**UNIVERSIDADE FEDERAL DE JUIZ DE FORA**
**INSTITUTO DE CIÊNCIAS EXATAS / FACULDADE DE ENGENHARIA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL**

**Túlio Valadão dos Santos**

# Métodos de Ciência de Dados em *People Analytics* para Previsão de *Turnover*

Juiz de Fora
2026

**Túlio Valadão dos Santos**

# Métodos de Ciência de Dados em *People Analytics* para Previsão de *Turnover*

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Orientadora: Profa. Dra. Priscila Vanessa Zabala Capriles Goliatt

Juiz de Fora

2026

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Santos, Túlio Valadão dos.
Métodos de Ciência de Dados em *People Analytics* para Previsão de *Turnover* / Túlio Valadão dos Santos. – 2026.
95 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Goliatt
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas / Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2026.

1. Turnover. 2. Ciência de Dados. 3. People Analytics. 4. Desbalanceamento de Classes. 5. Seleção de Atributos. I. Capriles, Priscila, orient. II. Título.

**Túlio Valadão dos Santos**

**Métodos de Ciência de Dados em People Analytics para Previsão de Turnover**

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 10 de setembro de 2026.

BANCA EXAMINADORA

**Prof.ª Dr.ª. Priscila Vanessa Zabala Capriles Goliatt** - Orientadora

Universidade Federal de Juiz de Fora

**Prof. Dr. Leonardo Goliatt da Fonseca**

Universidade Federal de Juiz de Fora

**Prof.ª Dr.ª. Mônica Carvalho Alves Cappelle**

Universidade Federal de Lavras

Juiz de Fora, 01/09/2026.

Documento assinado eletronicamente por **Priscila Vanessa Zabala Capriles Goliatt**, **Professor(a)**, em 13/09/2026, às 10:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 14/09/2026, às 08:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **MONICA CARVALHO ALVES CAPPELLE**, **Usuário Externo**, em 18/09/2026, às 17:19, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **3188666** e o código CRC **7658AECD**.

## AGRADECIMENTOS

Gostaria de agradecer a todos que fizeram parte, direta ou indiretamente, da minha caminhada no mestrado e da elaboração deste trabalho.

Agradeço, primeiramente, a Deus, pela oportunidade de ter vivido esse programa de pós-graduação, pela saúde, pela força e por me guiar por esse caminho.

À minha professora e orientadora, Priscila, que muito me ensinou e ajudou, tanto nas disciplinas cursadas ao longo da formação quanto no desenvolvimento deste trabalho, o qual não seria possível sem sua presença. Muito obrigado pelo apoio, paciência e confiança.

Agradeço à minha família, à minha namorada e aos meus amigos, que sempre estiveram presentes ao longo desta jornada, nos momentos felizes e nos mais difíceis, oferecendo bastante suporte e confiança.

Por fim, obrigado à UFJF e aos ótimos professores com quem tive a honra de aprender, tanto na graduação quanto no mestrado. Foi muito bom conhecer e trabalhar com vocês. Ao todo, foram 8 anos na universidade, que me deixaram admirado com esse ambiente! Que estrutura e ensino de altíssimas qualidades.

**RESUMO**

Esta pesquisa teve como objetivo avaliar e comparar o desempenho de diferentes modelos computacionais aplicados a três bases públicas de dados de funcionários para a previsão de *turnover*, considerando estratégias de tratamento do desbalanceamento de classes, diferentes representações dos atributos e análises estratificadas por sexo. Foram avaliados seis classificadores, uma condição sem reamostragem e as técnicas SMOTE, ADASYN, SMOTE-Tomek e SMOTE-ENN. Para cada população e representação experimental, foram realizadas 30 divisões estratificadas em 80% para treinamento e 20% para teste, com sementes distintas, mantendo as transformações, a seleção de atributos do Cenário 2 e o balanceamento restritos ao conjunto de treinamento. O experimento totalizou 24.300 execuções. O desempenho foi analisado por acurácia, precisão, *recall*, F1-score, MCC, AUC-ROC e *Log Loss*, e o Índice Multivariado Ponderado foi utilizado para ordenar as configurações dentro de cada contexto e examinar sua estabilidade relativa ao longo das repetições. Nas populações completas do Cenário 1, as configurações de referência foram LDA com a condição Desbalanceado no *Dataset* 1, XGBoost com SMOTE-Tomek no *Dataset* 2 e AdaBoost com SMOTE-ENN no *Dataset* 3. Dentre essas configurações, o *Dataset* 2 apresentou os maiores valores absolutos observados de F1-score, MCC e AUC-ROC e maior estabilidade relativa; o *Dataset* 1 apresentou desempenho intermediário e maior alternância; e o *Dataset* 3 permaneceu com MCC próximo de zero e AUC-ROC próxima de 0,5, de modo que sua primeira posição interna não representou desempenho absoluto elevado. As técnicas de reamostragem aumentaram principalmente o *recall* e o F1-score da classe positiva, sem melhora uniforme das demais métricas. A seleção automática de atributos realizada no treinamento não apresentou melhora média consistente em relação à representação completa, e os efeitos das representações reduzidas dependeram da base e da representação. Nenhum subgrupo por sexo apresentou desempenho consistentemente superior. O SHAP foi aplicado a uma execução representativa de cada configuração de referência do Cenário 1, e não à média de 30 execuções SHAP. Os principais atributos foram `OverTime` e `JobRole` no *Dataset* 1; `satisfaction_level`, `number_project` e `time_spend_company` no *Dataset* 2; e `Number_of_Companies_Worked`, `Relationship_with_Manager` e `Job_Involvement` no *Dataset* 3, estes últimos sob cautela adicional devido à baixa discriminação do modelo. As contribuições SHAP descrevem o comportamento preditivo e não estabelecem causalidade. Conclui-se que não existe modelo ou estratégia universalmente superior e que a estabilidade interna observada não substitui validação externa em organizações reais.

Palavras-chave: *turnover*; ciência de dados; *people analytics*; desbalanceamento de classes; seleção de atributos.

**ABSTRACT**

This research aimed to evaluate and compare the performance of different computational models applied to three public employee datasets for turnover prediction, considering class imbalance treatment strategies, different feature representations, and analyses stratified by sex. Six classifiers, one condition without resampling, and the SMOTE, ADASYN, SMOTE-Tomek, and SMOTE-ENN techniques were evaluated. For each population and experimental representation, 30 stratified 80% training and 20% testing splits were performed with different seeds, while transformations, Scenario 2 feature selection, and class balancing were restricted to the training set. The experiment comprised 24,300 model runs. Performance was assessed through accuracy, precision, recall, F1-score, MCC, AUC-ROC, and Log Loss, and the Weighted Multivariate Index was used to rank configurations within each context and examine their relative stability across repetitions. For the complete populations in Scenario 1, the reference configurations were LDA under the Imbalanced condition for Dataset 1, XGBoost with SMOTE-Tomek for Dataset 2, and AdaBoost with SMOTE-ENN for Dataset 3. Among these configurations, Dataset 2 achieved the highest observed absolute F1-score, MCC, and AUC-ROC values and greater relative stability; Dataset 1 showed intermediate performance and greater alternation among configurations; and Dataset 3 remained close to zero in MCC and close to 0.5 in AUC-ROC, so its first internal position did not indicate high absolute performance. Resampling mainly increased recall and the positive-class F1-score, without uniformly improving the other metrics. Feature selection performed on the training data did not show a consistent average improvement over the complete representation, and the effects of the reduced representations depended on the dataset and representation. Neither sex-based subgroup consistently outperformed the other. SHAP was applied to one representative run of each Scenario 1 reference configuration, rather than to an average of 30 SHAP runs. The main attributes were `OverTime` and `JobRole` in Dataset 1; `satisfaction_level`, `number_project`, and `time_spend_company` in Dataset 2; and `Number_of_Companies_Worked`, `Relationship_with_Manager`, and `Job_Involvement` in Dataset 3, with additional caution in the latter because of the model's low discriminative ability. SHAP contributions describe predictive behavior and do not establish causality. The results do not support a universally superior model or strategy, and the observed internal stability does not replace external validation in real organizations.

Keywords: turnover; data science; people analytics; class imbalance; feature selection.

## LISTA DE ILUSTRAÇÕES

Figura 1 – Fluxograma do processo de seleção dos estudos da revisão sistemática conforme o PRISMA 2020. . . . . 21
Figura 2 – Fluxograma do processo experimental e *pipeline* de modelagem preditiva de *turnover*. . . . . 39
Figura 3 – Distribuição dos valores SHAP das *features* transformadas no *Dataset* 1 para LDA com a condição Desbalanceado, semente 54. . . . . 68
Figura 4 – Distribuição dos valores SHAP das *features* transformadas no *Dataset* 2 para XGBoost com SMOTE-Tomek, semente 58. . . . . 69
Figura 5 – Distribuição dos valores SHAP das *features* transformadas no *Dataset* 3 para AdaBoost com SMOTE-ENN, semente 45. . . . . 70

## LISTA DE TABELAS


Tabela 1 – Síntese das respostas às questões secundárias da revisão sistemática. . 26
Tabela 2 – Comparação das vantagens e desvantagens dos métodos de balanceamento de dados utilizados neste estudo. . . . . 33
Tabela 3 – Caracterização estrutural do *Dataset* 1 em sua forma original. . . . . 41
Tabela 4 – Caracterização estrutural do *Dataset* 2 em sua forma original. . . . . 42
Tabela 5 – Caracterização estrutural do *Dataset* 3 em sua forma original. . . . . 42
Tabela 6 – Configurações e hiperparâmetros dos algoritmos de classificação. . . . . 48
Tabela 7 – Distribuição das classes nos conjuntos de treinamento dos *datasets* completos do Cenário 1 ao longo das 30 repetições. . . . . 54
Tabela 8 – Variação média pareada das métricas de teste após a aplicação das técnicas de balanceamento em relação à condição Desbalanceado nas 30 sementes. 55
Tabela 9 – Variação média pareada do F1-score da classe positiva por técnica de balanceamento e *dataset* nas populações completas. . . . . 56
Tabela 10 – Desempenho médio das configurações dos *datasets* completos no Cenário 1 considerando as 30 sementes. . . . . 58
Tabela 11 – Variações médias pareadas de desempenho entre as representações nas populações completas. . . . . 61
Tabela 12 – Desempenho médio e variabilidade dos modelos nos subgrupos por sexo ao longo das 30 sementes. . . . . 63
Tabela 13 – Comparação descritiva entre F1-score de treinamento e teste em configurações sem reamostragem externa no Cenário 1 ao longo das 30 sementes. . . 65
Tabela 14 – Atributos com maior contribuição global SHAP nas populações completas do Cenário 1. . . . . 66
Tabela 15 – Configurações de referência e robustez do Índice Multivariado nas populações completas. . . . . 72
Tabela 16 – Média dos componentes das matrizes de confusão das configurações de referência nas populações completas ao longo das 30 sementes. . . . . 72
Tabela 17 – Síntese temática do confronto entre os resultados desta pesquisa e a revisão sistemática. . . . . 75
Tabela 18 – Configurações de referência e robustez do Índice Multivariado nos subgrupos por sexo ao longo das 30 sementes. . . . . 94
Tabela 19 – Média dos componentes das matrizes de confusão das configurações de referência nos subgrupos por sexo ao longo das 30 sementes. . . . . 94

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AdaBoost | Aumento Adaptativo (do inglês, *Adaptive Boosting*) |
| ADASYN | Abordagem de Amostragem Sintética Adaptativa (do inglês, *Adaptive Synthetic Sampling Approach*) |
| AUC | Área sob a Curva (do inglês, *Area Under the Curve*) |
| AUC-ROC | Área sob a Curva ROC (do inglês, *Area Under the Receiver Operating Characteristic Curve*) |
| Bi-TCN | Rede Convolucional Temporal Bidirecional (do inglês, *Bidirectional Temporal Convolutional Network*) |
| C1 | Cenário 1 |
| C2 | Cenário 2 |
| C3 | Cenário 3 |
| CAGED | Cadastro Geral de Empregados e Desempregados |
| CAPES | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| CatBoost | Aumento Categórico (do inglês, *Categorical Boosting*) |
| CHAID | Detecção Automática de Interação por Qui-Quadrado (do inglês, *Chi-square Automatic Interaction Detection*) |
| D1 | Dataset 1 |
| D2 | Dataset 2 |
| D3 | Dataset 3 |
| DL | Aprendizado Profundo (do inglês, *Deep Learning*) |
| DNN | Redes Neurais Profundas (do inglês, *Deep Neural Networks*) |
| ENN | Vizinhos Mais Próximos Editados (do inglês, *Edited Nearest Neighbours*) |
| F1 | Escore F1 (do inglês, *F1-score*) |
| FN | Falso Negativo (do inglês, *False Negative*) |
| FP | Falso Positivo (do inglês, *False Positive*) |
| FPR | Taxa de Falsos Positivos (do inglês, *False Positive Rate*) |
| GANs | Redes Adversárias Generativas (do inglês, *Generative Adversarial Networks*) |
| HR | Recursos Humanos (do inglês, *Human Resources*) |
| IBM | Máquinas de Negócios Internacionais (do inglês, *International Business Machines*) |
| IC95% | Intervalo de Confiança de 95% |
| IEEE | Instituto de Engenheiros Eletricistas e Eletrônicos (do inglês, *Institute of Electrical and Electronics Engineers*) |
| IM | Índice Multivariado Ponderado |
| LDA | Análise Discriminante Linear (do inglês, *Linear Discriminant Analysis*) |
| LGPD | Lei Geral de Proteção de Dados Pessoais |
| LightGBM | Máquina Leve de Aumento de Gradiente (do inglês, *Light Gradient Boosting Machine*) |
| LR | Regressão Logística (do inglês, *Logistic Regression*) |

| | |
|---|---|
| MCC | Coeficiente de Correlação de Matthews (do inglês, *Matthews Correlation Coefficient*) |
| ML | Aprendizado de Máquina (do inglês, *Machine Learning*) |
| P&D | Pesquisa e Desenvolvimento |
| PA | Análise de Pessoas (do inglês, *People Analytics*) |
| PICO | Paciente, Intervenção, Comparação e Desfecho/Resultado (do inglês, *Patient, Intervention, Comparison and Outcome*) |
| p.p. | Pontos percentuais |
| PRISMA | Itens Preferenciais para Relato de Revisões Sistemáticas e Meta-Análises (do inglês, *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) |
| RF | Floresta Aleatória (do inglês, *Random Forest*) |
| RH | Recursos Humanos |
| ROC | Característica de Operação do Receptor (do inglês, *Receiver Operating Characteristic*) |
| ROC-AUC | Área sob a Curva ROC (do inglês, *Receiver Operating Characteristic – Area Under the Curve*) |
| ROI | Retorno sobre o Investimento (do inglês, *Return on Investment*) |
| ROSE | Sobreamostragem Aleatória de Exemplos (do inglês, *Random Over-Sampling Examples*) |
| RQ | Questão de Pesquisa (do inglês, *Research Question*) |
| SHAP | Explicações Aditivas de Shapley (do inglês, *Shapley Additive Explanations*) |
| SMOTE | Técnica de Sobreamostragem Sintética da Classe Minoritária (do inglês, *Synthetic Minority Over-sampling Technique*) |
| SMOTE-ENN | SMOTE com Vizinhos Mais Próximos Editados (do inglês, *Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours*) |
| SMOTE-Tomek | SMOTE com Links de Tomek (do inglês, *Synthetic Minority Over-sampling Technique with Tomek Links*) |
| SVM | Máquinas de Vetores de Suporte (do inglês, *Support Vector Machines*) |
| TN | Verdadeiro Negativo (do inglês, *True Negative*) |
| TP | Verdadeiro Positivo (do inglês, *True Positive*) |
| TPR | Taxa de Verdadeiros Positivos (do inglês, *True Positive Rate*) |
| UFJF | Universidade Federal de Juiz de Fora |
| XGBoost | Aumento de Gradiente Extremo (do inglês, *Extreme Gradient Boosting*) |

**SUMÁRIO**


**1 INTRODUÇÃO . . . . . 13**
1.1 APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO DO PROBLEMA 13
1.2 OBJETIVOS . . . . . 15
**1.2.1 Objetivo geral** . . . . . 15
**1.2.2 Objetivos específicos** . . . . . 15
**2 REVISÃO SISTEMÁTICA DA LITERATURA . . . . . 17**
2.1 RESPOSTAS DAS QUESTÕES SECUNDÁRIAS DE PESQUISA (RQ) 20
**2.1.1 RQ1: Quais modelos ou algoritmos de aprendizado de máquina têm sido mais utilizados para prever o *turnover* de profissionais?** . 22
**2.1.2 RQ2: Quais fatores apresentam maior influência nos modelos de previsão de *turnover*?** . . . . . 23
**2.1.3 RQ3: Quais atributos apresentam maior influência no *turnover* de profissionais?** . . . . . 23
**2.1.4 RQ4: Quais critérios ou métricas são mais utilizados para avaliar o desempenho dos modelos preditivos do *turnover*?** . . . . . 24
**2.1.5 RQ5: Quais lacunas de pesquisa têm sido identificadas no uso de modelos computacionais para previsão de *turnover* em contextos organizacionais?** . . . . . 25
2.2 TRABALHOS RELACIONADOS À PREVISÃO DE *TURNOVER* USANDO MODELOS COMPUTACIONAIS . . . . . 25
**2.2.1 Abordagens Preditivas com Bases Históricas de RH** . . . . . 26
**2.2.2 Abordagens Híbridas e *Survey-Driven*** . . . . . 27
2.3 RESPOSTA DA QUESTÃO PRIMÁRIA DE PESQUISA . . . . . 27
**3 REFERENCIAL TEÓRICO . . . . . 29**
3.1 *PEOPLE ANALYTICS*, RETENÇÃO DE COLABORADORES E *TURNOVER* . . . . . 29
3.2 *MACHINE LEARNING* APLICADO À CLASSIFICAÇÃO . . . . . 30
3.3 ABORDAGENS PARA MITIGAR O DESBALANCEAMENTO DE DADOS . . . . . 30
**3.3.1 SMOTE** . . . . . 31
**3.3.2 ADASYN** . . . . . 31
**3.3.3 Métodos Híbridos (SMOTE-ENN e SMOTE-Tomek)** . . . . . 31
3.4 MODELOS COMPUTACIONAIS . . . . . 32
**3.4.1 Regressão Logística** . . . . . 33
**3.4.2 Naive Bayes** . . . . . 33
**3.4.3 *Random Forest*** . . . . . 34
**3.4.4 XGBoost** . . . . . 34

**3.4.5** **AdaBoost** . . . 34
**3.4.6** ***Linear Discriminant Analysis* (LDA)** . . . 35
3.5 AVALIAÇÃO DOS MODELOS E MÉTRICAS DE DESEMPENHO . . 35
**4 MATERIAL E MÉTODOS . . . 38**
4.1 DESCRIÇÃO DAS BASES DE DADOS E CENÁRIOS EXPERIMENTAIS 40
4.2 PRÉ-PROCESSAMENTO E NORMALIZAÇÃO DOS DADOS . . . 43
4.3 SEPARAÇÃO DE DADOS E MITIGAÇÃO DO DESEQUILÍBRIO DAS CLASSES . . . 45
4.4 IMPLEMENTAÇÃO DOS MODELOS DE CLASSIFICAÇÃO E AVALIAÇÃO DE DESEMPENHO . . . 47
4.5 USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA COMO APOIO À ANÁLISE DOS RESULTADOS . . . 52
**5 ANÁLISE DOS RESULTADOS E DISCUSSÃO . . . 53**
5.1 RESULTADOS OBTIDOS COM A APLICAÇÃO DE MÉTODOS DE BALANCEAMENTO NOS CONJUNTOS DE TREINAMENTO . . . 53
**5.1.1** **Distribuição das classes e impacto do balanceamento nas métricas** 53
**5.1.2** **Heterogeneidade dos efeitos entre os datasets** . . . 56
5.2 DESEMPENHO GERAL DOS MODELOS E IMPACTO NOS DIFERENTES CENÁRIOS . . . 58
**5.2.1** **Desempenho global e impacto da seleção de atributos** . . . 58
**5.2.2** **Desempenho dos modelos nos subgrupos por sexo** . . . 62
**5.2.3** **Desempenho de treinamento e teste e análise exploratória dos *datasets*** . . . 64
5.3 SELEÇÃO DAS CONFIGURAÇÕES DE REFERÊNCIA . . . 70
**5.3.1** **Critérios, configurações selecionadas e robustez** . . . 70
**5.3.2** **Confronto com a literatura e síntese dos objetivos** . . . 75
**6 CONCLUSÃO . . . 80**
6.1 LIMITAÇÕES DA PESQUISA . . . 83
6.2 TRABALHOS FUTUROS . . . 86
**REFERÊNCIAS . . . 88**
**APÊNDICE A – Configurações selecionadas nos subgrupos por sexo . . . 94**

# 1 INTRODUÇÃO

## 1.1 APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO DO PROBLEMA

As organizações se encontram em um contexto caracterizado por mudanças constantes, em que a capacidade de compreender e antecipar fenômenos relacionados à gestão de pessoas se torna cada vez mais relevante para a tomada de decisão. Nesse cenário, o *turnover* de colaboradores se mostra um tema de interesse para empresas de diferentes áreas, uma vez que a previsão desse fenômeno pode auxiliar na elaboração de estratégias de recursos humanos (RH).

Segundo Mobley (1982), o *turnover* é um evento pontual que corresponde ao desligamento de um indivíduo de uma organização (43). Altas taxas de *turnover* geram custos adicionais relacionados à gestão de recursos humanos (52), que podem ser diretos ou indiretos e podem afetar o desempenho organizacional e o clima de trabalho.

Nos custos diretos associados ao *turnover*, destacam-se despesas relacionadas ao processo de desligamento e à substituição do colaborador, incluindo o tempo despendido por profissionais de RH e gestores, a cobertura temporária da posição, o recrutamento e a seleção de um substituto e sua posterior integração e capacitação (2):

- Custos administrativos associados ao desligamento, incluindo o tempo dedicado pela equipe de RH e pelos gestores ao processo;
- Custos de cobertura temporária da posição, como contratação de funcionários temporários ou pagamento de horas extras aos funcionários atuais;
- Custos de recrutamento e seleção, incluindo divulgação de vagas, processos seletivos e tempo dedicado pelos profissionais envolvidos;
- Custos de orientação, treinamento e socialização do novo funcionário até que ele esteja plenamente integrado e capacitado para desempenhar suas funções.

Além dos custos diretamente mensuráveis, o *turnover* pode produzir custos e efeitos indiretos relacionados à perda de conhecimento, às relações entre os membros da organização e ao desempenho das atividades. Entre eles, destacam-se (2, 60):

- Redução temporária da produtividade e da qualidade do trabalho durante o período em que a posição permanece vaga e enquanto o novo funcionário ainda não atingiu plena proficiência;
- Perda de memória organizacional, experiência acumulada e conhecimentos associados ao colaborador que deixa a organização;

- Rupturas no trabalho em equipe e nas redes de relacionamento e cooperação estabelecidas entre os funcionários;

A crescente disponibilidade de dados e o desenvolvimento de ferramentas computacionais ampliaram as possibilidades de análise em diferentes contextos, incluindo a gestão de pessoas e sua subárea de *people analytics*. Segundo Marler e Boudreau (2017), *people analytics* consiste no uso de análises e modelagem estatística com os dados do RH, buscando tomar decisões organizacionais suportadas por dados e melhorar os resultados (39).

Nesse sentido, modelos computacionais de previsão podem ser usados como uma alternativa para ajudar na compreensão das saídas de colaboradores, fornecendo subsídios para o planejamento e para a tomada de decisões nas organizações.

O recorte desta pesquisa está nos processos e nas decisões situados no interior das organizações, tendo como unidade de análise os registros de funcionários e os atributos associados ao vínculo de trabalho. O Novo Caged reúne registros administrativos de admissões e desligamentos e é utilizado para a produção de estatísticas do emprego formal (10). Embora essa fonte seja relevante para análises do mercado de trabalho e para políticas públicas, ela não reúne o mesmo conjunto de variáveis internas e contextuais utilizado nos estudos de *people analytics* aqui comparados. Sua adoção alteraria a unidade de análise e deslocaria o estudo para uma perspectiva macro do mercado de trabalho, distinta da comparação de modelos voltada ao ambiente organizacional proposta nesta dissertação.

A interpretação gerencial do *turnover* também exige reconhecer que as saídas podem resultar de antecedentes individuais, relacionais e organizacionais e que nem toda ocorrência é evitável ou necessariamente disfuncional. A retenção pode ser pertinente em determinados contextos, mas não constitui um objetivo universal e não deve ser aplicada de forma indiscriminada. Neste trabalho, os modelos estimam o rótulo de saída disponível nas bases; eles não identificam, por si sós, os motivos do desligamento nem definem quais pessoas deveriam permanecer na organização (29, 58).

Considerando as inúmeras possibilidades de análises dentro de uma organização, pode surgir o questionamento da relevância de *people analytics* ou se realmente é válido gastar tempo e recursos para tentar prever o *turnover*. Além dos custos diretos e indiretos ligados à saída de funcionários, o capital humano constitui um recurso estratégico relacionado ao desempenho organizacional. Evidências meta-analíticas demonstram que o capital humano, representado pelos conhecimentos, competências e experiências dos trabalhadores, está associado a melhores resultados das organizações, especialmente quando os conhecimentos e competências desenvolvidos possuem caráter específico à empresa, reforçando a importância de sua formação e retenção (17).

Diante disso, este trabalho teve como objetivo avaliar e comparar o desempenho

de diferentes modelos computacionais aplicados a distintas bases de dados de funcionários para a previsão de *turnover*, considerando técnicas de tratamento do desbalanceamento de classes, estratégias de seleção e redução de atributos e análises estratificadas por sexo, sob diferentes métricas de desempenho.

Muitos trabalhos desse tema focam apenas em avaliar, dentre alguns modelos, o que apresenta o melhor desempenho, usando apenas uma base de dados (55, 65, 7, 46). Este trabalho buscou demonstrar como decisões práticas do dia a dia (como aplicar técnicas de balanceamento, diminuir a quantidade de colunas da base e segmentar os dados por sexo) alteram o comportamento dos algoritmos e a capacidade deles de identificar quem vai sair da empresa.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo geral

O presente trabalho tem como objetivo avaliar e comparar o desempenho de diferentes modelos computacionais aplicados a bases de dados de funcionários para a previsão de *turnover*, considerando estratégias de tratamento do desbalanceamento de classes, diferentes representações dos atributos e análises estratificadas por sexo.

### 1.2.2 Objetivos específicos

- Comparar, de forma descritiva, o desempenho dos modelos nos subgrupos de funcionários registrados nas bases como masculino e feminino;
- Examinar, de forma exploratória, os atributos com maior relevância preditiva nas bases avaliadas;
- Avaliar estratégias para tratar o desbalanceamento de classes;
- Avaliar o impacto do uso de técnicas de seleção de atributos;
- Analisar diferentes métricas de desempenho dos modelos, considerando a complexidade do *turnover* no cotidiano das organizações.

Este trabalho foi organizado com base na estrutura a seguir: o capítulo 1 apresenta a introdução ao tema, a contextualização do problema de pesquisa e os objetivos geral e específicos que orientam o desenvolvimento do estudo.

O capítulo 2 apresenta a revisão sistemática da literatura sobre o uso de modelos computacionais na previsão de *turnover*, descrevendo o processo de busca e seleção dos estudos, as questões de pesquisa investigadas, os principais resultados encontrados e os trabalhos relacionados ao tema.

O capítulo 3 apresenta o referencial teórico que fundamenta a pesquisa, abordando conceitos relacionados a *people analytics*, retenção de colaboradores e *turnover*, aprendizado de máquina aplicado à classificação, tratamento do desbalanceamento de dados, modelos computacionais e métricas utilizadas para avaliação de desempenho.

O capítulo 4 descreve os materiais e métodos empregados na pesquisa, incluindo as bases de dados utilizadas, os cenários experimentais, as etapas de pré-processamento, a divisão dos dados, as técnicas de balanceamento e seleção de atributos, os modelos de classificação avaliados e os procedimentos adotados para análise de desempenho.

O capítulo 5 apresenta a análise e a discussão dos resultados obtidos nos experimentos, considerando os efeitos das técnicas de balanceamento, o desempenho dos modelos nos diferentes conjuntos de dados e cenários, a análise por sexo, os efeitos da seleção e redução de atributos e a seleção das configurações com melhor desempenho observado.

Por fim, o capítulo 6 apresenta as conclusões da pesquisa, retomando os objetivos propostos e sintetizando os principais achados do estudo, além de discutir suas limitações e indicar possibilidades para trabalhos futuros.

# 2 REVISÃO SISTEMÁTICA DA LITERATURA

Neste capítulo, é apresentada a revisão sistemática da literatura realizada, cujo objetivo foi identificar e sintetizar evidências sobre os modelos computacionais empregados entre 2020 e 2025 para prever *turnover* e comparar o desempenho dos mesmos, bem como apontar fatores preditivos mais influentes.

O início da revisão se deu com a estratégia PICO[1] (59), adaptada ao contexto desta pesquisa para estruturar a pergunta norteadora. No componente **P** (Problema/População/Paciente), foram considerados profissionais inseridos em contextos organizacionais. O componente **I** (Intervenção) correspondeu à aplicação de modelos computacionais para previsão de *turnover*. No componente **C** (Comparação), foi considerada a comparação entre os diferentes modelos e abordagens identificados na literatura. Por fim, o componente **O** (*Outcomes*) compreendeu o desempenho preditivo reportado pelos estudos e os fatores ou atributos relevantes para a previsão de *turnover*. A partir desses elementos, foi definida a seguinte pergunta norteadora da pesquisa: "Quais modelos computacionais têm sido utilizados para prever o *turnover* de profissionais em contextos organizacionais, em comparação a modelos atualmente utilizados, e quais fatores preditivos demonstram maior influência no desempenho desses modelos?"

Com base na pergunta norteadora, foram estabelecidas as cinco questões secundárias a seguir, com o objetivo de aprofundar aspectos específicos do tema investigado e complementar os resultados obtidos pela revisão:

- Quais modelos ou algoritmos de aprendizado de máquina têm sido mais utilizados para prever o *turnover* de profissionais?
- Quais fatores apresentam maior influência nos modelos de previsão de *turnover*?
- Quais atributos apresentam maior influência no *turnover* dos profissionais?
- Quais critérios ou métricas são mais utilizados para avaliar o desempenho dos modelos preditivos de *turnover*?
- Quais lacunas de pesquisa têm sido identificadas no uso de modelos computacionais para previsão de *turnover* em contextos organizacionais?

O Portal de Periódicos CAPES[2] foi usado como meio de acesso às bases de dados científicas, por meio das credenciais institucionais da Universidade Federal de Juiz de Fora (UFJF). A plataforma permitiu o acesso a periódicos e estudos científicos, garantindo a utilização de fontes de informação confiáveis e relevantes para esta revisão sistemática.

[1] Acrônimo para Problema/População/Paciente, Intervenção, Comparação e "*Outcomes*"(desfecho/resultado).

[2] https://www.periodicos.capes.gov.br. Acesso em 30 out. 2025.

As buscas foram realizadas nas quatro bases de dados eletrônicas que se seguem, selecionadas por sua relevância e abrangência na literatura científica:

- *Web of Science*[3] - 935 registros identificados;
- *Scopus*[4] - 35 registros identificados, mas que não foram usados por instabilidade da plataforma, o que impossibilitou a exportação dos trabalhos;
- *IEEE Xplore Digital Library*[5] - 7 registros identificados;
- *Elicit*[6] - 10 registros identificados.

As estratégias de busca foram adaptadas às características e à sintaxe de cada base consultada, mantendo-se, sempre que possível, os mesmos termos relacionados ao contexto organizacional, ao *turnover* e aos modelos preditivos. Na *Web of Science*, a busca foi realizada no campo de resumo (*Abstract*), representado pelo operador `AB`, por meio da seguinte expressão:

```
AB = ( (employee AND (turnover OR retention OR attrition)) AND
(intention OR predictors OR "predictive models"OR
"machine learning"OR "computational models") ).
```

Na *Web of Science*, os resultados foram refinados para publicações dos anos de 2020 a 2025, disponíveis em *Open Access*, classificadas como *Article* e publicadas nos idiomas inglês ou português. Após a aplicação desses critérios, foram identificados 935 registros.

Na *Scopus*, a busca também foi restrita ao campo de resumo (*Abstract*), representado pelo operador `ABS`. A expressão utilizada foi:

```
ABS ( ( EMPLOYEE AND ( TURNOVER OR RETENTION OR ATTRITION ) ) AND
( INTENTION OR PREDICTORS OR PREDICTIVE MODELS OR MACHINE LEARNING
OR COMPUTATIONAL MODELS ) ) AND PUBYEAR > 2019 AND PUBYEAR < 2026
AND ( LIMIT-TO ( SRCTYPE , "j") ) AND ( LIMIT-TO ( OA , "all") )
AND ( LIMIT-TO ( PUBSTAGE , "final") ) AND
( LIMIT-TO ( DOCTYPE , "ar") ) AND
( LIMIT-TO ( LANGUAGE , "English") ).
```

Dessa forma, na *Scopus*, a própria expressão incorporou os critérios de publicação entre 2020 e 2025, periódicos científicos, acesso aberto, estágio final de publicação, tipo documental *Article* e idioma inglês. A busca resultou em 35 registros. Entretanto, esses

[3] https://www.webofscience.com. Acesso em 30 out. 2025.
[4] https://www.scopus.com. Acesso em 04 nov. 2025.
[5] https://ieeexplore.ieee.org. Acesso em 04 nov. 2025.
[6] https://elicit.com. Acesso em 04 nov. 2025.

registros não foram incorporados à etapa de triagem devido à instabilidade da plataforma durante a realização da revisão, que impossibilitou sua exportação.

Na *IEEE Xplore Digital Library*, a busca foi realizada no campo de resumo (*Abstract*), utilizando a seguinte expressão:

```
("Abstract":employee) AND
("Abstract":turnover OR "Abstract":retention OR "Abstract":attrition) AND
("Abstract":intention OR "Abstract":predictors OR
"Abstract":"predictive models"OR "Abstract":"machine learning"OR
"Abstract":"computational models").
```

Nessa fonte, foram aplicados os filtros de publicação entre 2020 e 2025, conteúdo do tipo *Journals* e disponibilidade em *Open Access*. A combinação da expressão e desses filtros resultou em 7 registros.

Por fim, no *Elicit*, em razão das características próprias da plataforma, não foi utilizada uma expressão booleana equivalente às empregadas nas demais fontes. A busca foi realizada em linguagem natural por meio da seguinte pergunta:

> *"Which computational models have been used between 2020 and 2025 to predict employee turnover in organizational contexts, compared to the models currently applied, and which predictive factors demonstrate the greatest influence on these models' performance?"*

A delimitação temporal de 2020 a 2025 foi incorporada diretamente à pergunta submetida ao *Elicit*, que retornou 10 registros. Por utilizar busca em linguagem natural, essa fonte foi tratada como recurso complementar de identificação e seus resultados foram submetidos aos mesmos critérios de triagem e elegibilidade aplicados aos registros das demais fontes. A ausência de uma expressão booleana equivalente às utilizadas na *Web of Science* e na *IEEE Xplore Digital Library* constitui uma delimitação metodológica da revisão.

O recorte temporal da revisão compreendeu as publicações a partir do ano de 2020 até a data de realização das buscas em 2025. Como os filtros temporais disponíveis nas bases consultadas foram definidos por ano de publicação, não foi estabelecido um mês específico como parâmetro inicial ou final. Assim, foram contempladas as publicações de 2020 em diante disponíveis até 30 de outubro de 2025 na *Web of Science* e até 4 de novembro de 2025 na *Scopus*, na *IEEE Xplore Digital Library* e no *Elicit*.

Foram estabelecidos critérios de inclusão e exclusão para orientar a seleção dos estudos. Como critérios de inclusão, foram considerados: artigos publicados no período temporal definido para a revisão; disponíveis em *Open Access*; escritos em inglês ou português; classificados como *Article*; relacionados ao *turnover*, *attrition*, retenção ou

intenção de *turnover* de profissionais em contextos organizacionais; e capazes de contribuir para a resposta de pelo menos uma das questões secundárias de pesquisa.

Como critérios de exclusão, foram considerados: registros duplicados entre as fontes consultadas; estudos cujo título e resumo não apresentavam aderência ao escopo temático da revisão; e trabalhos que, após a avaliação de elegibilidade e leitura do texto, não apresentavam resultados capazes de contribuir para a resposta das questões secundárias estabelecidas. Os filtros relativos ao período de publicação, idioma, disponibilidade em *Open Access* e tipo de documento foram aplicados durante a etapa de busca, enquanto os demais critérios foram empregados nas etapas subsequentes de triagem e elegibilidade.

Após a obtenção dos 952 registros provenientes da *Web of Science*, da *IEEE Xplore Digital Library* e do *Elicit*, os artigos foram importados para a plataforma *Rayyan*[7], utilizada para apoiar a identificação de duplicatas e a triagem por títulos e resumos (50). Foram removidos 9 registros duplicados. Na triagem, 891 registros foram excluídos por não apresentarem aderência temática suficiente ao escopo e às questões secundárias da revisão, considerando conjuntamente o título e o resumo, e não apenas a presença literal de cada termo da estratégia de busca. Os 52 textos remanescentes foram avaliados quanto à elegibilidade; 16 foram excluídos por não fornecerem resultados capazes de responder às questões secundárias. Ao final, 36 artigos foram incluídos na revisão sistemática.

A Figura 1 resume o processo de seleção dos estudos incluídos na revisão, contemplando as etapas de identificação, triagem, elegibilidade e inclusão, conforme as diretrizes do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA 2020) (51).

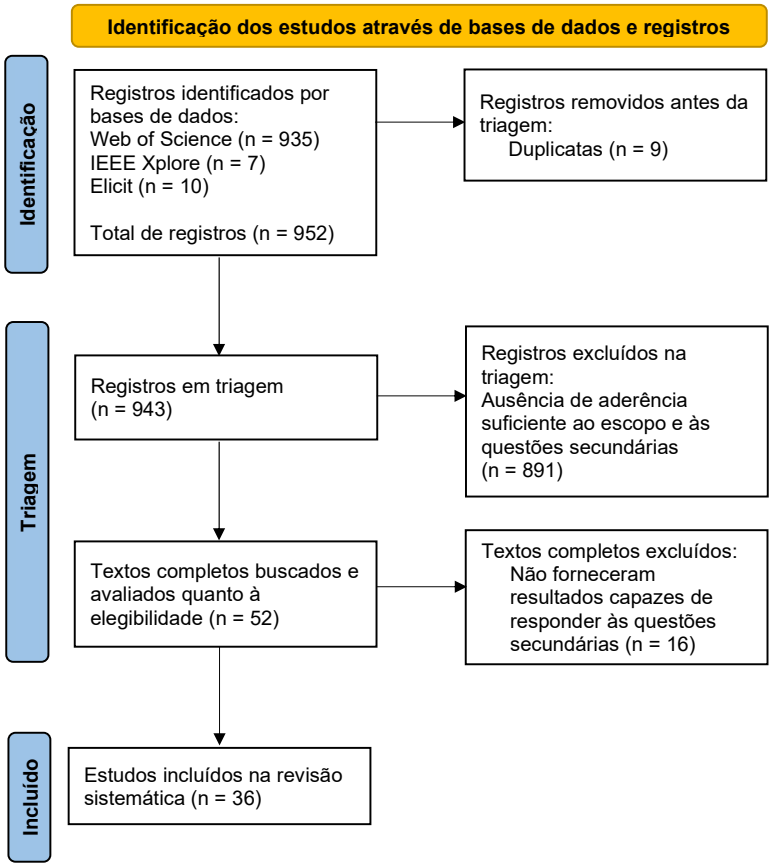
A análise dos 36 artigos selecionados para leitura na íntegra possibilitou a extração e a síntese das evidências necessárias para responder às perguntas secundárias de pesquisa apresentadas anteriormente. A partir dessa etapa, foram identificadas na literatura as abordagens adotadas pelos estudos e os resultados mais relevantes relacionados ao uso de modelos preditivos com foco no *turnover* de funcionários.

## 2.1 RESPOSTAS DAS QUESTÕES SECUNDÁRIAS DE PESQUISA (RQ)

Considerando conjuntamente os 36 estudos incluídos, a síntese mostrou que a literatura abrange tanto a previsão do *turnover* observado quanto análises da intenção de *turnover*, utilizando bases públicas, dados provenientes de organizações e informações obtidas por questionários. Entre os estudos que empregaram modelagem preditiva, foram recorrentes algoritmos clássicos, como *Random Forest*, Regressão Logística, Árvores de Decisão e *Support Vector Machines* (SVM), além de abordagens de *boosting* e aprendizado profundo (23, 55, 65, 68, 44, 66). Os resultados reportados indicaram que o desempenho dependeu de aspectos como o desbalanceamento das classes, a seleção de atributos e as

[7] https://www.rayyan.ai. Acesso em 04 nov. 2025.

Figura 1 – Fluxograma do processo de seleção dos estudos da revisão sistemática conforme o PRISMA 2020.




Fonte: Elaborado pelo autor (2026).

características das bases, enquanto satisfação no trabalho, tempo na organização, renda, idade, horas extras e fatores relacionados ao ambiente de trabalho apareceram entre os atributos ou fatores destacados nos estudos. A avaliação foi realizada por diferentes métricas, com recorrência de acurácia, *recall*, F1-score e AUC-ROC. Como os trabalhos diferem quanto ao desfecho analisado, à natureza dos dados e aos procedimentos de avaliação, seus resultados não devem ser interpretados como diretamente comparáveis em um único ranking de desempenho. As subseções seguintes detalham essas evidências em resposta a cada questão secundária de pesquisa.

### 2.1.1 RQ1: Quais modelos ou algoritmos de aprendizado de máquina têm sido mais utilizados para prever o *turnover* de profissionais?

A literatura demonstra uma grande diversidade nos algoritmos aplicados em *People Analytics*. Foi observado que a base das análises preditivas está no *Machine Learning* (ML) clássico. Modelos como Regressão Logística, *Random Forest*, Árvores de Decisão, *Support Vector Machines* (SVM) e *Naïve Bayes* foram testados em diversos cenários (23, 55, 65, 7, 68). Apesar de sua simplicidade teórica, a Regressão Logística frequentemente atinge desempenhos competitivos devido à sua alta interpretabilidade linear (23, 64, 65, 7). Contudo, o *Random Forest* foi o modelo que alcançou a maior acurácia em muitos estudos (32, 61, 68, 70).

Para cenários organizacionais complexos, os estudos usaram algoritmos de *Ensemble* e *Boosting*, tais como *XGBoost, CatBoost, AdaBoost, LightGBM* e *Extra Trees Classifier*. Tais modelos destacam-se ao processar *datasets* com interações não lineares complexas, alcançando métricas superiores e requerendo menor esforço no tratamento prévio de variáveis categóricas, como demonstrado pelo sucesso do *CatBoost* e *Extra Trees* em algumas bases de dados da plataforma Kaggle[8] (12, 56). Esses modelos podem capturar interações não lineares em dados tabelados, embora as exigências de pré-processamento dependam da implementação. Portanto, o desempenho observado não deve ser atribuído genericamente à menor necessidade de tratamento de variáveis categóricas.

Alguns métodos de *Deep Learning (DL)* também foram utilizados. Modelos baseados em Redes Neurais Profundas (DNN), Arquiteturas Convolucionais Temporais Bidirecionais (*Bi-TCN*) e *Deep Forest* alcançaram desempenho expressivo, frequentemente superando os 95% de acurácia (44, 66, 63). Ademais, técnicas de aprendizado profundo, como as Redes Adversárias Generativas (GANs), foram empregadas no pré-processamento de dados para gerar amostras sintéticas e solucionar o problema de desbalanceamento das classes minoritárias (44, 6).

[8] https://www.kaggle.com/.

### 2.1.2 RQ2: Quais fatores apresentam maior influência nos modelos de previsão de *turnover*?

De acordo com os artigos analisados nesta revisão sistemática da literatura, o desempenho preditivo reportado não está relacionado apenas à escolha do algoritmo de *Machine Learning*, mas também às características dos dados e às decisões metodológicas adotadas em sua preparação e avaliação.

Um dos aspectos recorrentes é o **desbalanceamento de classes**. Como o *turnover* frequentemente corresponde à classe minoritária, a utilização de conjuntos desbalanceados pode favorecer classificações direcionadas à classe majoritária e dificultar a identificação dos casos de desligamento. Nos estudos revisados, técnicas de reamostragem, como SMOTE e ADASYN, e abordagens baseadas em geração de dados sintéticos por GANs foram empregadas com o objetivo de aumentar a identificação da classe minoritária, com resultados dependentes da base, do modelo e da métrica utilizada (68, 56, 6, 44).

Outro aspecto identificado foi a **seleção e redução de atributos** (*feature selection*). Essas técnicas foram utilizadas para reduzir a dimensionalidade e priorizar características consideradas mais informativas para os modelos. Em estudos específicos, sua aplicação esteve associada a redução da complexidade do conjunto de entrada e a variações favoráveis de desempenho, embora esses resultados dependam do procedimento de seleção e da base analisada (46).

Por fim, a **origem e a natureza dos dados** também diferenciam os contextos experimentais encontrados na literatura. Estudos baseados em conjuntos públicos ou sintéticos, como o *dataset* da IBM[9], reportaram desempenhos elevados em diferentes configurações (46, 66, 67). Em estudos que utilizaram dados empíricos de organizações, também foram relatadas dificuldades relacionadas à disponibilidade, qualidade, completude e características demográficas dos registros, exigindo procedimentos adicionais de preparação dos dados (70, 69). Essas diferenças reforçam que métricas obtidas em bases distintas devem ser comparadas com cautela.

### 2.1.3 RQ3: Quais atributos apresentam maior influência no *turnover* de profissionais?

A análise dos 36 trabalhos da revisão indica que diferentes atributos e fatores foram reportados de forma recorrente como associados ao *turnover*, à intenção de saída ou ao desempenho dos modelos preditivos. Para organizar essas evidências, eles foram agrupados em três categorias principais: características demográficas e profissionais, fatores psicológicos e variáveis relacionadas à liderança e ao ambiente de trabalho. Como os estudos apresentam diferentes desenhos e desfechos, essas associações não devem ser interpretadas, de forma geral, como relações causais.

[9] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

As **características demográficas e profissionais** são frequentemente extraídas de bases de dados de RH, incluindo o amplamente utilizado *dataset* da IBM. Variáveis como renda mensal, idade, tempo na organização e distância entre residência e trabalho aparecem de forma recorrente entre os atributos associados ao *turnover* ou considerados relevantes pelos modelos (23, 44, 55, 56, 70). Horas extras, quantidade de projetos e avaliações de desempenho também foram reportadas como características relevantes em diferentes análises (1, 46, 63). A recorrência dessas variáveis indica relevância associativa ou preditiva nos contextos estudados, e não demonstra que elas provoquem o desligamento.

No âmbito dos **fatores psicológicos**, a satisfação no trabalho aparece de forma recorrente associada ao *turnover* ou à intenção de saída (15, 71). Exaustão emocional, *burnout* e estresse também foram relacionados a desfechos de saída ou intenção de saída nos estudos analisados (16, 3, 28). O comprometimento organizacional afetivo e a adequação pessoa-organização (*person-environment fit*) foram associados a menores intenções de saída (72, 34), enquanto práticas de *job crafting* foram relacionadas a maior engajamento e menor intenção de deixar a organização (71). Sobrecarga de informações e interferência negativa entre trabalho e vida pessoal também aparecem associadas a resultados menos favoráveis de retenção (35, 38, 49).

Quanto às **variáveis de liderança e ambiente de trabalho**, liderança despótica ou abusiva e ambiente de trabalho tóxico foram associados a maior intenção de *turnover* em estudos específicos (31). Em contrapartida, estilos de liderança servidora e autêntica foram relacionados a menores intenções de saída e a fatores como confiança e suporte percebido (33, 18, 45). Justiça organizacional, equidade e apoio emocional também aparecem entre os fatores associados à retenção e à intenção de permanência em grupos específicos, como profissionais da Geração Z e de Pesquisa e Desenvolvimento (64, 5).

### 2.1.4 RQ4: Quais critérios ou métricas são mais utilizados para avaliar o desempenho dos modelos preditivos do *turnover*?

Um achado importante presente em diversos estudos é que a acurácia não é uma métrica adequada e suficiente quando analisada isoladamente para avaliar o desempenho das predições (23, 68). A análise do *turnover* corporativo, na maioria das vezes, se depara com *datasets* fortemente desbalanceados (ex.: 80% dos funcionários permanecem e apenas 20% deixam a empresa). Nesse caso, um modelo que apontar a permanência de todos os colaboradores da base atingiria uma acurácia significativa, mesmo sem identificar nenhuma possível saída.

Diante disso, a eficácia do desempenho vem sendo validada de forma mais estruturada através de outras métricas, que são complementares à acurácia, como a área sob a curva ROC (AUC-ROC) e o *F1-score*. A curva ROC permite observar a capacidade do modelo em distinguir entre as duas classes (23, 44, 65). O F1-score corresponde à média

harmônica entre precisão e recall e atribui igual importância às duas medidas. Portanto, não minimiza especificamente os falsos negativos. Quando o custo de deixar de identificar um caso de desligamento for considerado superior ao custo de um falso alerta, o recall deve receber atenção direta (64, 68, 6).

### 2.1.5 RQ5: Quais lacunas de pesquisa têm sido identificadas no uso de modelos computacionais para previsão de *turnover* em contextos organizacionais?

A revisão de literatura permitiu identificar lacunas críticas que demandam atenção em trabalhos futuros. A principal limitação observada na literatura atual foi a **dependência excessiva de dados transversais e sintéticos**. Grande parte das pesquisas foca em demonstrar o desempenho dos algoritmos em bases simuladas e de domínio público, o que restringe a capacidade de generalização e validação desses modelos para ambientes corporativos reais, dinâmicos e de diferentes setores (23, 7, 66).

Em paralelo, nota-se uma carência significativa de **estudos longitudinais**. A maioria dos modelos analisa o *turnover* como uma "fotografia"estática no tempo. Porém a decisão de sair da empresa é, muitas vezes, um processo gradual. Há uma necessidade metodológica de desenvolver pesquisas que façam análises temporais e acompanhem o decaimento dos indicadores ao longo de meses, tornando os modelos preditivos mais sensíveis às condições reais (61, 68).

Outra lacuna relevante é a falta de abordagens que combinem diferentes fontes de dados. Atualmente, os modelos tendem a usar apenas dados corporativos e objetivos (como idade, cargo e salário) (55), ou se baseiam puramente em questionários psicológicos de intenção de *turnover* (64, 53). O desenvolvimento de **modelos híbridos**, capazes de avaliar tanto o histórico do funcionário quanto dados contínuos de clima organizacional, é uma necessidade nesse campo de estudo.

A Tabela 1 apresenta uma síntese das respostas às cinco questões secundárias, reunindo os achados recorrentes dos 36 estudos incluídos e sua relação com as decisões metodológicas desta dissertação.

## 2.2 TRABALHOS RELACIONADOS À PREVISÃO DE *TURNOVER* USANDO MODELOS COMPUTACIONAIS

A revisão de literatura demonstra duas vertentes principais nas pesquisas envolvendo *turnover*: a abordagem preditiva baseada em conjuntos de dados de RH (*Data-Driven*) e a abordagem exploratória fundamentada em questionários de autorrelato para prever a intenção de *turnover* (*Survey-Driven*).

Os estudos foram separados de acordo com o desfecho analisado. O *turnover* observado corresponde ao registro efetivo de saída, enquanto a intenção de *turnover*

Tabela 1 – Síntese das respostas às questões secundárias da revisão sistemática.

| Questão | Resposta principal | Evidências recorrentes | Implicação para a dissertação |
|---|---|---|---|
| RQ1 | Modelos clássicos permanecem frequentes, enquanto métodos de *boosting* e aprendizado profundo também são utilizados. | Regressão Logística, Random Forest, Árvores de Decisão, SVM e Naive Bayes aparecem em diferentes contextos; não há vencedor universal entre protocolos distintos. | Comparar seis classificadores de famílias distintas e evitar um ranking global desvinculado da base e do procedimento experimental. |
| RQ2 | O desempenho depende tanto do algoritmo quanto das decisões de preparação e avaliação dos dados. | Desbalanceamento, reamostragem, seleção de atributos, origem, dimensionalidade e qualidade da base alteram as métricas reportadas. | Avaliar condições de treinamento e representações de atributos por comparações controladas dentro de cada contexto. |
| RQ3 | Atributos demográficos, profissionais, psicológicos e organizacionais aparecem associados às predições ou ao desfecho. | Satisfação, tempo na organização, renda, idade, horas extras, liderança e ambiente de trabalho foram recorrentes. | Examinar contribuições preditivas com SHAP, sem convertê-las em efeitos causais ou importância prática automática. |
| RQ4 | A acurácia isolada é insuficiente, especialmente em bases desbalanceadas. | Recall, F1-score e AUC-ROC são recorrentes e descrevem dimensões complementares do desempenho. | Utilizar precisão, recall, F1-score, MCC, AUC-ROC e Log Loss, complementados pelo Índice Multivariado. |
| RQ5 | Persistem limitações de validade externa, temporalidade e diversidade das fontes de dados. | Dependência de bases públicas ou sintéticas, escassez de estudos longitudinais e de combinações entre registros e medidas organizacionais. | Delimitar a generalização dos resultados. |

A síntese considera exclusivamente os 36 estudos incluídos na revisão sistemática.

Fonte: Elaborado pelo autor (2026).

representa uma manifestação autorreferida de disposição para deixar a organização. Como a intenção não se converte necessariamente em desligamento, os resultados desses dois grupos foram sintetizados separadamente e não foram tratados como estimativas diretamente equivalentes.

### 2.2.1 Abordagens Preditivas com Bases Históricas de RH

Uma parcela significativa da literatura apoia-se em *datasets* públicos para comparar o desempenho dos modelos. O *dataset* da IBM se destaca como a principal base de dados utilizada. Trabalhos como os de Guerranti e Dimitri (2022), Najafi-Zangeneh et al. (2021), Vaiyapuri e Sbai (2025), Taner et al. (2024), Tokmak (2025) e Benabou et al. (2025) usaram esta base (composta por características demográficas fictícias) para avaliar diversos modelos preditivos (23, 46, 67, 65, 66, 7). Nestes estudos, foram alcançados resultados em torno de 88%-90% de acurácia (23), (65) ao usar modelos tradicionais, como regressão logística e *random forest* e percentuais de acurácia que beiram os 96% a 99% (67, 66), aplicando hiperparâmetros com Otimização Bayesiana, *Deep Forest Classifier* ou Redes Neurais no mesmo conjunto de dados.

Outros autores buscaram aplicar o aprendizado de máquina em contextos reais de

empresas. Varkiani et al. (2025) previram o *turnover* utilizando dados de uma organização financeira italiana (68). Similarmente, Veglio et al. (2025) aplicaram árvores de decisão CHAID em uma multinacional europeia/asiática de telecomunicações (69), e Wu e Fukui (2024) previram saídas com base em dados de 621 profissionais reais do setor de saúde mental (70), demonstrando alta capacidade de interpretação ao lidar com o problema no mundo real.

### 2.2.2 Abordagens Híbridas e *Survey-Driven*

A segunda frente metodológica não se concentra nos dados passados de quem efetivamente saiu da empresa, mas utiliza *machine learning* e análises estatísticas sobre respostas de escalas de *Likert* para modelar a intenção de *turnover*. Srivastava et al. (2025) convertem a média das respostas de indivíduos em firmas de P&D indianas em uma variável binária (*attrition vs. non-attrition*), criando *labels* baseadas na percepção psicológica contemporânea dos trabalhadores (64). Essa abordagem também é adotada no setor governamental sul-africano (38), na indústria energética (72) e em universitários recém-formados na Coreia (53). Embora não prevejam o *turnover* com base em dados históricos, esses estudos mostram a vantagem de focar em prevenção, ao identificar a intenção de sair enquanto o funcionário ainda não pediu desligamento.

## 2.3 RESPOSTA DA QUESTÃO PRIMÁRIA DE PESQUISA

A partir da síntese da literatura revisada, é possível responder de forma fundamentada à pergunta norteadora deste trabalho: **"Quais modelos computacionais têm sido utilizados para prever o *turnover* de profissionais em contextos organizacionais, em comparação a modelos atualmente utilizados, e quais fatores preditivos demonstram maior influência no desempenho desses modelos?"**

Os resultados demonstraram que uma gama diversificada de modelos computacionais tem sido empregada com sucesso, desde algoritmos tradicionais (*Random Forest* e Regressão Logística) até técnicas de aprendizado profundo de última geração. De modo geral, essas abordagens apresentam desempenho preditivo superior quando comparadas à abordagens convencionais de tomada de decisão. Essa mudança é especialmente relevante diante do uso tradicional de heurísticas, experiência profissional e julgamentos intuitivos nas decisões de gestão de pessoas (47).

Os fatores preditivos de maior influência englobam aspectos financeiros, características da estrutura do trabalho (horas extras, tempo na organização) e o estado emocional e de satisfação do colaborador. No entanto, a performance efetiva dos modelos depende fortemente da qualidade do pré-processamento dos dados. A literatura consultada evidencia que a busca pela maior acurácia, de forma isolada, é uma prática ultrapassada. Os modelos também devem ser avaliados com base no *recall*, AUC-ROC e *F1-score*, além de

serem apoiados por técnicas adequadas de tratamento do desbalanceamento de classes e otimização das variáveis.

# 3 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos que fundamentam a metodologia deste trabalho, bem como as referências e definições de termos que serão utilizados nos capítulos seguintes, com o intuito de fornecer uma base teórica comum aos leitores, promovendo o nivelamento do conhecimento acerca do tema e facilitando a compreensão do conteúdo apresentado ao longo do trabalho.

## 3.1 *PEOPLE ANALYTICS*, RETENÇÃO DE COLABORADORES E *TURNOVER*

O campo de *People Analytics* (PA) representa a transição da gestão de recursos humanos baseada em intuição e heurísticas para um modelo estratégico fundamentado em dados e evidências empíricas. Conforme definido por Marler e Boudreau (2017), PA envolve a aplicação de processos analíticos e modelagem estatística aos dados de recursos humanos, com o objetivo de aprimorar o retorno sobre o investimento em capital humano e alinhar as decisões organizacionais aos resultados de negócios (39).

Nesse contexto, a retenção de colaboradores pode ser relevante para reduzir custos de substituição e perdas de conhecimento, especialmente quando a saída envolve competências importantes para a continuidade das atividades (29). Entretanto, a retenção não constitui um objetivo universal. Em determinadas circunstâncias, a rotatividade pode ser funcional para a organização ou para o trabalhador, enquanto políticas indiscriminadas de retenção podem conservar relações inadequadas, condições de trabalho desfavoráveis ou segmentações entre pessoas consideradas mais e menos valiosas.

A literatura de gestão de pessoas distingue o *turnover* voluntário, iniciado pelo colaborador, do involuntário, iniciado pela organização. Também diferencia saídas evitáveis e inevitáveis e, conforme suas consequências, funcionais e disfuncionais (29, 58). Essas modalidades possuem antecedentes, responsabilidades e possibilidades de intervenção distintas. Neste trabalho, porém, a variável-alvo é tratada de forma genérica como ocorrência registrada de saída, pois as bases não permitem distinguir consistentemente essas modalidades. O experimento prevê o rótulo disponível nos dados, e não a decisão humana de deixar uma organização em toda a sua complexidade.

Os antecedentes do *turnover* podem ser organizados em diferentes níveis. No nível individual, incluem satisfação, comprometimento, saúde, exaustão e alternativas profissionais; no nível relacional e de equipe, incluem liderança, suporte, reconhecimento e conflitos; e, no nível organizacional, envolvem justiça, remuneração, desenho do trabalho, cultura, mudanças, políticas de carreira e condições de trabalho (29, 58). Essa distinção evita atribuir a saída exclusivamente às características do indivíduo e permite interpretar padrões preditivos também como sinais de condições produzidas ou modificáveis pela organização.

Por fim, a importância estimada por um modelo, uma associação estatística ou uma contribuição SHAP não equivale, por si só, à importância prática do atributo nem demonstra causalidade. A aplicação de *people analytics* constitui uma prática sociotécnica: os resultados dependem dos dados e dos modelos, mas seus efeitos dependem também de como gestores interpretam e utilizam as previsões. Por isso, modelos devem apoiar diagnóstico e julgamento responsável, sem substituir a compreensão do contexto, a participação dos trabalhadores ou a revisão humana das decisões (21, 37).

## 3.2 *MACHINE LEARNING* APLICADO À CLASSIFICAÇÃO

O Aprendizado de Máquina (*Machine Learning* - ML) é um campo da Inteligência Artificial dedicado ao desenvolvimento de algoritmos que permitem aos sistemas computacionais reconhecer padrões e tomar decisões a partir de dados, sem a necessidade de programação explícita para cada cenário. Em ML, os problemas são categorizados principalmente pela natureza da variável de saída. No Aprendizado Supervisionado, o modelo é treinado com exemplos para os quais a resposta que se deseja prever já é conhecida (8).

No contexto da previsão de *turnover*, a modelagem assume a forma de um problema de classificação binária supervisionada. Formalmente, dado um conjunto de dados composto por $N$ observações, cada colaborador $i$ é representado por um vetor de características (atributos preditores) denotado por $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, que contém informações demográficas, financeiras e organizacionais. A variável resposta (rótulo) é definida por $Y_i \in \{0, 1\}$, onde a classe 0 representa a retenção (classe negativa) e a classe 1 representa o desligamento (classe positiva) (25).

O objetivo do modelo de classificação é mapear o espaço de características para a variável resposta, estimando uma função $f(X)$ ou uma probabilidade $P(Y = 1|X)$ de modo que o erro de classificação seja minimizado em dados não vistos durante o treinamento (8).

## 3.3 ABORDAGENS PARA MITIGAR O DESBALANCEAMENTO DE DADOS

A modelagem preditiva de fenômenos organizacionais como o *turnover* frequentemente lida com conjuntos de dados fortemente desbalanceados. Em situações reais, a proporção de funcionários que permanecem na organização (classe majoritária) é expressivamente superior à proporção de funcionários que são desligados (classe minoritária). Esse desequilíbrio provoca um viés no processo de aprendizado dos algoritmos tradicionais, que tendem a classificar a maior parte das instâncias na classe majoritária, resultando no chamado paradoxo da acurácia — onde o modelo atinge uma alta taxa de acerto global falhando sistematicamente na predição da classe minoritária, que é, de fato, a classe de interesse (27).

Para mitigar este problema, são utilizadas técnicas de reamostragem ao nível dos

dados, buscando reequilibrar a distribuição das classes antes da etapa de treinamento do modelo.

### 3.3.1 SMOTE

O *Synthetic Minority Over-sampling Technique* (SMOTE) é um método de sobreamostragem (*oversampling*) proposto por Chawla et al. (2002) que, em vez de simplesmente duplicar os registros da classe minoritária, cria novas amostras sintéticas no espaço de características (13).

O funcionamento do SMOTE baseia-se na interpolação linear entre instâncias da classe minoritária que estão próximas no espaço dimensional. Para cada amostra $x_i$ da classe minoritária, o algoritmo identifica seus $k$ vizinhos mais próximos (*k-nearest neighbors*) pertencentes à mesma classe. Um desses vizinhos, denotado por $x_{zi}$, é selecionado aleatoriamente e a nova instância sintética ($x_{novo}$) é gerada pela seguinte equação:

$$x_{novo} = x_i + \lambda \times (x_{zi} - x_i)$$

onde $\lambda$ é um valor aleatório pertencente ao intervalo [0,1]. Essa interpolação expande a região de decisão da classe minoritária sem introduzir repetições exatas, ajudando a evitar o superajuste (*overfitting*) que normalmente ocorre no *oversampling* aleatório simples (13).

### 3.3.2 ADASYN

O *Adaptive Synthetic Sampling Approach* (ADASYN), proposto por He et al. (2008), é uma evolução conceitual do SMOTE (26). Enquanto o SMOTE gera amostras sintéticas a partir das instâncias da classe minoritária, o ADASYN emprega uma distribuição de densidade como critério dinâmico, direcionando uma maior geração de amostras sintéticas para as regiões em que as instâncias da classe minoritária apresentam maior dificuldade de classificação (26).

Na prática, o algoritmo avalia a vizinhança de cada exemplo minoritário e calcula uma proporção de exemplos da classe majoritária presentes nesse raio. As regiões do espaço de características onde a classe minoritária está cercada por instâncias da classe majoritária (fronteiras de decisão complexas) recebem uma carga maior de criação de dados sintéticos. O objetivo do ADASYN não é apenas balancear matematicamente as classes, mas forçar a fronteira de decisão do classificador a se afastar da classe minoritária em áreas de sobreposição (26).

### 3.3.3 Métodos Híbridos (SMOTE-ENN e SMOTE-Tomek)

Embora os métodos de *oversampling* aumentem a representatividade da classe minoritária, eles podem gerar ruído, invadindo o espaço de características da classe

majoritária e criando limites de decisão ambíguos. Para solucionar isso, Batista et al. (2004) sugerem o uso de métodos híbridos, que combinam técnicas de *oversampling* com a subamostragem (*undersampling*) aplicada como uma etapa de limpeza posterior (4).

As duas abordagens mais consolidadas na literatura são o SMOTE-ENN e o SMOTE-Tomek:

- **SMOTE-Tomek:** Após a geração de dados sintéticos pelo SMOTE, o algoritmo busca pares de instâncias de classes opostas que são vizinhos mais próximos um do outro (denominados *Tomek Links*). A existência de um *Tomek Link* indica que as instâncias constituem ruído ou estão situadas exatamente na borda de decisão. Ambas as instâncias do par são então removidas, aumentando a separabilidade das classes (4).
- **SMOTE-ENN:** Utiliza a técnica de *Edited Nearest Neighbours* (ENN) após a aplicação do SMOTE. O ENN remove qualquer observação (independentemente de sua classe) cujo rótulo se diferencie da maioria de seus $k$ vizinhos mais próximos. O SMOTE-ENN tende a remover mais exemplos do que o SMOTE-Tomek, promovendo uma limpeza mais agressiva no espaço amostral e resultando em clusters de dados mais definidos e coesos (4).

As técnicas apresentadas se diferenciam tanto no mecanismo empregado para reduzir o desbalanceamento entre as classes, quanto nas vantagens e limitações decorrentes de sua aplicação. Enquanto SMOTE e ADASYN atuam por meio da geração de amostras sintéticas, SMOTE-ENN e SMOTE-Tomek combinam a sobreamostragem com procedimentos de limpeza ou remoção de instâncias. A Tabela 2 sintetiza as principais vantagens e desvantagens dos quatro métodos de balanceamento empregados nesta pesquisa, a partir da comparação apresentada por Nogueira (2025) (48).

## 3.4 MODELOS COMPUTACIONAIS

A seleção de algoritmos para a modelagem preditiva exige o equilíbrio entre a capacidade de generalização e a interpretabilidade dos resultados. No contexto de recursos humanos, modelos mais simples oferecem maior transparência sobre a influência de cada variável na decisão. Por outro lado, algoritmos mais complexos, especialmente aqueles baseados na combinação de múltiplos estimadores (*Ensemble Learning*), podem capturar relações mais complexas entre os atributos, embora sua interpretação direta seja geralmente mais desafiadora (36).

Os modelos apresentados nas subseções a seguir representam diferentes abordagens de classificação, incluindo métodos lineares, probabilísticos, discriminantes e baseados em *Ensemble Learning*, possibilitando uma visão abrangente das estratégias de modelagem consideradas neste estudo.

Tabela 2 – Comparação das vantagens e desvantagens dos métodos de balanceamento de dados utilizados neste estudo.

| Método | Vantagens | Desvantagens | Rótulo |
|---|---|---|---|
| *SMOTE* (13) | – Eficiente para aumentar o tamanho do conjunto de dados;<br>– Gera amostras sintéticas diversas;<br>– Melhora a separação das classes em regiões densas dos dados. | – Pode aumentar o risco de *overfitting*;<br>– Sensível a *outliers*. | B |
| *ADASYN* (26) | – Gera amostras sintéticas adaptativas específicas para instâncias minoritárias;<br>– Trata desequilíbrios de classe em diferentes regiões do espaço de características. | – Pode ser computacionalmente intensivo em grandes conjuntos de dados;<br>– Sensível a *outliers*. | B |
| *SMOTE-ENN* (4) | – Reduz o risco de gerar amostras sintéticas não representativas;<br>– Melhora a qualidade das amostras geradas. | – Pode não remover completamente amostras ruidosas;<br>– Sensível a desequilíbrios extremos. | C |
| *SMOTE-Tomek* (4) | – Melhora a separabilidade entre as classes;<br>– Pode contribuir para a redução do desequilíbrio do conjunto de dados. | – Pode remover amostras relevantes da classe majoritária;<br>– Sensível a amostras ruidosas. | C |

**Legenda:** B – Cria amostras artificiais. C – Combina a criação de amostras artificiais com procedimentos de remoção ou limpeza de dados.

Fonte: Adaptado de Nogueira (2025) (48).

### 3.4.1 Regressão Logística

A Regressão Logística é um modelo estatístico utilizado para modelar a probabilidade de ocorrência de um evento binário. Apesar do nome, trata-se de um algoritmo de classificação que aplica a função logística (ou sigmoide) a uma combinação linear das variáveis preditoras, restringindo o resultado ao intervalo $[0, 1]$ (30).

Diferentemente da regressão linear clássica, a Regressão Logística estima a probabilidade $P(Y = 1|X)$ de uma instância pertencer à classe positiva (desligamento), dada pela equação:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}$$

onde $\beta_0$ é o intercepto e $\beta_1, \ldots, \beta_n$ são os coeficientes associados a cada variável preditora $X_n$. O sinal e a magnitude dos coeficientes podem auxiliar na interpretação das associações condicionais entre os preditores e o logaritmo das chances da classe positiva. Essa interpretação depende da codificação, da escala e da correlação entre os atributos e não estabelece efeitos causais sobre o *turnover* (30).

### 3.4.2 Naive Bayes

O classificador *Naive Bayes* fundamenta-se no Teorema de Bayes, incorporando a suposição "ingênua" (*naive*) de que os atributos preditores são condicionalmente independentes entre si, dada a classe alvo (57).

Para classificar um novo colaborador, o algoritmo calcula a probabilidade *a posteriori* de pertencimento a cada classe e atribui a instância àquela com a maior probabilidade. O cálculo segue a formulação:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Apesar da forte suposição de independência – que raramente se sustenta de forma estrita em dados reais de RH (como a correlação entre cargo e salário) – o *Naive Bayes* é computacionalmente eficiente e frequentemente apresenta desempenho satisfatório em problemas de classificação complexos (57).

### 3.4.3 *Random Forest*

O *Random Forest*, proposto por Breiman (2001), é um algoritmo de aprendizado baseado em *Ensemble* que constrói uma "floresta"de múltiplas árvores de decisão durante a fase de treinamento (11).

Para reduzir a variância e evitar o *overfitting* (superajuste) comum em árvores de decisão individuais, o algoritmo emprega a técnica de *Bagging* (*Bootstrap Aggregating*). O *Random Forest* treina cada árvore em um subconjunto aleatório dos dados e, a cada divisão de nó, seleciona um subconjunto aleatório de atributos. A predição final em problemas de classificação é obtida pela votação majoritária entre todas as árvores geradas (11).

### 3.4.4 XGBoost

O *eXtreme Gradient Boosting* (XGBoost) é uma implementação avançada e altamente escalável do método de *Gradient Boosting* (14). O conceito de *Boosting* baseia-se na construção sequencial de preditores fracos (geralmente árvores de decisão rasas), onde cada nova árvore é treinada especificamente para corrigir os erros residuais das árvores anteriores (25).

O diferencial do XGBoost é a sua formulação matemática que incorpora de forma explícita uma função de penalização (regularização $L1$ e $L2$) no cálculo do ganho de informação das folhas das árvores, controlando a complexidade do modelo e prevenindo o superajuste, o que frequentemente o torna um dos algoritmos com maior capacidade preditiva em dados tabulares (14).

### 3.4.5 AdaBoost

O *Adaptive Boosting* (AdaBoost), introduzido por Freund e Schapire (1997), é outro algoritmo clássico da família *Boosting* (20). Diferente do *Gradient Boosting* que minimiza erros residuais, o AdaBoost funciona ajustando iterativamente os "pesos"das instâncias do conjunto de treinamento.

Na primeira iteração, todos os colaboradores possuem pesos iguais. Após o treinamento do primeiro classificador fraco, o algoritmo aumenta o peso das instâncias que foram classificadas incorretamente e reduz o daquelas classificadas corretamente. O próximo classificador é forçado a focar nas observações "difíceis"(por exemplo, funcionários que saíram da empresa, mas cujas características indicavam permanência). O modelo final é uma combinação linear ponderada de todos os classificadores iterativos (20).

### 3.4.6 *Linear Discriminant Analysis* (LDA)

A *Linear Discriminant Analysis* (LDA) é uma técnica clássica de estatística multivariada utilizada para classificação e redução de dimensionalidade. O objetivo da LDA é projetar um conjunto de dados de alta dimensionalidade em um espaço de menor dimensão, maximizando a separação entre as diferentes classes (42).

Matematicamente, a LDA busca o vetor de projeção $w$ que maximiza o Critério de Fisher, ou seja, a razão entre a matriz de dispersão entre as classes ($S_B$) e a matriz de dispersão dentro das classes ($S_W$):

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Ao maximizar essa função objetivo, a LDA encontra uma fronteira de decisão linear que torna os grupos (retenção e *turnover*) o mais compactos e distantes entre si quanto possível (42).

Na sua formulação clássica de classificação, a LDA assume que os preditores têm distribuições que são aproximadamente normais condicionais às classes, além de compartilharem uma matriz de covariância comum. A qualidade da fronteira linear estimada pode ser comprometida quando esses pressupostos são violados, particularmente em dados organizacionais que possuem variáveis indicadoras e distribuições assimétricas.

## 3.5 AVALIAÇÃO DOS MODELOS E MÉTRICAS DE DESEMPENHO

A avaliação do desempenho dos modelos preditivos se baseia nos resultados extraídos da Matriz de Confusão, que sumariza a contagem das predições corretas e incorretas realizadas pelo modelo em comparação com as classes reais. A matriz é composta por quatro elementos principais: Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN) (19).

Em um cenário de *People Analytics*, a classe positiva (1) representa o desligamento e a negativa (0), a permanência. Consequentemente, o erro do tipo Falso Negativo ocorre quando o modelo prevê que um colaborador irá permanecer, mas ele se desliga da empresa. Esse erro pode assumir maior relevância gerencial quando o custo de deixar de identificar antecipadamente uma ocorrência de saída for considerado superior ao custo de um falso

alerta. Em outros contextos, entretanto, falsos positivos também podem representar custos relevantes, especialmente quando as intervenções de retenção forem dispendiosas ou a capacidade de atuação da organização for limitada.

Com base nesses elementos, são calculadas as seguintes métricas de desempenho (62):

**Acurácia:** Mensura a proporção total de predições corretas em relação ao total de instâncias avaliadas.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Embora intuitiva, a acurácia é uma métrica insuficiente para bases de dados desbalanceadas, pois um classificador pode obter alta acurácia prevendo sistematicamente apenas a classe majoritária, negligenciando completamente os casos reais de desligamento (27).

**Precisão:** Indica a proporção de classificações positivas corretas sobre o total de predições apontadas como positivas pelo modelo.

$$Precision = \frac{TP}{TP + FP}$$

***Recall* (ou Revocação / Sensibilidade):** Representa a proporção de instâncias positivas (funcionários que de fato saíram) que foram corretamente identificadas pelo algoritmo. No contexto do *turnover*, essa métrica assume maior relevância quando a prioridade organizacional é identificar a maior proporção possível dos casos de saída, reconhecendo-se que o aumento do *recall* pode ser acompanhado por maior quantidade de falsos positivos.

$$Recall = \frac{TP}{TP + FN}$$

**F1-Score:** Como existe um *trade-off* natural entre Precisão e *Recall*, o F1-Score é utilizado para fornecer uma métrica única que balanceia ambas. Ele é calculado como a média harmônica entre a Precisão e o *Recall*:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

**Coeficiente de Correlação de Matthews (MCC):** O *Matthews Correlation Coefficient* (MCC) é uma métrica de avaliação para problemas de classificação que considera simultaneamente os quatro componentes da matriz de confusão: Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN) (40). Seu valor varia entre $-1$ e $+1$, onde $+1$ representa uma classificação perfeita, 0 indica

ausência de correlação entre as classes previstas e observadas, e $-1$ corresponde a uma discordância completa entre predições e valores reais.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Por considerar conjuntamente os acertos e erros das duas classes, o MCC fornece uma medida equilibrada do desempenho global do classificador, sendo particularmente útil em problemas com classes desbalanceadas. Diferentemente de métricas que podem apresentar valores elevados devido ao bom desempenho apenas sobre a classe majoritária, o MCC exige que o modelo apresente resultados consistentes tanto para a classe positiva quanto para a negativa para alcançar valores elevados.

**Curva ROC e AUC:** A *Receiver Operating Characteristic* (ROC) é uma ferramenta gráfica que plota a Taxa de Verdadeiros Positivos ($TPR = Recall$) contra a Taxa de Falsos Positivos ($FPR = \frac{FP}{FP+TN}$) para diferentes limiares de probabilidade de decisão. A *Area Under the Curve* (AUC) resume o gráfico em um único valor escalar. A AUC representa a probabilidade de o classificador atribuir um ranqueamento de probabilidade maior a uma instância positiva escolhida aleatoriamente do que a uma instância negativa aleatória, fornecendo uma medida robusta da capacidade geral de separação do modelo, independentemente de um ponto de corte específico ou de distribuições desbalanceadas (19).

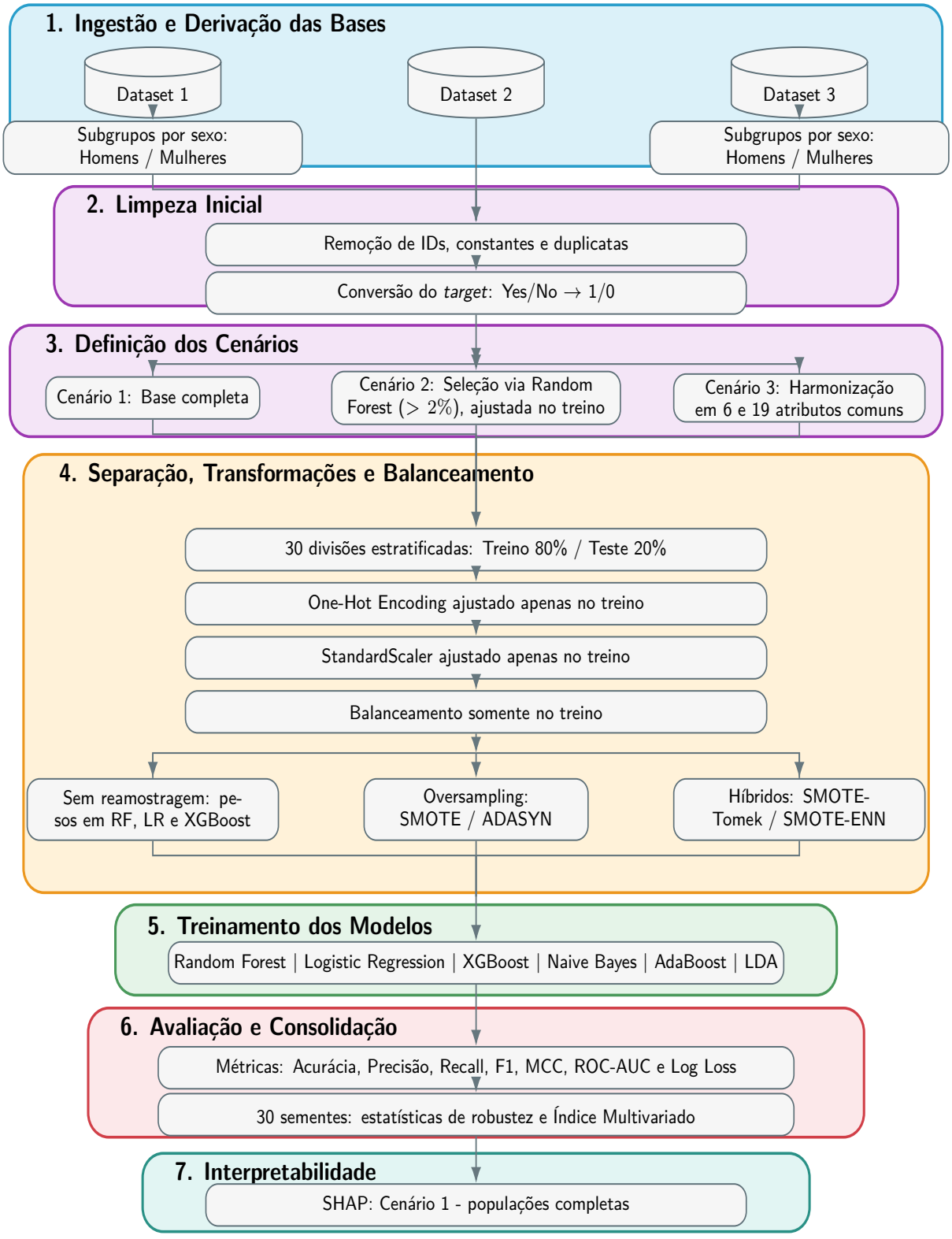
Apesar de a AUC-ROC não depender de um único limiar de classificação, sua interpretação deve ser feita juntamente com métricas focadas na classe positiva. Em problemas desbalanceados, a curva de precisão-recall e sua área fornecem uma visualização direta do compromisso entre a precisão das predições positivas e a recuperação dos casos positivos.

# 4 MATERIAL E MÉTODOS

Este capítulo tem como objetivo detalhar o percurso metodológico adotado para o desenvolvimento desta pesquisa. Nas seções a seguir, são apresentadas as etapas que compõem o *pipeline* analítico, desde a descrição e limpeza das bases de dados até o treinamento e a avaliação dos modelos. Também são descritos os procedimentos utilizados para a execução dos experimentos com 30 sementes, a construção dos cenários experimentais, o tratamento do desbalanceamento das classes, a consolidação das repetições e a análise de interpretabilidade com SHAP. As etapas que exigiram ajuste a partir dos dados foram realizadas utilizando somente o conjunto de treinamento, mantendo o conjunto de teste separado para a avaliação dos modelos.

Para proporcionar uma visão geral das etapas conduzidas, a Figura 2 apresenta o processo metodológico utilizado nesta pesquisa. Após a limpeza inicial dos dados, para cada população e cada uma das 30 sementes foi realizada uma divisão estratificada em 80% para treinamento e 20% para teste. Os mesmos registros de treinamento e teste foram mantidos nas configurações comparáveis dos três cenários. As transformações aprendidas foram ajustadas somente no treinamento, assim como a seleção de atributos do Cenário 2. O balanceamento também foi aplicado exclusivamente ao conjunto de treinamento. Em seguida, os seis classificadores foram treinados e avaliados no conjunto de teste original. Por fim, os resultados das 30 repetições foram consolidados para as análises de robustez e para o Índice Multivariado, e a interpretabilidade foi avaliada por meio do SHAP.

Figura 2 – Fluxograma do processo experimental e *pipeline* de modelagem preditiva de *turnover*.




Fonte: Elaborado pelo autor (2026).

## 4.1 DESCRIÇÃO DAS BASES DE DADOS E CENÁRIOS EXPERIMENTAIS

Para a condução dos experimentos preditivos propostos nesta pesquisa, foram utilizadas três bases de dados distintas contendo registros de recursos humanos e informações demográficas, organizacionais e comportamentais de colaboradores. As bases foram obtidas a partir da plataforma pública Kaggle. O *Dataset* 1 corresponde ao *IBM HR Analytics Employee Attrition & Performance*[1], conjunto artificial desenvolvido por cientistas de dados da IBM, contendo originalmente 1.470 registros e 35 colunas. O *Dataset* 2 é o *HR Analytics Data Set*[2], também de natureza artificial, contando com 14.999 registros e 10 colunas. Por fim, o *Dataset* 3 corresponde ao *Employee Attrition Prediction Dataset*[3], apresentado pelo disponibilizador como uma base de dados real e composto por 10.000 registros e 26 colunas.

As Tabelas 3, 4 e 5 detalham a estrutura original das três bases, incluindo a quantidade e o tipo das variáveis, as categorias textuais observadas, a variável-alvo e aspectos relevantes identificados na inspeção inicial dos dados.

Após os procedimentos iniciais de limpeza, os conjuntos efetivamente utilizados na modelagem apresentaram tamanhos distintos. O *Dataset* 1 permaneceu com 1.470 registros, dos quais 237 pertenciam à classe positiva de *turnover* (16,12%). O *Dataset* 2 continha originalmente 14.999 registros, sendo identificadas 3.008 duplicatas exatas, removidas antes da construção dos conjuntos utilizados na modelagem e, consequentemente, antes da divisão entre treinamento e teste. Após essa remoção, o *Dataset* 2 passou a conter 11.991 registros, dos quais 1.991 pertenciam à classe positiva (16,60%). O *Dataset* 3 permaneceu com 10.000 registros, sendo 1.997 classificados na classe positiva (19,97%).

Como apresentado nas tabelas anteriores, as três bases diferem tanto na quantidade de atributos quanto nos tipos e nas categorias das variáveis disponíveis. Para fins de modelagem, as variáveis-alvo originais foram tratadas para forma binária e com a mesma interpretação: o valor 1 representa a ocorrência de *turnover*, enquanto o valor 0 representa a permanência do funcionário. Nos *Datasets* 1 e 3, as categorias `Yes` e `No` da variável `Attrition` foram convertidas, respectivamente, para 1 e 0. No *Dataset* 2, a variável `left` já se encontrava originalmente nesse formato.

Para organizar os experimentos, foram definidos três cenários de representação dos atributos. No Cenário 1 (C1), foram utilizados todos os atributos preditores disponíveis após os procedimentos iniciais de limpeza de cada base. No Cenário 2 (C2), partiu-se da mesma representação utilizada no Cenário 1, mas foi aplicada uma etapa de seleção automática de atributos por *Random Forest*. Para cada população e cada semente, essa

[1] `https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset`
[2] `https://www.kaggle.com/datasets/raminhuseyn/hr-analytics-data-set`
[3] `https://www.kaggle.com/datasets/ziya07/employee-attrition-prediction-dataset`

Tabela 3 – Caracterização estrutural do *Dataset* 1 em sua forma original.

| **Característica** | **Descrição** |
|---|---|
| Dimensão original | 1.470 registros e 35 colunas. |
| Tipos de dados | 26 variáveis numéricas inteiras e 9 variáveis categóricas armazenadas como texto. |
| Variáveis numéricas | `Age`, `DailyRate`, `DistanceFromHome`, `Education`, `EmployeeCount`, `EmployeeNumber`, `EnvironmentSatisfaction`, `HourlyRate`, `JobInvolvement`, `JobLevel`, `JobSatisfaction`, `MonthlyIncome`, `MonthlyRate`, `NumCompaniesWorked`, `PercentSalaryHike`, `PerformanceRating`, `RelationshipSatisfaction`, `StandardHours`, `StockOptionLevel`, `TotalWorkingYears`, `TrainingTimesLastYear`, `WorkLifeBalance`, `YearsAtCompany`, `YearsInCurrentRole`, `YearsSinceLastPromotion` e `YearsWithCurrManager`. |
| Variáveis categóricas e categorias observadas | `Attrition`: `No`, `Yes`;<br>`BusinessTravel`: `Non-Travel`, `Travel_Frequently`, `Travel_Rarely`;<br>`Department`: `Human Resources`, `Research & Development`, `Sales`;<br>`EducationField`: `Human Resources`, `Life Sciences`, `Marketing`, `Medical`, `Other`, `Technical Degree`;<br>`Gender`: `Female`, `Male`;<br>`JobRole`: `Healthcare Representative`, `Human Resources`, `Laboratory Technician`, `Manager`, `Manufacturing Director`, `Research Director`, `Research Scientist`, `Sales Executive`, `Sales Representative`;<br>`MaritalStatus`: `Divorced`, `Married`, `Single`;<br>`Over18`: `Y`;<br>`OverTime`: `No`, `Yes`. |
| Variável-alvo | `Attrition`, originalmente representada pelas categorias `Yes` e `No`. |
| Qualidade inicial | Não foram identificados valores ausentes nem registros exatamente duplicados. |
| Colunas removidas antes da modelagem | `EmployeeNumber`, por atuar como identificador, e `EmployeeCount`, `StandardHours` e `Over18`, por apresentarem valor constante nos registros. |

Fonte: Elaborado pelo autor a partir do *Dataset* 1 (2026).

seleção foi ajustada exclusivamente sobre o conjunto de treinamento já codificado e ainda não balanceado, sendo mantidas as *features* com importância superior a 2%. A seleção foi realizada uma única vez para cada combinação de população e semente, e as mesmas *features* selecionadas foram utilizadas nos diferentes métodos de balanceamento e classificadores daquela repetição.

O Cenário 3 (C3) teve como objetivo construir representações reduzidas com atributos comuns entre as bases. Diferentemente dos Cenários 1 e 2, essa representação foi construída diretamente a partir dos dados após a limpeza inicial, antes da aplicação do *One-Hot Encoding*. Foi definido um núcleo de seis atributos disponível para os três *datasets*: departamento, satisfação no trabalho, tempo na empresa, renda ou salário, avaliação de desempenho e promoção nos últimos cinco anos. Para os *Datasets* 1 e 3, também foi utilizada uma representação ampliada com 19 atributos, formada pelos seis anteriores e por idade, distância da residência, gênero, valor por hora, envolvimento no trabalho, nível do cargo, cargo, estado civil, número de empresas trabalhadas, hora extra, treinamentos no último ano, equilíbrio entre trabalho e vida pessoal e tempo no cargo atual.

Tabela 4 – Caracterização estrutural do *Dataset* 2 em sua forma original.

| **Característica** | **Descrição** |
|---|---|
| Dimensão original | 14.999 registros e 10 colunas. |
| Tipos de dados | 8 variáveis numéricas, sendo 6 inteiras e 2 decimais, e 2 variáveis categóricas armazenadas como texto. |
| Variáveis numéricas | `satisfaction_level` e `last_evaluation`, armazenadas como valores decimais; e `number_project`, `average_montly_hours`, `time_spend_company`, `Work_accident`, `left` e `promotion_last_5years`, armazenadas como valores inteiros. As variáveis `Work_accident`, `left` e `promotion_last_5years` assumem os valores 0 e 1. |
| Variáveis categóricas e categorias observadas | `Department`: `accounting`, `hr`, `IT`, `management`, `marketing`, `product_mng`, `RandD`, `sales`, `support`, `technical`;<br>`salary`: `high`, `low`, `medium`. |
| Variável-alvo | `left`, originalmente representada de forma binária, com valores 0 e 1. |
| Qualidade inicial | Não foram identificados valores ausentes. Foram identificados 3.008 registros exatamente duplicados, posteriormente removidos antes da modelagem. |
| Registros após a limpeza | 11.991 registros após a remoção das 3.008 duplicatas exatas. |

Fonte: Elaborado pelo autor a partir do *Dataset* 2 (2026).

Algumas correspondências utilizadas no Cenário 3 exigiram tratamento específico. Nos *Datasets* 1 e 3, o atributo referente à promoção nos últimos cinco anos foi derivado, respectivamente, de `YearsSinceLastPromotion` e `Years_Since_Last_Promotion`, considerando como ocorrência de promoção os registros com valor menor ou igual a cinco anos.

Tabela 5 – Caracterização estrutural do *Dataset* 3 em sua forma original.

| **Característica** | **Descrição** |
|---|---|
| Dimensão original | 10.000 registros e 26 colunas. |
| Tipos de dados | 20 variáveis numéricas inteiras e 6 variáveis categóricas armazenadas como texto. |
| Variáveis numéricas | `Employee_ID`, `Age`, `Job_Level`, `Monthly_Income`, `Hourly_Rate`, `Years_at_Company`, `Years_in_Current_Role`, `Years_Since_Last_Promotion`, `Work_Life_Balance`, `Job_Satisfaction`, `Performance_Rating`, `Training_Hours_Last_Year`, `Project_Count`, `Average_Hours_Worked_Per_Week`, `Absenteeism`, `Work_Environment_Satisfaction`, `Relationship_with_Manager`, `Job_Involvement`, `Distance_From_Home` e `Number_of_Companies_Worked`. |
| Variáveis categóricas e categorias observadas | `Gender`: `Female`, `Male`;<br>`Marital_Status`: `Divorced`, `Married`, `Single`;<br>`Department`: `Finance`, `HR`, `IT`, `Marketing`, `Sales`;<br>`Job_Role`: `Analyst`, `Assistant`, `Executive`, `Manager`;<br>`Overtime`: `No`, `Yes`;<br>`Attrition`: `No`, `Yes`. |
| Variável-alvo | `Attrition`, originalmente representada pelas categorias `Yes` e `No`. |
| Qualidade inicial | Não foram identificados valores ausentes nem registros exatamente duplicados. |
| Coluna removida antes da modelagem | `Employee_ID`, utilizada como identificador dos registros. |

Fonte: Elaborado pelo autor a partir do *Dataset* 3 (2026).

No *Dataset* 2, foi utilizada diretamente a variável `promotion_last_5years`. Além disso, a variável `salary` do *Dataset* 2 foi convertida das categorias `low`, `medium` e `high` para os valores ordinais 1, 2 e 3. A representação com 19 atributos não foi aplicada ao *Dataset* 2, devido à ausência de atributos correspondentes em quantidade suficiente para compor esse conjunto ampliado.

Os atributos utilizados no Cenário 3 foram considerados correspondentes para fins da comparação experimental, mas não representam necessariamente medidas idênticas entre as três bases. Diferenças de definição, escala e forma de registro permanecem possíveis e devem ser consideradas na interpretação das comparações entre os *datasets*.

Com o objetivo de examinar possíveis diferenças no comportamento preditivo dos modelos por sexo, foram realizadas análises segmentadas nos *Datasets* 1 e 3. Dessa forma, foram consideradas sete populações experimentais: o *Dataset* 1 completo e seus subgrupos masculino e feminino; o *Dataset* 2 completo; e o *Dataset* 3 completo e seus subgrupos masculino e feminino. Após a limpeza, o *Dataset* 1 apresentou 882 registros classificados como `Male` e 588 como `Female`, enquanto o *Dataset* 3 apresentou 4.958 e 5.042 registros, respectivamente. O *Dataset* 2 não foi segmentado por não possuir variável correspondente a sexo ou gênero entre os atributos disponíveis.

Nos experimentos realizados separadamente por sexo, a variável `Gender` foi retirada da matriz de preditores, uma vez que apresentava valor constante dentro de cada subgrupo. Para a construção da representação de 19 atributos do Cenário 3, o valor correspondente ao sexo do próprio subgrupo foi reinserido apenas na etapa de harmonização dos atributos, permitindo manter a mesma estrutura conceitual utilizada na população completa.

Essa estrutura permitiu comparar o desempenho dos modelos entre as bases completas e seus respectivos subgrupos masculino e feminino, bem como observar se os padrões de desempenho se modificavam após a segmentação. Porém essa comparação não consiste em avaliação formal de equidade algorítmica (*fairness*), uma vez que não foram calculadas métricas específicas de equidade, paridade ou discriminação entre grupos, nem realizados testes formais destinados a avaliar vieses algorítmicos.

## 4.2 PRÉ-PROCESSAMENTO E NORMALIZAÇÃO DOS DADOS

A etapa de pré-processamento é crítica para garantir a qualidade dos dados de entrada e a convergência adequada dos algoritmos de aprendizado de máquina (24). Todos os procedimentos de transformação de dados foram implementados na linguagem Python, utilizando extensivamente a biblioteca Pandas (41) para a manipulação dos *dataframes* e a biblioteca Scikit-Learn (54) para a padronização das variáveis preditoras e outras operações necessárias à preparação dos dados para a modelagem.

O processo de limpeza inicial consistiu na remoção de variáveis identificadoras e de

atributos constantes, que não apresentavam variabilidade útil para a modelagem. Foram excluídas colunas como `EmployeeNumber`, utilizada como identificador único, `StandardHours`, constante nos registros, e `Over18`, também constante. Em seguida, a variável-alvo foi convertida para representação binária, de modo que o valor 1 correspondeu à ocorrência de *turnover* (classe positiva) e o valor 0 à permanência do empregado (classe negativa).

O tratamento das variáveis categóricas foi realizado por meio de *One-Hot Encoding*, utilizando a classe `OneHotEncoder` da biblioteca Scikit-Learn. A codificação foi realizada somente após a separação dos dados em treinamento e teste. Para cada divisão, o codificador foi ajustado exclusivamente sobre o conjunto de treinamento e, em seguida, utilizado para transformar os dois conjuntos. Foi utilizada a opção `drop='first'`, retirando uma categoria de referência de cada variável categórica. Também foi utilizada a opção `handle_unknown='ignore'`, permitindo que categorias eventualmente presentes no teste e ausentes no treinamento fossem transformadas sem que o codificador precisasse ser reajustado.

As variáveis numéricas foram mantidas em sua forma numérica e reunidas às colunas geradas pelo *One-Hot Encoding*. Dessa forma, as representações de treinamento e teste passaram a possuir as mesmas colunas, definidas a partir das informações disponíveis no treinamento. O conjunto de teste não foi utilizado para definir categorias nem para ajustar qualquer parâmetro dessa etapa.

Após a codificação, foi aplicada a padronização por Escore-Z (*Z-Score Standardization*) por meio da classe `StandardScaler`. Nos Cenários 1 e 3, o estimador foi ajustado exclusivamente sobre o conjunto de treinamento codificado e, posteriormente, utilizado para transformar os dados de treinamento e teste. A transformação produz variáveis centradas em torno de média nula e escaladas em unidades de desvio padrão, conforme a Equação 4.1:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{4.1}$$

onde $x_i$ é o valor original do atributo, $\mu$ é a média aritmética da variável no conjunto de treinamento e $\sigma$ é o seu desvio padrão no mesmo conjunto.

No Cenário 2, foi adotada uma ordem específica para preservar a separação entre treinamento e teste durante a seleção de atributos. Inicialmente, o *One-Hot Encoding* foi ajustado somente no treinamento. Em seguida, a seleção de atributos por *Random Forest* foi realizada sobre o conjunto de treinamento codificado e ainda não padronizado, utilizando a variável-alvo apenas desse conjunto. As mesmas colunas selecionadas foram então aplicadas ao conjunto de teste. Somente após essa etapa, um novo `StandardScaler` foi ajustado sobre os atributos selecionados do treinamento e utilizado para transformar os dois conjuntos. O procedimento de seleção é detalhado posteriormente, juntamente com as configurações dos modelos.

Assim, em todos os cenários, os parâmetros aprendidos durante o pré-processamento foram estimados exclusivamente a partir do conjunto de treinamento. O conjunto de teste recebeu apenas as transformações definidas a partir do treinamento, não participando do ajuste do *One-Hot Encoding*, da padronização ou da seleção supervisionada de atributos. Esse procedimento reduz o risco de vazamento de dados (*data leakage*) entre as etapas de treinamento e avaliação (36).

## 4.3 SEPARAÇÃO DE DADOS E MITIGAÇÃO DO DESEQUILÍBRIO DAS CLASSES

Neste estudo, a partição dos conjuntos de dados foi realizada antes da aplicação de qualquer técnica de balanceamento, adotando-se a proporção de 80% para treinamento e 20% para teste (*hold-out validation*). A divisão foi efetuada por meio de amostragem estratificada, utilizando-se o parâmetro `stratify` com base na variável-alvo, de forma a preservar a proporção entre as classes. O procedimento foi repetido com 30 sementes predefinidas, correspondentes aos valores de 42 a 71. Dessa forma, para cada população foram obtidas 30 divisões estratificadas distintas entre treinamento e teste.

Dentro de cada combinação de população e semente, os mesmos registros de treinamento e teste foram mantidos entre os Cenários 1, 2 e 3 e entre as configurações comparáveis de balanceamento e classificação. Assim, as diferenças observadas entre essas configurações não decorreram da utilização de amostras distintas dentro da mesma repetição. As 30 divisões constituem repetições de *hold-out* estratificado e não correspondem a um procedimento de validação cruzada. Entre sementes diferentes, os conjuntos de treinamento e teste também se modificaram.

O balanceamento foi realizado somente após essa separação e exclusivamente sobre o conjunto de treinamento. O conjunto de teste permaneceu isolado durante essa etapa e não foi submetido a reamostragem. Nos procedimentos estocásticos de balanceamento, a semente correspondente à respectiva repetição foi utilizada como `random_state`, permitindo a reprodução das operações realizadas. Essa separação foi mantida nas 30 repetições do experimento.

Como já foi mencionado, um desafio inerente aos dados de recursos humanos, especificamente na predição de *turnover*, é o agudo desequilíbrio de classes, onde o número de colaboradores que permanecem na empresa excede consideravelmente o número daqueles que saem. Para avaliar diferentes formas de tratamento desse problema, foram consideradas cinco condições de treinamento: uma condição sem reamostragem dos dados, denominada *Desbalanceado*, e quatro condições com reamostragem, correspondentes aos métodos SMOTE, ADASYN, SMOTE-Tomek e SMOTE-ENN. Na condição desbalanceada, foram empregados mecanismos internos de ponderação apenas nos classificadores que os utilizaram na implementação adotada.

Na condição denominada *Desbalanceado*, nenhuma técnica de reamostragem foi

aplicada ao conjunto de treinamento. Entretanto, o tratamento do desequilíbrio não foi idêntico entre os seis classificadores. Para Random Forest e Regressão Logística, foi utilizado o parâmetro `class_weight='balanced'`, por meio do qual os pesos das classes são definidos automaticamente de forma inversamente proporcional às respectivas frequências no conjunto de treinamento. Para uma classe $j$, esse procedimento pode ser representado pela Equação 4.2:

$$w_j = \frac{N}{K \times N_j} \tag{4.2}$$

em que $N$ representa o número total de observações do conjunto de treinamento, $K$ corresponde ao número de classes e $N_j$ representa a quantidade de observações pertencentes à classe $j$.

No XGBoost, a compensação foi realizada por meio do parâmetro `scale_pos_weight`, calculado a partir da razão entre a quantidade de observações da classe negativa e da classe positiva no conjunto de treinamento, conforme a Equação 4.3:

$$w_{positivo} = \frac{N_{negativo}}{N_{positivo}} \tag{4.3}$$

em que $N_{negativo}$ e $N_{positivo}$ representam, respectivamente, as quantidades de observações das classes negativa e positiva no conjunto de treinamento.

Por outro lado, Naive Bayes, AdaBoost e LDA foram ajustados, nessa condição, sem mecanismo equivalente de ponderação interna das classes. Assim, a denominação *Desbalanceado* indica a ausência de reamostragem dos dados de treinamento, mas não representa exatamente a mesma estratégia de compensação do desequilíbrio para todos os classificadores.

Além da condição sem reamostragem descrita anteriormente, foram avaliadas técnicas de reamostragem no nível dos dados (*Data-level approaches*), executadas exclusivamente sobre o conjunto de treinamento após a divisão entre treinamento e teste. Empregou-se o método SMOTE (*Synthetic Minority Over-sampling Technique*) (13), que sintetiza novas instâncias interpolando pontos vizinhos da classe minoritária no espaço de características, conforme a Equação 4.4:

$$x_{novo} = x_i + \lambda \times (x_{zi} - x_i) \tag{4.4}$$

onde $x_i$ é uma instância da classe minoritária, $x_{zi}$ é um de seus vizinhos mais próximos (*k-nearest neighbors*), e $\lambda$ é um valor aleatório pertencente ao intervalo entre 0 e 1, que determina a posição da nova instância ao longo do segmento entre $x_i$ e $x_{zi}$. Adicionalmente ao SMOTE, testou-se o ADASYN (*Adaptive Synthetic Sampling Approach*) (26), que

direciona a geração de dados sintéticos para as regiões de fronteira de decisão mais complexas.

Além disso, empregaram-se abordagens híbridas que combinam o SMOTE com técnicas de subamostragem (*undersampling*) para a remoção de ruídos. Foram implementadas as variantes SMOTE-Tomek, que remove pares de instâncias de classes distintas que são vizinhas mais próximas entre si, e o SMOTE-ENN (*Edited Nearest Neighbours*), que elimina instâncias cuja classe difere da maioria de seus vizinhos (4).

Após as etapas de transformação previstas em cada cenário, incluindo a codificação das variáveis categóricas e a padronização dos preditores, as técnicas de reamostragem foram aplicadas exclusivamente ao conjunto de treinamento. Como resultado, a interpolação realizada pelo SMOTE e ADASYN, mesmo nas fases de sobreamostragem das estratégias híbridas, aconteceu no espaço numérico codificado. Esse processo pode gerar valores intermediários em colunas derivadas de variáveis categóricas, os quais não se relacionam diretamente com as categorias observadas originalmente. Assim, as instâncias sintéticas devem ser vistas como pontos numéricos auxiliares para ajustar as fronteiras de decisão, em vez de perfis de funcionários com interpretação organizacional direta. Essa consideração não se aplica ao conjunto de teste, que continuou sendo composto apenas por observações originais e não passou por reamostragem.

Para a avaliação do ajuste dos modelos ao conjunto de treinamento, o F1-score de treinamento foi calculado sobre `X_train_final` e `y_train_final`, isto é, sobre os dados efetivamente utilizados para o ajuste de cada classificador. Na condição desbalanceada, esses conjuntos correspondem aos dados originais de treinamento após a padronização. Nas condições submetidas a técnicas de balanceamento, entretanto, correspondem ao conjunto resultante da reamostragem, podendo incluir instâncias sintéticas geradas pelos métodos de *oversampling* e, nas abordagens híbridas, também a remoção de observações. O conjunto de teste, por sua vez, não passou por nenhum procedimento de reamostragem, preservando a distribuição observada após a *hold-out validation*. Consequentemente, nas condições balanceadas, a diferença entre o F1-score de treinamento e o F1-score de teste deve ser interpretada com cautela, uma vez que essas métricas são calculadas sobre conjuntos com composições e distribuições de classe distintas.

## 4.4 IMPLEMENTAÇÃO DOS MODELOS DE CLASSIFICAÇÃO E AVALIAÇÃO DE DESEMPENHO

A experimentação preditiva se baseou no treinamento de seis algoritmos de aprendizado de máquina, selecionados por cobrirem distintos paradigmas matemáticos: modelos baseados em árvores e *ensembles* (*Random Forest*, XGBoost, *AdaBoost*), probabilísticos (*Naive Bayes* Gaussiano), equações lineares (*Logistic Regression*) e discriminantes (*Linear Discriminant Analysis* - LDA).

O *Random Forest* (11) foi empregado tanto como um dos algoritmos de classificação avaliados quanto como método de apoio à seleção de atributos no Cenário 2. Para cada combinação de população e semente, foi ajustado um `RandomForestClassifier` com `n_estimators`=100 e `random_state` correspondente à semente da repetição. O modelo foi ajustado exclusivamente sobre o conjunto de treinamento após o *One-Hot Encoding* e antes da padronização e do balanceamento. A seleção foi realizada a partir de `feature_importances_`, mantendo-se as *features* com importância estritamente superior a 2%. Esse procedimento foi executado uma única vez para cada população e semente, e o mesmo conjunto de atributos selecionados foi utilizado nos cinco tratamentos de classe e nos seis classificadores daquela repetição.

As parametrizações dos classificadores variaram conforme a utilização ou não de reamostragem. Na condição sem reamostragem, denominada Desbalanceado, *Random Forest* e *Logistic Regression* utilizaram ponderação interna das classes por meio de `class_weight=balanced`. No XGBoost, foi utilizado `scale_pos_weight`, calculado a partir da razão entre o número de observações negativas e positivas no conjunto de treinamento. Nas condições submetidas a SMOTE, ADASYN, SMOTE-Tomek ou SMOTE-ENN, essas compensações internas não foram utilizadas, uma vez que o treinamento ocorreu sobre os conjuntos resultantes da reamostragem. A semente de cada repetição foi propagada aos algoritmos e procedimentos estocásticos que disponibilizavam o parâmetro `random_state`. Não foi realizada busca sistemática de hiperparâmetros. A Tabela 6 apresenta os parâmetros explicitamente definidos no procedimento experimental.

Tabela 6 – Configurações e hiperparâmetros dos algoritmos de classificação.

| **Algoritmo** | **Condição sem reamostragem** | **Condições com reamostragem** |
|---|---|---|
| *Random Forest* | `random_state`=seed; `class_weight=balanced` | `random_state`=seed |
| *Logistic Regression* | `random_state`=seed; `max_iter`=1000; `class_weight=balanced` | `random_state`=seed; `max_iter`=1000 |
| *XGBoost Classifier* | `random_state`=seed; `eval_metric=logloss`; `scale_pos_weight`=$w_{positivo}$ | `random_state`=seed; `eval_metric=logloss` |
| *AdaBoost Classifier* | `random_state`=seed | `random_state`=seed |
| *Naive Bayes (Gaussian)* | Demais parâmetros conforme valores padrão | Demais parâmetros conforme valores padrão |
| *Linear Discriminant Analysis* | Parâmetros conforme valores padrão | Parâmetros conforme valores padrão |
| **Divisão dos dados** | `test_size`=0.2; `stratify`=y; `random_state`=seed, com sementes de 42 a 71 | |
| **Seleção de atributos no C2** | `RandomForestClassifier`; `n_estimators`=100; `random_state`=seed; importância > 0,02 | |

Fonte: Elaborado pelo autor (2026).

Considerando as 7 populações do Cenário 1, as 7 representações do Cenário 2 e as 13 representações do Cenário 3, foram avaliadas 27 representações experimentais em cada semente. Para cada uma delas, foram combinadas as cinco condições de treinamento e os seis classificadores, totalizando 810 execuções por semente. Com as 30 sementes, o experimento resultou em 24.300 execuções de modelos.

A avaliação de desempenho foi realizada sobre o conjunto de teste, que permaneceu sem reamostragem. Foram registradas a acurácia, a precisão da classe positiva, o *recall* da classe positiva, o F1-score da classe positiva, o Macro-F1, o Coeficiente de Correlação de Matthews (MCC), a AUC-ROC e o Log Loss. Também foram registrados os componentes

da matriz de confusão, correspondentes a verdadeiros negativos (TN), falsos positivos (FP), falsos negativos (FN) e verdadeiros positivos (TP). As métricas dependentes dos rótulos preditos foram calculadas com limiar de probabilidade fixo em 0,5, mantido em todas as configurações.

Adicionalmente, foi registrado o F1-score calculado sobre o conjunto efetivamente utilizado para o ajuste de cada classificador. Na condição sem reamostragem, esse conjunto corresponde ao treinamento original após as transformações previstas no respectivo cenário. Nas condições balanceadas, o valor é calculado sobre o conjunto resultante da reamostragem. Por esse motivo, esse indicador foi utilizado apenas como informação diagnóstica do ajuste e não como medida principal de desempenho fora da amostra.

A função de perda adotada para avaliar as predições probabilísticas foi a *Binary Cross-Entropy* (ou *Log Loss*), formulada segundo a Equação 4.5:

$$L(y, p) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{4.5}$$

onde $y_i$ é a classe real (0 ou 1) e $p_i$ é a probabilidade predita pelo modelo para a classe positiva.

O Coeficiente de Correlação de Matthews (MCC) (40) foi adotado como uma das principais métricas de avaliação do desempenho dos modelos, por considerar simultaneamente os quatro componentes da matriz de confusão. Seu valor varia entre -1 e +1. A Equação 4.6 descreve sua formulação:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.6}$$

Para avaliar a estabilidade dos resultados entre as 30 repetições, cada uma das 810 configurações experimentais foi resumida separadamente. Para cada métrica, foram calculados média, desvio-padrão amostral, mediana, primeiro quartil (Q1), terceiro quartil (Q3), intervalo interquartil (IQR), mínimo e máximo. Dessa forma, a comparação não ficou restrita ao resultado obtido em uma única divisão entre treinamento e teste.

As comparações entre representações e entre condições de balanceamento também foram realizadas de forma pareada. Para cada contraste, os dois resultados comparados correspondiam à mesma população, à mesma semente e, consequentemente, à mesma divisão entre treinamento e teste. Foram calculadas as diferenças entre as configurações em cada uma das 30 sementes e, posteriormente, sua média, desvio-padrão, mediana, Q1, Q3, IQR, mínimo e máximo. Também foram registradas as quantidades e os percentuais de sementes em que a primeira configuração apresentou resultado superior, equivalente ou inferior à segunda. Para o Log Loss, o sentido da diferença foi invertido na medida

orientada, de modo que valores positivos indicassem vantagem da primeira configuração, assim como nas métricas em que valores maiores representam melhor desempenho.

No Cenário 2, a estabilidade da própria seleção de atributos também foi registrada ao longo das 30 sementes. Foram resumidos o número de *features* selecionadas em cada população e a frequência com que cada *feature* apareceu nas diferentes repetições. Também foram registradas sua posição média e mediana entre os atributos selecionados.

Para consolidar o desempenho das configurações, foi utilizado um Índice Multivariado Ponderado ($IM$). As métricas MCC, F1-score da classe positiva, AUC-ROC e Log Loss foram inicialmente convertidas em Escores-Z ($Z$). Essa transformação foi realizada separadamente para cada semente e para cada contexto comparável, definido pela combinação de *dataset*, população, cenário e representação. Dentro de cada contexto foram comparadas 30 configurações, correspondentes às cinco condições de treinamento combinadas com os seis classificadores.

O Índice Multivariado Ponderado foi calculado conforme a Equação 4.7, atribuindo maior peso ao MCC e ao F1-score e penalizando valores elevados de Log Loss:

$$IM = 0.40 \times Z_{MCC} + 0.30 \times Z_{F1} + 0.20 \times Z_{AUC-ROC} - 0.10 \times Z_{LogLoss} \tag{4.7}$$

Em cada semente, as configurações foram ordenadas de forma decrescente pelo valor do $IM$. Como o índice foi calculado separadamente nas 30 sementes, cada configuração passou a possuir 30 valores de $IM$ e 30 posições no ranking. Para avaliar sua estabilidade, foram calculados média, desvio-padrão, mediana, Q1, Q3, IQR, mínimo e máximo do $IM$, além do *rank* médio e mediano, melhor e pior posição e frequência de ocorrência na primeira posição, entre as três primeiras e entre as cinco primeiras posições.

A configuração de referência de cada uma das 27 representações foi definida inicialmente pelo maior valor médio do Índice Multivariado nas 30 sementes. Em caso de empate, foram utilizados, nessa ordem, o menor *rank* médio, o maior MCC médio e o maior F1-score médio. O índice permaneceu restrito a cada contexto de comparação e não foi utilizado para produzir um ranking global entre diferentes *datasets*, populações, cenários ou representações.

Por utilizar as métricas do próprio conjunto de teste no cálculo do Índice Multivariado e na escolha das configurações de referência, não foi reservado um segundo conjunto de teste independente após essa seleção. Dessa forma, o ranking foi utilizado para organizar de forma descritiva as configurações avaliadas e deve ser interpretado em conjunto com as métricas originais e com a estabilidade observada nas 30 repetições.

A análise de interpretabilidade por SHAP foi realizada somente nas populações completas e no Cenário 1. Para cada *dataset*, foi utilizada a configuração de referência obtida a partir do procedimento descrito anteriormente. As configurações analisadas foram

LDA com a condição Desbalanceado no Dataset 1, XGBoost com SMOTE-Tomek no Dataset 2 e AdaBoost com SMOTE-ENN no Dataset 3.

Como cada configuração possuía resultados para 30 sementes, a execução utilizada no SHAP não foi escolhida pelo maior desempenho individual. Foi selecionada a semente mais próxima do comportamento mediano da configuração nas quatro métricas utilizadas pelo Índice Multivariado: F1-score, MCC, AUC-ROC e Log Loss. Para isso, foi calculada a distância entre o resultado de cada semente e a mediana das 30 repetições em cada métrica, dividindo-se cada diferença pelo respectivo desvio-padrão. A menor distância multivariada indicou a execução considerada mais representativa. Em caso de empate, foi utilizada a menor semente. Esse procedimento resultou nas sementes 54 para o Dataset 1, 58 para o Dataset 2 e 45 para o Dataset 3.

Antes da aplicação do SHAP, os três modelos selecionados foram reconstruídos a partir dos dados originais, repetindo o mesmo *split*, pré-processamento, balanceamento e treinamento da execução correspondente. A reprodução foi aceita somente quando as métricas de desempenho, os componentes da matriz de confusão e as *features* utilizadas coincidiram com os registros da execução original.

A análise utilizou a versão 0.52.0 da biblioteca SHAP, por meio de `shap.Explainer`, com algoritmo de permutação e `shap.maskers.Independent`. As explicações foram calculadas sobre a probabilidade predita da classe positiva. Como conjunto de referência (*background*), foram utilizadas 50 observações estratificadas da representação de treinamento já transformada, antes da reamostragem. As observações explicadas foram retiradas exclusivamente do conjunto de teste. No Dataset 1, foram utilizadas todas as 294 observações do teste. Nos Datasets 2 e 3, foram utilizadas amostras estratificadas e determinísticas de 400 observações. O número máximo de avaliações utilizado pelo método de permutação foi definido como $3 \times (2p + 1)$, em que $p$ representa o número de *features* transformadas.

Após o cálculo, foi verificada a propriedade de aditividade dos valores SHAP, comparando-se a soma do valor de referência e das contribuições dos atributos com a probabilidade produzida pelo modelo. Para a análise da importância global, as contribuições das colunas indicadoras originadas de uma mesma variável categórica foram somadas para cada indivíduo antes do cálculo da importância do atributo original. Em seguida, a importância global foi obtida pela média do valor absoluto dessas contribuições.

Foram considerados até dez atributos originais com maior contribuição global em cada *dataset*; no Dataset 2, que possuía nove atributos preditores originais, todos foram considerados. Os gráficos do tipo *beeswarm* foram apresentados no nível das *features* transformadas, permitindo observar o sinal das contribuições e os valores associados a cada *feature*. Para atributos numéricos com mais de cinco valores distintos, a análise descritiva da direção das contribuições incluiu correlação de Spearman, comparação entre os quartis inferior e superior e agrupamento em quintis. Para atributos categóricos ou discretos,

foram calculadas a média, a mediana e a média absoluta dos valores SHAP em cada nível.

Os valores SHAP foram utilizados exclusivamente para descrever como os atributos contribuíram para as predições dos modelos analisados. Portanto, uma contribuição positiva ou negativa para a predição não foi interpretada como evidência de que o atributo cause ou evite o *turnover*.

Por fim, a execução foi estruturada de forma reprodutível e retomável. Cada combinação de semente, população, cenário, representação, condição de treinamento e classificador recebeu um identificador único de execução. Os resultados foram salvos por meio de *checkpoints* durante o processamento e posteriormente consolidados por semente e em um arquivo final. Também foram registrados os identificadores dos conjuntos de treinamento e teste, as *features* utilizadas, os tamanhos das amostras, as matrizes de confusão e as versões principais do ambiente computacional. O arquivo consolidado apresentou 24.300 identificadores de execução distintos, correspondentes às 30 repetições das 810 configurações experimentais.

## 4.5 USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA COMO APOIO À ANÁLISE DOS RESULTADOS

Durante a etapa de análise e consolidação dos resultados, especialmente para a elaboração do Capítulo 5, foi utilizada a ferramenta de inteligência artificial generativa ChatGPT, desenvolvida pela OpenAI, como recurso auxiliar na análise dos resultados obtidos e na organização de sua forma de apresentação. A ferramenta também foi empregada como apoio às análises exploratórias complementares conduzidas para investigar particularidades observadas nos *Datasets* 2 e 3, a partir dos dados, planilhas e resultados produzidos no desenvolvimento da pesquisa. As evidências, cálculos e interpretações utilizados no trabalho foram posteriormente verificados pelo autor em relação às respectivas fontes primárias. A ferramenta de inteligência artificial generativa não foi utilizada no treinamento dos modelos nem na execução das etapas computacionais do experimento.

# 5 ANÁLISE DOS RESULTADOS E DISCUSSÃO

Este capítulo apresenta e discute os resultados dos experimentos de previsão de turnover a partir de três eixos: os efeitos do balanceamento, o desempenho dos modelos nos diferentes cenários e a seleção das configurações com melhor desempenho observado. Foram realizadas 30 repetições com diferentes sementes, permitindo analisar a estabilidade dos resultados entre diferentes divisões estratificadas de treinamento e teste. A análise também incorpora a interpretação das contribuições dos atributos para as predições por meio do SHAP. A classe positiva corresponde à ocorrência registrada de *attrition*, `left` ou saída do funcionário nos *datasets*, sem diferenciar desligamentos voluntários de involuntários. A interpretação dos resultados considera as diferentes métricas de forma conjunta, evitando conclusões baseadas em um único indicador de desempenho.

## 5.1 RESULTADOS OBTIDOS COM A APLICAÇÃO DE MÉTODOS DE BALANCEAMENTO NOS CONJUNTOS DE TREINAMENTO

### 5.1.1 Distribuição das classes e impacto do balanceamento nas métricas

A divisão dos conjuntos de treino e teste preservou, nos conjuntos de treinamento, a predominância da classe 0. Nas bases completas do Cenário 1, a classe positiva representou 16,16% das 1.176 instâncias de treinamento do *Dataset* 1, 16,61% das 9.592 instâncias do *Dataset* 2 e 19,98% das 8.000 instâncias do *Dataset* 3. Como essas quantidades são determinadas pela divisão estratificada, permaneceram constantes nas 30 repetições. A Tabela 7 sintetiza a composição dos conjuntos de treinamento após a aplicação das quatro técnicas de reamostragem. Nos procedimentos cujo resultado variou entre sementes, são apresentadas a média e a amplitude observada nas 30 repetições.

O método SMOTE produziu igualdade numérica entre as classes nos três conjuntos em todas as repetições. O SMOTE-Tomek também manteve proporções de 50%, embora o número de observações remanescentes variasse ligeiramente entre as sementes em razão da etapa de limpeza. O ADASYN aproximou as distribuições do equilíbrio, com pequenas variações na quantidade de instâncias sintéticas geradas. O SMOTE-ENN apresentou o comportamento mais distinto: a proporção média da classe positiva foi de 66,56% no *Dataset* 1, 52,16% no *Dataset* 2 e 80,86% no *Dataset* 3.

A diferença entre as técnicas não se limita ao percentual final das classes. SMOTE e ADASYN aumentam a presença da classe minoritária por geração de instâncias sintéticas, enquanto SMOTE-Tomek e SMOTE-ENN combinam superamostragem com procedimentos de limpeza. Por isso, duas técnicas podem produzir proporções semelhantes e, ainda assim, entregar aos classificadores conjuntos de treinamento com tamanhos e fronteiras locais diferentes. O caso do SMOTE-ENN é particularmente expressivo no *Dataset* 3: o conjunto resultante apresentou, em média, 7.546,9 instâncias, variando de 7.467 a 7.618, enquanto

Tabela 7 – Distribuição das classes nos conjuntos de treinamento dos *datasets* completos do Cenário 1 ao longo das 30 repetições.

| Dataset | Procedimento | Classe 0 | Classe 0 (%) | Classe 1 | Classe 1 (%) | Total |
|---|---|---|---|---|---|---|
| *Dataset* 1 | Desbalanceado | 986 | 83,84 | 190 | 16,16 | 1.176 |
| *Dataset* 1 | SMOTE-Tomek | 985,7 [984–986] | 50,00 | 985,7 [984–986] | 50,00 | 1.971,3 [1.968–1.972] |
| *Dataset* 1 | SMOTE-ENN | 485,5 [453–512] | 33,44 [31,77–34,81] | 965,8 [955–975] | 66,56 [65,19–68,23] | 1.451,3 [1.426–1.480] |
| *Dataset* 1 | SMOTE | 986 | 50,00 | 986 | 50,00 | 1.972 |
| *Dataset* 1 | ADASYN | 986 | 50,90 [49,30–51,87] | 951,6 [915–1.014] | 49,10 [48,13–50,70] | 1.937,6 [1.901–2.000] |
| *Dataset* 2 | Desbalanceado | 7.999 | 83,39 | 1.593 | 16,61 | 9.592 |
| *Dataset* 2 | SMOTE-Tomek | 7.966,1 [7.953–7.979] | 50,00 | 7.966,1 [7.953–7.979] | 50,00 | 15.932,1 [15.906–15.958] |
| *Dataset* 2 | SMOTE-ENN | 7.071,3 [7.022–7.120] | 47,84 [47,71–48,00] | 7.709,9 [7.666–7.739] | 52,16 [52,00–52,29] | 14.781,2 [14.713–14.851] |
| *Dataset* 2 | SMOTE | 7.999 | 50,00 | 7.999 | 50,00 | 15.998 |
| *Dataset* 2 | ADASYN | 7.999 | 49,92 [49,58–50,49] | 8.023,6 [7.843–8.135] | 50,08 [49,51–50,42] | 16.022,6 [15.842–16.134] |
| *Dataset* 3 | Desbalanceado | 6.402 | 80,03 | 1.598 | 19,98 | 8.000 |
| *Dataset* 3 | SMOTE-Tomek | 6.395,6 [6.391–6.401] | 50,00 | 6.395,6 [6.391–6.401] | 50,00 | 12.791,2 [12.782–12.802] |
| *Dataset* 3 | SMOTE-ENN | 1.444,7 [1.367–1.523] | 19,14 [18,28–19,99] | 6.102,2 [6.073–6.131] | 80,86 [80,01–81,72] | 7.546,9 [7.467–7.618] |
| *Dataset* 3 | SMOTE | 6.402 | 50,00 | 6.402 | 50,00 | 12.804 |
| *Dataset* 3 | ADASYN | 6.402 | 49,81 [49,62–50,05] | 6.451,1 [6.389–6.499] | 50,19 [49,95–50,38] | 12.853,1 [12.791–12.901] |

Valores entre colchetes representam o mínimo e o máximo observados nas 30 repetições; quando não há colchetes, o valor foi idêntico em todas as sementes.

Fonte: Elaborado pelo autor (2026).

a classe positiva representou, em média, 80,86% dos registros. Essa alteração mostra que o balanceamento não é apenas uma correção numérica do desequilíbrio, mas também uma transformação da amostra utilizada no ajuste dos modelos.

É importante observar que essa mudança aconteceu na representação numérica codificada dos dados. Por esse motivo, as distribuições mostradas na Tabela 7 representam a composição computacional dos conjuntos usados no treinamento, mas não sugerem que todas as instâncias sintéticas sejam combinações categóricas diretamente identificáveis no âmbito de recursos humanos. A análise a seguir examina o impacto dessas representações no comportamento preditivo dos modelos no conjunto de teste original.

Os *Datasets* 1 e 2 apresentavam cerca de 16% de registros da classe positiva no conjunto de treinamento, enquanto o *Dataset* 3 apresentava aproximadamente 20%. Como essas proporções são relativamente próximas, o desbalanceamento inicial, isoladamente, não explica as diferenças de desempenho observadas posteriormente entre as bases. Apesar de partirem de distribuições de classe semelhantes, os três datasets apresentaram comportamentos preditivos distintos, aspecto retomado nas análises seguintes.

O efeito sobre as métricas foi avaliado por comparações pareadas entre cada configuração reamostrada e a respectiva condição Desbalanceado, mantendo-se a mesma semente, o mesmo *dataset*, população, cenário, representação e modelo. Para cada técnica de reamostragem, foram consideradas 162 configurações comparáveis, cada uma avaliada nas mesmas 30 sementes da condição de referência, totalizando 4.860 diferenças pareadas por técnica. A Tabela 8 apresenta a média dessas diferenças no conjunto de teste.

Um padrão relevante observado foi o aumento do *recall* da classe positiva, acompanhado por ganhos médios de F1-score. SMOTE e SMOTE-Tomek apresentaram resultados

Tabela 8 – Variação média pareada das métricas de teste após a aplicação das técnicas de balanceamento em relação à condição Desbalanceado nas 30 sementes.

| Técnica | Δ Acurácia | Δ Precisão | Δ Recall | Δ F1-score | Δ MCC | Δ AUC-ROC | Δ Log Loss |
|---|---|---|---|---|---|---|---|
| SMOTE-Tomek | -7,93 p.p. | +0,81 p.p. | +18,37 p.p. | +8,75 p.p. | +0,013 | +0,001 | +0,125 |
| SMOTE-ENN | -24,22 p.p. | -3,21 p.p. | +46,53 p.p. | +13,73 p.p. | +0,003 | -0,002 | +0,424 |
| SMOTE | -7,93 p.p. | +0,91 p.p. | +18,29 p.p. | +8,69 p.p. | +0,013 | +0,001 | +0,122 |
| ADASYN | -8,40 p.p. | +0,16 p.p. | +18,31 p.p. | +8,18 p.p. | +0,006 | -0,002 | +0,113 |

Δ representa a diferença entre o valor da métrica após a aplicação da técnica de balanceamento e o valor obtido na respectiva condição Desbalanceado. Cada média considera 4.860 diferenças pareadas, correspondentes a 162 configurações avaliadas nas mesmas 30 sementes.

Fonte: Elaborado pelo autor (2026).

muito próximos, com acréscimos de aproximadamente 18,3 pontos percentuais no *recall* e 8,7 pontos no F1-score. O SMOTE-ENN apresentou a alteração mais intensa, elevando o *recall* em 46,53 pontos percentuais e o F1-score em 13,73 pontos, mas reduzindo a acurácia em 24,22 pontos e a precisão em 3,21 pontos, além de elevar o Log Loss em 0,424. O ADASYN também aumentou o *recall* e o F1-score, enquanto apresentou variações pequenas no MCC e na AUC-ROC.

A direção do efeito sobre o *recall* também se mostrou estável entre as repetições. Considerando, para cada uma das 162 configurações, a proporção das 30 sementes em que a técnica apresentou *recall* superior à condição Desbalanceado, a mediana dessa frequência foi de 100% para as quatro técnicas. Para o F1-score, as medianas foram de 73,33% para SMOTE-Tomek, 100% para SMOTE-ENN, 73,33% para SMOTE e 68,33% para ADASYN. Assim, o aumento da recuperação da classe positiva foi mais consistente entre as sementes do que os efeitos observados em MCC e AUC-ROC, cujas variações médias permaneceram próximas de zero.

É importante observar que a reamostragem altera a prevalência observada durante o treinamento, fazendo com que as probabilidades estimadas não reflitam mais a frequência da classe positiva no conjunto de teste. Dessa forma, o aumento do Log Loss pode estar ligado tanto a falhas probabilísticas quanto à falta de recalibração após o balanceamento. Este estudo não incluiu uma etapa de calibração específica, portanto, as probabilidades devem ser entendidas como escores gerados pelos modelos, e não como riscos organizacionais calibrados.

Na condição Desbalanceado, considerando conjuntamente as 162 configurações comparáveis e suas 30 repetições, as médias de teste foram 78,32% de acurácia, 32,33% de precisão, 28,72% de *recall* e 26,01% de F1-score; MCC e AUC-ROC foram 0,196 e 0,660, respectivamente. Esses valores fornecem a escala sobre a qual os deltas devem ser interpretados. Um acréscimo de aproximadamente 18 pontos percentuais no *recall*, por exemplo, representa uma alteração relevante na quantidade de casos positivos recuperados, enquanto variações próximas de 0,001 ou 0,002 na AUC-ROC são pequenas em comparação.

Por isso, a melhora em uma métrica não significa necessariamente melhora equivalente do desempenho global.

Esses resultados evitam que a reamostragem seja vista como uma melhoria consistente na capacidade preditiva. No pipeline empregado, os conjuntos reamostrados numericamente apresentaram maior detecção de casos positivos, porém também resultaram em diminuição da acurácia e aumento do Log Loss. Adicionalmente, como as instâncias foram geradas após a codificação das variáveis categóricas, os resultados refletem o efeito computacional da reamostragem no espaço transformado, e não indicam que as amostras sintéticas correspondem a perfis organizacionais semanticamente realistas. Assim, os ganhos observados devem ser limitados às bases, representações, classificadores e procedimentos que foram efetivamente avaliados.

Diante disso, a escolha da técnica de balanceamento deve considerar, em conjunto, o modelo preditivo utilizado e as características da base de dados. O balanceamento altera a distribuição, mas a forma como isso se converte em decisão depende do modelo. Por isso, a análise dos melhores resultados preserva a combinação completa Modelo + Balanceamento, em vez de selecionar primeiro uma técnica de reamostragem e aplicá-la para todos os classificadores.

### 5.1.2 Heterogeneidade dos efeitos entre os datasets

A média agregada apresentada anteriormente pode ocultar diferenças relevantes entre as bases. Para examinar especificamente essa heterogeneidade sem combinar os resultados das populações completas com os subgrupos por sexo, a Tabela 9 apresenta a variação do F1-score da classe positiva somente nas populações completas. As comparações foram realizadas de forma pareada em relação à condição Desbalanceado, mantendo-se a mesma semente, cenário, representação e modelo. Para cada técnica, foram consideradas 24 configurações no *Dataset* 1, 18 no *Dataset* 2 e 24 no *Dataset* 3, cada uma avaliada nas mesmas 30 sementes.

Tabela 9 – Variação média pareada do F1-score da classe positiva por técnica de balanceamento e *dataset* nas populações completas.

| **Dataset** | **SMOTE-Tomek** | **SMOTE-ENN** | **SMOTE** | **ADASYN** |
|---|---|---|---|---|
| *Dataset* 1 | +7,22 p.p. | +6,90 p.p. | +7,11 p.p. | +6,55 p.p. |
| *Dataset* 2 | +3,47 p.p. | +3,00 p.p. | +3,40 p.p. | +0,39 p.p. |
| *Dataset* 3 | +11,09 p.p. | +22,78 p.p. | +11,13 p.p. | +11,37 p.p. |

Os valores representam a média das diferenças pareadas entre cada condição reamostrada e a respectiva condição Desbalanceado nas 30 sementes. Foram consideradas 24 configurações comparáveis nos *Datasets* 1 e 3 e 18 no *Dataset* 2.

Fonte: Elaborado pelo autor (2026).

No *Dataset* 1, as quatro técnicas apresentaram aumentos médios de F1-score re-

lativamente próximos, entre 6,55 e 7,22 pontos percentuais. Entretanto, o efeito não foi uniforme entre as configurações: o delta médio de F1-score foi positivo em 18 das 24 configurações para SMOTE e SMOTE-Tomek, em 15 para SMOTE-ENN e em 13 para ADASYN. No *Dataset* 2, os ganhos médios foram menores. SMOTE-Tomek, SMOTE e SMOTE-ENN aumentaram o F1-score em 3,47, 3,40 e 3,00 pontos percentuais, respectivamente, enquanto o ADASYN apresentou variação média de apenas +0,39 ponto.

O comportamento do ADASYN no *Dataset* 2 evidencia a importância de interpretar as métricas conjuntamente. Embora o *recall* tenha aumentado, em média, 15,62 pontos percentuais, a precisão diminuiu 11,37 pontos, o MCC variou -0,011 e a AUC-ROC -0,005. Além disso, somente 6 das 18 configurações apresentaram aumento médio de F1-score ao longo das 30 sementes. Assim, a pequena variação positiva observada no F1-score agregado não caracteriza um benefício consistente da técnica nessa base.

O efeito mais expressivo ocorreu no *Dataset* 3. O SMOTE-ENN aumentou o F1-score médio em 22,78 pontos percentuais e o *recall* em 66,40 pontos. Em contraste, as variações médias no MCC e na AUC-ROC foram de apenas +0,003 e +0,005, respectivamente. O aumento do F1-score foi observado, em média, em 23 das 24 configurações avaliadas, mostrando que o deslocamento em direção à identificação da classe positiva foi recorrente nas diferentes combinações de cenário, representação e modelo. A magnitude muito menor das alterações de MCC e AUC-ROC, porém, indica que esse ganho não foi acompanhado por aumento equivalente da capacidade discriminativa.

A magnitude dos efeitos no *Dataset* 3 é útil para mostrar por que o F1-score não pode ser isolado das demais medidas. Quando a técnica desloca fortemente a decisão para a classe positiva, o *recall* pode aumentar mesmo em um problema cuja ordenação probabilística entre classes permanece fraca. O MCC e a AUC-ROC fornecem uma contraposição necessária a esse ganho. No *Dataset* 3, as alterações muito pequenas nessas métricas, em comparação com o aumento de F1-score e *recall*, mostram que a maior identificação da classe positiva não é suficiente para afirmar uma melhora equivalente da capacidade preditiva geral dos modelos.

No *Dataset* 2, o desempenho de referência já é elevado, e as técnicas de reamostragem produziram alterações médias menores de F1-score. O ADASYN ilustra particularmente esse comportamento: o aumento do *recall* foi acompanhado por redução da precisão, do MCC e da AUC-ROC, resultando em apenas +0,39 ponto percentual de F1-score na média. Nesse contexto, ampliar a sensibilidade à classe positiva pode envolver perdas em outras dimensões do desempenho de um problema que já apresenta boa capacidade discriminativa. Portanto, a utilidade do balanceamento depende também das características preditivas da própria base e da métrica considerada.

Nas populações completas, também houve heterogeneidade entre os modelos preditivos. O Random Forest apresentou, em média, aumento de F1-score com as quatro técnicas,

enquanto a regressão logística mostrou alterações menores e redução com ADASYN. Para XGBoost, algumas técnicas reduziram ligeiramente o F1-score médio ao mesmo tempo em que elevaram acurácia, precisão ou AUC-ROC. Naive Bayes apresentou aumentos de *recall* e F1-score acompanhados por variações negativas de MCC e AUC-ROC. Portanto, os efeitos do balanceamento estão ligados à combinação entre *dataset*, modelo e métrica. Os resultados não sustentam a aplicação de uma única estratégia de reamostragem como solução geral para os contextos avaliados.

## 5.2 DESEMPENHO GERAL DOS MODELOS E IMPACTO NOS DIFERENTES CENÁRIOS

### 5.2.1 Desempenho global e impacto da seleção de atributos

A apresentação a seguir toma os cenários como primeira camada de leitura. O Cenário 1, que utiliza todos os atributos disponíveis após os procedimentos iniciais de preparação de cada base, fornece a referência para a comparação. Em seguida, os resultados do Cenário 2 e das duas representações do Cenário 3 são examinados em relação a essa referência. Dentro de cada cenário, as diferenças entre os *datasets* são discutidas sem pressupor equivalência de tamanho, estrutura ou dificuldade preditiva. No Cenário 1, cada *dataset* completo reúne 30 configurações, correspondentes à combinação dos seis modelos com as cinco condições de treinamento, e cada configuração foi avaliada nas 30 sementes. A Tabela 10 apresenta as médias das métricas obtidas no conjunto de teste ao longo dessas repetições.

Tabela 10 – Desempenho médio das configurações dos *datasets* completos no Cenário 1 considerando as 30 sementes.

| **Dataset** | **Acurácia (%)** | **Precisão (%)** | **Recall (%)** | **F1-score (%)** | **MCC** | **AUC-ROC** | **Log Loss** |
|---|---|---|---|---|---|---|---|
| *Dataset* 1 | 76,26 | 46,93 | 58,17 | 44,83 | 0,365 | 0,796 | 0,884 |
| *Dataset* 2 | 84,56 | 63,14 | 85,09 | 69,32 | 0,639 | 0,905 | 0,445 |
| *Dataset* 3 | 59,30 | 17,01 | 34,90 | 18,84 | 0,002 | 0,503 | 0,714 |

Cada valor corresponde à média das 30 configurações do respectivo *dataset*, avaliadas em 30 sementes, totalizando 900 resultados por base.

Fonte: Elaborado pelo autor (2026).

Os valores da Tabela 10 caracterizam o comportamento de cada base no Cenário 1 e não constituem um ranking global entre *datasets*, uma vez que eles diferem em tamanho, atributos e processo de construção. No *Dataset* 2, as médias de F1-score, MCC e AUC-ROC foram 69,32%, 0,639 e 0,905, respectivamente. Random Forest e XGBoost apresentaram desempenho elevado nessa base: considerando conjuntamente as cinco condições de treinamento e as 30 sementes, seus F1-scores médios foram 93,84% e 93,53%, com MCC de 0,927 e 0,923. O *Dataset* 1 apresentou F1-score médio de 44,83%, MCC de 0,365 e AUC-ROC de 0,796, enquanto o *Dataset* 3 apresentou F1-score médio de

18,84%, MCC próximo de zero e AUC-ROC de 0,503. Esses resultados mostram que o desempenho observado depende da combinação entre os dados disponíveis e o procedimento de modelagem, não podendo ser tratado como propriedade isolada do algoritmo.

No *Dataset* 1, a ordem dos modelos variou conforme a métrica considerada. O AdaBoost apresentou a maior média de F1-score, de 50,29%, e a maior média de MCC, de 0,433, enquanto o LDA apresentou a maior AUC-ROC média, de 0,820. A presença de resultados distintos conforme a métrica reforça que a identificação da configuração mais adequada depende do critério utilizado e que nenhuma medida isolada resume todas as dimensões do desempenho.

No *Dataset* 3, a Logistic Regression apresentou o maior F1-score médio entre os seis classificadores, de 29,44%, mas seu MCC médio foi de apenas 0,003 e a AUC-ROC média foi 0,501. Assim, mesmo o modelo com maior F1-score médio permaneceu com capacidade de discriminação próxima ao acaso segundo as demais métricas, mostrando que uma melhora relativa dentro da própria base não implica necessariamente forte separação entre as classes.

O Log Loss também não reproduziu necessariamente a mesma ordenação observada nas métricas baseadas em classificação ou discriminação. Como acurácia, F1-score, MCC, AUC-ROC e Log Loss avaliam dimensões distintas — decisão de classe, equilíbrio entre erros, associação entre rótulos e previsões, ordenação probabilística e qualidade das probabilidades estimadas — não se espera que todas conduzam à mesma avaliação das configurações. Sendo assim, nenhuma dessas métricas deve ser considerada isoladamente como medida suficiente de desempenho.

As bases não são equivalentes em tamanho, estrutura ou dimensionalidade. Foram utilizados 1.470 registros no *Dataset* 1, 11.991 no *Dataset* 2 após a remoção das duplicatas e 10.000 no *Dataset* 3. Após a codificação das variáveis categóricas, as matrizes completas continham 44, 18 e 30 colunas preditoras, respectivamente. Essas diferenças permitem comparar como os mesmos algoritmos se comportaram em contextos distintos, mas exigem cautela na comparação entre os *datasets*, pois as diferenças de desempenho também podem estar relacionadas às características de cada base.

A origem dos conjuntos de dados acrescenta contexto, mas não deve ser transformada em explicação causal. Os *Datasets* 1 e 2 são artificiais, enquanto o *Dataset* 3 é apresentado pelo disponibilizador como real sem documentação independente suficiente do processo de coleta. O melhor desempenho do *Dataset* 2 não demonstra que bases artificiais sejam mais previsíveis, assim como a baixa discriminação do *Dataset* 3 não pode ser atribuída ao fato de ser apresentado como real. O que se pode concluir é apenas que os atributos presentes em cada base parecem oferecer diferentes níveis de informação preditiva para os modelos empregados.

As diferenças entre as bases também permaneceram visíveis nas representações

reduzidas. No Cenário 2, o número de features selecionadas no *Dataset* 1 variou entre 21 e 24 ao longo das 30 sementes, com média de 22,33 e mediana de 22. No *Dataset* 2, foram selecionadas cinco features em todas as sementes, enquanto o *Dataset* 3 permaneceu com 19. A redução mais intensa observada no *Dataset* 2 não impediu que a base mantivesse desempenho elevado, enquanto o *Dataset* 3 continuou apresentando baixa capacidade discriminativa com 19 features. Dessa forma, a quantidade de atributos, isoladamente, não determina a dificuldade preditiva do problema.

No Cenário 2, considerando somente as populações completas e as 30 sementes, os F1-scores médios foram 43,61% no *Dataset* 1, 68,83% no *Dataset* 2 e 18,78% no *Dataset* 3. Os respectivos valores médios de MCC foram 0,347, 0,631 e -0,002, e as AUC-ROC foram 0,785, 0,902 e 0,498. Na representação de seis atributos do Cenário 3, os F1-scores médios foram 30,11%, 66,95% e 20,83%, com MCC de 0,168, 0,608 e -0,001 e AUC-ROC de 0,672, 0,890 e 0,497. Esses valores devem ser interpretados dentro de cada base, pois as representações e populações não são diretamente equivalentes entre os diferentes *datasets*.

A comparação dentro de cada *dataset* mostra que o efeito da redução de atributos não foi uniforme. No *Dataset* 2, as representações reduzidas conservaram grande parte do desempenho do Cenário 1. No *Dataset* 3, nenhuma das representações avaliadas produziu aumento relevante e consistente de MCC ou AUC-ROC. No *Dataset* 1, por sua vez, a representação com apenas seis atributos esteve associada a uma redução mais acentuada do desempenho, enquanto o conjunto de 19 atributos recuperou grande parte dessa diferença. Esses resultados indicam que o efeito da representação depende do conteúdo informacional das variáveis preservadas em cada base.

O efeito do Cenário 2 foi avaliado por comparações pareadas com o Cenário 1, mantendo-se, em cada contraste, a mesma semente, população, modelo e condição de balanceamento. Nas populações completas, o F1-score médio variou -1,22 ponto percentual no *Dataset* 1, -0,49 ponto no *Dataset* 2 e -0,06 ponto no *Dataset* 3. As variações correspondentes no MCC foram -0,018, -0,008 e -0,004, e na AUC-ROC foram -0,011, -0,002 e -0,005. Assim, a seleção automática reduziu o número de atributos sem produzir melhora média consistente das métricas principais.

A Tabela 11 apresenta os principais contrastes entre representações separadamente para cada *dataset*. Essa desagregação evita combinar, em uma única média, bases com estruturas e dificuldades preditivas distintas. Cada linha corresponde a 30 combinações de modelo e balanceamento avaliadas nas mesmas 30 sementes, totalizando 900 diferenças pareadas.

Nas 30 sementes, a seleção automática do Cenário 2 foi ajustada exclusivamente sobre o conjunto de treinamento de cada repetição. No *Dataset* 1, essa seleção apresentou alguma variação no número de features mantidas, enquanto nos *Datasets* 2 e 3 o total selecionado permaneceu constante. A redução foi particularmente expressiva no *Dataset*

Tabela 11 – Variações médias pareadas de desempenho entre as representações nas populações completas.

| Contraste | Dataset | Δ Acurácia | Δ Precisão | Δ Recall | Δ F1-score | Δ MCC | Δ AUC-ROC |
|---|---|---|---|---|---|---|---|
| C2 – C1 | D1 | -0,06 p.p. | -2,93 p.p. | -0,97 p.p. | -1,22 p.p. | -0,018 | -0,011 |
| C2 – C1 | D2 | +2,07 p.p. | +0,82 p.p. | -5,19 p.p. | -0,49 p.p. | -0,008 | -0,002 |
| C2 – C1 | D3 | -0,31 p.p. | -0,51 p.p. | -0,15 p.p. | -0,06 p.p. | -0,004 | -0,005 |
| C3, 6 atributos – C1 | D1 | -8,29 p.p. | -21,25 p.p. | -8,89 p.p. | -14,72 p.p. | -0,196 | -0,123 |
| C3, 6 atributos – C1 | D2 | -0,92 p.p. | -2,61 p.p. | -2,67 p.p. | -2,37 p.p. | -0,031 | -0,014 |
| C3, 6 atributos – C1 | D3 | -0,92 p.p. | +0,93 p.p. | +0,90 p.p. | +1,99 p.p. | -0,004 | -0,006 |
| C3, 19 atributos – C1 | D1 | -1,38 p.p. | -3,09 p.p. | +0,32 p.p. | -1,04 p.p. | -0,019 | -0,010 |
| C3, 19 atributos – C1 | D3 | +0,89 p.p. | +1,06 p.p. | -0,98 p.p. | -0,03 p.p. | +0,004 | +0,003 |
| C3, 19 atributos – 6 atributos | D1 | +6,91 p.p. | +18,16 p.p. | +9,21 p.p. | +13,68 p.p. | +0,177 | +0,113 |
| C3, 19 atributos – 6 atributos | D3 | +1,81 p.p. | +0,13 p.p. | -1,88 p.p. | -2,02 p.p. | +0,007 | +0,009 |

Δ representa a primeira representação indicada no contraste menos a segunda. Cada linha resume 900 diferenças pareadas, correspondentes às 30 combinações de modelo e balanceamento avaliadas nas mesmas 30 sementes.

Fonte: Elaborado pelo autor (2026).

2, que passou de 18 features codificadas no Cenário 1 para cinco no Cenário 2, sem queda expressiva das métricas discriminativas. Esse resultado não implica que as demais variáveis sejam irrelevantes em sentido substantivo, pois o critério de seleção corresponde à importância estimada pelo Random Forest dentro do treinamento e tem finalidade preditiva.

A ausência de uma direção uniforme também aparece quando cada combinação de modelo e balanceamento é examinada ao longo das 30 sementes. Considerando a média do F1-score de cada uma das 30 configurações das populações completas, o Cenário 2 superou o Cenário 1 em 9 configurações no *Dataset* 1, em 12 no *Dataset* 2 e em 12 no *Dataset* 3. No primeiro e no segundo conjunto, a maioria das configurações apresentou redução média de F1-score; no terceiro, as diferenças permaneceram especialmente próximas de zero. Dessa forma, o Cenário 2 reduziu a dimensionalidade, mas não apresentou evidência de benefício preditivo geral.

O Cenário 3 utilizou representações definidas a partir de atributos conceitualmente correspondentes entre as bases. No *Dataset* 1, a redução do Cenário 1 para o núcleo de seis atributos esteve associada a uma queda média de 14,72 pontos percentuais no F1-score, 0,196 no MCC e 0,123 na AUC-ROC. Esse padrão foi consistente entre as configurações: somente 1 das 30 combinações de modelo e balanceamento apresentou F1-score médio superior ao Cenário 1 ao longo das sementes. A ampliação de seis para 19 atributos elevou o F1-score médio em 13,68 pontos, o MCC em 0,177 e a AUC-ROC em 0,113, com ganho médio de F1-score em 29 das 30 configurações. Em relação diretamente ao Cenário 1, entretanto, a representação de 19 atributos ainda apresentou diferenças de -1,04 ponto no F1-score, -0,019 no MCC e -0,010 na AUC-ROC.

No *Dataset* 2, disponível apenas no núcleo de seis atributos do Cenário 3, as diferenças em relação ao Cenário 1 foram menores: -2,37 pontos percentuais no F1-score,

-0,031 no MCC e -0,014 na AUC-ROC. Assim, a redução para seis atributos preservou grande parte do desempenho observado com a representação completa, embora sem produzir ganho médio nas principais métricas.

No *Dataset* 3, o comportamento foi distinto. A representação de seis atributos elevou o F1-score médio em 1,99 ponto percentual em relação ao Cenário 1, mas apresentou variações de -0,004 no MCC e -0,006 na AUC-ROC. A representação de 19 atributos ficou ainda mais próxima do Cenário 1, com diferença de -0,03 ponto no F1-score, +0,004 no MCC e +0,003 na AUC-ROC. Portanto, as alterações de F1-score observadas entre as representações não foram acompanhadas por mudanças relevantes na capacidade de discriminação, que permaneceu próxima ao acaso.

As diferenças entre as representações do Cenário 3 não devem ser atribuídas apenas à quantidade de atributos. Os atributos considerados correspondentes foram harmonizados conceitualmente antes da codificação, mas não representam necessariamente medidas idênticas entre as bases. Podem existir diferenças de definição, escala e forma de registro; por exemplo, alguns atributos exigiram regras de derivação para estabelecer uma representação comum. Assim, os contrastes apresentados descrevem o comportamento dos modelos nas representações construídas para cada base e não isolam um efeito causal do número de atributos sobre o desempenho.

Portanto, ao avaliar os três cenários, é necessário diferenciar redução de dimensionalidade de melhoria preditiva. O Cenário 2 realizou seleção automática de atributos dentro do conjunto de treinamento e produziu, em média, alterações pequenas e heterogêneas em relação ao Cenário 1. O Cenário 3 empregou representações comuns reduzidas e apresentou efeitos mais dependentes da base: a restrição a seis atributos afetou de forma mais intensa o *Dataset* 1, enquanto produziu alterações menores no *Dataset* 2 e não aumentou de forma relevante a discriminação no *Dataset* 3.

Em conjunto, os resultados mostram que menos atributos não implicam necessariamente melhor ou pior desempenho. O *Dataset* 2 manteve desempenho elevado mesmo sob representações reduzidas; o *Dataset* 1 perdeu parte substancial do desempenho com seis atributos e recuperou grande parte dessa diferença com 19; e o *Dataset* 3 permaneceu com MCC próximo de zero e AUC-ROC próxima de 0,5 nas diferentes representações. Dessa forma, o efeito da redução depende da informação preditiva preservada em cada conjunto de variáveis e do contexto específico da base analisada.

### 5.2.2 Desempenho dos modelos nos subgrupos por sexo

A segmentação por sexo foi possível apenas nos *Datasets* 1 e 3. Os modelos foram treinados separadamente em cada subgrupo, de modo que a comparação não corresponde à aplicação de um único classificador comum aos dois grupos. No *Dataset* 1, havia 882 registros masculinos, com 150 casos positivos, e 588 femininos, com 87 positivos. Em cada

uma das 30 divisões estratificadas, os conjuntos de teste continham 30 casos positivos no subgrupo masculino e 17 no feminino. No *Dataset* 3, os subgrupos eram maiores e mais equilibrados, com 4.958 homens e 5.042 mulheres; os conjuntos de teste continham, respectivamente, 201 e 199 casos positivos em cada repetição. Por isso, as métricas obtidas para o subgrupo feminino do *Dataset* 1 devem ser interpretadas com maior cautela, devido ao menor número de registros e de casos positivos disponíveis para avaliação.

A Tabela 12 sintetiza o F1-score, o MCC e a AUC-ROC nos diferentes cenários. Para cada semente, foi calculada a média das 30 configurações formadas pelas combinações entre os seis classificadores e as cinco condições de treinamento. A tabela apresenta a média e o desvio-padrão dessas médias ao longo das 30 sementes. No Cenário 3, as representações de 6 e 19 atributos são apresentadas separadamente.

Tabela 12 – Desempenho médio e variabilidade dos modelos nos subgrupos por sexo ao longo das 30 sementes.

| **Dataset** | **Cenário/representação** | **Subgrupo** | **F1-score (%)** | **MCC** | **AUC-ROC** |
|---|---|---|---|---|---|
| *Dataset* 1 | Cenário 1 | Masculino | $44,45 \pm 4,06$ | $0,349 \pm 0,056$ | $0,791 \pm 0,041$ |
| | | Feminino | $42,90 \pm 4,15$ | $0,353 \pm 0,051$ | $0,799 \pm 0,045$ |
| | Cenário 2 | Masculino | $42,54 \pm 4,25$ | $0,322 \pm 0,058$ | $0,774 \pm 0,042$ |
| | | Feminino | $40,77 \pm 4,78$ | $0,324 \pm 0,060$ | $0,783 \pm 0,046$ |
| | Cenário 3 – 6 atributos | Masculino | $30,96 \pm 2,75$ | $0,162 \pm 0,041$ | $0,659 \pm 0,034$ |
| | | Feminino | $27,76 \pm 3,20$ | $0,155 \pm 0,050$ | $0,684 \pm 0,050$ |
| | Cenário 3 – 19 atributos | Masculino | $43,69 \pm 3,83$ | $0,333 \pm 0,051$ | $0,784 \pm 0,038$ |
| | | Feminino | $42,81 \pm 5,06$ | $0,345 \pm 0,065$ | $0,797 \pm 0,048$ |
| *Dataset* 3 | Cenário 1 | Masculino | $18,97 \pm 0,64$ | $0,003 \pm 0,010$ | $0,503 \pm 0,011$ |
| | | Feminino | $18,58 \pm 0,64$ | $0,003 \pm 0,010$ | $0,503 \pm 0,012$ |
| | Cenário 2 | Masculino | $19,26 \pm 0,65$ | $0,001 \pm 0,011$ | $0,503 \pm 0,011$ |
| | | Feminino | $18,83 \pm 0,61$ | $0,004 \pm 0,010$ | $0,506 \pm 0,012$ |
| | Cenário 3 – 6 atributos | Masculino | $21,52 \pm 0,95$ | $0,005 \pm 0,014$ | $0,504 \pm 0,014$ |
| | | Feminino | $19,38 \pm 0,70$ | $-0,015 \pm 0,010$ | $0,482 \pm 0,011$ |
| | Cenário 3 – 19 atributos | Masculino | $19,38 \pm 0,60$ | $0,009 \pm 0,009$ | $0,509 \pm 0,011$ |
| | | Feminino | $18,37 \pm 0,76$ | $0,003 \pm 0,011$ | $0,503 \pm 0,011$ |

Valores apresentados como média ± desvio-padrão entre as 30 sementes. Em cada semente, o valor corresponde à média das 30 configurações formadas pelos seis classificadores e pelas cinco condições de treinamento.

Fonte: Elaborado pelo autor (2026).

No *Dataset* 1, não foi observada uma direção única de diferença entre os subgrupos. No Cenário 1, o F1-score médio foi de $44,45 \pm 4,06\%$ no subgrupo masculino e de $42,90 \pm 4,15\%$ no feminino. O MCC foi de $0,349 \pm 0,056$ e $0,353 \pm 0,051$, enquanto a AUC-ROC foi de $0,791 \pm 0,041$ e $0,799 \pm 0,045$, respectivamente. Como os modelos foram ajustados separadamente em populações distintas, essas diferenças têm caráter descritivo e não constituem uma comparação pareada entre os subgrupos.

O mesmo comportamento dependente da métrica aparece nas demais representações. No Cenário 3 com seis atributos, por exemplo, o F1-score médio foi de 30,96% no subgrupo masculino e 27,76% no feminino, enquanto a AUC-ROC foi maior no subgrupo feminino, com 0,684 frente a 0,659. Na representação de 19 atributos, os F1-scores foram próximos,

com 43,69% e 42,81%, enquanto MCC e AUC-ROC foram numericamente maiores no subgrupo feminino. Dessa forma, os resultados não sustentam uma vantagem consistente de um dos subgrupos nas diferentes métricas e representações.

No *Dataset* 3, a segmentação manteve o padrão geral de baixa capacidade discriminativa. No Cenário 1, os F1-scores médios foram 18,97% no subgrupo masculino e 18,58% no feminino, enquanto o MCC permaneceu próximo de zero em ambos e a AUC-ROC foi aproximadamente 0,503 nos dois grupos. O Cenário 2 apresentou comportamento semelhante, com F1-score de 19,26% e 18,83% e AUC-ROC de 0,503 e 0,506 para os subgrupos masculino e feminino, respectivamente.

A maior diferença numérica entre os subgrupos do *Dataset* 3 ocorreu na representação de seis atributos do Cenário 3. Nessa condição, o subgrupo masculino apresentou F1-score médio de $21,52 \pm 0,95\%$, MCC de $0,005 \pm 0,015$ e AUC-ROC de $0,504 \pm 0,014$, enquanto o feminino apresentou $19,38 \pm 0,70\%$, $-0,015 \pm 0,010$ e $0,482 \pm 0,011$, respectivamente. Entretanto, os valores absolutos de MCC permaneceram próximos de zero e as AUC-ROC próximas de 0,5, de modo que essa diferença descritiva não representa forte capacidade discriminativa em nenhum dos subgrupos.

Em conjunto, a análise dos subgrupos mostra que a direção e a magnitude das diferenças dependem da representação e da métrica considerada. Além disso, os modelos foram ajustados sobre populações distintas, com tamanhos amostrais e quantidades de casos positivos diferentes, especialmente no *Dataset* 1. Por esse motivo, as comparações têm caráter descritivo e não permitem interpretar o sexo como causa das diferenças observadas no *turnover*, nem constituem uma avaliação formal de equidade algorítmica.

### 5.2.3 Desempenho de treinamento e teste e análise exploratória dos *datasets*

A comparação entre o desempenho no treinamento e no teste foi utilizada como uma indicação descritiva de quanto o desempenho se manteve fora dos dados utilizados para ajustar os modelos. Como o F1-score da classe positiva foi a única métrica registrada nas duas etapas, essa comparação foi realizada com base nessa medida. Nas configurações com reamostragem, entretanto, o F1-score de treinamento foi calculado após a aplicação de SMOTE, ADASYN, SMOTE-Tomek ou SMOTE-ENN, enquanto o conjunto de teste manteve sua distribuição original. Por esse motivo, a comparação entre treinamento e teste é mais direta nas configurações sem reamostragem.

A Tabela 13 apresenta exemplos das populações completas do Cenário 1 na condição Desbalanceado, isto é, sem reamostragem externa do conjunto de treinamento. Para cada combinação de *dataset* e modelo, são apresentadas a média e a variabilidade observadas nas 30 sementes. O objetivo da comparação não é estimar generalização externa, mas observar conjuntamente o ajuste aos dados de treinamento e o desempenho obtido nos respectivos conjuntos de teste.

Tabela 13 – Comparação descritiva entre F1-score de treinamento e teste em configurações sem reamostragem externa no Cenário 1 ao longo das 30 sementes.

| Base | Modelo | F1 treino (%) | F1 teste (%) | Gap absoluto (p.p.) | AUC-ROC teste |
|---|---|---|---|---|---|
| D1 | Random Forest | $100,00 \pm 0,00$ | $19,68 \pm 6,70$ | $80,32 \pm 6,70$ | $0,788 \pm 0,040$ |
| D1 | Logistic Regression | $54,35 \pm 1,27$ | $49,31 \pm 3,40$ | $5,53 \pm 3,40$ | $0,824 \pm 0,027$ |
| D2 | Random Forest | $99,99 \pm 0,02$ | $94,62 \pm 0,64$ | $5,38 \pm 0,64$ | $0,980 \pm 0,004$ |
| D2 | Logistic Regression | $54,75 \pm 0,37$ | $54,71 \pm 1,19$ | $1,15 \pm 0,83$ | $0,840 \pm 0,006$ |
| D3 | Random Forest | $100,00 \pm 0,00$ | $0,00 \pm 0,00$ | $100,00 \pm 0,00$ | $0,503 \pm 0,014$ |
| D3 | Logistic Regression | $31,17 \pm 0,43$ | $28,41 \pm 1,23$ | $2,76 \pm 1,38$ | $0,498 \pm 0,012$ |

Valores apresentados como média ± desvio-padrão entre as 30 sementes. O gap absoluto foi calculado, em cada semente, como o valor absoluto da diferença entre o F1-score de treinamento e o F1-score de teste.

Fonte: Elaborado pelo autor (2026).

Os resultados mostram que a magnitude da diferença entre treinamento e teste não deve ser interpretada isoladamente. O Random Forest apresentou F1-score de treinamento de 100% em todas as sementes nos *Datasets* 1 e 3 e média de 99,99% no *Dataset* 2. No conjunto de teste, entretanto, as médias de F1-score foram de 19,68%, 94,62% e 0%, respectivamente. No *Dataset* 1, o gap médio foi de 80,32 pontos percentuais, com desvio-padrão de 6,70 pontos, enquanto no *Dataset* 3 o F1-score de teste permaneceu em zero nas 30 repetições. Dessa forma, um ajuste elevado aos dados de treinamento pode ser acompanhado por comportamentos de teste muito distintos conforme os dados analisados.

O comportamento da Logistic Regression reforça que um pequeno gap também não representa, por si só, bom desempenho preditivo. No *Dataset* 3, a diferença absoluta média entre F1-score de treinamento e teste foi de apenas 2,76 pontos percentuais, mas o F1-score de teste foi de $28,41 \pm 1,23\%$ e a AUC-ROC de $0,498 \pm 0,012$. Portanto, a proximidade entre as métricas de treinamento e teste deve ser interpretada juntamente com o nível absoluto de desempenho obtido.

No *Dataset* 2, o desempenho elevado no conjunto de teste ocorreu em diferentes classificadores e condições experimentais, indicando que os atributos disponíveis permitiram elevada discriminação das classes dentro dessa base. Como o conjunto possui natureza artificial, esse comportamento não deve ser extrapolado diretamente para contextos organizacionais reais. As 3.008 duplicatas exatas existentes na base original foram removidas antes da divisão entre treinamento e teste, evitando que a repetição literal desses registros entre os dois conjuntos explicasse o desempenho observado. A contribuição dos atributos para as predições é examinada a seguir por meio da análise SHAP.

Para complementar a análise de desempenho, o SHAP foi aplicado somente às populações completas no Cenário 1. Para cada *dataset*, foi analisada a configuração de referência definida a partir dos resultados das 30 sementes: LDA com a condição Desbalanceado no *Dataset* 1, XGBoost com SMOTE-Tomek no *Dataset* 2 e AdaBoost com SMOTE-ENN no *Dataset* 3. As sementes 54, 58 e 45, respectivamente, foram selecionadas

por sua proximidade ao comportamento multivariado mediano de cada configuração, e não pelo maior desempenho individual.

As explicações foram calculadas sobre a probabilidade predita da classe positiva. No *Dataset* 1, foram explicadas as 294 observações do conjunto de teste; nos *Datasets* 2 e 3, foram utilizadas amostras estratificadas e determinísticas de 400 observações do teste. Em todos os casos, foram utilizadas 50 observações do treinamento como conjunto de referência. Para obter a contribuição dos atributos originais, as contribuições das *features* codificadas provenientes de uma mesma variável foram agregadas antes do cálculo da média absoluta dos valores SHAP.

A Tabela 14 apresenta, de forma sintética, os cinco atributos com maior contribuição global para as predições em cada modelo analisado. O ordenamento é interno a cada *dataset* e não deve ser utilizado para estabelecer um ranking entre as bases.

Tabela 14 – Atributos com maior contribuição global SHAP nas populações completas do Cenário 1.

| **Dataset** | **Posição** | **Atributo** | **Média de $\|SHAP\|$** | **Participação (%)** |
|---|---|---|---|---|
| D1 | 1 | `OverTime` | 0,08527 | 12,38 |
| D1 | 2 | `JobRole` | 0,06267 | 9,10 |
| D1 | 3 | `EnvironmentSatisfaction` | 0,03828 | 5,56 |
| D1 | 4 | `MaritalStatus` | 0,03804 | 5,52 |
| D1 | 5 | `JobSatisfaction` | 0,03458 | 5,02 |
| D2 | 1 | `satisfaction_level` | 0,09863 | 35,00 |
| D2 | 2 | `number_project` | 0,04942 | 17,53 |
| D2 | 3 | `time_spend_company` | 0,04587 | 16,28 |
| D2 | 4 | `last_evaluation` | 0,04015 | 14,25 |
| D2 | 5 | `average_montly_hours` | 0,02791 | 9,90 |
| D3 | 1 | `Number_of_Companies_Worked` | 0,01562 | 15,00 |
| D3 | 2 | `Relationship_with_Manager` | 0,01285 | 12,34 |
| D3 | 3 | `Job_Involvement` | 0,01117 | 10,72 |
| D3 | 4 | `Work_Life_Balance` | 0,00932 | 8,95 |
| D3 | 5 | `Work_Environment_Satisfaction` | 0,00900 | 8,64 |

A participação corresponde à proporção da média absoluta de SHAP do atributo em relação à soma das contribuições absolutas de todos os atributos originais do respectivo modelo. D1: LDA + Desbalanceado, semente 54; D2: XGBoost + SMOTE-Tomek, semente 58; D3: AdaBoost + SMOTE-ENN, semente 45. Os valores são específicos a cada modelo e não constituem comparação de importância entre *datasets*.

Fonte: Elaborado pelo autor (2026).

No *Dataset* 1, `OverTime` apresentou a maior contribuição global para as predições do modelo analisado, correspondendo a 12,38% da soma das contribuições absolutas, seguido por `JobRole`, com 9,10%. Na execução representativa, a categoria `OverTime = Yes` apresentou contribuição SHAP média positiva para a probabilidade predita de *turnover*, enquanto `OverTime = No` apresentou contribuição média negativa. Esse resultado descreve o comportamento do modelo e não permite concluir que a realização de horas extras seja causa do desligamento.

No *Dataset* 2, `satisfaction_level` respondeu por 35,00% das contribuições absolutas, seguido por `number_project`, `time_spend_company`, `last_evaluation` e `average_montly_hours`. Em conjunto, esses cinco atributos corresponderam a 92,96% das contribuições absolutas do modelo analisado. Para `satisfaction_level`, valores mais elevados tenderam a produzir contribuições mais negativas para a probabilidade prevista da classe positiva; para `time_spend_company`, valores mais elevados tenderam a contribuições mais positivas. Novamente, essas relações dizem respeito à forma como o modelo utilizou os atributos em suas predições, sem estabelecer relações causais com o *turnover*.

No *Dataset* 3, as maiores contribuições globais foram associadas a `Number_of_Companies_Worked`, `Relationship_with_Manager` e `Job_Involvement`. Entretanto, esses resultados exigem cautela adicional: a configuração analisada apresentou, nas 30 sementes, MCC médio de 0,005 e AUC-ROC média de 0,509. Dessa forma, o SHAP descreve como um modelo com baixa capacidade discriminativa distribuiu suas predições entre os atributos, não fornecendo evidência de que esses atributos constituam padrões organizacionais confiáveis de *turnover*.

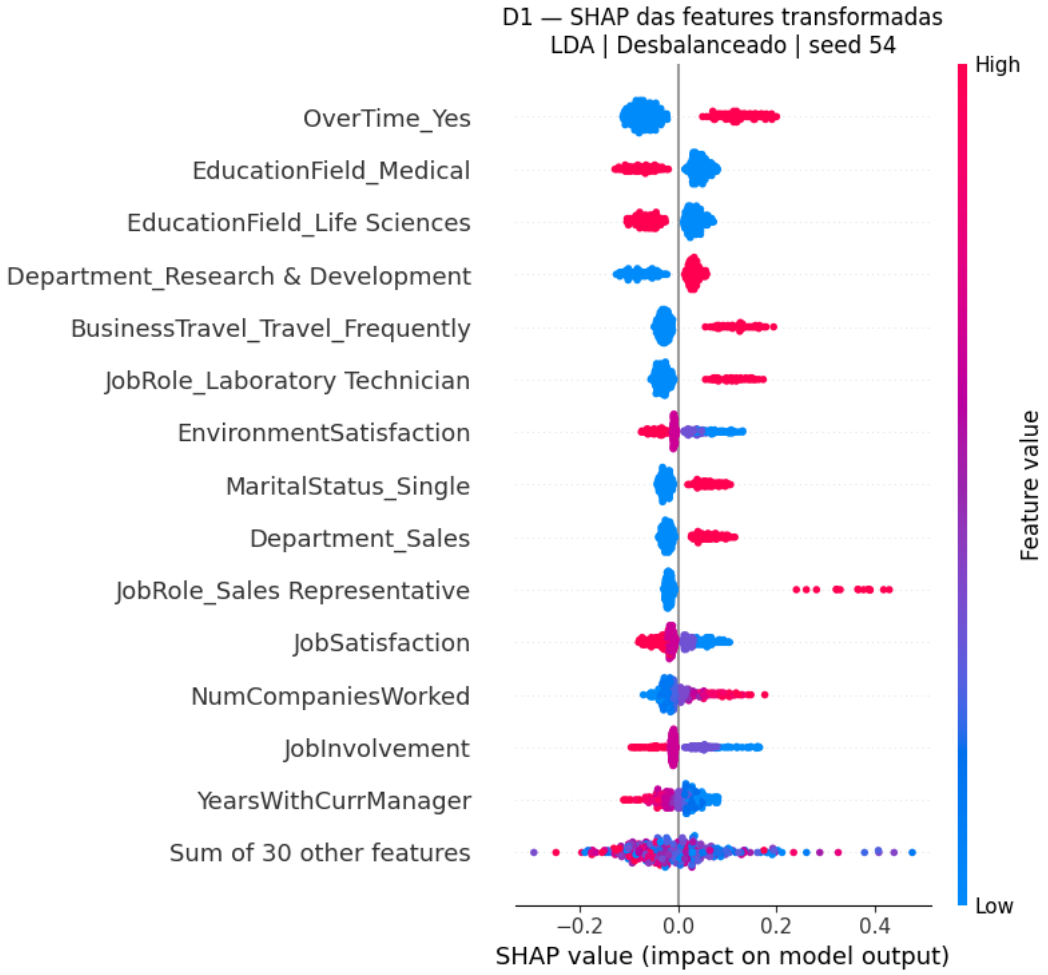
Os gráficos *beeswarm* das Figuras 3, 4 e 5 complementam a importância global ao mostrar a distribuição e o sinal das contribuições no nível das *features* transformadas. Valores SHAP positivos deslocam a predição do modelo em direção a uma maior probabilidade da classe positiva, enquanto valores negativos atuam no sentido oposto. As cores representam os valores observados das respectivas *features*. Esses sinais descrevem exclusivamente o comportamento preditivo dos modelos.

A comparação entre os resultados das três bases também evidencia que o desempenho não pode ser atribuído somente ao tamanho da amostra. O *Dataset* 2 utilizado na modelagem contém 11.991 registros após a remoção das duplicatas, enquanto o *Dataset* 3 contém 10.000 registros. Como o desenho experimental não isolou o efeito do tamanho amostral, não é possível determinar sua contribuição específica para a diferença de desempenho observada entre as bases.

No *Dataset* 3, o padrão predominante foi de baixa capacidade de discriminação. Embora as técnicas de reamostragem tenham aumentado o *recall* e o F1-score em determinadas configurações, esses ganhos não foram acompanhados de melhora uniforme no MCC e na AUC-ROC. Consequentemente, o aumento da identificação da classe positiva após o balanceamento não deve ser interpretado automaticamente como aumento equivalente da capacidade geral de distinguir as duas classes.

Os resultados disponíveis não permitem determinar a causa do baixo desempenho do *Dataset* 3. Entre as explicações possíveis estão a ausência de variáveis relevantes, ruído ou erros de medição, características da definição da variável-alvo e relações temporais não representadas pelos registros tabulares disponíveis. Essas possibilidades constituem hipóteses e não foram confirmadas pelo desenho experimental. A documentação disponível

Figura 3 – Distribuição dos valores SHAP das *features* transformadas no *Dataset* 1 para LDA com a condição Desbalanceado, semente 54.



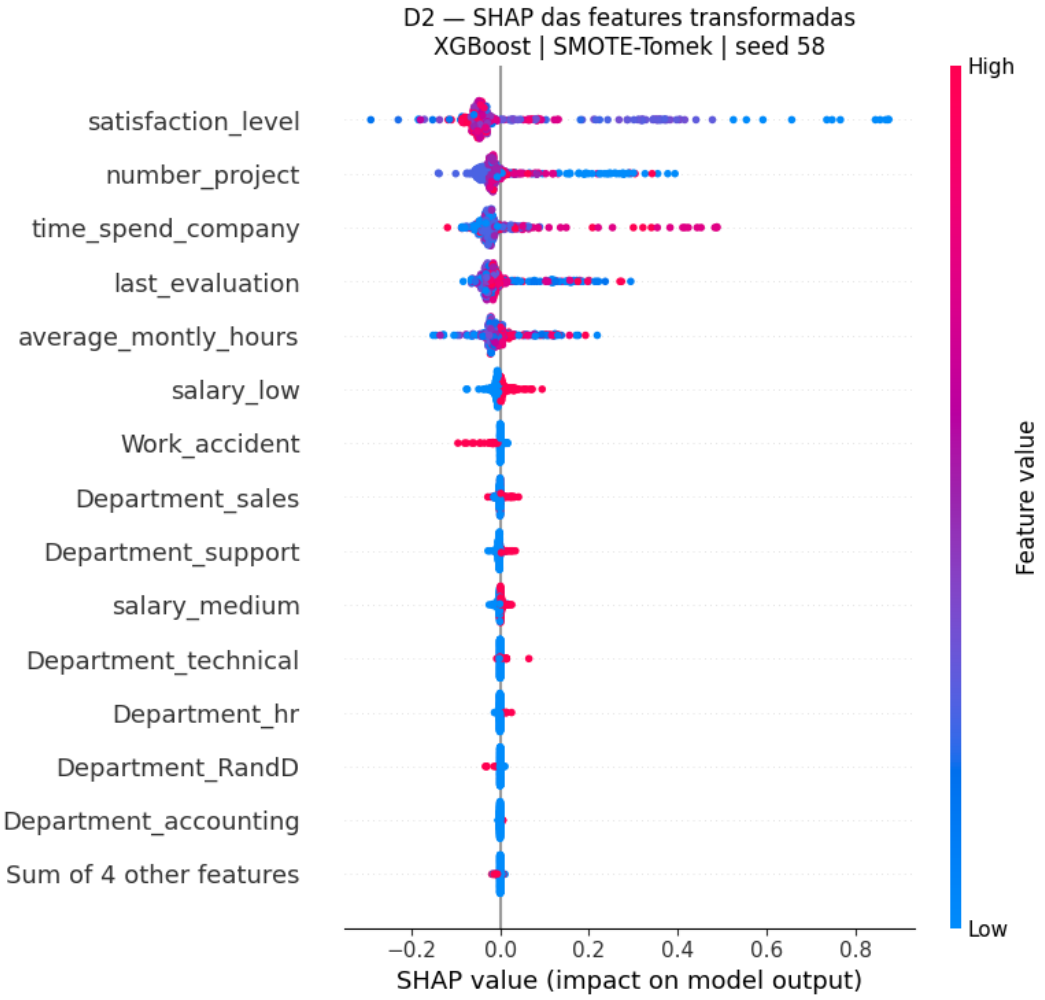

Fonte: Elaborado pelo autor (2026).

para essa base também não permite verificar de forma independente todos os aspectos de sua coleta e da construção da variável `Attrition`.

No *Dataset* 1, o desempenho mostrou maior sensibilidade à combinação entre modelo, balanceamento e representação utilizada. As diferentes configurações apresentaram variações relevantes de F1-score, MCC e AUC-ROC, e a redução mais restritiva do conjunto de atributos também afetou o desempenho observado. Esse comportamento reforça que os resultados constituem uma propriedade da combinação entre dados, representação, procedimento de treinamento e classificador, e não apenas do algoritmo considerado isoladamente.

Do ponto de vista de *people analytics*, esses resultados reforçam a necessidade de interpretar conjuntamente as métricas e os perfis de erro. Um aumento de *recall* pode ser relevante quando a prioridade é identificar mais ocorrências da classe positiva, mas

Figura 4 – Distribuição dos valores SHAP das *features* transformadas no *Dataset* 2 para XGBoost com SMOTE-Tomek, semente 58.

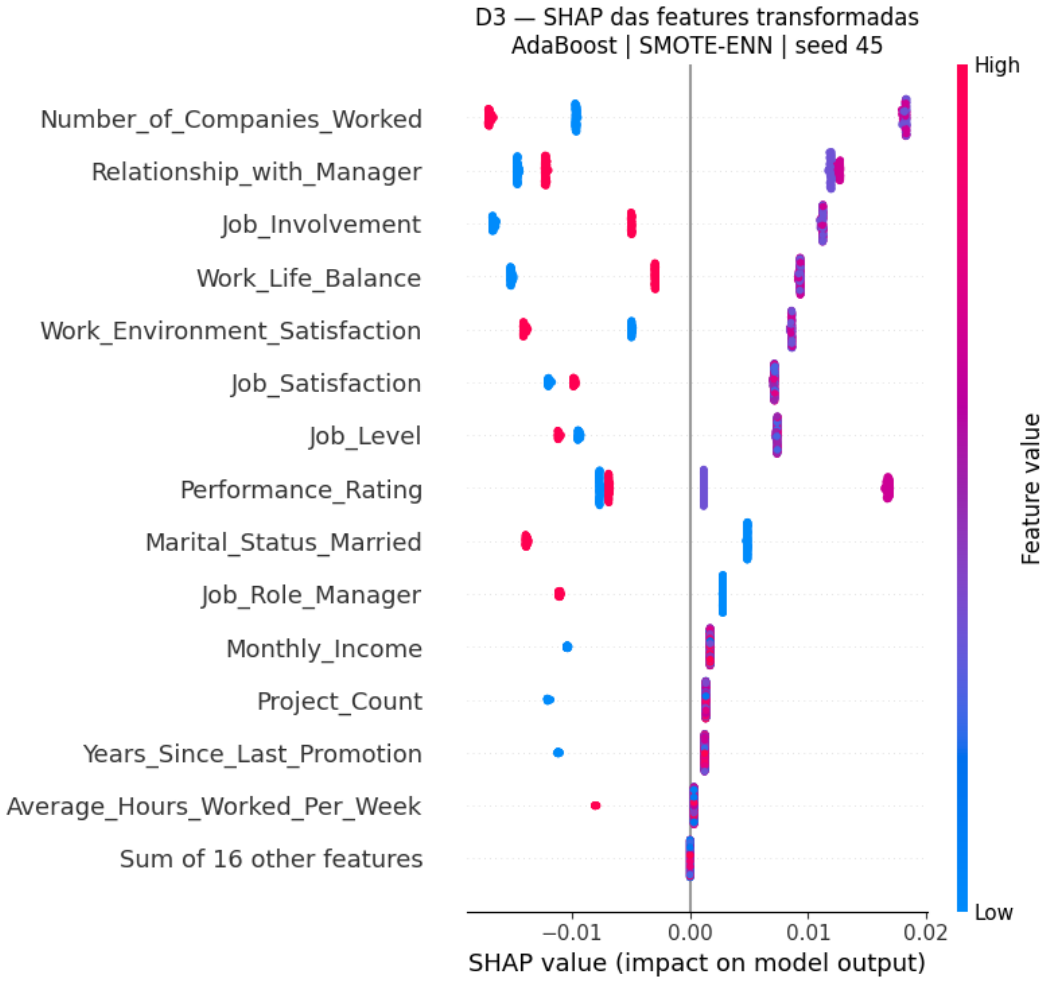



Fonte: Elaborado pelo autor (2026).

deve ser analisado juntamente com a precisão, a quantidade de falsos positivos, o MCC, a AUC-ROC e o Log Loss. As contribuições identificadas pelo SHAP acrescentam informação sobre a forma como os modelos utilizaram os atributos para produzir suas predições. Elas podem apoiar a formulação de perguntas para um diagnóstico organizacional coletivo, mas não demonstram causas, não estabelecem automaticamente importância prática e não prescrevem uma intervenção. Qualquer interpretação gerencial deve ser confrontada com o contexto, com outras fontes de informação e com a escuta dos trabalhadores.

Em uma aplicação real, os modelos devem apoiar, e não substituir, decisões humanas. Uma previsão individual não deve justificar punição, vigilância, restrição de oportunidades ou tratamento discriminatório, nem ser utilizada para definir perfis de trabalhadores desejáveis e indesejáveis. Falsos positivos podem produzir estigmatização mesmo quando não ocorre desligamento, e alertas corretos também podem ser empregados

Figura 5 – Distribuição dos valores SHAP das *features* transformadas no *Dataset* 3 para AdaBoost com SMOTE-ENN, semente 45.




Fonte: Elaborado pelo autor (2026).

de forma inadequada. Por isso, o uso organizacional exige finalidade legítima, adequação, necessidade e minimização dos dados, acesso restrito, segurança, transparência, prevenção de discriminação, documentação, possibilidade de revisão e supervisão humana, além de anonimização ou pseudonimização quando compatível com a finalidade do tratamento (9, 21, 37). Políticas de retenção também devem considerar que nem toda saída precisa ser evitada e que intervenções podem ser direcionadas às condições de trabalho, à liderança ou aos processos organizacionais, em vez de ao controle do indivíduo.

## 5.3 SELEÇÃO DAS CONFIGURAÇÕES DE REFERÊNCIA

### 5.3.1 Critérios, configurações selecionadas e robustez

O critério de ordenação utilizado foi o Índice Multivariado Ponderado (IM) definido no Capítulo 4, que combina Escores-Z de MCC, F1-score, AUC-ROC e Log Loss com pesos

de 0,40, 0,30, 0,20 e -0,10, respectivamente. A padronização foi realizada separadamente em cada semente e em cada contexto comparável, definido pela combinação de *dataset*, população, cenário e representação. Dentro de cada contexto foram comparadas 30 configurações, correspondentes às cinco condições de treinamento combinadas com os seis classificadores.

Em cada semente, o Índice Multivariado foi calculado conforme a expressão definida no Capítulo 4:

$$IM = 0{,}40Z_{\mathrm{MCC}} + 0{,}30Z_{\mathrm{F1}} + 0{,}20Z_{\mathrm{AUC}} - 0{,}10Z_{\mathrm{LogLoss}}.$$

Como a padronização foi realizada separadamente dentro de cada contexto e semente, os valores do IM não devem ser utilizados para comparar diretamente diferentes *datasets*, populações, cenários ou representações. O índice expressa o desempenho relativo de uma configuração em relação às outras 29 avaliadas no mesmo contexto experimental.

Em cada uma das 30 sementes, as configurações foram ordenadas de forma decrescente pelo IM. Dessa forma, cada configuração possui 30 valores do índice e 30 posições no *ranking*. Para avaliar a robustez dessa posição, foram considerados o IM médio e seu desvio-padrão, o *rank* médio e mediano e as frequências de ocorrência na primeira posição, entre as três primeiras e entre as cinco primeiras posições. Ao todo, foram consolidadas 810 configurações, cada uma avaliada nas 30 sementes.

A configuração de referência de cada contexto foi definida pelo maior IM médio nas 30 sementes. Em caso de empate, foram considerados, nesta ordem, o menor *rank* médio, o maior MCC médio e o maior F1-score médio. Precisão, *recall* e outras métricas são utilizadas como informações complementares para a interpretação das configurações, mas não integram diretamente o cálculo do índice.

Os três cenários participaram desse procedimento. No Cenário 2, a seleção de atributos foi ajustada exclusivamente sobre o conjunto de treinamento de cada semente, permitindo que suas configurações fossem avaliadas pelo mesmo procedimento aplicado aos demais cenários. Considerando todas as populações, cenários e representações, foram definidas 27 configurações de referência.

A Tabela 15 apresenta as configurações de referência das 11 representações correspondentes às populações completas. Além das métricas médias de desempenho, são apresentados indicadores do comportamento do Índice Multivariado ao longo das 30 sementes, permitindo interpretar conjuntamente desempenho absoluto e estabilidade relativa dentro de cada contexto.

Nas populações completas do Cenário 1, as configurações de referência foram LDA + Desbalanceado no *Dataset* 1, XGBoost + SMOTE-Tomek no *Dataset* 2 e AdaBoost + SMOTE-ENN no *Dataset* 3. Essas três configurações também foram utilizadas na análise

Tabela 15 – Configurações de referência e robustez do Índice Multivariado nas populações completas.

| Dataset | Cenário / representação | Modelo + condição | F1 (%) | MCC | AUC-ROC | IM | Rank médio | Rank 1 | Top 3 (%) |
|---|---|---|---|---|---|---|---|---|---|
| D1 | C1 – completo | LDA + Desbalanceado | $52,73 \pm 5,52$ | $0,481 \pm 0,065$ | $0,828 \pm 0,028$ | $0,970 \pm 0,352$ | 4,27 | 11/30 | 60,0 |
| D1 | C2 – seleção de atributos | AdaBoost + SMOTE-Tomek | $51,02 \pm 4,38$ | $0,413 \pm 0,054$ | $0,801 \pm 0,029$ | $0,738 \pm 0,406$ | 6,93 | 5/30 | 43,3 |
| D1 | C3 – 6 atributos | AdaBoost + SMOTE | $40,21 \pm 3,83$ | $0,266 \pm 0,054$ | $0,715 \pm 0,040$ | $0,993 \pm 0,287$ | 4,00 | 7/30 | 53,3 |
| D1 | C3 – 19 atributos | AdaBoost + SMOTE-Tomek | $51,60 \pm 5,02$ | $0,420 \pm 0,060$ | $0,806 \pm 0,033$ | $0,817 \pm 0,381$ | 6,33 | 4/30 | 36,7 |
| D2 | C1 – completo | XGBoost + SMOTE-Tomek | $94,15 \pm 0,72$ | $0,930 \pm 0,009$ | $0,984 \pm 0,003$ | $1,171 \pm 0,021$ | 2,20 | 12/30 | 86,7 |
| D2 | C2 – seleção de atributos | XGBoost + SMOTE | $94,09 \pm 0,75$ | $0,930 \pm 0,009$ | $0,983 \pm 0,003$ | $1,190 \pm 0,019$ | 2,13 | 13/30 | 86,7 |
| D2 | C3 – 6 atributos | XGBoost + SMOTE | $91,86 \pm 0,78$ | $0,903 \pm 0,009$ | $0,978 \pm 0,003$ | $1,213 \pm 0,025$ | 1,43 | 20/30 | 100,0 |
| D3 | C1 – completo | AdaBoost + SMOTE-ENN | $32,79 \pm 0,46$ | $0,005 \pm 0,021$ | $0,509 \pm 0,016$ | $0,448 \pm 0,558$ | 9,17 | 4/30 | 23,3 |
| D3 | C2 – seleção de atributos | AdaBoost + SMOTE-ENN | $32,72 \pm 0,53$ | $0,003 \pm 0,023$ | $0,509 \pm 0,018$ | $0,504 \pm 0,488$ | 8,97 | 2/30 | 30,0 |
| D3 | C3 – 6 atributos | AdaBoost + SMOTE-ENN | $31,40 \pm 1,04$ | $0,020 \pm 0,017$ | $0,517 \pm 0,013$ | $0,947 \pm 0,467$ | 4,10 | 8/30 | 63,3 |
| D3 | C3 – 19 atributos | Naive Bayes + ADASYN | $27,19 \pm 1,42$ | $0,017 \pm 0,021$ | $0,510 \pm 0,014$ | $0,475 \pm 0,538$ | 9,23 | 3/30 | 36,7 |

F1-score, MCC, AUC-ROC e IM são apresentados como média ± desvio-padrão nas 30 sementes. Rank 1 indica o número de sementes em que a configuração ocupou a primeira posição dentro de seu próprio contexto. Top 3 indica a frequência de permanência entre as três primeiras posições. Os valores do IM não são comparáveis entre contextos diferentes.

Fonte: Elaborado pelo autor (2026).

SHAP apresentada na Subseção 5.2.3, após a escolha de uma semente representativa do comportamento das 30 repetições.

A Tabela 16 complementa a análise das configurações de referência apresentando os valores médios dos componentes das matrizes de confusão ao longo das 30 sementes. Como os conjuntos de teste possuem o mesmo tamanho e a mesma quantidade de casos positivos em cada repetição de um determinado *dataset*, esses valores permitem observar o perfil médio de erros associado a cada configuração.

Tabela 16 – Média dos componentes das matrizes de confusão das configurações de referência nas populações completas ao longo das 30 sementes.

| Dataset | Cenário / representação | Modelo + condição | TP n (%) | TN n (%) | FP n (%) | FN n (%) | Precisão (%) | Recall (%) | $n_{\text{teste}}$ | $n_+$ |
|---|---|---|---|---|---|---|---|---|---|---|
| D1 | C1 – completo | LDA + Desbalanceado | 20,1 (6,8%) | 237,9 (80,9%) | 9,1 (3,1%) | 26,9 (9,1%) | 69,5 | 42,8 | 294 | 47 |
| D1 | C2 – seleção de atributos | AdaBoost + SMOTE-Tomek | 25,7 (8,7%) | 218,9 (74,5%) | 28,1 (9,6%) | 21,3 (7,2%) | 48,3 | 54,7 | 294 | 47 |
| D1 | C3 – 6 atributos | AdaBoost + SMOTE | 27,8 (9,5%) | 183,4 (62,4%) | 63,6 (21,6%) | 19,2 (6,5%) | 30,5 | 59,2 | 294 | 47 |
| D1 | C3 – 19 atributos | AdaBoost + SMOTE-Tomek | 25,9 (8,8%) | 219,6 (74,7%) | 27,4 (9,3%) | 21,1 (7,2%) | 48,9 | 55,1 | 294 | 47 |
| D2 | C1 – completo | XGBoost + SMOTE-Tomek | 367,1 (15,3%) | 1986,2 (82,8%) | 14,8 (0,6%) | 30,9 (1,3%) | 96,1 | 92,2 | 2399 | 398 |
| D2 | C2 – seleção de atributos | XGBoost + SMOTE | 367,6 (15,3%) | 1985,2 (82,8%) | 15,8 (0,7%) | 30,4 (1,3%) | 95,9 | 92,4 | 2399 | 398 |
| D2 | C3 – 6 atributos | XGBoost + SMOTE | 363,3 (15,1%) | 1971,4 (82,2%) | 29,6 (1,2%) | 34,7 (1,4%) | 92,5 | 91,3 | 2399 | 398 |
| D3 | C1 – completo | AdaBoost + SMOTE-ENN | 361,5 (18,1%) | 157,0 (7,8%) | 1444,0 (72,2%) | 37,5 (1,9%) | 20,0 | 90,6 | 2000 | 399 |
| D3 | C2 – seleção de atributos | AdaBoost + SMOTE-ENN | 360,1 (18,0%) | 159,3 (8,0%) | 1441,7 (72,1%) | 38,9 (1,9%) | 20,0 | 90,3 | 2000 | 399 |
| D3 | C3 – 6 atributos | AdaBoost + SMOTE-ENN | 267,6 (13,4%) | 565,8 (28,3%) | 1035,2 (51,8%) | 131,4 (6,6%) | 20,5 | 67,1 | 2000 | 399 |
| D3 | C3 – 19 atributos | Naive Bayes + ADASYN | 156,3 (7,8%) | 1006,6 (50,3%) | 594,4 (29,7%) | 242,7 (12,1%) | 20,8 | 39,2 | 2000 | 399 |

TP, TN, FP e FN correspondem às médias obtidas nas 30 sementes; por esse motivo, seus valores podem não ser inteiros. Os percentuais entre parênteses correspondem à proporção média em relação ao total do conjunto de teste. Precisão e recall correspondem às médias das métricas calculadas separadamente em cada semente.

Fonte: Elaborado pelo autor (2026).

Os perfis das matrizes de confusão ajudam a interpretar diferenças que não são evidentes apenas pelo F1-score. No *Dataset* 3, por exemplo, AdaBoost + SMOTE-ENN apresentou *recall* médio superior a 90% nos Cenários 1 e 2, mas acompanhado por mais de 1.400 falsos positivos em média e precisão próxima de 20%. Esse comportamento é compatível com os valores muito baixos de MCC e com AUC-ROC próxima de 0,5, reforçando que a elevada recuperação da classe positiva não representa, nesse caso, elevada capacidade geral de discriminação.

As configurações de referência correspondentes aos subgrupos masculino e feminino

dos *Datasets* 1 e 3 são apresentadas no Apêndice A. No corpo do capítulo são apresentadas apenas as populações completas para preservar a concisão da apresentação.

No *Dataset* 1, a configuração de referência variou conforme a representação. No Cenário 1, LDA + Desbalanceado apresentou F1-score médio de $52,73 \pm 5,52\%$, MCC de $0,481 \pm 0,065$ e AUC-ROC de $0,828 \pm 0,028$. Essa configuração ocupou a primeira posição em 11 das 30 sementes, apresentou *rank* médio de 4,27 e permaneceu entre as três primeiras em 60,0% das repetições. Portanto, apresentou desempenho relativo recorrente, mas sem dominar o contexto em todas as divisões de treinamento e teste.

Nas demais representações do *Dataset* 1, as referências foram AdaBoost + SMOTE-Tomek no Cenário 2, AdaBoost + SMOTE no Cenário 3 com seis atributos e AdaBoost + SMOTE-Tomek no Cenário 3 com 19 atributos. As frequências de primeira posição foram de 16,7%, 23,3% e 13,3%, respectivamente. A alternância entre configurações mostra que a identidade da alternativa mais bem posicionada depende da representação utilizada, não sendo adequado definir uma única combinação como superior para todo o *Dataset* 1.

No *Dataset* 2, o XGBoost apareceu nas três configurações de referência das populações completas. No Cenário 1, XGBoost + SMOTE-Tomek apresentou F1-score médio de $94,15 \pm 0,72\%$, MCC de $0,930 \pm 0,009$ e AUC-ROC de $0,984 \pm 0,003$. A configuração ocupou a primeira posição em 12 das 30 sementes e permaneceu entre as três primeiras em 86,7% das repetições. No Cenário 2, XGBoost + SMOTE apresentou comportamento semelhante, com *rank* médio de 2,13 e presença no Top 3 em 86,7% das sementes. No Cenário 3 com seis atributos, XGBoost + SMOTE apresentou *rank* médio de 1,43, mediana igual a 1 e permaneceu entre as três primeiras configurações em todas as 30 sementes. Esse padrão mostra recorrência do classificador dentro do *Dataset* 2, embora a condição de treinamento de referência varie entre os cenários.

No *Dataset* 3, a primeira posição relativa deve ser interpretada principalmente em conjunto com as métricas absolutas. No Cenário 1, AdaBoost + SMOTE-ENN apresentou F1-score médio de $32,79 \pm 0,46\%$, mas MCC de apenas $0,005 \pm 0,021$ e AUC-ROC de $0,509 \pm 0,016$. A configuração ocupou a primeira posição em 4 das 30 sementes e apresentou *rank* médio de 9,17. No Cenário 2, a mesma combinação foi selecionada, também com MCC próximo de zero e AUC-ROC próxima de 0,5.

No Cenário 3 do *Dataset* 3, AdaBoost + SMOTE-ENN foi a configuração de referência na representação de seis atributos, com *rank* médio de 4,10 e presença no Top 3 em 63,3% das sementes. Entretanto, o MCC médio permaneceu em 0,020 e a AUC-ROC em 0,517. Na representação de 19 atributos, Naive Bayes + ADASYN foi a configuração de referência, com F1-score médio de 27,19%, MCC de 0,017 e AUC-ROC de 0,510. Assim, a existência de uma configuração de referência não modifica o padrão geral de baixa capacidade discriminativa observado nessa base.

Essa distinção entre posição no *ranking* e desempenho absoluto é importante para

interpretar a seleção. O Índice Multivariado foi utilizado para ordenar as 30 configurações dentro de cada contexto comparável, e não para impor um limiar mínimo de desempenho. Assim, mesmo em um contexto no qual todas as configurações apresentam AUC-ROC próxima ao acaso, haverá uma configuração de referência. O resultado deve, portanto, ser interpretado em conjunto com as métricas originais e com a estabilidade observada nas 30 sementes.

A robustez foi avaliada diretamente para todas as 810 configurações. Cada uma foi examinada nas 30 divisões estratificadas, permitindo observar tanto a variabilidade das métricas quanto a estabilidade do Índice Multivariado e das posições no *ranking*. Dessa forma, a avaliação das configurações não ficou restrita ao resultado de uma única divisão entre treinamento e teste.

Essa robustez interna deve ser diferenciada de generalização externa. As 30 repetições modificam os registros utilizados nos conjuntos de treinamento e teste, mas permanecem restritas aos mesmos *datasets*. Portanto, permitem avaliar a estabilidade das configurações diante de diferentes divisões da amostra, mas não demonstram que o desempenho será mantido em outras organizações, períodos ou populações.

Há uma limitação associada ao próprio procedimento de seleção. As métricas obtidas nos conjuntos de teste das 30 sementes foram utilizadas para calcular o Índice Multivariado e definir as configurações de referência. Não foi reservado um segundo conjunto independente para uma avaliação adicional após essa seleção. Por esse motivo, as configurações de referência correspondem às alternativas com maior Índice Multivariado médio dentro de cada contexto comparável, acompanhadas pelos respectivos indicadores de estabilidade nas 30 sementes, sem constituir demonstração de superioridade definitiva em dados externos.

A forma de normalização do índice também exige esse cuidado. Como os Escores-Z são calculados separadamente em cada semente e contexto comparável, uma configuração pode obter posição elevada no *ranking* mesmo quando seus valores absolutos de desempenho são baixos, como observado no *Dataset* 3. Da mesma forma, pequenas diferenças de IM dentro de um contexto com bom desempenho não representam necessariamente vantagens práticas relevantes. Assim, o Índice Multivariado é útil para resumir e ordenar as configurações dentro de cada contexto, mas deve ser interpretado conjuntamente com as métricas originais e com os indicadores de estabilidade.

Também não há fundamento para construir um ranking global dos seis classificadores a partir das configurações de referência. A frequência com que um modelo aparece entre os primeiros colocados envolve populações, representações e condições de treinamento distintas. O XGBoost apresentou elevada recorrência nas representações do *Dataset* 2, enquanto o *Dataset* 1 apresentou maior alternância entre configurações e o *Dataset* 3 manteve baixa discriminação mesmo nas configurações de referência. A interpretação

metodologicamente adequada permanece, portanto, restrita a cada contexto comparável.

### 5.3.2 Confronto com a literatura e síntese dos objetivos

Uma comparação com a revisão sistemática foi feita para contextualizar os resultados. Os estudos diferem em *dataset*, definição do desfecho, pré-processamento, balanceamento, validação e conjunto no qual as métricas foram calculadas. Mesmo quando a base é a mesma, diferenças numéricas não podem ser atribuídas diretamente ao algoritmo utilizado. A Tabela 17 organiza a comparação por temas, reduzindo a descrição artigo a artigo. Os três cenários experimentais são considerados na interpretação dos resultados desta pesquisa, respeitando-se as diferenças de representação e os limites de comparabilidade entre os contextos.

Tabela 17 – Síntese temática do confronto entre os resultados desta pesquisa e a revisão sistemática.

| **Eixo de comparação** | **Evidência na literatura** | **Relação com os resultados desta pesquisa** |
|---|---|---|
| Mesma base IBM | Resultados heterogêneos com Logistic Regression, Random Forest, redes neurais, Deep Forest, Extra Trees e modelos profundos (23, 1, 55, 65, 66, 56, 44, 6). | No D1, a configuração de referência variou conforme a representação. No C1, LDA + Desbalanceado apresentou F1-score de $52,73 \pm 5,52\%$, MCC de $0,481 \pm 0,065$ e AUC-ROC de $0,828 \pm 0,028$. Diferenças em relação à literatura não constituem superioridade direta devido às diferenças de protocolo. |
| Modelos recorrentes | Random Forest e Logistic Regression aparecem com frequência entre os melhores modelos, ao lado de XGBoost e métodos mais complexos (23, 55, 65, 67, 66, 7). | Nas populações completas, XGBoost integrou as configurações de referência do D2 nos três cenários, enquanto D1 e D3 apresentaram maior alternância entre classificadores. O padrão reforça a dependência do *dataset* e da representação avaliada. |
| Balanceamento | Reamostragem, SMOTE, ADASYN, ROSE e estratégias gerativas são recorrentes (56, 44, 6, 68). | Nas comparações pareadas das 30 sementes, a reamostragem elevou principalmente *recall* e F1-score, acompanhada por perdas de acurácia e aumento de Log Loss em parte das condições, com efeitos heterogêneos entre bases, modelos e métricas. |
| Diferentes bases no mesmo estudo | Estudos que combinam IBM artificial com *datasets* apresentados como reais obtêm desempenhos distintos entre as bases (44, 6). | Os três *datasets* desta pesquisa também apresentaram comportamentos preditivos distintos nas diferentes representações. A natureza declarada da base, isoladamente, não é suficiente para explicar essas diferenças. |
| Target e protocolo | Parte da literatura usa ocorrências de saída; outros trabalhos analisam intenção de *turnover* ou adotam validações distintas (53, 32). | O *target* desta pesquisa corresponde à saída registrada nas bases e não pode ser tratado como exclusivamente voluntário. O desempenho foi avaliado em 30 repetições de *hold-out* estratificado 80/20, procedimento distinto de protocolos de validação cruzada utilizados em parte da literatura. |

Fonte: Elaborado pelo autor com base nos estudos citados e nos resultados desta pesquisa (2026).

O *Dataset* 1 oferece o confronto mais amplo, pois a base artificial da IBM aparece em 11 estudos da revisão. Ainda assim, os valores publicados variam consideravelmente. Alguns estudos utilizaram arquiteturas neurais profundas, como DNNs e Bi-TCNs, enquanto outros avaliaram modelos profundos baseados em ensembles de árvores, como o Deep Forest. Embora apresente organização em camadas, o Deep Forest não corresponde a uma rede neural e deve ser distinguido das arquiteturas de Deep Learning baseadas em retropropagação.

Guerranti e Dimitri (2022) reportaram acurácia de 88% e AUC de 85% para Logistic Regression (23). Taner et al. (2024) registraram acurácia de 89,8%, F1-score de 56,14% e AUC de 90,2% para o mesmo tipo de modelo (65). Tokmak (2025) reportou 98,8% para

acurácia, precisão, recall e F1-score com Deep Forest (66). Ahmed (2021) apresentou acurácia de 84% e AUC de 74% para uma rede neural (1). A amplitude desses resultados no mesmo *dataset* reforça que a comparação depende do protocolo e da métrica, e não permite ordenar os estudos apenas pelo maior valor publicado.

Pratt et al. (2021) também destacaram Random Forest, com acurácia de 85,12% (55), enquanto Mortezapour Shiri et al. (2025) reportaram, para o conjunto IBM, F1-score de 61,61% e AUC de 77,58% com Bi-TCN (44). No Cenário 1 desta pesquisa, a configuração de referência do D1, LDA + Desbalanceado, apresentou F1-score médio de $52,73 \pm 5,52\%$ e AUC-ROC de $0,828 \pm 0,028$ nas 30 sementes. A comparação ilustra como a ordenação relativa pode mudar conforme a métrica considerada e reforça que valores obtidos sob protocolos distintos não devem ser interpretados como evidência direta de superioridade entre modelos.

Baydili e Tasci (2025) reportaram, no *dataset* da IBM, F1-score de 91,58% e AUC de 96,32%, avaliando também estratégias de tratamento do desbalanceamento (6). Na própria literatura revisada, existem outros estudos que utilizaram essa mesma base e encontraram resultados bem menores. A literatura, portanto, oferece evidência de que a preparação dos dados, o protocolo de validação e o modelo influenciam fortemente o nível reportado.

A mesma cautela se aplica a Mortezapour Shiri et al. (2025) e Baydili e Tasci (2025), que utilizaram tanto a base artificial da IBM quanto outro *dataset* apresentado como real e reportaram desempenhos distintos entre as bases (44, 6). Raza et al. (2022), por sua vez, empregaram reamostragem e validação cruzada de dez *folds*, reportando 93% em acurácia, precisão, *recall* e F1-score com Extra Trees (56). Nesta pesquisa foram realizadas 30 repetições de *hold-out* estratificado 80/20, com diferentes sementes, e não validação cruzada. Portanto, mesmo resultados obtidos sobre o mesmo *dataset* não constituem estimativas diretamente equivalentes.

Entre os estudos incluídos na revisão, Jain et al. (2020) utilizaram a mesma fonte de dados do *Dataset* 2 e destacaram Random Forest como o melhor entre os modelos avaliados (32). Nesta pesquisa, XGBoost integrou as configurações de referência das três representações avaliadas na população completa do D2: associado a SMOTE-Tomek no Cenário 1 e a SMOTE nos Cenários 2 e 3. No Cenário 1, XGBoost + SMOTE-Tomek apresentou F1-score médio de $94,15 \pm 0,72\%$, MCC de $0,930 \pm 0,009$ e AUC-ROC de $0,984 \pm 0,003$. A diferença entre os modelos destacados nos dois trabalhos reforça a influência do protocolo experimental e das configurações avaliadas. Nenhum dos estudos incluídos na revisão utilizou a mesma base do *Dataset* 3, de modo que sua baixa AUC-ROC não possui comparação direta com a mesma fonte de dados.

Essa assimetria de cobertura da literatura também limita a força do confronto. O *Dataset* 1 possui ampla presença em trabalhos anteriores, permitindo observar grande

variação de modelos e resultados; o *Dataset* 2 possui referência direta muito mais restrita; e o *Dataset* 3 não possui comparação com a mesma fonte. Assim, o grau de contextualização externa não é uniforme entre as três bases.

A seleção automática de atributos do Cenário 2 também pode ser contextualizada por estudos que empregam procedimentos de seleção de características, como Najafi-Zangeneh et al. (2021) e Benabou et al. (2025) (46, 7). Nesta pesquisa, a seleção foi ajustada exclusivamente sobre o conjunto de treinamento de cada semente. Nas populações completas, as diferenças pareadas de F1-score entre C2 e C1 foram de $-1,22$ ponto percentual no D1, $-0,49$ ponto no D2 e $-0,06$ ponto no D3. Portanto, a redução automática da representação não apresentou uma direção geral de melhora nas três bases. As diferenças de procedimentos entre os estudos impedem, contudo, uma comparação quantitativa direta do efeito da seleção de atributos.

A revisão também converge com a pesquisa quanto à relevância do tratamento do desbalanceamento. Métodos de reamostragem aparecem em diferentes estudos (56, 44, 6, 68), e os resultados desta dissertação mostram por que a técnica deve ser avaliada em conjunto com a métrica: aumentos de recall e F1-score coexistiram com reduções de acurácia, pequenas alterações de MCC e AUC e piora do Log Loss. O conjunto de evidências revisado e os experimentos realizados, portanto, não sustentam um algoritmo ou estratégia de balanceamento universalmente superior.

Também há convergência quanto à recorrência de famílias de modelos tradicionais. Random Forest e Logistic Regression aparecem repetidamente entre os melhores resultados de diferentes estudos, ao lado de XGBoost, CatBoost, Deep Forest e arquiteturas profundas (23, 46, 55, 65, 7, 67). Nesta pesquisa, as configurações de referência também variaram conforme o contexto. LDA e AdaBoost apareceram nas populações completas do D1, XGBoost foi recorrente nas três representações do D2 e, no D3, AdaBoost e Naive Bayes ocuparam as posições de referência conforme a representação. Esse resultado é compatível com a conclusão de que não há fundamento para indicar um classificador universalmente superior.

A própria definição do desfecho restringe comparações. Park et al. (2024) analisam intenção de turnover por meio de questionário (53), enquanto esta pesquisa utiliza ocorrências registradas de saída. Intenção e desligamento observado são alvos relacionados, mas não equivalentes. Mesmo entre trabalhos de *attrition*, a identificação da classe positiva e o conjunto de dados avaliado nem sempre são reportados com o detalhamento necessário para comparar métricas de desempenho. Por isso, a discussão prioriza padrões metodológicos e resultados comparáveis, evitando converter métricas calculadas em condições diferentes em um ranking único.

A natureza dos dados também precisa ser mencionada. Os *Datasets* 1 e 2 desta pesquisa são artificiais; o *Dataset* 3 é apresentado pelo disponibilizador como real. Na

literatura, Mortezapour Shiri et al. (2025) e Baydili e Tasci (2025) combinam o conjunto da IBM com *datasets* apresentados como reais (44, 6), enquanto a base de Chaudhary et al. (2022) é artificial (12). Essas classificações descrevem proveniência, mas não demonstram representatividade, ausência de viés ou maior dificuldade preditiva.

Em relação aos objetivos da dissertação, o objetivo geral de avaliar diferentes modelos em bases distintas e comparar seus resultados foi atendido por meio das 30 repetições experimentais. Os seis classificadores apresentaram comportamentos distintos conforme o *dataset*, a representação, a condição de treinamento e a métrica considerada. As configurações de referência também variaram entre os contextos, sem evidência de um modelo ou estratégia universalmente superior.

O objetivo de comparar descritivamente os subgrupos registrados como masculino e feminino pôde ser avaliado apenas nos *Datasets* 1 e 3. No D1, a direção das diferenças variou conforme a métrica e a representação. No D3, os dois subgrupos permaneceram predominantemente com MCC próximo de zero e AUC-ROC próxima de 0,5. Dessa forma, não foi observada vantagem consistente de um dos subgrupos nos diferentes contextos avaliados. Essas comparações descrevem o desempenho preditivo dos modelos e não constituem avaliação formal de equidade algorítmica.

O objetivo relacionado à interpretação dos atributos foi examinado por meio do SHAP nas populações completas do Cenário 1. Para cada *dataset*, a análise foi realizada em uma execução representativa da respectiva configuração de referência, correspondente às sementes 54 no D1, 58 no D2 e 45 no D3. Portanto, os valores SHAP apresentados a seguir referem-se a essas execuções representativas e não à média de 30 execuções SHAP. No D1, os atributos com maior contribuição global para as predições foram `OverTime`, com 12,38% das contribuições absolutas, e `JobRole`, com 9,10%, seguidos por `EnvironmentSatisfaction`. No D2, houve maior concentração das contribuições em `satisfaction_level`, com 35,00%, `number_project`, com 17,53%, e `time_spend_company`, com 16,28%; os cinco primeiros atributos responderam conjuntamente por 92,96% das contribuições absolutas. No D3, destacaram-se `Number_of_Companies_Worked`, `Relationship_with_Manager` e `Job_Involvement`. Entretanto, a configuração analisada nessa base apresentou MCC médio de 0,005 e AUC-ROC média de 0,509 nas 30 sementes, exigindo cautela adicional na interpretação. Em todos os casos, os valores SHAP descrevem a contribuição dos atributos para as predições dos modelos e não demonstram relações causais com o *turnover*.

O objetivo relacionado ao tratamento do desbalanceamento foi atendido pela comparação pareada da condição Desbalanceado com SMOTE, ADASYN, SMOTE-Tomek e SMOTE-ENN nas 30 sementes. SMOTE e SMOTE-Tomek apresentaram comportamentos próximos, com aumentos médios de aproximadamente 18,3 pontos percentuais no *recall* e 8,7 pontos no F1-score. O SMOTE-ENN produziu a alteração mais intensa, com aumento médio de 46,53 pontos percentuais no *recall* e de 13,73 pontos no F1-score,

acompanhado por redução de 24,22 pontos na acurácia e aumento de 0,424 no Log Loss. O ADASYN também elevou, no agregado, *recall* e F1-score, mas seus efeitos variaram entre os contextos. No D2, por exemplo, a variação média do F1-score foi de apenas $+0,39$ ponto percentual. Assim, os resultados não sustentam uma técnica de balanceamento universalmente superior.

O objetivo referente à seleção e à redução de atributos também apresentou resultados dependentes da base. No Cenário 2, a seleção automática foi ajustada exclusivamente sobre os dados de treinamento e reduziu a dimensionalidade sem produzir melhora média consistente em relação ao Cenário 1. Nas populações completas, as diferenças pareadas de F1-score entre C2 e C1 foram de $-1,22$, $-0,49$ e $-0,06$ ponto percentual nos Datasets 1, 2 e 3, respectivamente. No Cenário 3, a representação de seis atributos produziu uma redução mais acentuada no D1, uma redução menor no D2 e não recuperou a capacidade discriminativa do D3. A representação de 19 atributos recuperou grande parte do desempenho no D1, enquanto no D3 as métricas discriminativas permaneceram próximas das observadas nas demais representações. Dessa forma, os resultados não sustentam que a redução do número de atributos, isoladamente, determine melhora preditiva.

O objetivo de avaliar múltiplas métricas também se mostrou relevante. Acurácia, precisão, *recall*, F1-score, MCC, AUC-ROC e Log Loss evidenciaram aspectos distintos do comportamento dos classificadores. O Índice Multivariado Ponderado complementou essa análise ao integrar F1-score, MCC, AUC-ROC e Log Loss e ordenar, em cada semente, as 30 configurações pertencentes ao mesmo contexto comparável. A consolidação das 30 posições por configuração permitiu considerar simultaneamente desempenho médio e estabilidade do *ranking*, sem utilizar o índice para comparar diretamente diferentes *datasets*, populações, cenários ou representações.

Em conjunto, os resultados mostraram que o desempenho dos classificadores depende da combinação entre dados, representação, tratamento do desbalanceamento, modelo e critério de avaliação. As 30 repetições permitiram verificar que algumas configurações permaneceram entre as mais bem posicionadas com maior frequência, enquanto outras apresentaram maior alternância entre sementes. A análise SHAP complementou essa avaliação ao mostrar quais atributos contribuíram mais para as predições das configurações analisadas no Cenário 1, sem transformar essas associações preditivas em explicações causais do *turnover*. As implicações gerais desses resultados, as limitações da pesquisa e as possibilidades de continuidade são apresentadas no capítulo seguinte.

# 6 CONCLUSÃO

Esta pesquisa avaliou e comparou seis classificadores aplicados a três conjuntos públicos de dados de funcionários para a previsão de *turnover*, considerando diferentes condições de reamostragem, representações dos atributos e análises estratificadas por sexo. O objetivo geral foi atendido pela comparação sistemática dessas configurações, cujos resultados mostraram que o desempenho variou conforme a base analisada, a representação dos atributos, o tratamento do desbalanceamento e a métrica utilizada para avaliação. Dessa forma, os resultados não sustentam a existência de um algoritmo ou de uma estratégia universalmente superior para as diferentes condições estudadas.

Nas populações completas do Cenário 1, as configurações de referência definidas pelo maior Índice Multivariado médio nas 30 sementes foram LDA com a condição Desbalanceado no *Dataset* 1, XGBoost com SMOTE-Tomek no *Dataset* 2 e AdaBoost com SMOTE-ENN no *Dataset* 3. No D1, a configuração apresentou F1-score de $52,73 \pm 5,52\%$, MCC de $0,481 \pm 0,065$ e AUC-ROC de $0,828 \pm 0,028$; ocupou a primeira posição em 11 das 30 sementes e permaneceu no Top 3 em 60,0% das repetições. No D2, os valores foram $94,15 \pm 0,72\%$, $0,930 \pm 0,009$ e $0,984 \pm 0,003$, com primeira posição em 12 sementes e presença no Top 3 em 86,7%. No D3, apesar da primeira posição relativa da configuração de referência, o F1-score foi de $32,79 \pm 0,46\%$, o MCC de $0,005 \pm 0,021$ e a AUC-ROC de $0,509 \pm 0,016$, com *rank* médio de 9,17. Assim, o D2 apresentou o melhor desempenho absoluto, o D1 mostrou desempenho intermediário e maior alternância, e o D3 permaneceu próximo ao acaso nas métricas discriminativas. A primeira posição em um *ranking* interno não implica, portanto, desempenho absoluto elevado.

O primeiro objetivo específico, referente à comparação descritiva dos subgrupos registrados nas bases como masculino e feminino, pôde ser avaliado nos *Datasets* 1 e 3. No Cenário 1 do D1, o F1-score médio foi de $44,45 \pm 4,06\%$ no subgrupo masculino e de $42,90 \pm 4,15\%$ no feminino, enquanto MCC e AUC-ROC permaneceram próximos entre os grupos. A direção das diferenças variou conforme a métrica e a representação: no Cenário 3 com seis atributos, por exemplo, o F1-score foi numericamente maior no subgrupo masculino, mas a AUC-ROC foi maior no feminino. No D3, ambos os subgrupos permaneceram predominantemente com MCC próximo de zero e AUC-ROC próxima de 0,5. Dessa forma, não foi observada vantagem consistente de um dos subgrupos nos diferentes contextos avaliados. Essas comparações descrevem o comportamento preditivo de modelos ajustados separadamente em populações distintas e não permitem interpretar o sexo como causa do *turnover*, identificar discriminação ou caracterizar uma avaliação formal de equidade algorítmica.

O objetivo de examinar, de forma exploratória, os atributos com maior relevância preditiva foi atendido por meio do SHAP nas populações completas do Cená-

rio 1. Para cada *dataset*, foi analisada uma execução representativa da configuração de referência, correspondente às sementes 54 no D1, 58 no D2 e 45 no D3; esses valores não representam a média de 30 execuções SHAP. No D1, `OverTime` e `JobRole` responderam por 12,38% e 9,10% das contribuições absolutas, respectivamente. No D2, houve maior concentração em `satisfaction_level`, com 35,00%, `number_project`, com 17,53%, e `time_spend_company`, com 16,28%; os cinco primeiros atributos reuniram 92,96% das contribuições absolutas. No D3, destacaram-se `Number_of_Companies_Worked`, `Relationship_with_Manager` e `Job_Involvement`; entretanto, o MCC médio de 0,005 e a AUC-ROC média de 0,509 da configuração exigem cautela adicional. Em todos os casos, o SHAP descreve como os atributos contribuíram para as predições dos modelos analisados e não demonstra relações causais com o *turnover*.

O objetivo relacionado ao tratamento do desbalanceamento foi atendido pela comparação pareada da condição Desbalanceado com SMOTE, ADASYN, SMOTE-Tomek e SMOTE-ENN nas 30 sementes. SMOTE e SMOTE-Tomek apresentaram comportamentos próximos, com aumentos médios de aproximadamente 18,3 pontos percentuais no *recall* e 8,7 pontos no F1-score. O SMOTE-ENN produziu a alteração mais intensa, com aumento médio de 46,53 pontos percentuais no *recall* e de 13,73 pontos no F1-score, acompanhado por redução de 24,22 pontos na acurácia e aumento de 0,424 no Log Loss. O ADASYN também elevou, no agregado, *recall* e F1-score, mas seus efeitos variaram entre os contextos; no D2, por exemplo, a variação média do F1-score foi de apenas $+0,39$ ponto percentual. A reamostragem modificou principalmente o compromisso entre a identificação da classe positiva e os demais componentes do desempenho, sem sustentar uma técnica universalmente superior.

A avaliação da seleção e da redução de atributos apresentou resultados dependentes da base e da representação. No Cenário 2, a seleção automática foi ajustada exclusivamente sobre o conjunto de treinamento de cada semente. Nas populações completas, as diferenças pareadas de F1-score entre C2 e C1 foram de $-1,22$, $-0,49$ e $-0,06$ ponto percentual nos *Datasets* 1, 2 e 3, respectivamente, sem evidência de benefício preditivo geral. No Cenário 3 com seis atributos, o F1-score variou $-14,72$ pontos no D1, $-2,37$ no D2 e $+1,99$ no D3 em relação ao Cenário 1; no D3, esse aumento não foi acompanhado por melhora relevante de MCC ou AUC-ROC. A representação de 19 atributos ficou próxima da representação completa no D1 e no D3, com diferenças de F1-score de $-1,04$ e $-0,03$ ponto, respectivamente. Como a harmonização também envolve diferenças de definição, escala e registro entre as bases, os resultados não isolam um efeito causal da quantidade de atributos e não sustentam que reduzir ou ampliar o conjunto de variáveis produza melhora geral do desempenho.

O objetivo de analisar diferentes métricas mostrou-se particularmente relevante para a interpretação dos experimentos. Quando a prioridade em um contexto de RH for identificar a maior proporção possível de ocorrências da classe positiva, o *recall* deve

receber maior atenção, reconhecendo-se o possível aumento de falsos positivos. Quando a capacidade de intervenção for limitada ou o custo associado a um falso alerta for elevado, a precisão assume maior relevância. O F1-score permite avaliar o equilíbrio entre precisão e *recall*, enquanto o MCC sintetiza o desempenho considerando os quatro componentes da matriz de confusão. A AUC-ROC avalia a capacidade de ordenação entre as classes sem depender de um único limiar de classificação, e o Log Loss complementa a avaliação ao considerar as probabilidades produzidas, embora estas não tenham sido calibradas no estudo como riscos organizacionais. Assim, a definição de uma configuração adequada depende também das consequências práticas associadas aos diferentes tipos de erro.

Nesse contexto, o Índice Multivariado Ponderado constituiu uma contribuição metodológica exploratória para integrar F1-score, MCC, AUC-ROC e Log Loss e ordenar as configurações dentro de cada combinação de cenário e *dataset*. Sua utilização permitiu evitar que a seleção das configurações com melhor desempenho observado fosse baseada exclusivamente em uma única métrica. No *Dataset* 2, o índice selecionou configurações que também apresentaram desempenho absoluto elevado nas métricas originais; no *Dataset* 1, evidenciou a alternância das configurações mais bem colocadas conforme a representação e a população; e, no *Dataset* 3, tornou evidente que uma primeira posição relativa poderia coexistir com valores absolutos baixos de MCC e AUC-ROC. O índice auxilia, portanto, a organização da comparação multivariada, mas não substitui a definição dos custos de falsos positivos e falsos negativos nem deve ser interpretado isoladamente dos valores das métricas que o compõem.

A principal contribuição desta pesquisa consiste na análise comparativa de como diferentes decisões do *pipeline* modificam a interpretação do desempenho dos modelos de previsão de *turnover*. Ao utilizar três bases com comportamentos distintos, diferentes representações de atributos, quatro técnicas de reamostragem, análises estratificadas por sexo, múltiplas métricas e 30 divisões estratificadas, o estudo mostrou que uma configuração pode apresentar vantagem em determinado critério sem manter essa posição sob outra métrica, representação, população ou semente. O Índice Multivariado permitiu organizar a comparação dentro de cada contexto, enquanto o SHAP acrescentou uma leitura das contribuições dos atributos nas execuções representativas selecionadas. A contribuição do estudo não está, portanto, na indicação de um classificador único para aplicação generalizada, mas nos critérios e nas cautelas necessários para comparar, verificar a estabilidade interna e interpretar modelos preditivos de *turnover*.

Para a gestão de pessoas, as previsões podem apoiar a formulação de perguntas, a identificação de padrões coletivos e a priorização de diagnósticos, mas não substituem a escuta dos trabalhadores, a compreensão dos processos organizacionais ou o julgamento responsável. As contribuições SHAP não provam causas; falsos positivos podem gerar estigma e tratamento inadequado; e nem toda saída precisa ser impedida. Os modelos não devem ser utilizados para punir indivíduos, medir lealdade ou definir perfis de trabalhadores

desejáveis e indesejáveis. Uma eventual aplicação exige governança, proteção dos dados, finalidade legítima, minimização, transparência, prevenção de discriminação, revisão humana e avaliação contínua das consequências das decisões (9, 21, 37).

O alcance dessas conclusões permanece delimitado pelas bases públicas, pelos rótulos genéricos de saída e pelo procedimento experimental utilizado. As 30 repetições permitem examinar a estabilidade diante de diferentes divisões da mesma amostra, mas não constituem validação externa. As métricas dos conjuntos de teste também participaram da seleção das configurações de referência, sem um segundo conjunto independente reservado para avaliação posterior. Além disso, o SHAP foi aplicado a uma execução representativa de cada configuração de referência, e não a 30 execuções por *dataset*. Assim, os resultados não correspondem à validação prospectiva de um sistema pronto para utilização em organizações reais e devem ser interpretados dentro do contexto estudado.

## 6.1 LIMITAÇÕES DA PESQUISA

A primeira limitação está relacionada à natureza, à documentação e à representatividade das bases de dados. Os *Datasets* 1 e 2 são artificiais, enquanto o *Dataset* 3 é apresentado pelo disponibilizador como real, mas não possui documentação independente suficiente sobre a organização de origem, o processo de coleta, o período, a amostragem e o contexto em que os registros foram produzidos. Os três conjuntos públicos também diferem em tamanho, dimensionalidade, distribuição das classes, estrutura e qualidade dos atributos, o que restringe comparações perfeitamente controladas entre eles. A utilização frequente de determinado conjunto na literatura não comprova sua validade externa ou sua representatividade em relação a organizações reais. Além disso, as variáveis-alvo utilizadas registram *attrition*, `left` ou saída do empregado, mas não permitem distinguir com segurança turnover voluntário de involuntário. As conclusões, portanto, referem-se ao *turnover* na forma genérica operacionalizada pelos dados e não devem ser interpretadas como previsão exclusiva de desligamentos voluntários. As bases também não permitem distinguir saídas evitáveis e inevitáveis nem classificá-las como funcionais ou disfuncionais para a organização e para o trabalhador.

A própria representação do fenômeno organizacional é limitada pelo tipo de informação disponível. As bases utilizadas são públicas e possuem estrutura predominantemente transversal, sem acompanhamento sistemático da trajetória dos funcionários ao longo do tempo. Dessa forma, não capturam integralmente mudanças organizacionais, acontecimentos recentes, alterações na gestão, relações interpessoais, clima organizacional em tempo real ou motivos qualitativos individuais que podem participar de uma decisão de saída. Essa limitação não elimina a utilidade da análise quantitativa, mas restringe o alcance substantivo das variáveis observadas: um modelo pode identificar padrões presentes nos registros disponíveis sem representar toda a dinâmica que antecede o turnover em uma

organização.

O desenho experimental também impõe restrições à interpretação dos resultados. Foram realizadas 30 divisões estratificadas entre treinamento e teste, com sementes de 42 a 71, o que permitiu observar a variabilidade das métricas e das posições no *ranking* diante de diferentes particionamentos da mesma amostra. Essas repetições constituem *hold-outs* estratificados e não correspondem a validação cruzada nem a validação externa. As técnicas de balanceamento foram aplicadas somente ao treinamento; entretanto, a condição denominada Desbalanceado não empregou exatamente o mesmo mecanismo de compensação em todos os classificadores, pois alguns modelos utilizaram pesos internos de classe e outros não. Nas condições reamostradas, o F1-score de treinamento foi calculado sobre a amostra modificada, de modo que o *gap* entre treinamento e teste combina ajuste do modelo e mudança na composição dos dados. Adicionalmente, as métricas dos conjuntos de teste participaram do Índice Multivariado utilizado para selecionar as configurações de referência, sem uma segunda amostra independente reservada para avaliação posterior à seleção. Por isso, os primeiros lugares devem ser entendidos como resultados observados e internamente estáveis no procedimento executado, e não como estimativas independentes de desempenho futuro.

Uma limitação adicional diz respeito à utilização das técnicas de reamostragem após a realização do One-Hot Encoding e da padronização dos preditores. SMOTE, ADASYN, SMOTE-Tomek e SMOTE-ENN atuaram sobre a representação numérica resultante, o que significa que a interpolação em colunas originadas de variáveis categóricas pode ter gerado valores que não correspondem diretamente às categorias inicialmente observadas. Enquanto o conjunto de teste foi mantido sem reamostragem, essas instâncias permaneceram limitadas ao treinamento. Dessa forma, as métricas de teste ainda refletem o desempenho do pipeline em relação aos registros originais, porém os resultados não confirmam que as amostras sintéticas representem perfis organizacionais semanticamente válidos. Além disso, não é garantido que os efeitos observados sejam replicados com diferentes métodos de codificação e tratamento de variáveis categóricas.

As limitações específicas dos Cenários 2 e 3 também condicionam parte das conclusões. No Cenário 2, a seleção foi ajustada exclusivamente sobre o conjunto de treinamento de cada semente, evitando acesso ao conjunto de teste. Permanecem, contudo, a dependência de um único método de seleção, o limiar fixo de importância superior a 2% e a natureza preditiva e específica ao Random Forest das importâncias utilizadas. Por isso, uma *feature* não selecionada não deve ser considerada substantivamente irrelevante. No Cenário 3, as representações de seis e 19 atributos foram construídas a partir dos dados após a limpeza inicial e antes do *One-Hot Encoding*, com correspondências conceituais entre as bases. Algumas variáveis exigiram regras de derivação ou conversão, e diferenças de definição, escala e forma de registro permaneceram possíveis. Dessa forma, os contrastes entre cenários descrevem simultaneamente mudanças de dimensionalidade e de representação e

não isolam rigorosamente o efeito do número de atributos sobre o desempenho.

A segmentação por sexo deve ser interpretada dentro de um escopo igualmente restrito. Ela foi possível apenas nos *Datasets* 1 e 3 e reduziu o tamanho das amostras analisadas, com especial sensibilidade nos subgrupos com menor quantidade de ocorrências positivas. As diferenças de desempenho entre homens e mulheres variaram conforme o cenário e o modelo, não apresentando direção estável. Como os modelos foram ajustados separadamente para os subgrupos e não foram calculadas métricas específicas de equidade, a análise realizada não corresponde a uma avaliação completa de viés algorítmico. Tampouco é possível interpretar diferenças preditivas como efeitos causais do sexo sobre o turnover, pois o desenho não foi construído para inferência causal.

O estudo também não avaliou a implementação dos modelos em uma organização, o modo como gestores reagiriam aos alertas nem os efeitos dessas classificações sobre os trabalhadores. Em um contexto real, falsos positivos podem produzir estigmatização, vigilância, exclusão de oportunidades ou intervenções desnecessárias, enquanto previsões corretas também podem ser utilizadas de forma inadequada. As bases públicas não permitiram testar procedimentos de governança, segurança, acesso restrito, anonimização, transparência, revisão de decisões ou participação dos trabalhadores. Assim, a conformidade com a LGPD e a prevenção de usos discriminatórios constituem requisitos para uma eventual aplicação, mas não foram empiricamente avaliadas nesta pesquisa (9, 21, 37).

A generalização para contextos organizacionais reais permanece, assim, limitada. O desempenho observado reflete as populações, atributos, targets e procedimentos presentes nos conjuntos utilizados, e não garante resultado equivalente em outras organizações, setores, regiões ou períodos. Essa cautela também se aplica ao escopo dos algoritmos: a pesquisa foi delimitada aos seis classificadores efetivamente avaliados. A ausência de modelos de *Deep Learning* constitui uma delimitação do estudo, e não necessariamente uma falha metodológica, pois abordagens mais complexas exigem volume, estrutura e qualidade de dados compatíveis e não há fundamento para presumir que produziriam desempenho superior nas bases utilizadas. A comparação também está restrita às implementações e parametrizações adotadas no experimento, sem busca sistemática de hiperparâmetros para cada combinação de modelo e base. Essa restrição é particularmente relevante para a interpretação do baixo desempenho observado no *Dataset* 3: os resultados demonstram o comportamento das parametrizações avaliadas, mas não permitem concluir que representam o melhor desempenho alcançável pelas respectivas famílias de algoritmos. Assim, diferenças entre classificadores e bases não devem ser interpretadas como limites absolutos de desempenho das técnicas avaliadas. A aplicabilidade mais ampla dependeria não apenas de novos algoritmos, mas sobretudo de dados adequados e validação independente no contexto de uso.

Por fim, é necessário distinguir reprodutibilidade computacional, replicabilidade

e validade externa. A disponibilidade do *pipeline*, das sementes, dos identificadores de execução, dos conjuntos de treinamento e teste, das *features*, das matrizes de confusão e das versões principais do ambiente computacional favorece a reprodução do procedimento com os mesmos dados. Entretanto, o ambiente completo, incluindo sistema operacional, hardware e todas as dependências transitivas, não foi congelado em uma imagem computacional única; por isso, mudanças de implementação ou de parâmetros padrão ainda podem afetar a reprodução numérica exata. Mesmo uma reprodução bem-sucedida não demonstra replicabilidade em uma nova amostra, e a replicação em outra amostra ainda não garante generalização para populações organizacionais distintas. Assim, a rastreabilidade registrada fortalece a reprodutibilidade do procedimento, mas não substitui validação externa nem elimina a dependência dos resultados em relação à amostra, à prevalência da classe positiva e ao conjunto de variáveis disponível em cada base.

## 6.2 TRABALHOS FUTUROS

Uma continuidade prioritária consiste em avaliar os procedimentos desenvolvidos em bases organizacionais reais, anonimizadas e documentalmente rastreáveis, provenientes de diferentes organizações e contextos de trabalho. Além de ampliar a validade externa, esses conjuntos deveriam registrar de forma mais precisa a natureza do desligamento, permitindo distinguir turnover voluntário e involuntário e, quando possível, caracterizar o período e o contexto em que a saída ocorreu. Essa ampliação permitiria verificar se os padrões observados nas bases públicas se mantêm em dados produzidos por processos reais de gestão de pessoas e reduziria a dependência de conjuntos artificiais ou de proveniência insuficientemente documentada.

O uso de dados longitudinais representa uma extensão complementar. Acompanhamentos ao longo do tempo permitiriam analisar mudanças no comportamento dos funcionários, alterações nas condições organizacionais e a estabilidade dos modelos em períodos distintos. Esse desenho também possibilitaria investigar mudança de distribuição e *concept drift*, avaliando se relações preditivas identificadas em um período permanecem válidas depois de mudanças na organização ou no mercado de trabalho. Em paralelo, validações externas entre organizações, setores, regiões e janelas temporais distintas permitiriam separar desempenho interno da capacidade efetiva de transferência para novos contextos.

Embora a variabilidade entre 30 sementes tenha sido incorporada nesta pesquisa, o desenho de avaliação pode ser aprofundado por procedimentos complementares, como validação cruzada repetida, validação aninhada e separação de um conjunto final independente para avaliação posterior à seleção de atributos e das configurações de referência. Comparações estatísticas futuras também podem adotar unidades experimentais e modelos compatíveis com a estrutura de repetição, incorporando, quando necessário, efeitos de

*dataset*, cenário e interações. Esses procedimentos permitiriam estimar a incerteza sob outros protocolos e reduzir a reutilização dos conjuntos de teste no processo de seleção.

Outra linha de continuidade é ampliar a investigação do desbalanceamento e do uso das probabilidades preditas. Estratégias de aprendizagem sensível a custos podem ser comparadas às técnicas de reamostragem, enquanto calibração de probabilidades e ajuste de *thresholds* podem aproximar o modelo das diferentes consequências associadas a falsos positivos e falsos negativos. Como o SHAP foi aplicado a uma execução representativa de cada configuração do Cenário 1, trabalhos futuros podem avaliar a estabilidade das contribuições entre sementes, configurações e períodos, além de examinar explicações locais e a forma como gestores e trabalhadores compreendem essas informações, sempre sem convertê-las em relações causais. Análises de subgrupos também podem ser ampliadas com métricas específicas e amostras adequadas quando o objetivo for investigar equidade. Quando houver acesso ao contexto organizacional, entrevistas, questionários, pesquisa de clima ou outras fontes qualitativas podem complementar os registros estruturados.

Modelos mais complexos, incluindo arquiteturas de *Deep Learning*, podem ser investigados quando houver volume de dados suficiente ou quando a estrutura temporal, textual ou não estruturada justificar sua utilização. Essa extensão deve ser acompanhada de comparação com modelos mais simples sob o mesmo protocolo de validação, evitando assumir que maior complexidade implica melhor desempenho. Em aplicações organizacionais, trabalhos futuros também podem acompanhar os modelos após sua implementação, considerando monitoramento de desempenho, atualização periódica, mudanças de distribuição, governança e proteção dos dados e os efeitos do uso das previsões sobre decisões de gestão de pessoas. Dessa forma, a continuidade da pesquisa pode avançar da comparação experimental para a avaliação de sistemas preditivos em condições mais próximas de seu uso real, mantendo como requisito a validação contínua de sua utilidade e de seus limites.

# REFERÊNCIAS


1 AHMED, Tarig Mohamed. A Novel Classification Model for Employees Turnover Using Neural Network to Enhance Job Satisfaction in Organizations. **Journal of Information and Organizational Sciences**, v. 45, n. 2, p. 361–374, 2021.

2 ALLEN, D. G.; BRYANT, P. C.; VARDAMAN, J. M. Retaining talent: replacing misconceptions with evidence-based strategies. **Academy of Management Perspectives**, v. 24, n. 2, p. 48–64, 2010.

3 ANNAMALAI, Sumathi. Influence of perceived equity, job enrichment, and burnout among educators in Indian private universities on job satisfaction and the desire to quit. **Frontiers in Education**, v. 7, p. 991068, 2022.

4 BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20–29, 2004.

5 BAUTISTA, Paula Zeah N.; CAHIGAS, Maela Madel L. Exploring Employee Retention among Generation Z Engineers in the Philippines Using Machine Learning Techniques. **Sustainability**, v. 16, n. 12, p. 5207, 2024.

6 BAYDILI, İrem Tanyıldızı; TASCI, Burak. Predicting Employee Attrition: XAI-Powered Models for Managerial Decision-Making. **Systems**, Basel, v. 13, n. 7, p. 583, 2025.

7 BENABOU, Adil; TOUHAMI, Fatima; ABDELOUAHED SABRI, My. Predicting Employee Turnover Using Machine Learning Techniques. **Acta Informatica Pragensia**, v. 14, n. 1, p. 112–127, 2025.

8 BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006.

9 BRASIL. Presidência da República. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Acesso em: 20 out. 2026.

10 BRASIL. Ministério do Trabalho e Emprego. **O que é o Novo Caged?** Brasília, DF: Ministério do Trabalho e Emprego, [s.d.]. Disponível em: `https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/estatisticas-trabalho/o-pdet/o-que-e-o-novo-caged`. Acesso em: 20 out. 2026.

11 BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

12 CHAUDHARY, Meenu; GAUR, Loveleen; JHANJHI, N. Z.; MASUD, Mehedi; ALJAHDALI, Sultan. Envisaging Employee Churn Using MCDM and Machine Learning. **Intelligent Automation & Soft Computing**, v. 33, n. 2, p. 1009–1024, 2022.

13 CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002.

14 CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. **Proceedings...** New York: ACM, 2016. p. 785–794.

15 CHEN, Xuelin; AL MAMUN, Abdullah; HUSSAIN, Wan Mohd Hirwani Wan; JINGZU, Gao; YANG, Qing; SHAMI, Sayed Samer Ali Al. Envisaging the job satisfaction and turnover intention among the young workforce: Evidence from an emerging economy. **PLOS ONE**, v. 18, n. 6, p. e0287284, 2023.

16 CPS, Heno Merlin; KRISHNAN, Jayasree. Analyzing Employee Attrition Drivers: The Impact of Burnout Through Predictive Models. **International Journal on Science and Technology**, v. 16, n. 2, p. 6248, 2025.

17 CROOK, T. R.; TODD, S. Y.; COMBS, J. G.; WOEHR, D. J.; KETCHEN JR., D. J. Does human capital matter? A meta-analysis of the relationship between human capital and firm performance. **Journal of Applied Psychology**, v. 96, n. 3, p. 443–456, 2011.

18 DŽAMBIĆ, Amra; HADZIAHMETOVIC, Nereida. Retention in the IT industry: exploring the mediating role of knowledge-sharing in the relationship between authentic leadership style and employee turnover. **Central European Management Journal**, 2025.

19 FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

20 FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, 1997.

21 GIERMINDL, Lisa Marie; STRICH, Franz; CHRIST, Oliver; LEICHT-DEOBALD, Ulrich; REDZEPI, Abdullah. The dark sides of people analytics: reviewing the perils for organisations and employees. **European Journal of Information Systems**, v. 31, n. 3, p. 410–435, 2022. DOI: 10.1080/0960085X.2021.1927213.

22 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016.

23 GUERRANTI, Filippo; DIMITRI, Giovanna Maria. A Comparison of Machine Learning Approaches for Predicting Employee Attrition. **Applied Sciences**, v. 13, n. 1, p. 267, 2022.

24 HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. Waltham, MA: Morgan Kaufmann, 2011.

25 HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

26 HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In*: IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2008, Hong Kong. **Proceedings...** IEEE, 2008. p. 1322–1328.

27 HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009.

28 HINGE, Punamkumar; SALUNKHE, Harshal A.; BORALKAR, Mohit; BANG, Sanjay; THAKUR, Abhijeet; ADHEGAONKAR, Vikas R. Measuring employee attrition intention in an auto-component manufacturing organisation. **SA Journal of Human Resource Management**, v. 23, p. a3003, 2025.

29 HOM, P. W.; LEE, T. W.; SHAW, J. D.; HAUSKNECHT, J. P. One hundred years of employee turnover theory and research. **Journal of Applied Psychology**, v. 102, n. 3, p. 530–545, 2017.

30 HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. 3. ed. Hoboken: John Wiley & Sons, 2013.

31 IQBAL, Javed; ASGHAR, Ali; ASGHAR, Muhammad Zaheer. Effect of Despotic Leadership on Employee Turnover Intention: Mediating Toxic Workplace Environment and Cognitive Distraction in Academic Institutions. **Behavioral Sciences**, v. 12, n. 5, p. 125, 2022.

32 JAIN, Praphula Kumar; JAIN, Madhur; PAMULA, Rajendra. Explaining and predicting employees' attrition: a machine learning approach. **SN Applied Sciences**, v. 2, n. 4, p. 757, 2020.

33 JUN, Kiho; HU, Zhehua; SUN, Yi. Impact of authentic leadership on employee turnover intention: Perceived supervisor support as mediator and organizational identification as moderator. **Frontiers in Psychology**, v. 14, p. 1009639, 2023.

34 KERSE, Gökhan; KOÇAK, Daimi; BABADAG, Mustafa. Relationship between Person–Environment Fit Types and Turnover Intention: A Moderated Mediation Model. **Organizations and Markets in Emerging Economies**, v. 13, n. 2, p. 384–405, 2022.

35 KOPPERUD, Karoline Hofslett; NERSTAD, Christina G. L.; DYSVIK, Anders. Should I Stay or Should I Go? The Role of Motivational Climate and Work–Home Spillover for Turnover Intentions. **Frontiers in Psychology**, v. 11, p. 1107, 2020.

36 KUHN, M.; JOHNSON, K. **Applied predictive modeling**. New York: Springer, 2013.

37 LEICHT-DEOBALD, Ulrich; BUSCH, Thorsten; SCHANK, Christoph; WEIBEL, Antoinette; SCHAFHEITLE, Simon; WILDHABER, Isabelle; KASPER, Gabriel. The challenges of algorithm-based HR decision-making for personal integrity. **Journal of Business Ethics**, v. 160, p. 377–392, 2019. DOI: 10.1007/s10551-019-04204-w.

38 MAHLASELA, Samkelisiwe; CHINYAMURINDI, Willie T. Technology-related factors and their influence on turnover intentions: A case of government employees in South Africa. **The Electronic Journal of Information Systems in Developing Countries**, v. 86, n. 3, p. e12126, 2020.

39 MARLER, J. H.; BOUDREAU, J. W. An evidence-based review of HR Analytics. **The International Journal of Human Resource Management**, v. 28, n. 1, p. 3–26, 2017.

40 MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochimica et Biophysica Acta (BBA)-Protein Structure**, v. 405, n. 2, p. 442-451, 1975.

41 MCKINNEY, W. Data Structures for Statistical Computing in Python. In: **Proceedings of the 9th Python in Science Conference**. Austin, TX, 2010. p. 51-56.

42 MCLACHLAN, G. J. **Discriminant analysis and statistical pattern recognition**. Hoboken: John Wiley & Sons, 2004.

43 MOBLEY, W. H. Some unanswered questions in turnover and withdrawal research. **Academy of Management Review**, Briarcliff Manor, v. 7, n. 1, p. 111–116, 1982.

44 MORTEZAPOUR SHIRI, Farhad; YAMAGUCHI, Shingo; AHMADON, Mohd Anuaruddin Bin. A Deep Learning Model Based on Bidirectional Temporal Convolutional Network (Bi-TCN) for Predicting Employee Attrition. **Applied Sciences**, v. 15, n. 6, p. 2984, 2025.

45 MUSTAMIL, Norizah; NAJAM, Usama. The Impact of Servant Leadership on Follower Turnover Intentions: Mediating Role of Resilience. **Asian Journal of Business and Accounting**, v. 13, n. 2, p. 125–146, 2020.

46 NAJAFI-ZANGENEH, Saeed; SHAMS-GHARNEH, Naser; ARJOMANDI-NEZHAD, Ali; HASHEMKHANI ZOLFANI, Sarfaraz. An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. **Mathematics**, v. 9, n. 11, p. 1226, 2021.

47 NEGT, Philip; HAUNSCHILD, Axel. Exploring the gap between research and practice in human resource management (HRM): a scoping review and agenda for future research. **Management Review Quarterly**, v. 75, n. 1, p. 837–879, 2025. DOI: 10.1007/s11301-023-00397-7.

48 NOGUEIRA, Thallys da Silva. **Observatório do consumidor: uma ferramenta de mineração de dados de redes sociais para avaliação de tendências de consumo de derivados lácteos no Brasil**. 2025. Tese (Doutorado em Modelagem Computacional) – Universidade Federal de Juiz de Fora, Juiz de Fora, 2025. Disponível em: `https://repositorio.ufjf.br/jspui/handle/ufjf/18795`. Acesso em: 8 ago. 2026.

49 OHU, Eugene A. Too much information: When work-family conflict empowers senior managers to stay. **SA Journal of Industrial Psychology**, v. 49, p. 2033, 2023.

50 OUZZANI, Mourad; HAMMADY, Hossam; FEDOROWICZ, Zbys; ELMAGARMID, Ahmed. Rayyan: a web and mobile app for systematic reviews. **Systematic Reviews**, London, v. 5, n. 1, p. 210, 2016.

51 PAGE, Matthew J.; McKENZIE, Joanne E.; BOSSUYT, Patrick M.; BOUTRON, Isabelle; HOFFMANN, Tammy C.; MULROW, Cynthia D.; et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, London, v. 372, p. n71, 2021.

52 PARK, T.-Y.; SHAW, J. D. Turnover rates and organizational performance: a meta-analysis. **Journal of Applied Psychology**, Washington, v. 98, n. 2, p. 268–309, 2013.

53 PARK, Jungryeol; FENG, Yituo; JEONG, Seon-Phil. Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques. **Scientific Reports**, v. 14, n. 1, p. 1221, 2024.

54 PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

55 PRATT, Madara; BOUDHANE, Mohcine; CAKULA, Sarma. Employee Attrition Estimation Using Random Forest Algorithm. **Baltic Journal of Modern Computing**, v. 9, n. 1, p. 49–66, 2021.

56 RAZA, Ali; MUNIR, Kashif; ALMUTAIRI, Mubarak; YOUNAS, Faizan; FAREED, Mian Muhammad Sadiq. Predicting Employee Attrition Using Machine Learning Approaches. **Applied Sciences**, v. 12, n. 13, p. 6424, 2022.

57 RISH, I. An empirical study of the naive Bayes classifier. *In*: IJCAI WORKSHOP ON EMPIRICAL METHODS IN ARTIFICIAL INTELLIGENCE, 3., 2001, Seattle. **Proceedings...** Seattle: IJCAI, 2001. p. 41–46.

58 RUBENSTEIN, Alex L.; EBERLY, Marion B.; LEE, Thomas W.; MITCHELL, Terence R. Surveying the forest: a meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. **Personnel Psychology**, v. 71, n. 1, p. 23–65, 2018. DOI: 10.1111/peps.12226.

59 SANTOS, C. M. C.; PIMENTA, C. A. M.; NOBRE, M. R. C. The PICO strategy for the research question construction and evidence search. **Revista Latino-Americana de Enfermagem**, Ribeirão Preto, v. 15, n. 3, p. 508–511, 2007.

60 SHAW, J. D.; DUFFY, M. K.; JOHNSON, J. L.; LOCKHART, D. E. Turnover, social capital losses, and performance. **Academy of Management Journal**, v. 48, n. 4, p. 594–606, 2005.

61 SHLASH MOHAMMAD, Anber Abraheem; ALKHAZALI, Zeyad; SHELASH MOHAMMAD, Suleiman Ibrahim; AL ORAINI, Badrea; VASUDEVAN, Asokan; ALQAHTANI, Menahi Mosallam; ALSHURIDEH, Muhammad Turki. Machine Learning Models for Predicting Employee Attrition: A Data Science Perspective. **Data and Metadata**, v. 4, p. 669, 2025.

62 SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009.

63 SRIVASTAVA, Praveen Ranjan; EACHEMPATI, Prajwal. Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach. **Journal of Global Information Management**, v. 29, n. 6, p. 1–29, 2021.

64 SRIVASTAVA, Govind Nath; SHARMA, Himanshu; AGARWAL, Richa N.; JAIN, Ajay K. Analyzing employee attrition of research and development firms using mixed methods. **Cogent Business & Management**, v. 12, n. 1, p. 2565439, 2025.

65 TANER, Zeynep; HIZIROĞLU, Ourania Areta; HIZIROĞLU, Kadir. Leveraging Machine Learning Methods for Predicting Employee Turnover Within the Framework of Human Resources Analytics. **Journal of Intelligent Systems: Theory and Applications**, v. 7, n. 2, p. 145–158, 2024.

66 TOKMAK, Mahmut. DFCEA: Deep Forest Classifier-Based Employee Attrition Prediction. **Acta Infologica**, v. 9, n. 1, p. 19–33, 2025.

67 VAIYAPURI, Thavavel; SBAI, Zohra. Bayesian Optimized Boosted Ensemble models for HR Analytics - Adopting Green Human Resource Management Practices. **International Journal of Technology**, v. 16, n. 2, p. 561, 2025.

68 VARKIANI, Shahin Manafi; PATTARIN, Francesco; FABBRI, Tommaso; FANTONI, Gualtiero. Predicting employee attrition and explaining its determinants. **Expert Systems with Applications**, v. 272, p. 126575, 2025.

69 VEGLIO, Valerio; ROMANELLO, Rubina; PEDERSEN, Torben. Employee turnover in multinational corporations: a supervised machine learning approach. **Review of Managerial Science**, v. 19, n. 3, p. 687–728, 2025.

70 WU, Wei; FUKUI, Sadaaki. Using Human Resources Data to Predict Turnover of Community Mental Health Employees: Prediction and Interpretation of Machine Learning Methods. **International Journal of Mental Health Nursing**, v. 33, n. 6, p. 2180–2192, 2024.

71 ZHANG, Taotao; LI, Bingxiang. Job crafting and turnover intention: The mediating role of work engagement and job satisfaction. **Social Behavior and Personality: an international journal**, v. 48, n. 2, p. 1–9, 2020.

72 ŽIVKOVIĆ, Ana; PAP VORKAPIĆ, Ana; FRANJKOVIĆ, Jelena. Charting a Path to Sustainable Workforce: Exploring Influential Factors behind Employee Turnover Intentions in the Energy Industry. **Sustainability**, v. 16, n. 19, p. 8511, 2024.

# APÊNDICE A – Configurações selecionadas nos subgrupos por sexo

Este apêndice complementa a Tabela 15, apresentada no Capítulo 5 para as populações completas. A Tabela 18 reúne as configurações de referência dos subgrupos masculino e feminino dos *Datasets* 1 e 3 nos três cenários. Cada linha resume a mesma configuração avaliada nas 30 sementes; portanto, os valores não correspondem ao melhor resultado de uma única divisão entre treinamento e teste.

Tabela 18 – Configurações de referência e robustez do Índice Multivariado nos subgrupos por sexo ao longo das 30 sementes.

| Dataset | Cenário / representação | Subgrupo | Modelo + condição | F1 (%) | MCC | AUC-ROC | Rank médio | Rank 1 | Top 3 (%) |
|---|---|---|---|---|---|---|---|---|---|
| D1 | C1 – completo | Masculino | LDA + Desbalanceado | $52,54 \pm 9,49$ | $0,459 \pm 0,110$ | $0,830 \pm 0,042$ | 5,63 | 10/30 | 50,0 |
| D1 | C2 – seleção | Masculino | LDA + Desbalanceado | $45,13 \pm 7,99$ | $0,404 \pm 0,097$ | $0,804 \pm 0,044$ | 7,00 | 7/30 | 40,0 |
| D1 | C3 – 6 atributos | Masculino | AdaBoost + SMOTE | $38,45 \pm 4,02$ | $0,229 \pm 0,057$ | $0,701 \pm 0,039$ | 6,73 | 5/30 | 43,3 |
| D1 | C3 – 19 atributos | Masculino | LDA + Desbalanceado | $47,97 \pm 8,65$ | $0,426 \pm 0,094$ | $0,818 \pm 0,042$ | 7,57 | 9/30 | 46,7 |
| D1 | C1 – completo | Feminino | XGBoost + ADASYN | $48,29 \pm 10,09$ | $0,433 \pm 0,100$ | $0,829 \pm 0,051$ | 9,13 | 4/30 | 23,3 |
| D1 | C2 – seleção | Feminino | Naive Bayes + Desbalanceado | $48,82 \pm 9,56$ | $0,410 \pm 0,115$ | $0,774 \pm 0,071$ | 8,37 | 7/30 | 33,3 |
| D1 | C3 – 6 atributos | Feminino | AdaBoost + SMOTE-ENN | $35,95 \pm 5,97$ | $0,232 \pm 0,087$ | $0,720 \pm 0,061$ | 6,83 | 4/30 | 33,3 |
| D1 | C3 – 19 atributos | Feminino | Random Forest + SMOTE-ENN | $51,95 \pm 8,07$ | $0,434 \pm 0,099$ | $0,814 \pm 0,058$ | 7,20 | 3/30 | 26,7 |
| D3 | C1 – completo | Masculino | AdaBoost + SMOTE | $21,16 \pm 2,59$ | $0,016 \pm 0,023$ | $0,517 \pm 0,016$ | 8,47 | 2/30 | 36,7 |
| D3 | C2 – seleção | Masculino | AdaBoost + SMOTE | $22,81 \pm 2,25$ | $0,017 \pm 0,018$ | $0,515 \pm 0,014$ | 7,50 | 1/30 | 33,3 |
| D3 | C3 – 6 atributos | Masculino | AdaBoost + SMOTE-ENN | $32,41 \pm 1,39$ | $0,033 \pm 0,031$ | $0,519 \pm 0,021$ | 6,13 | 10/30 | 43,3 |
| D3 | C3 – 19 atributos | Masculino | AdaBoost + SMOTE | $21,25 \pm 2,84$ | $0,030 \pm 0,024$ | $0,522 \pm 0,016$ | 8,17 | 5/30 | 43,3 |
| D3 | C1 – completo | Feminino | Random Forest + SMOTE-ENN | $31,38 \pm 1,24$ | $0,011 \pm 0,030$ | $0,506 \pm 0,020$ | 10,13 | 6/30 | 23,3 |
| D3 | C2 – seleção | Feminino | Random Forest + SMOTE-ENN | $30,26 \pm 1,33$ | $0,016 \pm 0,025$ | $0,512 \pm 0,016$ | 8,17 | 6/30 | 30,0 |
| D3 | C3 – 6 atributos | Feminino | AdaBoost + SMOTE-ENN | $29,43 \pm 1,72$ | $-0,004 \pm 0,030$ | $0,496 \pm 0,018$ | 8,53 | 4/30 | 43,3 |
| D3 | C3 – 19 atributos | Feminino | Random Forest + SMOTE-ENN | $29,60 \pm 1,99$ | $0,012 \pm 0,036$ | $0,508 \pm 0,022$ | 10,53 | 6/30 | 33,3 |

F1-score, MCC e AUC-ROC são apresentados como média ± desvio-padrão entre as 30 sementes. Rank 1 indica o número de sementes em que a configuração ocupou a primeira posição dentro do respectivo contexto. Top 3 indica a frequência de permanência entre as três primeiras posições. Os valores do Índice Multivariado não são comparáveis entre contextos diferentes.

Fonte: Elaborado pelo autor (2026).

A Tabela 19 acrescenta os componentes médios das matrizes de confusão das configurações de referência, permitindo interpretar os valores das métricas em conjunto com o tamanho efetivo dos conjuntos de teste.

Tabela 19 – Média dos componentes das matrizes de confusão das configurações de referência nos subgrupos por sexo ao longo das 30 sementes.

| Dataset | Cenário / representação | Subgrupo | Modelo + condição | TP n (%) | TN n (%) | FP n (%) | FN n (%) | Precisão (%) | Recall (%) | $n_{teste}$ | $n_+$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | C1 – completo | Masculino | LDA + Desbalanceado | 13,6 (7,7%) | 138,9 (78,5%) | 8,1 (4,6%) | 16,4 (9,2%) | 63,8 | 45,4 | 177 | 30 |
| D1 | C2 – seleção | Masculino | LDA + Desbalanceado | 10,5 (5,9%) | 141,0 (79,7%) | 6,0 (3,4%) | 19,5 (11,0%) | 65,3 | 35,0 | 177 | 30 |
| D1 | C3 – 6 atributos | Masculino | AdaBoost + SMOTE | 17,2 (9,7%) | 105,0 (59,3%) | 42,0 (23,7%) | 12,8 (7,3%) | 29,1 | 57,2 | 177 | 30 |
| D1 | C3 – 19 atributos | Masculino | LDA + Desbalanceado | 11,6 (6,6%) | 140,5 (79,4%) | 6,5 (3,7%) | 18,4 (10,4%) | 65,1 | 38,8 | 177 | 30 |
| D1 | C1 – completo | Feminino | XGBoost + ADASYN | 6,9 (5,9%) | 96,6 (81,9%) | 4,4 (3,7%) | 10,1 (8,5%) | 61,4 | 40,8 | 118 | 17 |
| D1 | C2 – seleção | Feminino | Naive Bayes + Desbalanceado | 8,5 (7,2%) | 91,6 (77,6%) | 9,4 (8,0%) | 8,5 (7,2%) | 50,9 | 50,0 | 118 | 17 |
| D1 | C3 – 6 atributos | Feminino | AdaBoost + SMOTE-ENN | 10,5 (8,9%) | 70,4 (59,7%) | 30,6 (25,9%) | 6,5 (5,5%) | 25,6 | 61,6 | 118 | 17 |
| D1 | C3 – 19 atributos | Feminino | Random Forest + SMOTE-ENN | 10,5 (8,9%) | 88,0 (74,6%) | 13,0 (11,0%) | 6,5 (5,5%) | 45,3 | 62,0 | 118 | 17 |
| D3 | C1 – completo | Masculino | AdaBoost + SMOTE | 42,5 (4,3%) | 635,9 (64,1%) | 155,1 (15,6%) | 158,5 (16,0%) | 21,6 | 21,1 | 992 | 201 |
| D3 | C2 – seleção | Masculino | AdaBoost + SMOTE | 49,5 (5,0%) | 610,0 (61,5%) | 181,0 (18,2%) | 151,5 (15,3%) | 21,5 | 24,6 | 992 | 201 |
| D3 | C3 – 6 atributos | Masculino | AdaBoost + SMOTE-ENN | 138,7 (14,0%) | 275,2 (27,7%) | 515,8 (52,0%) | 62,3 (6,3%) | 21,2 | 69,0 | 992 | 201 |
| D3 | C3 – 19 atributos | Masculino | AdaBoost + SMOTE | 40,5 (4,1%) | 654,4 (66,0%) | 136,6 (13,8%) | 160,5 (16,2%) | 22,9 | 20,2 | 992 | 201 |
| D3 | C1 – completo | Feminino | Random Forest + SMOTE-ENN | 145,5 (14,4%) | 227,4 (22,5%) | 582,6 (57,7%) | 53,5 (5,3%) | 20,0 | 73,1 | 1009 | 199 |
| D3 | C2 – seleção | Feminino | Random Forest + SMOTE-ENN | 119,1 (11,8%) | 341,2 (33,8%) | 468,8 (46,5%) | 79,9 (7,9%) | 20,3 | 59,8 | 1009 | 199 |
| D3 | C3 – 6 atributos | Feminino | AdaBoost + SMOTE-ENN | 118,7 (11,8%) | 322,6 (32,0%) | 487,4 (48,3%) | 80,3 (8,0%) | 19,6 | 59,7 | 1009 | 199 |
| D3 | C3 – 19 atributos | Feminino | Random Forest + SMOTE-ENN | 111,1 (11,0%) | 369,6 (36,6%) | 440,4 (43,7%) | 87,9 (8,7%) | 20,2 | 55,8 | 1009 | 199 |

TP, TN, FP e FN correspondem às médias obtidas nas 30 sementes e, por isso, podem assumir valores não inteiros. Os percentuais entre parênteses correspondem à proporção média em relação ao total do respectivo conjunto de teste. Precisão e recall são as médias das métricas calculadas separadamente em cada semente.

Fonte: Elaborado pelo autor (2026).

Os resultados mostram que a configuração de referência variou entre subgrupos, cenários e representações, sem estabelecer uma direção estável de vantagem entre masculino

e feminino. No *Dataset* 1, o menor número de casos positivos no subgrupo feminino aumenta a sensibilidade das métricas às divisões da amostra. No *Dataset* 3, posições relativas de destaque coexistiram com valores absolutos baixos de MCC e AUC-ROC e, em algumas configurações, com grande quantidade média de falsos positivos. As tabelas devem, portanto, ser interpretadas como detalhamento do procedimento de seleção e dos perfis de erro, e não como evidência de superioridade populacional, causalidade ou avaliação formal de equidade algorítmica.