Classificação Preditiva de Níveis de Absenteísmo com Regressão Logística Multiclasse

Helena S. Balbino[1]

[1]Universidade Estadual de Londrina (UEL)
Centro de Ciências Exatas
Departamento de Computação
Londrina, PR – Brasil

helena.balbino@uel.br

Abstract. *This study investigates workplace absenteeism using a Multiclass Logistic Regression model applied to a public dataset from the UCI repository. The objective is to classify levels of absence based on organizational, personal, and work-related variables. The methodology included exploratory data analysis, variable transformation, and class balancing, with emphasis on the use of SMOTE. Two scenarios were compared: the original imbalanced data and synthetically balanced data. The model trained with SMOTE showed improved recall and F1-score metrics for minority classes, although the overall accuracy remained moderate (61%). The results demonstrate the feasibility of using interpretable models in organizational contexts, even in the presence of challenges such as class imbalance. This work reinforces the importance of well-structured pipelines for predictive tasks and highlights their potential to generate actionable insights for people management.*

Resumo. *Este trabalho investiga o absenteísmo no ambiente de trabalho por meio de um modelo de Regressão Logística Multiclasse, aplicado a dados públicos da UCI. O objetivo é classificar níveis de ausência com base em variáveis organizacionais, pessoais e laborais. Foram conduzidas etapas de análise exploratória, transformação de variáveis e balanceamento de classes, com destaque para o uso do SMOTE. Dois cenários foram comparados: dados originais e dados balanceados. O modelo com SMOTE apresentou melhora nas métricas de recall e F1-score para as classes minoritárias, ainda que a acurácia geral tenha se mantido moderada (61%). Os resultados demonstram a viabilidade de modelos interpretáveis em contextos organizacionais, mesmo diante de desafios como o desbalanceamento de classes. O estudo reforça a importância de pipelines bem estruturados para tarefas preditivas, com potencial para gerar insights aplicáveis à gestão de pessoas.*

1. Introdução

O absenteísmo emerge como um desafio crítico para organizações de diversos setores, com implicações diretas sobre a produtividade, o clima organizacional e os custos operacionais. Conceituado como a ausência de um dever ou obrigação, o absenteísmo funciona como um indicador de desempenho fundamental. As ausências de colaboradores, sejam elas pontuais ou prolongadas, impactam severamente o planejamento logístico e operacional das equipes, a distribuição de tarefas e, consequentemente, a qualidade dos serviços

entregues. Além disso, tais ausências podem ser sintomas de questões mais profundas, como condições de saúde física ou mental, insatisfação com o ambiente de trabalho ou dificuldades de deslocamento.

No âmbito da gestão de pessoas e da análise organizacional, embora se espere que os empregados cumpram integralmente suas jornadas laborais, diversas circunstâncias podem afetar a assiduidade. Nesse contexto, a compreensão dos padrões associados ao absenteísmo é imperativa para a formulação de políticas de prevenção e intervenção mais eficazes. Com o avanço das técnicas de análise de dados e aprendizado de máquina, tornou-se plenamente viável o desenvolvimento de modelos preditivos capazes de identificar colaboradores com maior predisposição a ausências, o que permite a implementação de ações proativas de suporte.

É neste cenário que a aplicação de métodos estatísticos e computacionais para analisar o absenteísmo se apresenta como uma estratégia altamente promissora, capaz de fundamentar decisões de gestão baseadas em evidências e, assim, contribuir diretamente para a eficiência e a sustentabilidade das operações organizacionais.

Considerando o exposto, o presente trabalho tem como objetivo central desenvolver um modelo preditivo utilizando Regressão Logística Multiclasse. Este modelo visará classificar o nível de absenteísmo de colaboradores, tomando como base um conjunto de variáveis que abrangem o ambiente organizacional, o perfil dos funcionários e as condições de trabalho. Através da aplicação de técnicas de análise exploratória, pré-processamento e balanceamento de dados, busca-se construir uma solução não apenas interpretável, mas também estrategicamente alinhada ao problema de negócio, proporcionando uma compreensão aprofundada dos fatores subjacentes as ausências.

Espera-se elucidar padrões significativos no comportamento de absenteísmo, mesmo diante de desafios inerentes como o desbalanceamento de classes e a variabilidade dos dados. Para além da acurácia preditiva, almeja-se gerar insights práticos e acionáveis para a gestão organizacional, incluindo a identificação das variáveis mais influentes no fenômeno de ausência. O estudo busca evidenciar os notáveis benefícios da aplicação estruturada e fundamentada de abordagens estatísticas, mesmo as mais tradicionais como a Regressão Logística.

2. Materiais e Métodos

O presente estudo fundamenta-se na análise do conjunto de dados "Absenteeism at Work", disponível no repositório da *University of California, Irvine (UCI)*[1]. Este conjunto de dados compreende 740 registros de colaboradores de uma empresa de entregas no Brasil, coletados entre julho de 2007 e julho de 2010. Para o carregamento dos dados de forma rápida e padronizada, utilizou-se a biblioteca *ucimlrepo* [Dua and Graff 2019], que fornece uma interface programática ao repositório da UCI, facilitando sua integração ao ambiente de programação através do ID do repositório.

Cada registro detalha um evento de presença ou ausência laboral, consolidando informações sobre o perfil do empregado, suas condições físicas, aspectos familiares e

[1]O UCI Machine Learning Repository é uma coleção de conjuntos de dados públicos amplamente utilizados em pesquisas de aprendizado de máquina. Disponível em: `https://archive.ics.uci.edu/ml/datasets`.

sociais, elementos organizacionais e, quando pertinente, os motivos declarados para as ausências.

Todas as análises e procedimentos foram conduzidos em ambiente interativo baseado em notebooks, utilizando a linguagem de programação *Python*. Para a implementação das diversas etapas do fluxo de processamento, que abrangem desde o carregamento e manipulação de dados até a visualização, pré-processamento, modelagem preditiva e avaliação de desempenho, foram empregadas bibliotecas amplamente reconhecidas e consolidadas. Os dados foram manipulados utilizando as bibliotecas *numpy* [Harris et al. 2020] e *pandas* [McKinney 2010], enquanto os gráficos foram gerados com *matplotlib* [Hunter 2007] e *seaborn* [Waskom 2021].

Originalmente, o conjunto de dados apresenta 21 variáveis, sendo 20 preditoras e uma variável-alvo. Para a modelagem preditiva e análise subsequente, optou-se por utilizar 19 variáveis preditoras, desconsiderando a variável de identificação do colaborador (ID) por não agregar valor estatístico ou informacional para a construção do modelo. As variáveis restantes podem ser agrupadas em:

(a.) **Características Sociodemográficas:** abrangem idade, escolaridade, número de filhos e de animais de estimação.
(b.) **Condições Físicas:** incluem peso, altura e índice de massa corporal.
(c.) **Hábitos Sociais e Comportamentais:** referem-se ao consumo de álcool e tabaco, e ao histórico disciplinar.
(d.) **Aspectos Organizacionais:** envolvem a carga de trabalho, tempo de serviço, distância de casa até o trabalho, cumprimento de metas e gastos com transporte.
(e.) **Informações Contextuais da Ausência:** detalham o motivo atestado, a estação do ano, o dia da semana e o mês de ocorrência.

Após o carregamento do conjunto de dados, realizou-se uma análise descritiva das variáveis para compreender suas escalas e distribuições, conforme a Tabela 1. Nota-se a integridade do conjunto de dados, confirmando a ausência de valores faltantes. Variáveis como custo de transporte (*transportation_expense*), idade (*age*), distância de casa ao trabalho (*distance_from_residence_to_work*) e a carga de trabalho média por dia (*work_load_average_day*) possuem escalas significativamente distintas. Tal disparidade justificou a aplicação de técnicas de normalização para equalizar a influência dessas variáveis no modelo.

A análise das informações estatísticas, contida na Tabela 1, revelou a presença de *outliers* em variáveis como custo de transporte (*transportation_expense*), idade (*age*) e carga de trabalho média por dia (*work_load_average_day*), contudo, optou-se por mantê-los. Essa decisão foi justificada pelo tamanho reduzido da amostra (740 registros) e pela forma com que a a regressão logística lida com valores extremos. Além disso, a normalização posterior das variáveis contribuiu para reduzir os impactos negativos desses valores.

Para permitir a aplicação de métodos de classificação multiclasse, a variável-alvo original, *absenteeism_time_in_hours*, que compreende a contagem de ausência em horas, com valores contínuos de 0 a 120, foi convertida em três categorias de absenteísmo: *No Absence* (zero horas), *Moderate Absence* (até quatro horas) e *High Absence* (mais de quatro horas). [2] Essa transformação baseou-se em regras de negócio específicas, uma vez

[2]Os nomes das categorias foram mantidos em inglês para consistência com a nomenclatura presente no

Tabela 1. Estatísticas descritivas das variáveis do conjunto de dados

Variable Name	Data Type	Non-Null Count	Null Count	Unique Values	Mean	Median	Mode	Std	Min	Max
reason_for_absence	int64	740	0	28	19.22	23.00	23.00	8.43	0.00	28.00
month_of_absence	int64	740	0	13	6.32	6.00	3.00	3.44	0.00	12.00
day_of_the_week	int64	740	0	5	3.91	4.00	2.00	1.42	2.00	6.00
seasons	int64	740	0	4	2.54	3.00	4.00	1.11	1.00	4.00
transportation_expense	int64	740	0	24	221.33	225.00	179.00	66.95	118.00	388.00
distance_from_residence_to_work	int64	740	0	25	29.63	26.00	26.00	14.84	5.00	52.00
service_time	int64	740	0	18	12.55	13.00	18.00	4.38	1.00	29.00
age	int64	740	0	22	36.45	37.00	28.00	6.48	27.00	58.00
work_load_average_day	float64	740	0	38	271.49	264.25	222.20	39.06	205.92	378.88
hit_target	int64	740	0	13	94.59	95.00	93.00	3.78	81.00	100.00
disciplinary_failure	int64	740	0	2	0.05	0.00	0.00	0.23	0.00	1.00
education	int64	740	0	4	1.29	1.00	1.00	0.67	1.00	4.00
son	int64	740	0	5	1.02	1.00	0.00	1.10	0.00	4.00
social_drinker	int64	740	0	2	0.57	1.00	1.00	0.50	0.00	1.00
social_smoker	int64	740	0	2	0.07	0.00	0.00	0.26	0.00	1.00
pet	int64	740	0	6	0.75	0.00	0.00	1.32	0.00	8.00
weight	int64	740	0	26	79.04	83.00	89.00	12.88	56.00	108.00
height	int64	740	0	14	172.11	170.00	170.00	6.03	163.00	196.00
body_mass_index	int64	740	0	17	26.68	25.00	31.00	4.29	19.00	38.00
absenteeism_time_in_hours	int64	740	0	19	6.92	3.00	8.00	13.33	0.00	120.00

Nota: Os nomes das variáveis foram mantidos em inglês para preservar a consistência com o conjunto de dados original.

que as regras estatísticas de Sturges e Freedman-Diaconis mostraram-se insuficientes ou excessivamente abrangentes para este contexto.

A Figura 1 apresenta a distribuição dessa variável após a discretização. Observa-se que a maior concentração está na categoria intermediária (*Moderate Absence*). Essa distribuição desproporcional motivou a adoção de técnicas de balanceamento de classes, utilizando a técnica SMOTE, implementada por meio da biblioteca *imbalanced-learn* [Lemaître et al. 2017], permitindo a comparação entre o desempenho do modelo de regressão treinado com dados originais desbalanceados e com dados balanceados sinteticamente. A transformação da variável-alvo de um valor contínuo para categorias foi

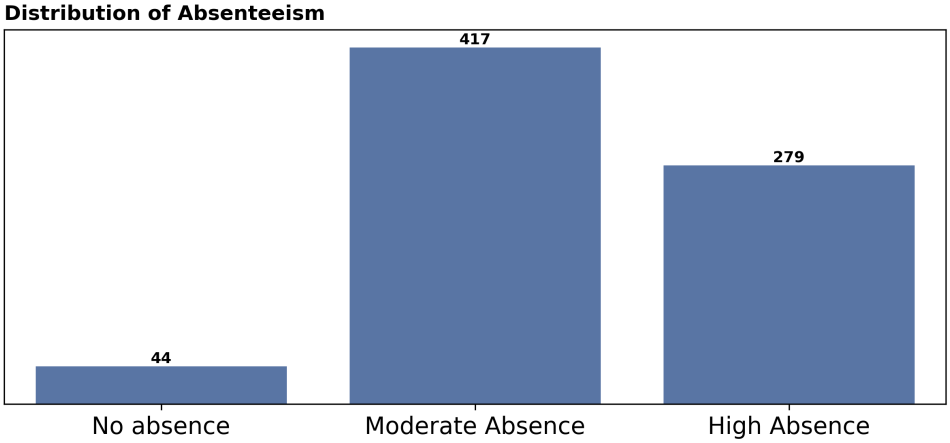


Figura 1. Distribuição das categorias de absenteísmo após a discretização da variável-alvo.

uma decisão metodológica e estratégica. Em vez de estimar o número exato de horas de ausência, o modelo classifica níveis de absenteísmo, o que se mostra mais aderente às necessidades práticas da gestão de pessoas. As categorias propóstas são mais com-

conjunto de dados original.

preensíveis e operacionalizáveis para os gestores, permitindo a formulação de diferentes estratégias de intervenção para cada nível de ausência. Abordagens dessa natureza favorece decisões mais ágeis e estratégicas, como a priorização de atendimentos médicos, ações de prevenção ou ajustes no planejamento de equipe, algo que previsões baseadas apenas em horas exatas não viabilizariam com a mesma clareza.

O conjunto de dados foi particionado em conjuntos de treino e teste, sendo destinado 20% das amostras para os testes do modelo. Para um treinamento eficiente a estratificação foi aplicada com base na variável-alvo, assegurando a preservação proporcional das classes em ambos os subconjuntos (treino e teste). Foram definidos número máximo de iterações e a tolerância para o critério de parada.

Para evitar o vazamento de informações, após a separação dos dados, foram aplicadas etapas de normalização. Assim, utilizou-se o método OneHotEncoder, da biblioteca *scikit-learn* [Pedregosa et al. 2011], para códificar a variável que continha os motivos de ausência, que antes foi reagrupada de 28 subcategorias para 6 macrocategorias de acordo com o Código Internacional de Doenças (CID). Foram aplicadas técnicas de normalização às variáveis numéricas por meio do StandardScaler, também da biblioteca *scikit-learn* [Pedregosa et al. 2011], transformando-as para apresentarem média zero e desvio padrão um, o que é particularmente relevante para o bom desempenho da regressão logística. Após essas etapas de codificação e normalização, os dados foram reestruturados por meio da concatenação das variáveis transformadas, garantindo a coerência estrutural e a compatibilidade dos conjuntos de treino e teste para a fase de modelagem preditiva.

3. Resultados

3.1. Análise Exploratória dos Dados

A análise exploratória do conjunto de dados forneceu informações valiosas sobre o perfil do absenteísmo, além de orientar as etapas de pré-processamento. Aproximadamente 20% das ausências foram atribuídas a consultas médicas, enquanto diagnósticos como neoplasias e malformações congênitas foram raros. Já a distribuição das faltas por dias da semana mostrou-se relativamente equilibrada, com uma leve concentração nas segundas.

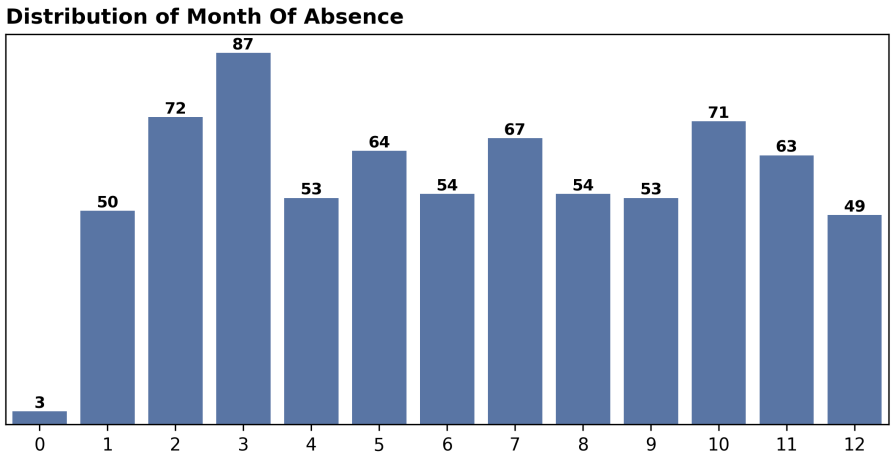


Figura 2. Distribuição mensal das ausências.

Como mostra a Figura 2 a concentração das ausências foi mais notável nos meses de fevereiro e março, período que pode ser influenciado por eventos sazonais, como o

Carnaval. Em relação aos atributos comportamentais, observou-se que a maioria dos colaboradores não possuía histórico disciplinar, apresentava ensino médio como grau de escolaridade predominante e declarou não fumar.

A análise exploratória das variáveis numéricas revelou padrões distributivos e a presença de alguns *outliers*. Variáveis como despesa com transporte (*transportation_expense*), idade (*age*) e tempo de serviço (*service_time*) exibiram distribuições assimétricas, com notável concentração de dados em faixas específicas e a ocorrência de alguns valores discrepantes. A distância de casa até o trabalho (*distance_from_residence_to_work*) demonstrou uma distribuição mais ampla, indicando heterogeneidade nos perfis dos colaboradores. Em relação ao alcance de metas (*hit_target*), observou-se um padrão de alto desempenho, com a maioria dos registros entre 90% e 100%, contudo, com a identificação de *outliers* que sugerem casos de menor eficácia. Variáveis de contagem discreta, como número de filhos (*son*) e animais de estimação (*pet*), apresentaram predominância de valores nulos, com ocorrências esporádicas de frequências elevadas. A altura (*height*) e peso (*weight*) demonstraram distribuições com concentrações definidas, mas também com a presença de valores atípicos nas extremidades.

Dentre as variáveis numéricas analisadas, destaca-se a variável *work_load_average_day*, que representa a carga de trabalho média por dia. Conforme ilustrado na Figura 3, observa-se uma concentração de valores entre 240 e 300 minutos, com maior densidade entre 250 e 270 minutos. Esses valores correspondem, aproximadamente, a jornadas diárias de 4 a 5 horas, o que está em conformidade com padrões observados em determinadas categorias profissionais no Brasil. No entanto, a mesma figura revela a existência de valores discrepantes acima de 360 minutos, sugerindo possíveis casos isolados de sobrecarga de trabalho no conjunto analisado.

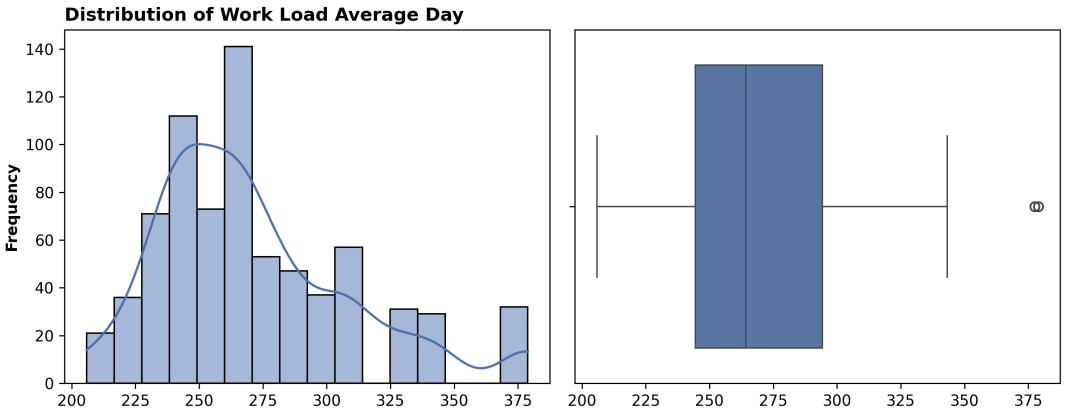


Figura 3. Distribuição da carga de trabalho média por dia.

A análise de correlação revelou associações relevantes entre variáveis numéricas, as quais foram reforçadas visualmente pelos gráficos de dispersão apresentados na Figura 4. Observa-se, por exemplo, uma forte correlação positiva entre idade (*age*) e tempo de serviço (*service_time*), evidenciada pela inclinação ascendente do gráfico. Esse comportamento é consistente com a expectativa de que colaboradores mais velhos tendem a possuir maior tempo de vínculo empregatício. De forma semelhante, nota-se uma correlação substancial entre peso (*weight*) e índice de massa corporal (*body_mass_index*),

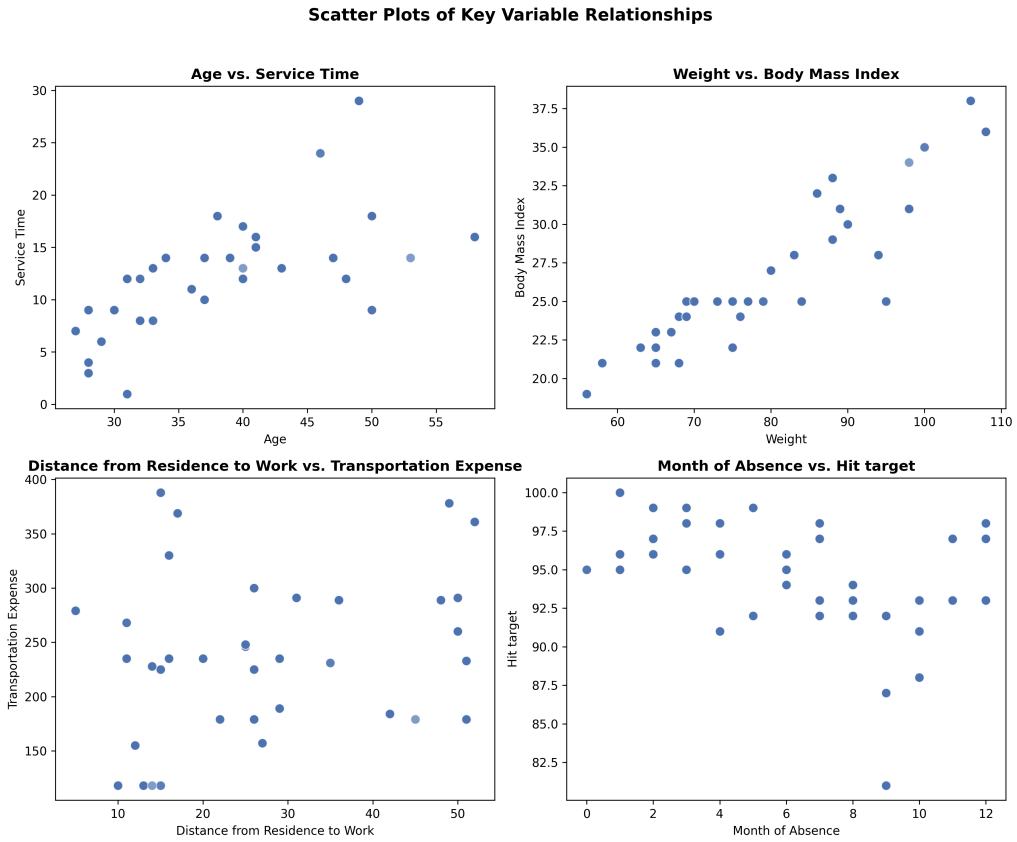


Figura 4. Correlação entre variáveis-chave para o conjunto de dados.

evidenciando uma relação direta e praticamente linear entre essas variáveis, o que também indica redundância no conjunto de variaveis. Desse modo, a variável de índice de massa corporal foi removida, mantendo-se *weight* e *height* para evitar superposição de informações.

Já a dispersão entre distância da residência ao trabalho (*distance_from_residence_to_work*) e o gasto com transporte (*transportation_expense*) revela uma relação mais difusa, com fraca associação linear, possivelmente influenciada por variáveis externas como o tipo de transporte ou subsídios aos trabalhadores. Por fim, a relação entre o indicador de desempenho (*hit_target*) e o mês de ausência (*month_of_absence*) sugere uma leve tendência de queda no desempenho em certos períodos do ano, surpreendentemente as quedas não estão nos meses com mais ausência como fevereiro e março, mas sim em meses como setembro e outubro. Essas representações gráficas oferece uma visão intuitiva das associações e auxilia na identificação de padrões e redundâncias que podem influenciar a modelagem preditiva.

3.2. Pré-processamento de Dados

O processo de pré-processamento dos dados foi iniciado com a remoção das colunas consideradas redundantes ou de menor relevância para a análise preditiva, especificamente o indice de massa corporal (*body_mass_index*) e variaveis secundárias criadas no momento de discretização. Essa etapa visou otimizar a dimensionalidade do conjunto de dados e mitigar potenciais problemas de multicolinearidade. Um tratamento significativo foi aplicado à variável com os motivos para as ausências (*reason_for_absence*), que consistiu em

sua recategorização.

As 29 categorias originais foram agrupadas em seis subgrupos mais abrangentes, que é ilustrado pela Figura 5. Esse agrupamento, realizado com o auxílio de um dicionário de dados predefinido, teve como objetivo reduzir a granularidade dos dados e simplificar a interpretação de modelos preditivos. Para preparar a variável para algoritmos de aprendizado de máquina, as categorias textuais da nova variável foram mapeadas para representações numéricas inteiras, utilizando o mesmo dicionário de dados, preservando a informação qualitativa da variável de forma quantificável.

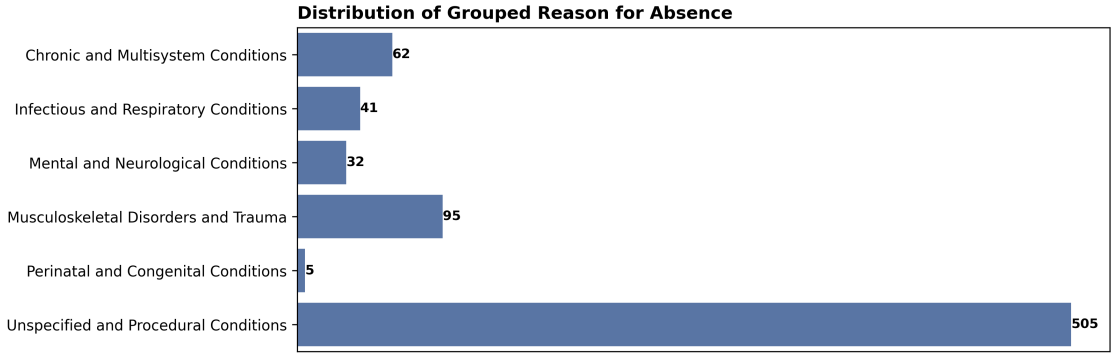


Figura 5. Distribuição dos motivos de ausência após a recategorização.

Visando o desenvolvimento e avaliação dos modelos preditivos, o conjunto de dados foi dividido em subconjuntos de treinamento e teste, utilizando a função *train_test_split* do *scikit-learn* [Pedregosa et al. 2011]. A variável com a classificação da ausência (*absenteeism_code*), que representa a classe alvo do modelo, foi definida como Y, enquanto as demais variáveis preditoras constituíram a matriz X. A divisão foi realizada com uma proporção de 80% dos dados para treinamento e 20% para teste. Para garantir a reprodutibilidade dos resultados, uma semente de aleatoriedade foi fixada em 42. É importante ressaltar que a separação foi estratificada pela variável alvo (stratify=y), assegurando que a distribuição das classes de ausência fosse mantida proporcionalmente em ambos os conjuntos.

As variáveis do dataset foram submetidas a processos de pré-processamento distintos, conforme suas tipologias. Para a variável categórica nominal *reason_for_absence_class*, aplicou-se o One-Hot Encoding (OHE) . Este método transforma cada categoria em uma nova coluna binária (0 ou 1) e, ao utilizar o parâmetro drop='first', evita a introdução de uma ordem artificial entre as classes, prevenindo a multicolinearidade [Pedregosa et al. 2011]. As colunas resultantes do OHE foram, então, incorporadas aos respectivos conjuntos de treino e teste.

Paralelamente, as variáveis numéricas foram submetidas ao escalonamento padrão através do StandardScaler. Este processo centraliza os dados, subtraindo a média e dividindo pelo desvio padrão, o que resulta em uma distribuição com média zero e desvio padrão um [Pedregosa et al. 2011]. É crucial destacar que o StandardScaler foi ajustado exclusivamente no conjunto de treinamento (fit_transform) e, em seguida, aplicado ao conjunto de teste (transform), a fim de evitar vazamento de dados.

Finalmente, as variáveis categóricas codificadas e as numéricas escalonadas foram concatenadas para reconstruir os conjuntos de dados de treinamento e teste processados.

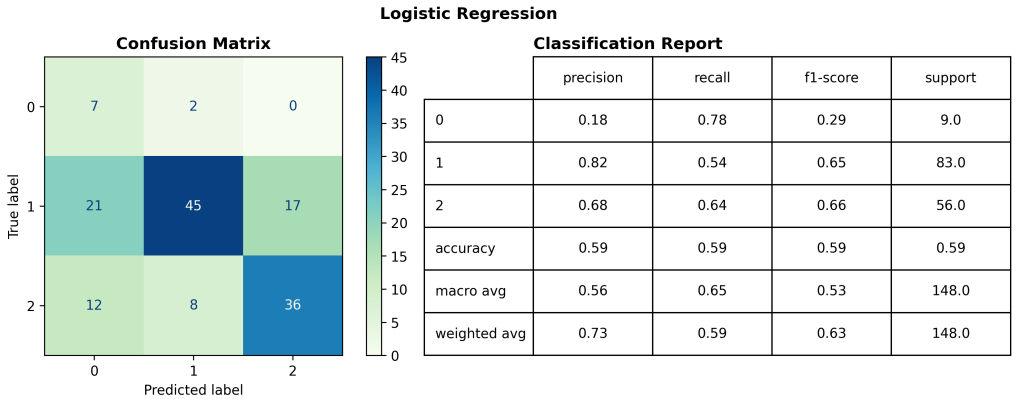
Este meticuloso processo de divisão e pré-processamento foi fundamental para garantir a validade e a aplicabilidade dos modelos de regresão logítica multiclasse a serem desenvolvidos, minimizando viés tendencioso e otimizando o desempenho preditivo.

3.3. Modelagem Preditiva

3.3.1. Regressão Logística e Análise de Desempenho

Para a etapa de modelagem preditiva, empregou-se o algoritmo de Regressão Logística. A fim de evitar viés do modelo foi utilizado o parâmetro class_weight='balanced', o que reduz o impacto de um possível desbalanceamento das classes na variável alvo [Pedregosa et al. 2011]. Além disso foi previamente definido um número máximo de 1000 interações com uma tolerância de erro em 0,000001 (1e-6) para garantir a convergência do algoritmo. Novamente, a semente de reprodutibilidade foi fixada em 42. E assim, o modelo foi treinado utilizando o conjunto de dados processado (X_train_processed, y_train) e, subsequentemente, suas predições (y_pred_reg) foram geradas sobre o conjunto de teste (X_test_processed).

A matriz de confusão resultante da aplicação do modelo de Regressão Logística sobre os dados de teste é apresentada na Figura 6. Esta visualização permite uma análise detalhada dos acertos e erros de classificação, discriminando os verdadeiros positivos, falsos positivos e falsos negativos em cada classe. Observa-se que o modelo teve maior desempenho na classe 1 (*Moderate Absence*), com 45 acertos (precisão $\approx 82\%$ e F1-score $\approx 65\%$). No entanto, houve confusões relevantes, principalmente com a classe 0 (*No Absence*), cuja precisão foi baixa ($\approx 18,9\%$), mostrando que muitos exemplos das classes 1 (*Moderate Absence*) e 2 (*High Absence*) foram incorretamente classificados como classe 0 (*No Absence*). A classe 2 (*High Absence*) apresentou desempenho intermediário, com pontuação F1 de aproximadamente 66%. A acurácia geral do modelo foi de 59,5% , o que indica desempenho razoável, mas com espaço para melhorias, especialmente na distinção entre as classes 0 e 1.



Classification Report

	precision	recall	f1-score	support
0	0.18	0.78	0.29	9.0
1	0.82	0.54	0.65	83.0
2	0.68	0.64	0.66	56.0
accuracy	0.59	0.59	0.59	0.59
macro avg	0.56	0.65	0.53	148.0
weighted avg	0.73	0.59	0.63	148.0

Figura 6. Matriz de Confusão e Métricas de Classificação para o modelagem com a regressão logistica.

Apesar da utilização do parâmetro class_weight='balanced' para lidar com o desbalanceamento entre as classes, as métricas de desempenho obtidas indicam que o modelo ainda apresentou dificuldades na classificação das classes minoritárias, especialmente a

classe 0, que representa as instâncias sem ausências e corresponde a aproximadamente 6% do conjunto de dados.

Com o objetivo de aprimorar a capacidade preditiva e, em particular, a identificação das classes menos representadas, optou-se por realizar um novo treinamento do modelo utilizando a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), que permite balancear sinteticamente o conjunto de dados de treinamento por meio da geração de instâncias artificiais das classes minoritárias [Lemaître et al. 2017].

3.3.2. Otimização do Modelo com SMOTE

Com o objetivo de reduzir os efeitos do desbalanceamento das classes, identificado como um fator limitante na etapa de modelagem anterior, aplicou-se a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) exclusivamente sobre o conjunto de treinamento. Essa abordagem foi adotada para preservar a integridade dos dados de teste e evitar qualquer vazamento de informação. A técnica SMOTE atua gerando amostras sintéticas para as classes minoritárias com base em vizinhos próximos no espaço de atributos, o que promove um balanceamento mais homogêneo e auxilia o modelo a aprender padrões mais robustos para essas classes [Lemaître et al. 2017]. Após o balanceamento, todas as classes passaram a ter o mesmo número de observações (334) no conjunto de treino, permitindo uma avaliação mais justa e representativa do desempenho do modelo.

Nesta segunda fase de modelagem, a Regressão Logística foi novamente empregada, agora com o conjunto de dados de treino balanceado via SMOTE. O parâmetro class_weight='balanced' foi mantido como estratégia complementar. Embora o uso combinado de SMOTE e class_weight='balanced' possa ser considerado redundante em alguns cenários, sua manutenção visou reforçar a equidade na aprendizagem do modelo para todas as classes. Os parâmetros adicionais max_iter=1000 e tol=1e-6 também foram preservados para assegurar a convergência do algoritmo.

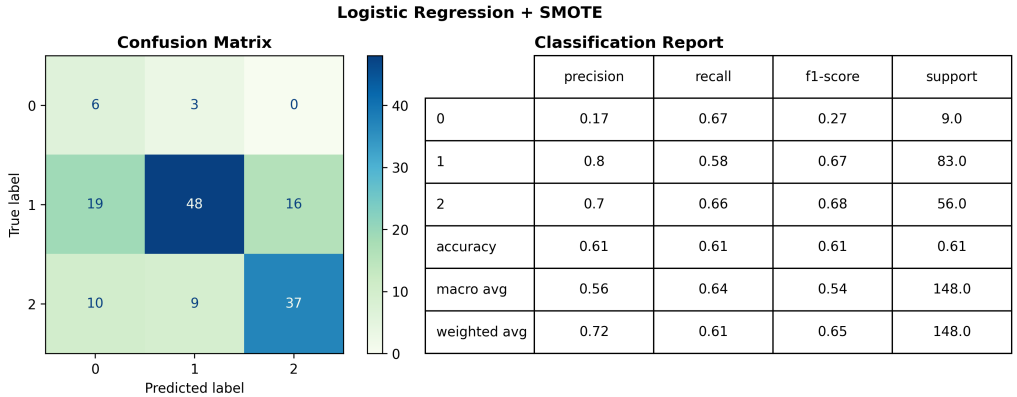


Figura 7. Matriz de Confusão e Métricas de Classificação para o modelagem com a regressão logistica após o balanceamento por SMOTE.

Os resultados da modelagem com o conjunto de treinamento balanceado por SMOTE são apresentados na Figura 7, que exibe a matriz de confusão e o relatório de classificação. A análise da matriz de confusão revelou uma melhora no desempenho geral da classificação multiclasse em comparação com o modelo anterior, especialmente na

identificação das classes minoritárias.

A acurácia global do modelo foi de aproximadamente 61%. O relatório de classificação detalhou os seguintes valores de *F1-score*: $0{,}27$ para a classe 0 (*No Absence*), $0{,}67$ para a classe 1 (*Moderate Absence*) e $0{,}68$ para a classe 2 (*High Absence*). Em relação ao *recall*, a classe 0 (*No Absence*) obteve $0{,}67$, indicando que a maior parte das instâncias dessa classe foi corretamente identificada, apesar de sua precisão ter sido baixa ($0{,}17$). A classe 1 (*Moderate Absence*), majoritária no conjunto de dados, obteve um *recall* de $0{,}58$. Em termos gerais, o modelo demonstrou desempenho consistente nas classes mais representadas, enquanto a classe minoritária, embora com *recall* aceitável, continuou a apresentar limitações em termos de precisão.

4. Conclusão

Este estudo explorou a modelagem preditiva dos padrões de ausência no ambiente de trabalho, avaliando o desempenho da Regressão Logística em dois cenários distintos: um modelo base, treinado com os dados originais, e um modelo alternativo, treinado após a aplicação da técnica SMOTE (*Synthetic Minority Over-sampling Technique*) para balanceamento das classes.

O modelo inicial, mesmo com o uso do parâmetro class_weight='balanced' para mitigar os efeitos do desbalanceamento, apresentou limitações notáveis na sua capacidade de generalização. Conforme evidenciado pela matriz de confusão e pelas métricas de desempenho, a classe 0 (*No Absence*) obteve um bom *recall*, mas com baixa precisão. As classes minoritárias, como um todo, exibiram desempenho inferior em comparação à classe majoritária. Essa dificuldade na correta identificação das instâncias menos frequentes impacta diretamente a eficácia do modelo em cenários práticos, especialmente quando o foco reside na detecção de padrões atípicos ou menos recorrentes.

A aplicação do SMOTE no conjunto de treinamento resultou em melhorias relevantes no desempenho do modelo. O balanceamento sintético das classes permitiu à Regressão Logística capturar padrões mais representativos, refletindo-se em maior estabilidade das métricas, especialmente para as classes 1(*Modera Absence*) e 2 (*High Absence*). Embora a acurácia geral tenha se mantido em torno de 61%, observou-se uma leve redução dos erros e aumento dos acertos em todas as classes. A Tabela 2 condensa o comparativo de acertos e erros por classe antes e após o balanceamento dos dados. Destaca-se, por exemplo, que a classe 2 (*High Absence*) apresentou 36 acertos no modelo original e 37 no modelo balanceado, enquanto os erros caíram de 20 para 19. Já na classe 1 (*Moderate Absence*), a mais frequente, os acertos passaram de 45 para 48, também com redução dos erros.

Esses resultados reforçam que a combinação de técnicas de pré-processamento com estratégias de balanceamento são fundamentais para a construção de modelos preditivos mais confiáveis, consistentes e equitativos, capazes de oferecer maior poder de generalização. A comparação entre os dois modelos evidencia que considerar a distribuição das classes na fase de treinamento é uma etapa crucial para alcançar uma classificação mais precisa e confiável de todos os níveis de ausência.

De forma geral, os resultados obtidos corroboram os objetivos propostos neste estudo. A construção e avaliação de um modelo preditivo baseado em Regressão Logística

Tabela 2. Comparativo de acertos e erros por classe antes e após o balanceamento dos dados

Classe	Acertos (Original)	Erros (Original)	Acertos (Balanceado)	Erros (Balanceado)
0	7	2	6	3
1	45	38	48	35
2	36	20	37	19

Multiclasse permitiu não apenas classificar os níveis de absenteísmo com desempenho satisfatório, mas também explorar de maneira interpretável os fatores associados ao fenômeno. A incorporação de variáveis relacionadas ao perfil dos colaboradores, ao ambiente organizacional e às condições de trabalho proporcionou uma visão abrangente das ausências. Além disso, a aplicação de técnicas de análise exploratória, pré-processamento e balanceamento de dados demonstrou ser essencial para lidar com os desafios de variabilidade e desbalanceamento das classes. Para além da acurácia, o modelo desenvolvido se mostrou útil como ferramenta estratégica de apoio à gestão, oferecendo subsídios para a identificação de padrões e possíveis intervenções em contextos organizacionais.

Ainda que o modelo tenha alcançado uma acurácia global de aproximadamente 61%, seu valor reside na capacidade de oferecer uma base interpretável e estatisticamente fundamentada para a análise do absenteísmo no ambiente organizacional. A Regressão Logística, por sua natureza, proporciona maior transparência na relação entre as variáveis preditoras e as classes de ausência, permitindo insights relevantes para gestores e analistas. No entanto, os resultados também evidenciam que há espaço para aprimoramentos substantivos. Futuras abordagens podem explorar algoritmos mais complexos, como árvores de decisão, *ensembles* ou redes neurais, além de técnicas de seleção e engenharia de atributos, ajustes finos de hiperparâmetros e até mesmo ampliação do conjunto de dados. Assim, embora o modelo atual ainda não ofereça precisão suficiente para aplicações automatizadas de alto impacto, ele cumpre um papel inicial importante como diagnóstico exploratório e apoio à tomada de decisão baseada em dados.

Referências

Dua, D. and Graff, C. (2019). Uci machine learning repository. Available at: https://archive.ics.uci.edu/ml.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with numpy. *Nature*, 585:357–362.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, pages 51–56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.