https://doi.org/10.47456/bjpe.v11i4.50222 ARTIGO ORIGINAL | OPEN ACCESS

NÚCLEO DE PESQUISA EM GESTÃO DE SISTEMAS DE PRODUÇÃO - NP GSP
Brazilian Journal of Production Engineering
BJPE - Revista Brasileira de Engenharia de Produção
Disponível em: https://periodicos.ufes.br/bjpe
ISSN: 2447-5580
Campus São Mateus
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Pró-Reitoria de Pesquisa e Pós-Graduação
Universidade Federal do Espírito Santo

Turnover em obras de montagem eletromecânica: aplicação de Machine Learning na gestão de pessoas

Turnover in electromechanical assembly projects: applying Machine Learning to people management

Rotación de personal en proyectos de montaje electromecánico: aplicación de Machine Learning en la gestión de personas

Gabriel Filipe Rebuiti Passos [1], & **Jailson dos Santos Silva** [2*]

[1] Universidade de São Paulo [2] Universidade Federal de Santa Catarina
[1] REBUITEG@gmail.com [2*] engjailsonsantos@outlook.com

ARTIGO INFO.

Recebido: 30.09.2025
Aprovado: 12.11.2025
Disponibilizado: 28.11.2025

PALAVRAS-CHAVE: rotatividade; RH; montagem eletromecânica.
KEYWORDS: turnover; HR; electromechanical assembly.
PALABRAS CLAVE: rotación de personal; RH; montaje electromecánico.

***Autor Correspondente**: **Silva, J. dos S.**

RESUMO

Este estudo investigou os fatores que influenciam a substituição de colaboradores em obras de montagem eletromecânica, com foco na aplicação de técnicas estatísticas e de aprendizado de máquina para apoiar a gestão de pessoas. A pesquisa utilizou uma base de 7.333 registros de desligamentos ocorridos entre janeiro de 2023 e julho de 2025 em uma empresa brasileira do setor, contemplando variáveis individuais, organizacionais e relacionadas à jornada de trabalho. Após o pré-processamento e a análise exploratória dos dados, foram aplicados três modelos de classificação: regressão logística binária (com seleção *stepwise*), árvore de decisão e *Random Forest* (com calibração por *grid search*). Os resultados mostraram que fatores como percentual de abonos, tempo de casa, tipo de obra e estado de residência apresentaram maior relevância para a reposição de vagas. O modelo de *Random Forest* apresentou o melhor desempenho preditivo, alcançando AUC-ROC de 0,878 e coeficiente de Gini de 0,756, enquanto a regressão logística destacou-se pela interpretabilidade dos coeficientes. Conclui-se que a integração de métodos estatísticos e de *machine learning* contribui para antecipar cenários de rotatividade, subsidiando estratégias de retenção e alocação mais eficientes na gestão de recursos humanos em projetos industriais.

ABSTRACT

This study investigated the factors influencing employee turnover in electromechanical assembly projects, focusing on the application of statistical techniques and machine learning to support people management. The research used a dataset of 7,333 terminations that occurred between January 2023 and July 2025 in a Brazilian company from the sector, covering individual, organizational, and work-related variables. After preprocessing and exploratory data analysis, three classification models were applied: binary logistic regression (with stepwise selection), decision tree, and Random Forest (with grid search calibration). The results showed that factors such as percentage of leave days, tenure, project type, and state of residence had greater relevance for position replacement. The Random Forest model achieved the best predictive performance, with an AUC-ROC of 0.878 and a Gini coefficient of 0.756, while logistic regression stood out for the interpretability of its coefficients. The findings indicate that integrating statistical methods and machine learning contributes to anticipating turnover scenarios, supporting more effective retention and allocation strategies in human resources management for industrial projects.

RESUMEN

Este estudio investigó los factores que influyen en la rotación de personal en proyectos de montaje electromecánico, con énfasis en la aplicación de técnicas estadísticas y de aprendizaje automático para apoyar la gestión de personas. La investigación utilizó una base de 7.333 registros de desvinculaciones ocurridas entre enero de 2023 y julio de 2025 en una empresa brasileña del sector, abarcando variables individuales, organizacionales y relacionadas con la jornada laboral. Tras el preprocesamiento y el análisis exploratorio de los datos, se aplicaron tres modelos de clasificación: regresión logística binaria (con selección stepwise), árbol de decisión y Random Forest (con calibración mediante grid search). Los resultados mostraron que factores como el porcentaje de ausencias justificadas, el tiempo en la empresa, el tipo de obra y el estado de residencia tuvieron mayor relevancia para la reposición de vacantes. El modelo Random Forest presentó el mejor desempeño predictivo, alcanzando un AUC-ROC de 0,878 y un coeficiente de Gini de 0,756, mientras que la regresión logística se destacó por la interpretabilidad de sus coeficientes. Se concluye que la integración de métodos estadísticos y de aprendizaje automático contribuye a anticipar escenarios de rotación, apoyando estrategias más eficientes de retención y asignación en la gestión de recursos humanos en proyectos industriales.

INTRODUÇÃO

A rotatividade de colaboradores, ou *turnover*, é um indicador bastante estudado na área da gestão de pessoas, pois afeta diretamente a produtividade e os custos organizacionais. Conforme destaca Chiavenato (2014), esse fenômeno afeta não apenas os resultados financeiros, mas também o clima e o ambiente de trabalho.

No setor de montagem eletromecânica, em especial, os desafios são maiores, tendo em vista que a substituição de profissionais qualificados nem sempre ocorre de forma ágil, gerando atrasos e riscos adicionais para os projetos. Dutra (2016) ressalta que a complexidade desse mercado decorre de características como a natureza temporária dos projetos, investimentos iniciais elevados, custos enxutos e condições de trabalho adversas, fatores que contribuem para uma maior rotatividade. A instabilidade do setor e a dinâmica própria das obras também afetam diretamente a permanência dos profissionais, exigindo estratégias eficazes de gestão de pessoas.

Nos últimos anos, a aplicação de *People Analytics* e de técnicas de *Machine Learning* tem se consolidado como uma prática estratégica para compreender padrões e prever tendências na gestão de recursos humanos. Essas metodologias permitem analisar dados relacionados aos perfis dos colaboradores e antecipar cenários de desligamento, possibilitando assim ações preventivas e alinhadas à uma estratégia organizacional (Dutra, 2016). No contexto da montagem eletromecânica, essa abordagem pode revelar de que forma variáveis individuais (como idade, tempo de empresa, escolaridade e estado civil) e organizacionais (como características da obra, regime de trabalho e jornada de trabalho) influenciam a decisão de permanência ou substituição dos profissionais.

Pesquisas recentes reforçam a relevância do tema. Aver et al. (2020), ao estudarem uma empresa do setor metalúrgico em Caxias do Sul, identificaram que o ambiente de trabalho é determinante para a ocorrência da substituição. Samuel (2024) destaca que ações voltadas para melhores condições operacionais e planejamento estratégico da força de trabalho são fundamentais para reduzir o *turnover* em projetos de construção. No entanto, foi observado uma escassez de estudos que utilizam técnicas avançadas de análise preditiva nesse contexto específico, o que justifica a realização desta pesquisa.

Diante do exposto, o presente estudo busca preencher essa lacuna, investigando quais são os fatores que influenciam na substituição de colaboradores em obras de montagem eletromecânica por meio de técnicas estatísticas e de *Machine Learning*. Para tanto, foram analisadas variáveis relacionadas ao perfil do colaborador, à jornada de trabalho e às características da obra, tendo como variável dependente o evento de reposição de vaga após o desligamento. O objetivo é identificar e analisar os principais determinantes da rotatividade, além de desenvolver modelos preditivos capazes de estimar a probabilidade de substituição do colaborador, fornecendo subsídios para uma gestão mais estratégica e eficiente da mão de obra.

REFERENCIAL TEÓRICO

A gestão de pessoas assume papel estratégico em organizações intensivas em mão de obra, como as empresas de montagem eletromecânica, nas quais o desempenho operacional está diretamente relacionado à disponibilidade e continuidade do trabalho realizado pelas equipes. Segundo Chiavenato (2014), a gestão de pessoas envolve o desenvolvimento, manutenção e monitoramento da força de trabalho por meio de práticas que conciliam

objetivos organizacionais e individuais. Em ambientes industriais, essa função torna-se ainda mais relevante devido à necessidade de adequação contínua da equipe às demandas produtivas, aos prazos do projeto e às condições de trabalho em campo.

Em obras de montagem eletromecânica, o caráter temporário dos projetos, a mobilização de equipes para diferentes regiões e a variabilidade das condições operacionais exigem processos estruturados de recrutamento, alocação e retenção. Dutra (2016) destaca que, nesses ambientes, a dinâmica de entrada e saída de colaboradores é influenciada por fatores como ritmo de produção, investimentos enxutos, prazos reduzidos e pressão por resultados. Assim, a gestão de pessoas não se limita a administrar admissões e desligamentos, mas a tomar decisões que afetam diretamente a continuidade produtiva, especialmente em atividades técnicas ou de difícil reposição.

Nesse cenário, a rotatividade de colaboradores (*turnover*) destaca-se como um fenômeno central e deve ser analisada além do simples desligamento. Chiavenato (2014) define rotatividade como o movimento de entradas e saídas de trabalhadores, cujos impactos incidem sobre custos, clima, cultura e produtividade. Entretanto, para o contexto da montagem eletromecânica, a rotatividade assume caráter específico quando associada à reposição de vagas, ou seja, à decisão gerencial de substituir o colaborador após seu desligamento. Nesse caso, não se estuda apenas a saída do profissional, mas a necessidade de recompor seu posto de trabalho, o que implica custos diretos e continuidade das operações.

Hom et al. (2017) reforçam que decisões envolvendo desligamento e substituição decorrem da interação entre fatores individuais, como desempenho, absenteísmo e vínculo com a empresa, e fatores organizacionais, como condições de trabalho, regime de jornada e metas estabelecidas. Quando a vaga precisa ser preenchida, a rotatividade passa a afetar diretamente o fluxo produtivo, pois novas contratações requerem ambientação, adaptação às normas do cliente e tempo até que o desempenho do novo trabalhador alcance o padrão exigido. De forma semelhante, Aver et al. (2020) e Samuel (2024) identificam que, em projetos industriais, características específicas do empreendimento (segmento, porte, localidade e pressão de cronograma) explicam parte significativa da necessidade de substituições.

Sob essa perspectiva, a reposição de colaboradores implica impactos econômicos diretos e indiretos. Os diretos incluem custos ligados a processos admissionais, treinamentos, exames, mobilização e logística, especialmente quando há deslocamento interestadual ou necessidade de alojamento. Os indiretos envolvem queda temporária de produtividade, redistribuição de tarefas, demora no aprendizado tácito e necessidade de supervisão adicional, aspectos essenciais em atividades de precisão ou de segurança operacional (Hom et al., 2017). Em setores regulados e com exigência de qualificação técnica, a perda desse capital humano acumulado impacta a curva de aprendizagem e o ritmo de execução, tornando a substituição particularmente onerosa.

Outro fator relevante refere-se aos efeitos do contexto geográfico e social sobre a permanência e a substituição. Projetos industriais mobilizam trabalhadores de diferentes regiões do país, e variáveis como distância da residência, necessidade de alojamento e deslocamento familiar influenciam o tempo de permanência e disponibilidade para a jornada. Samuel (2024) observa que a permanência em obras industriais está associada à capacidade

de conciliar trabalho e vida pessoal, especialmente quando o colaborador atua longe da residência. Assim, estado de origem e disponibilidade regional de mão de obra também interferem na decisão de repor a vaga, pois afetam custos e a facilidade de substituição.

Além disso, aspectos ligados à jornada de trabalho podem influenciar a decisão de substituição. Horas extras e absenteísmo representam indicadores operacionais que revelam disponibilidade produtiva. Chiavenato (2014) destaca que a consistência da presença e o comprometimento com metas organizacionais estão relacionados à permanência. No contexto das obras, enquanto horas extras podem indicar maior contribuição ao ritmo produtivo, elevadas proporções de absenteísmo podem comprometer a continuidade do trabalho, reforçando a necessidade de substituição para evitar atrasos e redistribuição de frentes.

Diante da complexidade do fenômeno, o uso de métodos quantitativos torna-se essencial para apoiar decisões na área de gestão de pessoas. A aplicação de *People Analytics* surge como abordagem capaz de analisar padrões comportamentais e prever eventos de desligamento e reposição com maior precisão. Para Fávero e Belfiore (2024), técnicas estatísticas e algoritmos de aprendizado de máquina permitem identificar fatores explicativos e estimar probabilidades futuras, substituindo decisões empíricas. Em ambientes industriais, essa abordagem possibilita antecipar substituições, dimensionar equipes, reduzir custos de rotatividade e melhorar a continuidade operacional.

Assim, compreender a rotatividade como reposição de vagas, analisar seus determinantes produtivos e sociais e utilizar métodos analíticos para prever esse fenômeno constitui uma evolução necessária para a gestão de pessoas em projetos industriais. Esse enquadramento teórico reforça a relevância prática e científica de pesquisas que, como a presente, buscam identificar fatores que influenciam a substituição de colaboradores em obras de montagem eletromecânica, por meio de técnicas estatísticas e de *machine learning*, produzindo conhecimento aplicável à tomada de decisão estratégica.

METODOLOGIA

Este estudo é uma pesquisa aplicada, com abordagem quantitativa e de natureza explicativa, pois busca entender como variáveis organizacionais e individuais influenciam a rotatividade de colaboradores em obras de montagem eletromecânica. Para isso, foram utilizadas técnicas de estatística e *Machine Learning*, permitindo identificar padrões e fatores que impactam a rotatividade. Como destacado por Hair et al. (2009), pesquisas explicativas ajudam a estabelecer conexões entre variáveis e fornecem uma base mais sólida para a tomada de decisão estratégica nas empresas. Assim, a metodologia foi dividida em três etapas principais: coleta de dados, processamento e análise dos dados e aplicação de modelos estatísticos e de *Machine Learning*.

Coleta de Dados

Os dados foram coletados de uma empresa brasileira de montagem eletromecânica de médio porte, que possui aproximadamente 2.700 funcionários. A base de dados inclui informações de profissionais desligados da organização nos últimos dois anos e sete meses (janeiro de 2023 a julho de 2025), diferenciando aqueles que sofreram rotatividade daqueles desligados por término de obra (final de contrato).

Neste estudo, considerou-se rotatividade a substituição de um colaborador por outro de mesmo cargo e mesma obra. Um aspecto relevante é que, no período analisado, a empresa executou cerca de 20 empreendimentos, cada um mobilizando, em média, 500 trabalhadores. Essa dinâmica explica o número de desligamentos acumulados, significativamente superior ao quadro ativo, totalizando mais de 7.000 registros no período.

As informações foram extraídas do banco de dados corporativo hospedado em *SQL Server*, utilizando consultas estruturadas (comandos SQL), conforme recomendado em Elmasri et al. (2005). A base consolidada resultou em 7.333 registros, contemplando nove variáveis relacionadas ao perfil dos colaboradores (cargo, salário, sexo, estado civil, idade, grau de escolaridade, tempo de empresa, tipo de mão de obra - direta ou indireta - e estado de residência), duas variáveis relacionadas à jornada de trabalho (percentual de horas extras trabalhadas e percentual de horas abonadas - absenteísmo) duas variáveis relacionadas às características da obra (tipo da obra e distância da obra em relação ao local de residência do colaborador). Além dessas variáveis explicativas, foi definida como variável dependente (*target*) uma variável binária, que assume valor 1 quando ocorreu substituição do colaborador desligado e valor 0 quando não ocorreu substituição (Tabela 1).

Tabela 1. Variáveis do *Dataset*

Categoria	Variável	Tipo de dado	Descrição
Perfil do colaborador (variáveis individuais)	Cargo	Qualitativo	Função exercida pelo colaborador na obra
	Salário	Quantitativo	Remuneração mensal
	Sexo	Qualitativo	Masculino / Feminino
	Estado_Civil	Qualitativo	Solteiro, casado, divorciado
	Idade	Quantitativo	Idade em anos do colaborador
	Grau_Instrucao	Qualitativo	Ensino fundamental, médio, técnico e superior
	_Tempo_Casa_Mes	Quantitativo	Tempo de vínculo do colaborador na organização (em meses)
	Tipo_MO	Qualitativo	Direta (Mão na massa) ou indireta (apoio à obra)
	Estado_Residencia	Qualitativo	Unidade federativa de residência do colaborador
Jornada de trabalho (variáveis organizacionais)	Horas_Extras	Quantitativo	Proporção de horas extras em relação à carga horária total
	Abono	Quantitativo	Proporção de horas abonadas (horas faltas justificadas) em relação à carga horária total
Obra (variáveis organizacionais)	Tipo_Obra	Qualitativo	Segmento da obra (mineração, siderurgia, petróleo etc.)
	Distancia_Casa_Obra	Quantitativo	Distância em quilômetros entre a obra e a residência do colaborador
Target (variável dependente)	Reposicao_Vaga	Qualitativo	Indica se houve substituição após o desligamento (1 = sim; 0 = não)

Fonte: Autores (2025).

As variáveis analisadas foram divididas em variáveis individuais e variáveis organizacionais, conforme discutido em Chiavenato (2014).

Processamento dos dados

Antes da modelagem, aplicaram-se técnicas de pré-processamento para garantir qualidade e consistência da base (Han et al., 2020). Foi realizado o agrupamento de categorias com baixa frequência. As variáveis categóricas com poucas ocorrências, como função, tipo de obra, grau de instrução e estado civil, tiveram suas categorias menos frequentes agrupadas em “Outras”. Após isso foi feito a “dummização” de variáveis categóricas. As variáveis categóricas foram transformadas em variáveis binárias, permitindo sua utilização em algoritmos de *Machine Learning*, que exigem dados numéricos (James et al., 2013). Após esse processo, a base passou a contar com 59 variáveis explicativas.

O processamento dos dados e as análises computacionais foram realizadas em *Python* v.3.12, utilizando as bibliotecas *Pandas, Numpy, Seaborn, Matplotlib, Statsmodels* e *Sklearn*.

Análise exploratória dos dados

A análise exploratória incluiu medidas descritivas, como média, mediana e dispersão para as variáveis numéricas, e frequências para as variáveis categóricas. Foram examinadas também as correlações entre a variável dependente e as variáveis explicativas, por meio de coeficientes adequados ao tipo de dado. Para as variáveis categóricas, aplicou-se o coeficiente de associação Cramér's V, calculado a partir do teste qui-quadrado, o qual varia de 0 (sem associação) a 1 (associação forte), conforme proposto por Cramér (1999) e amplamente utilizado em estudos contemporâneos de análise de associação.

Já para as variáveis quantitativas, adotou-se a correlação de Spearman (ρ), que mede a intensidade e a direção de relações monotônicas entre variáveis a partir da ordenação de seus valores, sendo especialmente recomendada quando não se pode assumir linearidade ou normalidade na distribuição dos dados (Spearman, 1961). Além disso, avaliou-se a distribuição da variável dependente e elaboraram-se gráficos bivariados, possibilitando examinar visualmente a relação entre a variável dependente e as variáveis explicativas.

Aplicação dos modelos

Neste estudo, foram explorados três métodos supervisionados para classificação binária (substituição: sim/não): Regressão Logística Binária, Árvore de Decisão de classificação e *Random Forest*. A escolha desses modelos permitiu uma abordagem comparativa, conciliando interpretabilidade (regressão logística) e poder preditivo (*Random Forest*).

Para os modelos de Arvore de decisão e *Random Forest*, o conjunto de dados foi dividido em subconjuntos de treinamento e teste, utilizando uma proporção de 70% e 30%, respectivamente. Essa divisão, recomendada por Hair et al. (2009), assegura que o modelo seja treinado em uma porção dos dados e avaliado de forma imparcial em outra, não vista durante o treino, garantindo assim uma estimativa robusta de sua capacidade de generalização.

O desempenho dos modelos foi avaliado com base em um conjunto de métricas apropriadas para problemas de classificação binária, incluindo Acurácia, Precisão, *Recall* (Sensibilidade), *F1-Score* e a Área sob a curva ROC (AUC-ROC) (Fávero & Belfiore, 2024).

A Regressão Logística foi escolhida por ser um modelo estatístico paramétrico clássico que, além de fornecer previsões, permite inferir sobre a relação probabilística entre as variáveis explicativas e a variável dependente binária (Fávero & Belfiore, 2024).

O modelo estima a probabilidade de um evento (Y=1, ou seja, reposição da vaga) ocorrer através da função logística (também conhecida como sigmoide), que transforma a saída de uma regressão linear em um valor entre 0 e 1. A forma funcional do modelo é dada pela Eq. 1.

$$P(Y=1|X)i = \frac{1}{1+e^{-(\alpha+\beta 1X1+\beta 2X2+\cdots+\beta kXk)i}} \quad (1)$$

Em que P(Y=1|X) é a probabilidade do evento; α é o intercepto; β1, β2, ..., βK são os coeficientes estimados para cada variável explicativa; X1, X2, ..., Xk são os valores das variáveis explicativas.

Os parâmetros β do modelo são estimados pelo método da Máxima Verossimilhança (*Maximum Likelihood Estimation* - MLE). Este método busca encontrar os valores dos coeficientes que maximizam a Função de *Log-Likelihood* (LLF), ou equivalentemente, minimizam o negativo do (LLF). A função de *log-Likelihood* para a regressão logística é expressa por meio da Eq. 2.

$$LL = \sum_{i=1}^{n}\{[(y_i) * \ln(P_i)] + [(1 - y_i) * \ln(1 - P_i)]\} \quad (2)$$

Onde Yi é o valor observado (0 ou 1) para a i-ésima observação e Pi é a probabilidade estimada pelo modelo para a i-ésima observação. Um valor de *Log-Likelihood* mais próximo de zero (menos negativo) indica um melhor ajuste do modelo aos dados observados (Fávero & Belfiore, 2024).

Após o modelo estimado, foi empregada também a técnica de seleção *stepwise*, a fim de obter um modelo parcimonioso e evitar a inclusão de variáveis irrelevantes. Esse procedimento consiste na inclusão e exclusão sequencial de variáveis explicativas com base em critérios estatísticos de significância e ajuste global do modelo, garantindo maior precisão e interpretabilidade dos resultados (Fávero & Belfiore, 2024).

Para complementar a análise iniciada com o modelo estatístico de Regressão Logística, foi realizado mais dois modelos de *Machine Learning*, Arvore de Decisão e *Random Forest*. A escolha desses modelos se deu para complementar o estudo, ter um maior poder preditivo e uma melhor relação entre interpretabilidade e *performance*.

Enquanto a Regressão Logística oferece transparência e inferência causal através dos coeficientes, as Árvores e *Ensembles* como *Random Forest* capturam relações não-lineares e de interação entre variáveis de forma automática, sem a necessidade de transformações complexas por parte do pesquisador (Hastie et al., 2009).

A *Random Forest*, em particular, é reconhecida por sua alta acurácia e robustez em problemas de classificação do mundo real, frequentemente superando modelos mais simples (Breiman, 2001).

Segundo James et al. (2013), o algoritmo de Árvore de Decisão é um método de aprendizado supervisionado não paramétrico que estrutura o processo de classificação em uma sequência de regras de decisão simples e intuitivas, similares a um fluxograma. O modelo é construído particionando recursivamente o conjunto de dados com base nos valores das variáveis explicativas criando nós de decisão que maximizam a pureza dos grupos resultantes.

Para auxiliar na interpretação do modelo de árvore de decisão, foi incluído no resultado a importância das variáveis. Nesse tipo de modelagem, a importância das variáveis é calculada a partir da redução da impureza gerada por cada divisão (*split*) ao longo da árvore, geralmente medida pelo índice de Gini ou pela entropia. Em cada nó, o algoritmo avalia quanto a divisão com determinada variável melhora a separação das classes, acumulando esse ganho em todos os pontos em que a variável é utilizada. Ao final, esses valores são normalizados para que a soma das importâncias seja igual a 1. Dessa forma, variáveis com maior importância são aquelas que mais contribuíram para a classificação correta, enquanto variáveis com valores próximos de zero tiveram influência irrelevante no processo decisório do modelo (James et al., 2013).

Já o *Random Forest* é um algoritmo de aprendizado por conjunto (*Ensemble Learning*) desenvolvido por Breiman (2001). O seu funcionamento baseia-se na construção de um grande número de Árvores de Decisão, cada uma treinada com um propósito específico. Este modelo foi escolhido para este estudo porque eleva significativamente a *performance* preditiva em relação a uma árvore única, mitigando o problema de *overfitting*.

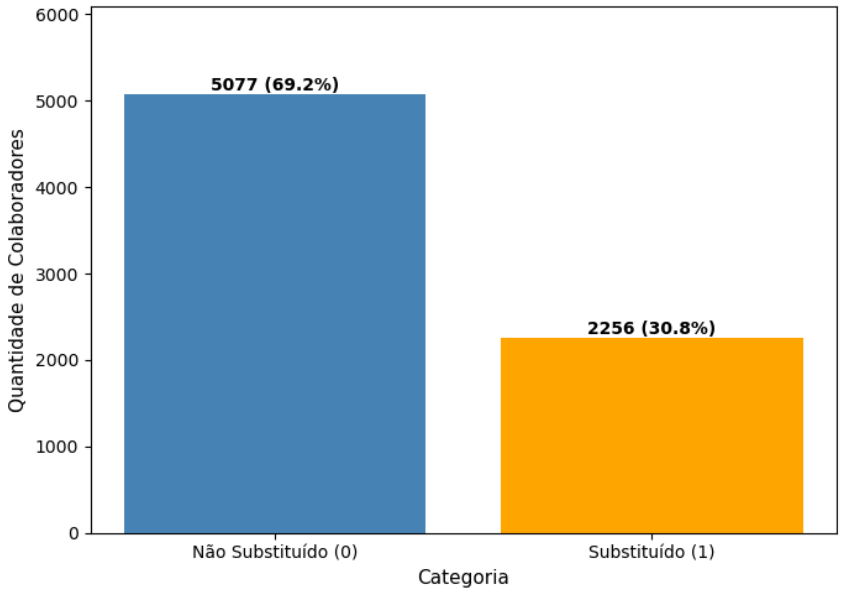
Para a calibração do modelo foi utilizado o método de busca em grade (*Grid Search*), com validação cruzada. Essa técnica consiste em testar combinações de hiperparâmetros pré-definidos e selecionar aquela que proporciona o melhor desempenho em métricas de validação. Os hiperparâmetros avaliados incluíram: número de árvores na floresta (*n_estimators*), profundidade máxima das árvores (*max_depth*), número mínimo de amostras para divisão interna (*min_samples_split*) e número mínimo de amostras em cada folha (*min_samples_leaf*). A busca foi conduzida com validação cruzada de *k-folds*, garantindo maior robustez na escolha do modelo final. Esse procedimento permitiu identificar a configuração que maximizou as métricas de acurácia balanceada e AUC-ROC, utilizadas como principais critérios de seleção (Pedregosa et al., 2011).

RESULTADOS E DISCUSSÃO

Análise exploratória dos dados

Diante da base de dados, iniciou-se o procedimento de análise descritiva das variáveis. A Figura 1 apresenta a distribuição da variável *target*, demonstrando a frequência dos colaboradores que foram substituídos e aqueles que foram desligados por término de obra (não foram substituídos).

Figura 1. Distribuição da variável *target* (Reposição de vaga)



Fonte: Autores (2025).

Observa-se que 69,2% dos desligamentos não resultaram em substituição do colaborador, enquanto 30,8% correspondem a casos em que houve substituição. Esse resultado evidencia uma predominância de situações de não reposição de vaga. Apesar de se observar um desbalanceamento entre as categorias, com maior incidência de registros em que não houve substituição do colaborador, o estudo prosseguiu com a análise exploratória e a modelagem preditiva, visto que a quantidade de dados disponível em ambas as classes se mostrou suficiente para a aplicação dos métodos propostos.

As variáveis categóricas da base de dados foram analisadas com o objetivo de identificar o perfil predominante dos colaboradores e possíveis padrões relacionados ao desligamento com ou sem substituição. A Tabela 2 apresenta a frequência das categorias das variáveis categóricas consideradas no estudo, permitindo observar o perfil predominante dos colaboradores quanto ao tipo de obra, sexo, estado civil, grau de instrução e tipo de mão de obra.

Tabela 2. Frequência das variáveis categóricas do modelo

Variável	Categoria	Frequência	Percentual (%)
Tipo_Obra	Dutovia	177	2%
	Manutenção	726	10%
	Mineração	1.887	26%
	Químico	610	8%
	Siderurgia	1.719	23%
	Tancagem	2.214	30%
Sexo	Feminino	178	2%
	Masculino	7.155	98%
Estado_Civil	Casado	1.965	27%
	Divorciado	74	1%
	Solteiro	5.294	72%
Grau_Instrução	Ensino Fundamental Completo	873	12%
	Ensino Fundamental Incompleto	228	3%
	Ensino Médio Completo	5.803	79%
	Ensino Médio Incompleto	229	3%
	Ensino Superior Completo	90	1%
	Técnico de Nível médio	110	2%
Tipo_MO	Mão-de-Obra Direta	6.679	91%
	Mão-de-Obra Indireta	654	9%

Fonte: Autores (2025).

Constata-se que a maior parte dos colaboradores atuava em projetos de tancagem (30%), seguidos de mineração (26%) e siderurgia (23%). A distribuição por sexo revela ampla predominância de trabalhadores do sexo masculino (98%), enquanto apenas 2% da amostra corresponde ao sexo feminino.

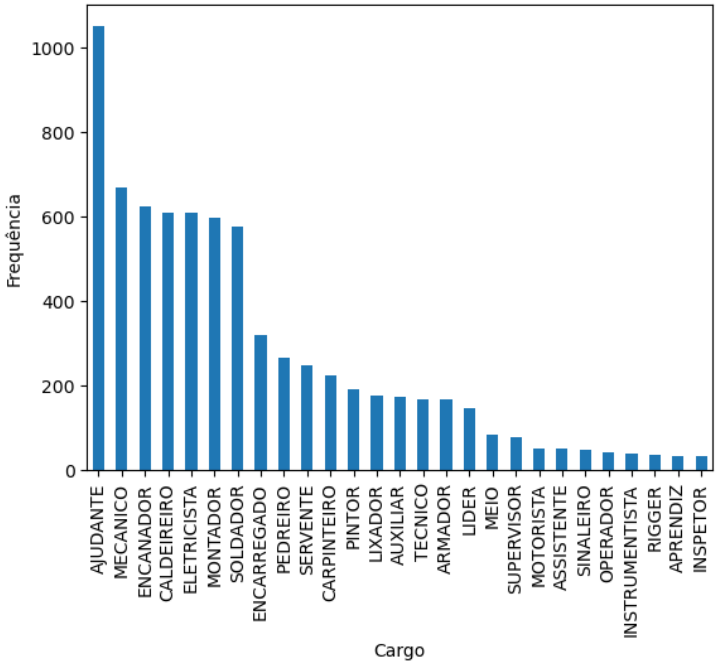
No que se refere ao estado civil, a maior parte dos colaboradores era solteiro (72%), seguida por casados (27%), perfil compatível com a mobilidade exigida em obras que demandam deslocamentos frequentes e períodos de afastamento da residência.

Quanto ao grau de instrução, destaca-se o ensino médio completo (79%), o que confirma a predominância de trabalhadores com escolaridade de nível intermediário. Apenas 1% da amostra possuía ensino superior, refletindo a menor representatividade de funções técnicas e administrativas na composição total.

Por fim, observa-se que 91% dos colaboradores estavam alocados em mão de obra direta, desempenhando funções operacionais ligadas à execução em campo, enquanto apenas 9% correspondiam à mão de obra indireta, voltada a atividades de apoio e supervisão.

Para complementar a análise das variáveis qualitativas, foram elaboradas as Figuras 2 e 3, demonstrando a frequência das variáveis Cargo e Estado de residência, respectivamente.

Figura 2. Frequência da variável Cargo

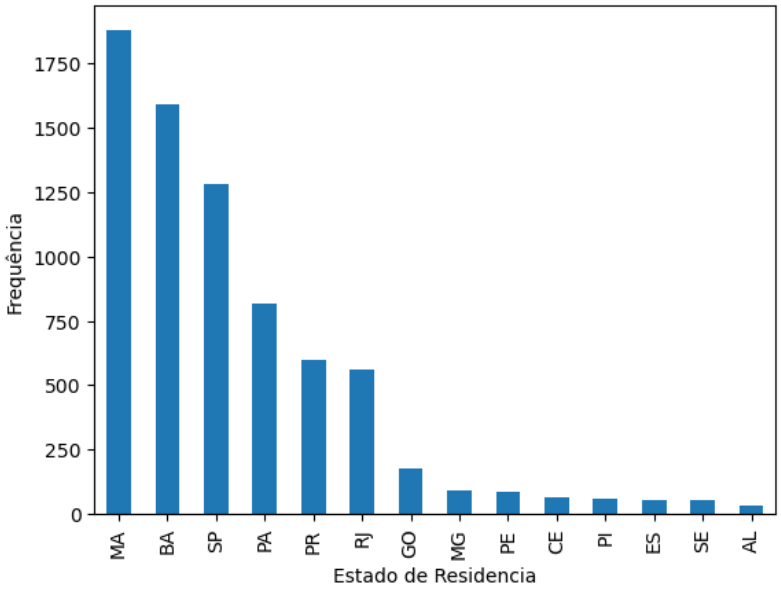


Fonte: Autores (2025).

Nota-se que há elevada concentração de funções operacionais, com destaque para ajudantes, mecânicos, encanadores, caldeireiros e eletricistas, que juntos representam a maior parcela da força de trabalho.

Outro aspecto relevante é a menor representatividade de funções de apoio, como supervisores, técnicos, motoristas e inspetores. Essas categorias aparecem em menor escala no gráfico, o que reforça a proporção observada anteriormente entre mão de obra direta e indireta.

Figura 3. Frequência da variável Estado de residência



Fonte: Autores (2025).

Observa-se uma forte concentração de colaboradores residentes do Maranhão (MA), Bahia (BA) e São Paulo (SP), que juntos representam a maior parcela da amostra. Em seguida, destacam-se colaboradores oriundos do Pará (PA), Paraná (PR) e Rio de Janeiro (RJ), que também apresentam participação expressiva, enquanto os demais estados aparecem em proporções bem menores, com menor representatividade no total da amostra.

A Tabela 3 demonstra as estatísticas univariadas das variáveis quantitativas do modelo.

Tabela 3. Estatísticas univariadas das variáveis explicativas quantitativas

	Salário (R$)	Idade (Anos)	Tempo Casa (Meses)	Abono (%)	Horas Extras (%)	Distância Casa Obra (Km)
Média	2.805,7	40,0	7,6	10,7%	17,4%	466,0
Desvio-Padrão	1.075,7	11,0	9,9	12,6%	12,4%	592,3
Min	569,4	18,0	0,0	0,0%	0,0%	0,0
25%	2.300,0	32,0	2,9	2,3%	8,6%	0,0
50%	2.651,0	40,0	5,3	6,3%	15,4%	369,4
75%	3.170,8	47,0	9,2	14,5%	23,9%	491,6
Max	19.500,0	77,0	235,6	79,5%	79,1%	2728,0

Fonte: Autores (2025).

A partir das informações obtidas na Tabela 3, é possível identificar o perfil dos colaboradores em relação à faixa salarial, idade e tempo de empresa. Também se verificam os fatores relacionados às condições de trabalho, como o percentual de abono e horas extras dos colaboradores, além da distância da residência dos colaboradores até a obra.

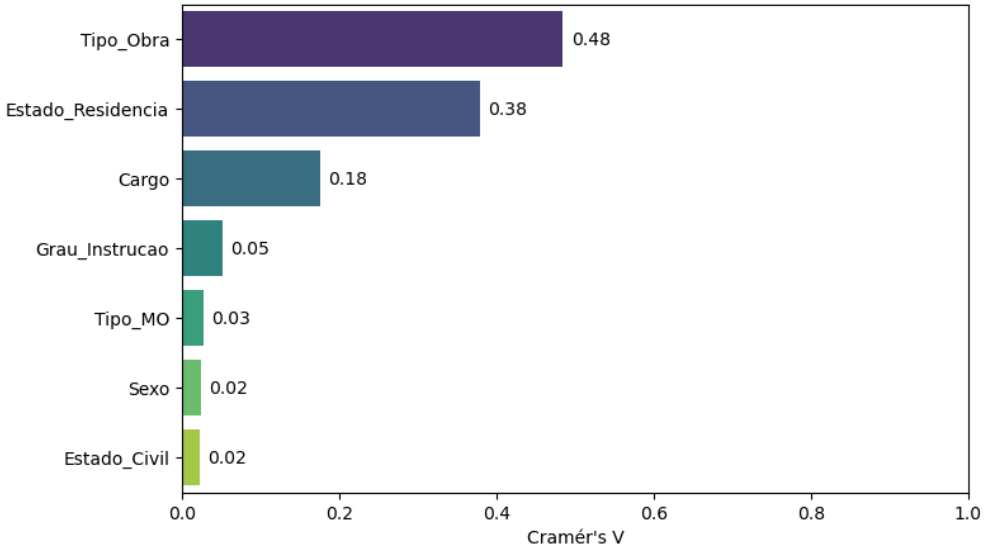
A média salarial da população do banco de dados é de R$ 2.805,70, com uma dispersão considerável (desvio padrão de R$ 1.075,70). A idade média dos funcionários era de 40 anos, com uma concentração entre 32 e 47 anos, no intervalo interquartil. Em relação ao tempo de casa, a média de permanência dos colaboradores, em meses, foi de 7,6, sugerindo que a maior parte dos colaboradores tem vínculo de curta duração.

No que se refere às variáveis da jornada de trabalho, destaca-se que o percentual médio de abono foi de 10,7%, com registros que chegaram a quase 80%, apontando para casos específicos de elevada ausência. De modo semelhante, o percentual de horas extras apresentou média de 17,4%, mas também com grande variabilidade, alcançando valores próximos de 80%, o que sugere que determinados grupos de colaboradores ou frentes de obra foram submetidos a jornadas intensificadas. Por fim, a distância média entre a residência do colaborador e a obra foi de 466 Km, com alguns casos que superam 2.700 Km, o que evidencia que parte dos colaboradores são alojados.

A análise de associação entre as variáveis explicativas e a variável dependente permitiu identificar quais fatores apresentam maior relevância para explicar o fenômeno de desligamento com ou sem reposição de vaga. Para as variáveis categóricas, aplicou-se o coeficiente de Cramér's V, e para as variáveis quantitativas, a correlação de Spearman, conforme descrito na metodologia.

A Figura 4 apresenta os resultados da associação das variáveis categóricas com a variável dependente. Observa-se que o Tipo de Obra (Cramér's V = 0,48) e o Estado de Residência (Cramér's V = 0,38) destacam-se como as variáveis de maior associação, sugerindo que tanto o contexto do projeto quanto a origem geográfica do trabalhador influenciam significativamente no padrão de reposição de vagas. Em contraste, variáveis como Sexo (Cramér's V = 0,04) e Estado Civil (Cramér's V = 0,03) apresentaram associações praticamente nulas, indicando baixa relevância para a explicação do fenômeno, quando analisadas de forma individualizadas.

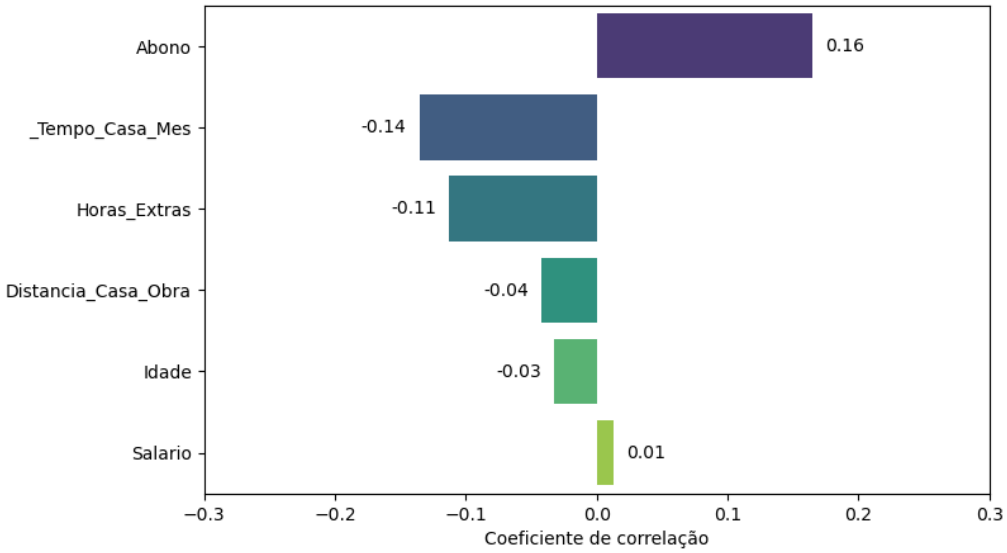
Figura 4. Associação entre as variáveis qualitativas e a substituição do colaborador



Fonte: Autores (2025).

No caso das variáveis quantitativas, (Figura 5), verifica-se que o Abono (ρ = 0,16) apresenta correlação positiva com a variável dependente, sugerindo que colaboradores com maior proporção de horas abonadas têm maior probabilidade de serem substituídos. Por outro lado, o Tempo de Casa (ρ = –0,14) mostrou correlação negativa, indicando que trabalhadores com vínculos mais longos tendem a apresentar menor probabilidade de substituição. Variáveis como Idade, Salário e Horas Extras exibiram correlações fracas, mas ainda assim podem contribuir marginalmente para a explicação do fenômeno quando consideradas em conjunto no modelo preditivo.

Figura 5. Associação entre as variáveis quantitativas e a substituição do colaborador



Fonte: Autores (2025).

Aplicação dos modelos

Após a análise exploratória e descritiva dos dados, iniciou-se a etapa de aplicação dos modelos com o objetivo de avaliar o poder explicativo e preditivo das variáveis em relação ao fenômeno de desligamento com ou sem reposição de vaga. No caso, o primeiro modelo foi a regressão logística binária após o processo de *stepwise* (Figura 6).

Figura 6. Informações do modelo de regressão logística binária

Logit Regression Results

Dep. Variable:	Reposicao_Vaga	No. Observations:	7333
Model:	Logit	Df Residuals:	7306
Method:	MLE	Df Model:	26
Date:	Tue, 09 Sep 2025	Pseudo R-squ.:	0.2610
Time:	21:51:36	Log-Likelihood:	-3344.7
converged:	True	LL-Null:	-4526.0
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5391	0.170	-3.180	0.001	-0.871	-0.207
Salario	-0.0002	4.52e-05	-3.800	0.000	-0.000	-8.33e-05
Idade	-0.0064	0.003	-2.074	0.038	-0.012	-0.000
_Tempo_Casa_Mes	-0.0211	0.004	-5.057	0.000	-0.029	-0.013
Abono	2.0358	0.264	7.701	0.000	1.518	2.554
Horas_Extras	-3.6342	0.284	-12.776	0.000	-4.192	-3.077
Distancia_Casa_Obra	0.0005	7.1e-05	7.071	0.000	0.000	0.001
Tipo_Obra_Manutenção	-0.6055	0.172	-3.524	0.000	-0.942	-0.269
Tipo_Obra_Químico	1.0780	0.150	7.197	0.000	0.784	1.372
Tipo_Obra_Siderurgia	2.0257	0.094	21.597	0.000	1.842	2.210
Tipo_Obra_Tancagem	-0.7960	0.130	-6.124	0.000	-1.051	-0.541
Cargo_APRENDIZ	-3.2449	0.781	-4.152	0.000	-4.777	-1.713
Cargo_AUXILIAR	-0.8934	0.262	-3.411	0.001	-1.407	-0.380
Cargo_CALDEIREIRO	0.5921	0.126	4.714	0.000	0.346	0.838
Cargo_ELETRICISTA	0.3228	0.124	2.593	0.010	0.079	0.567
Cargo_ENCANADOR	0.7700	0.122	6.290	0.000	0.530	1.010
Cargo_ENCARREGADO	0.7611	0.188	4.050	0.000	0.393	1.129
Cargo_INSTRUMENTISTA	1.3318	0.366	3.643	0.000	0.615	2.048
Cargo_MECANICO	0.4451	0.116	3.837	0.000	0.218	0.673
Cargo_MONTADOR	0.3407	0.124	2.742	0.006	0.097	0.584
Cargo_SOLDADOR	0.7304	0.138	5.275	0.000	0.459	1.002
Tipo_MO_Mão-de-Obra Indireta	0.6548	0.171	3.833	0.000	0.320	0.990
Estado_Residencia_BA	-0.8022	0.100	-8.015	0.000	-0.998	-0.606
Estado_Residencia_CE	-0.7606	0.359	-2.120	0.034	-1.464	-0.057
Estado_Residencia_MA	-0.6249	0.109	-5.739	0.000	-0.838	-0.411
Estado_Residencia_PR	0.5260	0.172	3.055	0.002	0.189	0.863
Estado_Residencia_SP	1.1241	0.148	7.592	0.000	0.834	1.414

Fonte: Autores (2025).

O modelo apresentou significância estatística global LLR *p-value* < 0,001 e um Pseudo R^2 de 0,2610, indicando que, apesar de não explicar toda a variabilidade do fenômeno, possui poder explicativo adequado para este tipo de análise (Fávero & Belfiore, 2024).

Entre as variáveis analisadas, algumas se destacaram pela magnitude de seus efeitos sobre a probabilidade de reposição de vaga. O abono apresentou *odds ratio* (OR) de 7,66, indicando que colaboradores com maior número de abonos possuem mais de sete vezes a chance de serem substituídos em comparação aos demais, evidenciando o impacto do absenteísmo justificado na continuidade das obras.

No sentido oposto, a variável horas extras reduziu significativamente a chance de reposição (OR = 0,026), sugerindo que colaboradores que realizam maior carga de horas extras são menos propensos a serem substituídos, possivelmente por desempenharem papéis estratégicos.

Em relação ao tipo de obra, a siderurgia se destacou (OR = 3,79), mostrando que esse segmento demanda maior reposição de mão de obra, enquanto as obras de manutenção (OR = 0,55) e tancagem (OR = 0,45) apresentaram efeito negativo, com menor probabilidade de substituição.

Por fim, o fator geográfico também se mostrou relevante: colaboradores residentes em São Paulo apresentaram três vezes mais chance de reposição (OR = 3,08), enquanto estados como Bahia (OR = 0,45) reduziram essa probabilidade. Esses resultados revelam a importância conjunta de aspectos individuais, funcionais e contextuais na determinação da reposição de vagas.

Para verificar a *performance* do modelo, foram mensurados os indicadores apresentados na Tabela 4.

Tabela 4. Métricas do modelo de regressão logística binária

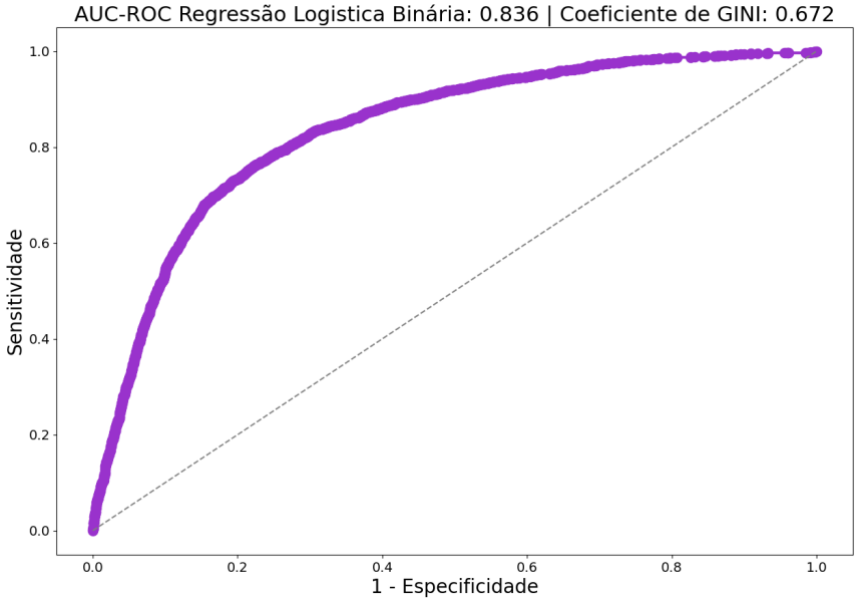
Métrica	Valor
Acurácia	0,792
Acurácia Balanceada	0,740
Precisão	0,685
Recall (Sensibilidade)	0,603
F1-score	0,641

Fonte: Autores (2025).

Os resultados mostram que o modelo de regressão logística apresentou bom desempenho geral, com acurácia média de 79,2% e acurácia balanceada de 74,0%, o que indica consistência mesmo diante do desbalanceamento da base. Embora a precisão (68,5%) tenha sido superior ao *recall* (60,3%), sugerindo que o modelo é mais conservador ao classificar reposições, o *F1-score* de 64,1% demonstra equilíbrio satisfatório entre as duas métricas. Esses valores estão em linha com pesquisas semelhantes que aplicaram modelos estatísticos e de *Machine Learning* para prever rotatividade ou substituição de colaboradores em setores intensivos em mão de obra, como construção civil e indústria (Hom et al., 2017).

Além das medidas informadas acima, também foi mensurado a área sob a curva ROC, um importante indicador que mostra a capacidade do modelo de distinguir entre as classes (quem foi substituído e quem não foi), independente de um ponto de corte (*cut-off*) (Figura 7).

Figura 7. AUC-ROC do modelo de regressão logística binária



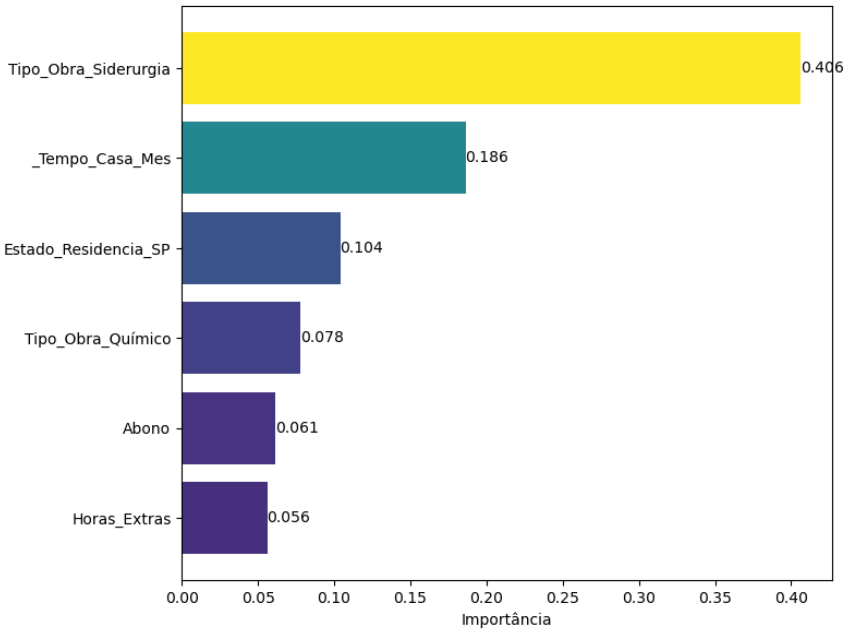
Fonte: Autores (2025).

A análise da curva ROC evidenciou um desempenho satisfatório do modelo de regressão logística, com AUC de 0,836 e coeficiente de Gini de 0,672. Esses resultados indicam bom poder discriminatório, ou seja, o modelo é capaz de distinguir adequadamente colaboradores que foram substituídos daqueles que não foram. De acordo com a literatura, valores de AUC entre 0,8 e 0,9 são considerados bons preditores (Hosmer et al., 2013), reforçando a adequação do modelo para fins de análise e previsão no contexto da reposição de vagas.

A título de comparação, outros dois modelos foram desenvolvidos. O modelo de árvore de decisão e *Random Forest*. Para a construção do modelo de árvore de decisão foram definidos os seguintes hiperparâmetros: profundidade máxima igual a 7, mínimo de 20 amostras por divisão e mínimo de 10 amostras por folha. A escolha desses valores foi realizada de forma iterativa, a partir de testes comparativos, buscando o equilíbrio entre desempenho preditivo e complexidade do modelo.

Após a modelagem, foi realizado as análises da importância das variáveis do modelo para a construção da árvore. A Figura 8 mostra as seis variáveis que mais contribuíram para reduzir a impureza (erro de classificação) ao longo de toda a árvore.

Figura 8. Importância das variáveis para a árvore de decisão



Fonte: Autores (2025).

A variável mais importante foi o tipo de obra em siderurgia (0,406), indicando que esse segmento concentra as condições mais determinantes para a reposição de colaboradores. Em seguida, o tempo de casa (0,186) mostrou-se fundamental, sugerindo que a permanência do colaborador na empresa exerce papel decisivo na manutenção ou substituição de vagas. O estado de residência em São Paulo (0,104) também apareceu como fator relevante, possivelmente associado à maior mobilidade e disponibilidade de mão de obra na região.

Variáveis como abono (0,061) e horas extras (0,056) mantiveram relevância, confirmando seu impacto já identificado na regressão logística. Por fim, o tipo de obra químico (0,078) complementou a lista das variáveis mais influentes. Esses resultados mostram que a árvore de decisão tende a valorizar variáveis estruturais (tipo de obra, tempo de casa, localização) em detrimento de variáveis demográficas ou contratuais, tendência já destacada por Chiavenato (2014) e observada também por Aver et al. (2020) em estudos sobre fatores de rotatividade (Tabela 5).

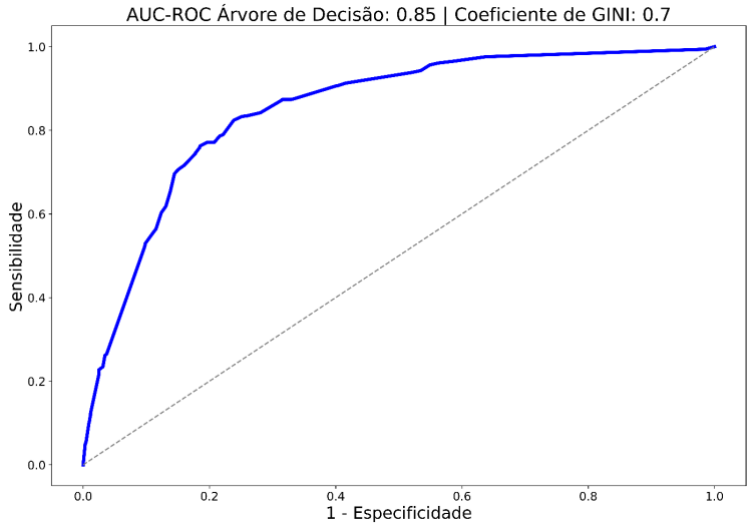
Tabela 5. Métricas do modelo de Árvore de Decisão

Métrica	Valor
Acurácia	0,805
Acurácia Balanceada	0,778
Precisão	0,668
Recall (Sensibilidade)	0,708
F1-score	0,688

Fonte: Autores (2025).

Observa-se que o modelo de árvore de decisão apresentou acurácia de 80,5% e acurácia balanceada de 77,8%, valores considerados satisfatórios para problemas de classificação em bases desbalanceadas. A sensibilidade de 70,8% indica boa capacidade do modelo em identificar colaboradores substituídos. O *F1-score* de 68,8% demonstra equilíbrio adequado entre precisão e *recall*, reforçando a robustez do modelo no contexto estudado. Para finalizar a análise da árvore de decisão, foi plotado a curva ROC da base de teste (Figura 9).

Figura 9. AUC-ROC do modelo de árvore de decisão



Fonte: Autores (2025).

O modelo alcançou AUC-ROC de 0,85 e coeficiente de Gini de 0,70, valores que indicam bom poder discriminatório. Isso significa que o modelo possui 85% de chance de classificar corretamente um colaborador substituído e um não substituído escolhidos aleatoriamente. Por fim, foi implementado o modelo de *Random Forest*. Foram selecionados os seguintes hiperparâmetros ótimos: *n_estimators* = 500, *max_depth* = 8, *max_features* = 35 e *min_samples_split* = 15 (Tabela 6).

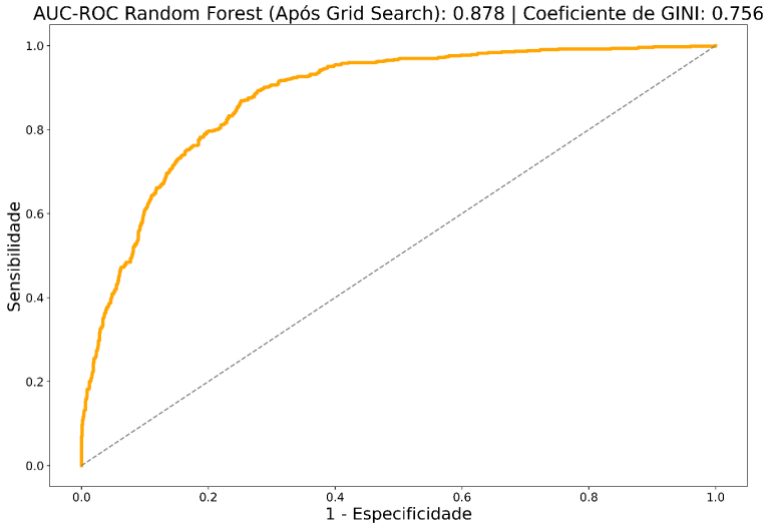
Tabela 6. Métricas do modelo de *Random Forest*

Métrica	Valor
Acurácia	0,810
Acurácia Balanceada	0,771
Precisão	0,693
Recall (Sensibilidade)	0,671
F1-score	0,681

Fonte: Autores (2025).

O modelo obteve acurácia de 81,0% e acurácia balanceada de 77,1%, reforçando sua capacidade de generalização em um cenário de dados desbalanceados. A precisão de 69,3% e o *recall* de 67,1% indicam um bom equilíbrio entre identificar corretamente os colaboradores substituídos e evitar falsos positivos, refletido no *F1-score* de 68,1%. Por fim, a Figura 10 representa a curva ROC do modelo.

Figura 10. AUC-ROC do modelo de *Random Forest*



Fonte: Autores (2025).

O modelo alcançou AUC-ROC de 0,878 e coeficiente de Gini de 0,756, resultados que evidenciam excelente poder discriminatório. Na prática, isso significa que o modelo tem quase 88% de chance de distinguir corretamente um colaborador substituído de um não substituído, quando selecionados aleatoriamente.

CONSIDERAÇÕES FINAIS

O presente estudo demonstrou que a aplicação de modelos estatísticos e de aprendizado de máquina são capazes de identificar de forma consistente os principais fatores associados à reposição de vagas em projetos de montagem eletromecânica. As análises evidenciaram que variáveis relacionadas ao perfil do colaborador, às características do trabalho e ao contexto regional exercem influência significativa sobre a decisão de substituição.

A utilização comparativa de regressão logística, árvore de decisão e *Random Forest* permitiu verificar que diferentes técnicas oferecem perspectivas complementares. Enquanto a regressão logística forneceu maior clareza interpretativa sobre o efeito das variáveis, os modelos baseados em árvores apresentaram desempenho preditivo superior e maior sensibilidade na identificação dos casos de substituição.

Constata-se que a integração dessas abordagens pode contribuir para uma gestão de pessoas mais eficaz, auxiliando na antecipação de situações críticas de rotatividade e no planejamento de estratégias voltadas à retenção de mão de obra. Os resultados obtidos reforçam a aplicabilidade de ferramentas analíticas no apoio à tomada de decisão em ambientes de alta complexidade, como os projetos industriais.

Este estudo apresenta como limitação o fato de ter sido desenvolvido a partir dos dados de uma única empresa, restringindo a generalização dos resultados. Além disso, foram consideradas apenas variáveis objetivas, sem contemplar aspectos subjetivos como clima organizacional ou satisfação dos colaboradores. Como sugestão para pesquisas futuras, recomenda-se expandir a análise para diferentes empresas e setores, bem como incluir variáveis comportamentais e psicossociais.

REFERÊNCIAS

Aver, G., Miri, D. H., Chais, C., Matte, J., Ganzer, P. P., & Olea, P. M. (2020). Fatores de rotatividade em uma empresa do segmento metalomecânico: Rotativity factors in a mechanical metal segment company. *Revista Visão: Gestão Organizacional*, 9(2), 168-186.

Breiman, L. (2001). *Random forests. Machine learning*, 45(1), 5-32.

Chiavenato, I. (2014). *Gestão de Pessoas: O Novo Papel dos Recursos Humanos nas Organizações*. 4 ed. Elsevier Brasil.

Cramér, H. (1999). *Mathematical methods of statistics* (Vol. 9). Princeton university press.

Dutra, J. S. (2016). *Gestão de pessoas: modelo, processos, tendências e perspectivas*. 3 ed. Atlas.

Elmasri, R., Navathe, S. B., & Pinheiro, M. G. (2005). *Sistemas de banco de dados*.

Fávero, L. P., & Belfiore, P. (2024). *Manual de análise de dados: Estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python®*. GEN, LTC.

Hair, J. F. (2009). *Multivariate data analysis*.

Han, J., Kamber, M., & Pei, J. (2020). *Data mining: Concepts and. Techniques*, Waltham: Morgan Kaufmann Publishers.

Hastie, T. (2009). *The elements of statistical learning: data mining, inference, and prediction*.

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of applied psychology*, 102(3), 530.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

James, G. (2013). *An introduction to statistical learning with applications in R*.

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., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12, 2825-2830.

Samuel, F. (2024). *Retention Strategies for Project-Based Construction Workers: The Role of Career-Family Balance Initiatives*.

Spearman, C. (1961). *"General Intelligence" Objectively Determined and Measured*.