O. Rahkonen, and E. Lahelma, "Explanations for gender differences in sickness absence: Evidence from middle-aged municipal employees from Finland," *Occupational and Environmental Medicine*, vol. 65, no. 5, pp. 325–330, 2008.

[17] P. Allebeck and A. Mastekaasa, "Risk factors for sick leave: General studies," *Scandinavian Journal of Public Health*, vol. 32, no. 5 suppl, pp. 49–108, 2004.

[18] A. Martiniano, R. P. Ferreira, R. J. Sassi, and C. Affonso, "Application of a neuro fuzzy network in prediction of absenteeism at work," in *7th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2012, pp. 1–4, conjunto de dados *Absenteeism at Work* disponível no UCI Machine Learning Repository, DOI: 10.24432/C5X882.

[19] K. Tewari, S. Vandita, and S. Jain, "Predictive analysis of absenteeism in MNCs using machine learning algorithm," in *Proceedings of ICRIC 2019*, ser. Lecture Notes in Electrical Engineering, P. K. Singh, A. K. Kar, Y. Singh, M. H. Kolekar, and S. Tanwar, Eds. Springer, Cham, 2020, vol. 597, pp. 3–14.

[20] P. Llamas Blázquez, "Predicting workplace absenteeism using machine learning: a pilot study in occupational health," *Journal of Occupational Medicine and Toxicology*, vol. 20, no. 38, 2025.

[21] R. Punnoose and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, no. 9, pp. 22–26, 2016.

[22] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, p. 86, 2020.

[23] C. Schröer, F. Kruse, and J. M. Gomez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[24] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013.

[25] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[26] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.
[27] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[28] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[29] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
[30] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[31] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015.
[32] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
[33] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, 1995, pp. 1137–1143. [Online]. Available: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf
[34] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967. [Online]. Available: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992
[35] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
[36] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
[37] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.
[38] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
[39] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[40] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York: Springer, 2013.
[41] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, CA: O’Reilly Media, 2018. [Online]. Available: https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/
[42] H. Blockeel and L. De Raedt, “Top-down induction of first-order logical decision trees,” *Artificial Intelligence*, vol. 101, no. 1–2, pp. 285–297, 1998.
[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://www.jmlr.org/papers/v12/pedregosa11a.html
[44] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,” *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
[45] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
[46] M. B. Suehara and M. C. P. d. Silva, “Prevalence of airborne fungi in Brazil and correlations with respiratory diseases and fungal infections,” *Ciência & Saúde Coletiva*, vol. 28, no. 11, pp. 3289–3300, 2023.