Industrialização, Digitalização, Desempenho

5º Simpósio Brasileiro de Tecnologia da Informação e Comunicação na Construção e 5º Workshop de Tecnologia de Processos e Sistemas Construtivos

FLORIANÓPOLIS-SC | 20 a 22 de agosto

[1]MODELO DE PREDIÇÃO DE ACIDENTES COM AFASTAMENTO USANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Prediction model for lost time accidents using machine learning techniques

Filipe dos Santos Freitas
Universidade Federal da Bahia | Salvador, Bahia | filipe.freitas@ufba.br

Mirian Caroline Farias Santos
Universidade Federal da Bahia | Salvador, Bahia | mirian.caroline@ufba.br

Roseneia Rodrigues Santos de Melo
Universidade Federal da Bahia | Salvador, Bahia | roseneia.engcivil@gmail.com

Paulo Henrique Ferreira
Universidade Federal da Bahia | Salvador, Bahia | paulohenri@ufba.br

Dayana Bastos Costa
Universidade Federal da Bahia | Salvador, Bahia | dayanabcosta@ufba.br

RESUMO

Na construção civil, as lesões com afastamentos implicam em um elevado custo para a saúde e previdência social, além dos custos indiretos para a empresa e os danos causados ao trabalhador. Nesse contexto, o aprendizado de máquina (AM) destaca-se como uma tecnologia promissora para agilizar a análise de grandes volumes de dados e prever eventos, auxiliando na gestão preventiva de acidentes. Portanto, este artigo busca desenvolver um modelo de predição de lesões com afastamento usando algoritmos de AM e dados históricos de acidentes. O conjunto de dados compreende 28.100 acidentes (ocorrências da construção civil – trabalhadores regulares), de 2018 a 2023, distribuídos em 15 variáveis binárias e categóricas. Após o pré-processamento, as previsões foram realizadas por meio de sete modelos com diferentes classificadores. O modelo *Random Forest* apresentou o melhor desempenho, alcançando acurácia de 0,775, precisão de 0,768 e recall de 0,776. Como contribuição, este estudo demostra o potencial da tecnologia para análise e predição de lesões com afastamento, além de possibilitar insights sobre os atributos que mais influenciam na ocorrência dos eventos. Entretanto, o estudo destaca a necessidade da inclusão de atributos relacionados ao contexto do acidente, visando aumentar a precisão e robustez do modelo.

Palavras-chave: Gestão de segurança; Indústria 4.0; Inteligência Artificial; Acidentes; Modelos preditivos.

ABSTRACT

In the construction industry, accidents resulting in leave of absence entail high costs for health and social security, as well as indirect costs for companies and damage to workers. In this context, machine learning (ML) stands out as a promising technology for accelerating the analysis of large volumes of data and predicting events, assisting in preventive accident management. Therefore, this article aims to develop a predictive model for leave-of-absence accidents using ML algorithms and historical accident data. The dataset comprises 28,100 accidents (occurrences in the construction industry – regular workers) from 2018 to 2023, distributed across 15 binary and categorical variables. After preprocessing, predictions were made using seven models with different classifiers. The Random Forest model achieved the best performance, reaching an accuracy of 0.775, a precision of 0.768, and a recall of 0.776. As a contribution, this study demonstrates the potential of technology for analyzing and predicting leave-of-absence accidents, as well as providing insights into the attributes that most influence the occurrence of such events. However, the study highlights the need to include attributes related to the accident context to increase the model's accuracy and robustness.

Keywords: Safety Management; Industry 4.0; Artificial Intelligence; Accidents; Predictive Models.

[1]FREITAS, F. S.; SANTOS, M. C. F.; MELO, R. R. S.; FERREIRA, P. H.; COSTA, D. B. Modelo de Predição de Acidentes com Afastamento usando Técnicas de Aprendizado de Máquina. In: 5º SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO NA CONSTRUÇÃO, 4., 2025, Florianópolis. **Anais [...].** Porto Alegre: ANTAC, 2025.

1 INTRODUÇÃO

Os afastamentos de trabalhadores no Brasil, especialmente no setor da construção civil, representam um desafio crescente, com impactos tanto para as empresas quanto para os sistemas de saúde e seguridade social. Nos últimos anos, houve um aumento significativo nos afastamentos concedidos pelo Instituto Nacional do Seguro Social (INSS), impulsionado por fatores como precarização do ambiente de trabalho e alta incidência de acidentes. De acordo com o Anuário Estatístico da Previdência Social - AEPS (2024), em 2023 houve um aumento de 14,4% na concessão de benefícios pelo INSS em relação a 2022. Foram concedidos 5,964 milhões de benefícios, sendo 5,159 milhões do Regime Geral de Previdência Social (RGPS) e 804,1 mil benefícios assistenciais. No âmbito da construção civil, em 2023 foram notificados 17.523 acidentes para os CNAE de construção de edifícios, construção de obras de arte especiais e construção de rodovias, representando um aumento de 19,2% (AEPS, 2024). A alta incidência de acidentes representa uma das principais causas dos afastamentos, gerando custos diretos com benefícios, além de prejuízos relacionados à perda de produtividade e à necessidade de reorganização das operações.

Segundo a lei 8.213 (1991), a empresa deverá comunicar o acidente de trabalho à Previdência Social por meio da emissão da Comunicação de Acidente do Trabalho (CAT). Embora a base de dados referente aos acidentes da CAT esteja disponível, as análises realizadas até o momento são superficiais, não possibilitando a formulação de estratégias e tomada de decisões no âmbito operacional, por parte das empresas. Além disso, como apontado por Borges, Vilaça e Laurindo (2021), em muitos casos a CAT não é emitida, o que dificulta ainda mais a análise, pois a quantidade de dados registrada não reflete o total real de acidentes, podendo ocultar casos importantes e comprometer a precisão das avaliações.

Dessa forma, tendo em vista que os acidentes do trabalho resultam da interação de múltiplos fatores (Manu *et al*., 2012) e volume de dados resultantes das comunicações de acidentes, o aprendizado de máquina (AM) tem se destacado como uma técnica poderosa para a análise de dados na indústria da construção civil (Zhu *et al*., 2021). Ao utilizar essas técnicas, é possível identificar padrões mesmo em conjuntos de dados reduzidos, detectar tendências e prever, com maior precisão, a probabilidade de afastamentos. Diferentemente dos métodos tradicionais, que requerem ajustes manuais, o AM é capaz de aprender e melhorar de maneira contínua, não dependendo exclusivamente do número absoluto de dados, mas da qualidade e das características intrínsecas dos indicadores principais disponíveis para uma organização ou projeto de construção, quaisquer que sejam eles (Jafari *et al*., 2019). Além disso, o AM pode identificar interações entre os fatores, revelando correlações e contribuições das variáveis que poderiam passar despercebidas em métodos convencionais, que tendem a analisar essas variáveis isoladamente (Choi *et al*., 2020; Lee *et al.*, 2020). Sob essa ótica, o desenvolvimento de modelos de aprendizado de máquina se mostra imprescindível para uma análise mais robusta dos fatores que contribuem para a ocorrência de acidentes, sejam eles fatais ou não.

Nesse contexto, a literatura aponta que a maioria das pesquisas nessa área está concentrada em países como Coreia do Sul (Choi et al., 2020; Kang, Koo e Ryu, 2022; Kang e Ryu, 2019; Lee et al., 2020), Turquia (Koc, Ekmekçioğlu e Gurgun, 2021, 2023a, 2023b, 2023c; Toptancı et al., 2023), e China (Zhu et al., 2021), indicando uma lacuna significativa de estudos de lesões com afastamentos. Ademais, é latente a necessidade de estudos relacionados ao desenvolvimento e aplicação de modelos de aprendizado de máquina no Brasil. Assim, este estudo tem como objetivos analisar as variáveis presentes na base de dados proveniente da CAT e construir um modelo para a classificação binária de afastamentos de trabalhadores devido a acidentes de trabalho, avaliando sua performance por meio de métricas como acurácia, precisão, *recall* e *F1-score*.

2 REVISÃO BIBLIOGRÁFICA

A literatura concentra diversos estudos voltados à previsão de acidentes fatais na construção civil, com ênfase na aplicação de técnicas de aprendizado de máquina. Choi *et al*. (2020) usaram aprendizado supervisionado para avaliar riscos, destacando variáveis como falhas em equipamentos de segurança. Kang e Ryu (2019) aplicaram algoritmos de clusterização para classificar acidentes com características semelhantes, facilitando ações preventivas. Poh *et al*. (2018) utilizaram modelos como árvores de decisão e floresta aleatória, com o método Boruta para seleção robusta de variáveis, aprimorando a interpretação dos riscos.

Enquanto isso, Jafari *et al*. (2019) propuseram o uso de indicadores proativos de desempenho de segurança, contribuindo para uma gestão mais adaptável. Baker *et al.* (2020) associaram a gravidade das

lesões à falta de treinamento e à exposição a riscos físicos. Koc, Ekmekcioğlu e Gurgun (2023a) investigaram partes do corpo mais suscetíveis a lesões, propondo medidas de mitigação e políticas mais direcionadas. Em seguida, os mesmos autores (2023b) apresentaram uma estrutura baseada em dados para prever acidentes fatais na Turquia, considerando fatores como clima, tempo e tipo de projeto. Ainda, Koc, Ekmekcioğlu e Gurgun (2023c) analisaram variáveis temporais - como ano, mês, dia da semana e hora do acidente -, identificando que o ano e o horário foram os principais preditores de fatalidade em acidentes.

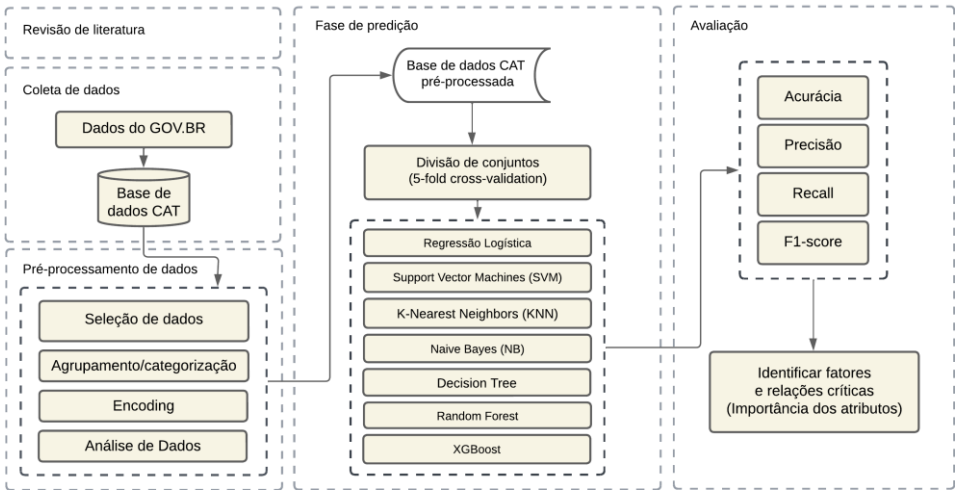
Estudos mais recentes ampliaram o foco para os afastamentos e suas consequências. Koc *et al.* (2021) usaram aprendizado de máquina para prever incapacidades permanentes, destacando os dias de trabalho perdidos como variável-chave. Kang, Koo e Ryu (2022) aplicaram *random forest* para classificar a gravidade das lesões, associando-a ao uso de Equipamento de Proteção Individual (EPI) e à experiência do trabalhador. Tözer, Gürcanlı e Yarkıner (2022) empregaram regressão para estimar perdas causadas por quedas de andaimes, considerando variáveis como altura da queda e idade. Toptancı *et al.* (2023) utilizaram regressão logística para prever a gravidade de acidentes com base em fatores como idade, escolaridade e tipo de ambiente. Por fim, Kim *et al.* (2024) propuseram um modelo de redes neurais profundas para estimar o tempo de recuperação dos trabalhadores, considerando a escala dos projetos.

Em contraste com esses estudos, o presente trabalho se concentra especificamente em dados do Brasil, focando na previsão dos afastamentos de trabalhadores no setor da construção civil. O modelo desenvolvido visa indicar, de forma binária, a ocorrência ou não de afastamentos, utilizando uma abordagem que visa melhorar a gestão da segurança do trabalho, considerando as características únicas do contexto do *dataset* brasileiro.

3 METODOLOGIA

A metodologia utilizada neste trabalho consiste em cinco etapas: (i) Revisão da literatura, (ii) Coleta de dados, (iii) Pré-processamento de dados, (iv) Predição, (v) Análise de dados e Avaliação de desempenho, conforme delineamento apresentado na Figura 1.

Figura 1: *Framework* de processamento dos dados na pesquisa.



Fonte: Os autores.

3.1 Coleta de dados

Esta pesquisa utiliza o conjunto de dados de acidentes do Ministério do Trabalho e Emprego (Brasil 2023), obtido pelo portal de dados abertos do Governo Federal. O período abrangente da pesquisa é de 2018 a 2023, compreendendo seis anos e tendo uma dimensão de 2.159.464 acidentes. O conjunto abrange todos os tipos de trabalhadores e está desbalanceado, tendo uma proporção de 5% de acidentes sem afastamento com lesão. Para que fosse possível manipular os dados, utilizou-se a linguagem *Python* e bibliotecas correlatas como *Pandas*, para manipulação de dados tabulares, *NumPy* para operações numéricas, *Matplotlib/Seaborn* para visualização de dados e *Scikit-Learn* para obter os algoritmos de aprendizado de máquina.

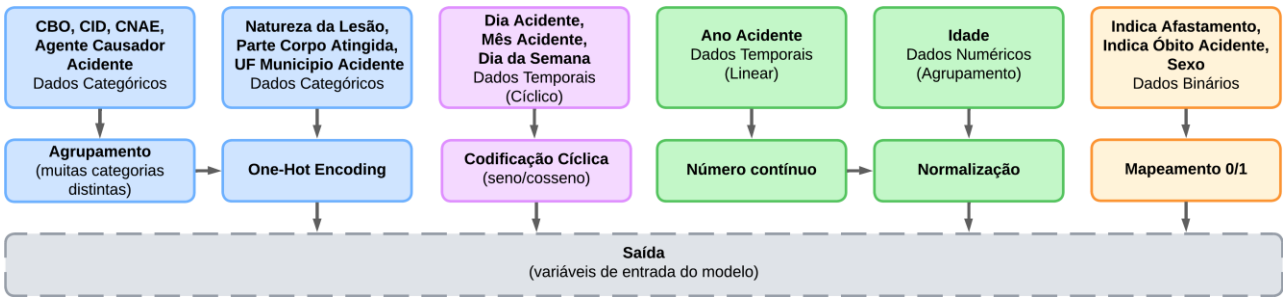
3.2 Pré-processamento

Primeiramente, foi decidido que seriam selecionadas apenas ocorrências da construção civil (abrangendo as divisões 41 a 43 da seção 'F' da CNAE - Classificação Nacional de Atividades Econômicas) e os profissionais empregados, fazendo com que o conjunto de dados resulte em 142.603 linhas (representando ocorrências de acidentes) e 14 colunas (representando características dos acidentes). Para mitigar o desbalanceamento entre ocorrências com e sem afastamento, foi aplicado *undersampling* nas instâncias sem afastamento, ajustando-as para uma proporção de 60% sem afastamento e 40% com afastamento.

Além disso, um dos maiores desafios no pré-processamento foi o grande número de colunas categóricas, o que tornou o processo de codificação mais complexo. Para isso, textos semelhantes ou fragmentados foram agrupados em categorias similares, a fim de reduzir o número possível de categorias de ocupações. Posteriormente, aplicou-se a codificação *One-Hot Encoder* para transformar as categorias em atributos binários, em que 1 representava a presença ou confirmação da variável e 0, sua ausência, escolhido por não impor ordem às categorias, ao contrário de outras técnicas como *Label Encoder.* Também, foi realizada uma análise exploratória de dados para determinar relações entre as variáveis.

Após isso, o conjunto de dados final consiste em **28.100 instâncias** e **14 atributos**, incluindo variáveis categóricas e binárias como variáveis de entrada, conforme apresentado na Figura 1. Assim, o conjunto de dados contém informações principais relacionadas a atributos baseados no trabalhador ("gênero", "idade" e "ocupação"), atributos baseados na organização ("subsetor da construção" e "região do acidente") e atributos relacionados ao tempo ("dia do mês", "mês" e "ano"), além de atributos relacionados ao acidente, como "mecanismo do acidente", "natureza do acidente", "Classificação Internacional de Doenças – CID", "tipo de lesão" e a ocorrência ou não de "lesões com afastamento". Um atributo adicional representando o "dia da semana" foi incluído, **totalizando 15 características**. A Figura 2 lista as variáveis usadas nos modelos.

Figura 2: Variáveis de entrada do modelo e seus pré-processamentos respectivos.



Fonte: Os autores.

3.3 Etapa de predição

Para evitar viés, os dados foram separados em conjuntos distintos, impedindo a mistura entre eles. Os dados para o modelo de previsão de acidentes foram divididos em dois conjuntos: treinamento (80%) e teste (20%). No treinamento, aplicou-se a validação cruzada com 5 divisões (5-*fold cross-validation*), uma técnica que reduz o risco de *overfitting* e aumenta a confiabilidade do modelo. Essa abordagem utiliza diferentes divisões internas e calcula a média dos resultados de cada ciclo de treinamento.

Para este estudo, foram escolhidos algoritmos de diferentes abordagens para garantir diversidade analítica. A Regressão Logística, como em Choi *et al*. (2020), é eficaz para prever eventos binários. O S*upport Vector Machine* (SVM) é adequado para separar classes em dados complexos. O K-nearest neighbors (KNN), embora simples, é intuitivo e útil em classificações multiclasse. *Decision Tree* organiza variáveis de forma hierárquica, enquanto o *Random Forest*, utilizado por Kang e Ryu (2019), melhora a precisão combinando árvores. O *Naive Bayes* é eficiente com dados de alta dimensionalidade. O *XGBoost*, aplicado por Baker *et al.* (2020), corrige erros iterativamente, sendo eficaz em problemas preditivos mais complexos.

3.4 Avaliação

Para medir o desempenho do modelo, identificando possíveis erros e permitindo ajustes, como a melhoria de parâmetros ou a escolha de outro algoritmo, é necessário utilizar métricas para os conjuntos. Isso é

possível criando uma tabela de contingência, utilizando métricas comuns como falsos positivos (FP, previsões incorretas da presença da classe), falsos negativos (FN, previsões incorretas da ausência da classe), verdadeiros positivos (VP, previsões corretas da presença da classe) e verdadeiros negativos (VN, previsões corretas da ausência da classe). Ao combinar essas métricas, é possível calcular indicadores importantes, como a **acurácia** (Equação 1), **precisão** (Equação 2), ***recall*** (Equação 3) e ***F1-score*** (Equação 4), utilizados para fornecer uma visão mais completa sobre a eficácia do modelo em termos de balanceamento entre erros e acertos (Choi *et al.*, 2020; Zhu *et al.*, 2021).

Acurácia = $\frac{VP+VN}{VP+VN+FP+FN}$ (Eq.1)

Precisão = $\frac{VP}{VP+FP}$ (Eq.2)

Recall = $\frac{VP}{VP+FN}$ (Eq.3)

F1-score = $\frac{2 \cdot precisao \cdot recall}{precisao+recall}$ (Eq.4)

Os algoritmos baseados em árvores, como o *Random Forest* e o *XGBoost*, incluem métodos intrínsecos que permitem calcular e visualizar a importância das variáveis no modelo final. Esses métodos nativos oferecem uma visão detalhada de quais variáveis têm maior impacto nas previsões. Quanto maior a importância de uma variável, mais significativa foi sua contribuição para dividir os dados e reduzir a incerteza dentro das árvores de decisão, desempenhando um papel crucial na capacidade do modelo de diferenciar entre as classes. No entanto, é importante ressaltar que essa importância não implica necessariamente em uma relação causal (Edwald et al., 2024).

4 RESULTADOS E DISCUSSÕES

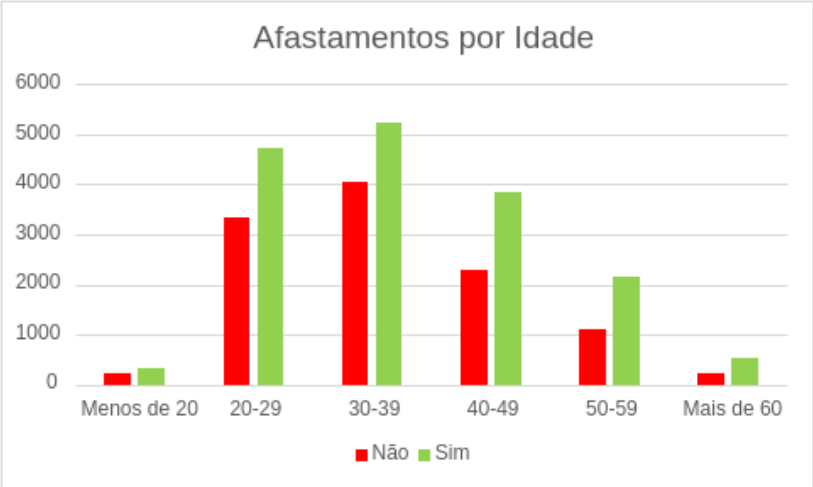
4.1 Visão geral dos dados

Esta subseção mostra como as variáveis se distribuem conforme o afastamento. Analisando os dados em relação à faixa etária (Figura 3), observa-se que a maior quantidade de afastamentos ocorreu entre a faixa de 30-39 anos (5.253 casos), seguida por 20-29 anos (4.740 casos), indicando que adultos jovens e trabalhadores de meia-idade, responsáveis por grande parte da força de trabalho, são os mais impactados. Apesar da faixa "Mais de 60 anos" apresentar um número absoluto menor de casos, isso sugere uma maior gravidade nos acidentes ou menor capacidade de retorno imediato. Já os menores de 20 anos têm uma proporção menor de afastamentos.

A análise da parte do corpo atingida revela que as lesões se concentram majoritariamente nos membros superiores (12.169 casos) e inferiores (8.062 casos), totalizando 72% dos casos, refletindo os riscos associados às atividades locomotoras. Além disso, em ambos, há uma predominância de afastamentos, o que sugere que essas lesões possuem um impacto funcional significativo, frequentemente incapacitante, exigindo períodos mais longos de recuperação e afastamento do trabalho. Outras áreas, como pescoço e tronco (2.923 casos) e cabeça/face (3.926 casos), apresentam números menores, mas ainda significativos, sugerindo riscos relacionados a quedas ou impactos. Por fim, casos envolvendo sistemas, aparelhos e partes múltiplas são menos frequentes, totalizando 1.020 registros; nessas categorias, há uma proporção inferior de afastamentos por casos totais, podendo indicar que as lesões não atendem aos critérios para afastamento prolongado ou até mesmo resultam em mortes (os dados revelam uma taxa de fatalidade superior a 90% nesses casos.)

Os dados sobre afastamentos por gênero revelam uma diferença significativa nas taxas entre homens e mulheres. Enquanto a proporção de mulheres que não se afastaram (cerca de 1%) é bem maior do que a das que se afastaram (aproximadamente 2%), o cenário é oposto entre os homens, com a maioria se afastando e uma proporção considerável permanecendo ativa.

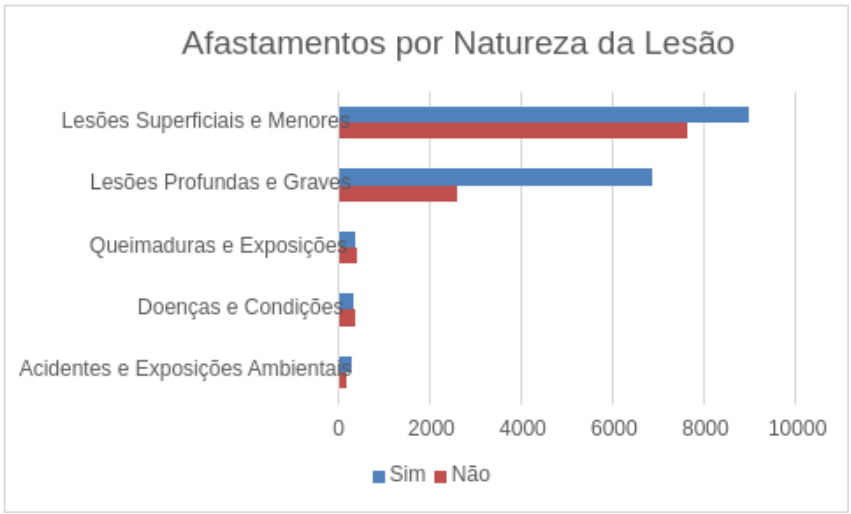
Figura 3: Gráfico de lesões com afastamentos por idade.



Fonte: Os autores.

A base de dados revela que ao comparar com a natureza da lesão (Figura 4), lesões superficiais e menores são as mais frequentes (59% do total), seguidas por lesões profundas e graves (34%). As lesões superficiais são comuns, enquanto as lesões profundas, embora menos frequentes, exigem afastamentos mais longos devido à gravidade. Por outro lado, doenças e condições, queimaduras e exposições, e acidentes ambientais são menos frequentes, com proporções mais equilibradas entre afastamentos.

Figura 4: Gráfico de lesões com afastamentos por natureza da lesão.



Fonte: Os autores.

4.2 Desempenho dos modelos

Após o pré-processamento e análise das variáveis que impactam no modelo, foram escolhidos sete algoritmos para treinar as variáveis a partir dos conjuntos de validação e teste. A Tabela 1 apresenta o desempenho dos algoritmos no treinamento.

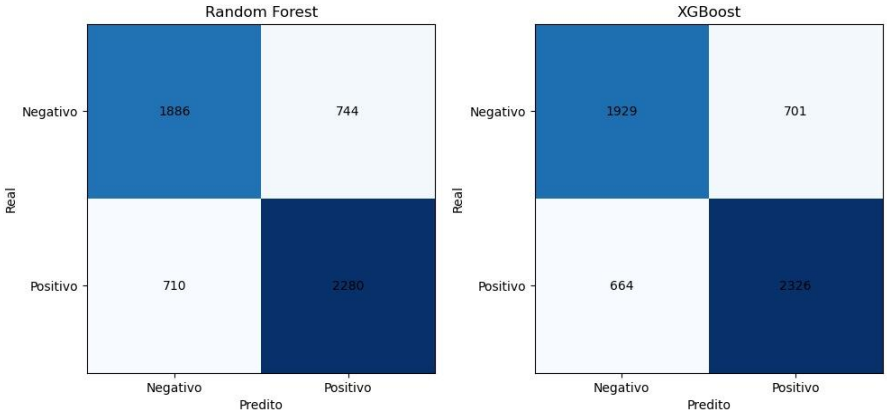
Tabela 1: Resultados das métricas obtidas para cada um dos algoritmos no conjunto de teste.

	ACURÁCIA	PRECISÃO	RECALL	F1-SCORE
Random Forest	**0,775667**	**0,768219**	**0,776456**	**0,770386**
XGBoost	**0,774422**	**0,767192**	**0,775549**	**0,769228**
SVM	0,752758	0,742451	0,741499	0,741932
Decision Tree	0,731717	0,720525	0,720107	0,720265
Regressão Logística	0,724733	0,712978	0,709333	0,710868
KNN	0,676735	0,662082	0,659264	0,660387
Naive Bayes	0,500044	0,648173	0,574578	0,456677

Fonte: Os autores.

Os resultados indicam que os algoritmos baseados em *ensemble*[2], como *Random Forest* e *XGBoost*, possuem as melhores métricas. Isso pode ser atribuído à natureza binária dos dados após a codificação, o que favorece a separação clara nesses modelos baseados em árvores. Por outro lado, algoritmos que são sensíveis à linearidade e às correlações independentes, como o *Naive Bayes*, apresentaram o pior desempenho, evidenciando sua inadequação para o conjunto de dados analisado. Esses resultados ressaltam que o formato dos dados e as características das variáveis influenciam diretamente a performance dos modelos. A Figura 5 contrasta as matrizes de confusão dos dois melhores modelos: o *XGBoost* (à direita) apresenta maior *recall*, com 2.326 verdadeiros-positivos, ainda que com 701 falsos-positivos. Já o *Random Forest* (à esquerda) demonstra 2.280 verdadeiros-positivos e 744 falsos positivos. No geral, o *XGBoost* apresenta desempenho ligeiramente superior em comparação ao *Random Forest*.

Figura 5: Matrizes de confusão obtidas para o *Random Forest* (à esquerda) e para o *XGBoost* (à direita).



Fonte: Os autores.

4.3 Análise da importância dos atributos

É importante notar que os algoritmos baseados em árvore fornecem visualizações dos atributos mais importantes no cálculo de pesos e sua utilidade no modelo como um todo, não necessariamente uma causalidade direta. Aqui, os algoritmos que tiveram o melhor desempenho, *Random Forest* e *XGBoost*, são obtidos e os sete atributos mais importantes selecionados, como apresentado na Tabela 2. Ambos os algoritmos identificaram o 'Ano do Acidente' como o fator com maior peso para a classificação. Embora esse atributo não cause diretamente o afastamento, ele funciona como um indicador temporal que reflete mudanças estruturais e contextuais ao longo dos anos, capturando tendências que influenciam as chances de afastamento.

[2] *Ensemble* = Método de aprendizado de máquina que combina múltiplos modelos para melhorar a precisão e a robustez da predição.

Tabela 2: Resultados das importâncias das características calculadas pelos modelos.

RANDOM FOREST		XGBOOST	
ATRIBUTO	IMPORTÂNCIA	ATRIBUTO	IMPORTÂNCIA
Ano Acidente	0.299381	Ano Acidente	0.317961
Dia do Mês	0.103612	Natureza da Lesão: Lesões Profundas e Graves	0.065133
Mes Acidente	0.082914	Agente Causador do Acidente: Seres Vivos e Animais	0.028886
Dia da Semana	0.065241	CNAE2.0 Empregador: Obras de infraestrutura	0.026722
CNAE2.0 Empregador: Obras de infraestrutura	0.025642	Indica Óbito Acidente	0.026629
Natureza da Lesão: Lesões Profundas e Graves	0.020686	UF do Município do Acidente: Sudeste	0.023695
Idade: 30-39	0.018692	CNAE2.0 Empregador: Serviços especializados para construção	0.020293

Fonte: Os autores.

Isso é corroborado pelos resultados de Koc *et al*. (2023c), que também identificaram variáveis temporais como o ano e o horário do acidente e dados apresentados no relatório do *Smartlab*, que indica uma variação considerável nas taxas de acidentes ao longo dos anos, com um aumento substancial nos afastamentos, especialmente após eventos como a pandemia de COVID-19. Outros atributos de destaque, presentes em ambos os modelos, incluem o 'Dia do Acidente', mesmo que não haja uma causa direta clara associada, pode revelar padrões sazonais e comportamentais relacionados à rotina de trabalho. Estudos apontam, por exemplo, pior desempenho nas segundas-feiras - o chamado “efeito segunda-feira” (*Monday effect*) (Fontaneda *et al.*, 2021) -, o que reforça a relevância das variáveis temporais no modelo. Também se destacam o 'CNAE 2.0 Empregador: Obras de Infraestrutura', indicando maior impacto em acidentes relacionados a obras de infraestrutura; e a 'Natureza da Lesão: Lesões Profundas e Graves', que apresentou alta relevância. A consistência na priorização desses atributos pelos dois algoritmos evidencia a robustez das relações extraídas dos dados. No entanto, a maior sensibilidade do *XGBoost* em capturar variações sutis pode justificar pequenas diferenças nos pesos atribuídos aos atributos de menor relevância.

4.4 **Discussão**

Ao examinar os atributos do banco de dados da CAT, observa-se que, embora alguns dados coincidam com os encontrados na literatura, como a distribuição etária, outros aspectos não estão tão bem representados. Alguns modelos preditivos encontrados na literatura sugerem a inclusão de variáveis mais amplas nos modelos de predição de acidentes (Koc *et al.*, 2021; Kang, Koo e Ryu, 2022), uma vez que os acidentes não ocorrem de forma isolada, mas sim como resultado da interação entre fatores individuais.

O conjunto de dados da CAT é limitado, concentrando-se em variáveis como gênero, idade, ocupação, mecanismo do acidente, tipo de lesão, CID e ocorrência de afastamento. Comparando com modelos da literatura, como o de Kang, Koo e Ryu (2022), nota-se ausência de variáveis relacionadas ao contexto do acidente. Esses autores utilizaram, além de atributos semelhantes, informações sobre o projeto e a atividade realizada - como tamanho da empresa, valor do projeto, processo construtivo, ato/condição insegura e dias de trabalho perdidos - o que pode enriquecer a predição e reduzir vieses. Koc *et al*. (2021) também incluíram variáveis como materiais utilizados, tipo de construção, histórico de acidentes e salário diário. Esses dados, além de melhorarem o desempenho dos modelos, aproximam-se de abordagens causais adotadas na literatura. Segundo Tixier *et al*. (2023), a inclusão de variáveis comuns entre diferentes estudos pode aprimorar a performance mesmo quando combinadas a abordagens específicas. A ausência de dados sobre o contexto do projeto e das atividades limita a capacidade do modelo proposto em capturar elementos críticos do ambiente de trabalho e prever riscos com maior precisão.

Além disso, a falta de dados categóricos mais específicos é um ponto crítico, pois a literatura sugere que, para um diagnóstico mais preciso, seria necessário segmentar detalhadamente os dados, com categorias que permitam distinguir diferentes tipos de afastamentos e causas mais específicas, como sugerido por Gong *et al*. (2020). Um exemplo disso, nesse trabalho, é o atributo 'Agente Causador do Acidente', que apresentou mais de 360 categorias diferentes, o que demandou tempo para encontrar uma abordagem adequada de agrupamento e simplificação. Isso implica que, para outras análises, será necessário lidar

com uma diversidade de categorias que exigirá métodos complexos para agrupá-las de maneira coerente, o que dificulta a interpretação.

Outro ponto relevante é a abrangência das variáveis, que são muitas vezes genéricas e não capturam a complexidade dos dados necessários, como, por exemplo, a categoria da data de afastamento. Na base de dados, essa informação frequentemente se limita ao mês e o ano, ou, em alguns casos, sequer é registrada. Dessa forma, não é possível determinar precisamente quando o trabalhador retornou ao trabalho ou se, de fato, retornou. Para aprimorar a qualidade da análise, seria recomendada uma revisão das variáveis presentes na CAT e uma busca por um maior detalhamento, além da consistência dos dados, como indicado por Lee *et al*. (2020), que sugere a inclusão de variáveis segmentadas e menos genéricas. Essa necessidade de refinamento também é mostrada por Zhu *et al.* (2021), que destacaram a limitação da classificação restrita de acidentes devido à escassez de dados, o que pode comprometer a aplicabilidade dos modelos e a precisão das análises.

Comparando com os resultados de Kang, Koo, Ryu (2022), que obtiveram uma acurácia de cerca de 79,0%, os modelos utilizados neste estudo apresentaram variações significativas, embora com desempenho inferior ao de Koc *et al*. (2021), com acurácia de 82,9%, utilizando variáveis como nível educacional, dias de trabalho perdidos, tipo de lesão, partes do corpo lesionadas, materiais utilizados, tipo de construção, número de acidentes anteriores do trabalhador e salário diário. O modelo *random forest* teve o melhor desempenho, com acurácia de 77,5%, seguido pelo *XGBoost* com 77,4%. Esses resultados indicam que, embora a base de dados utilizada tenha limitações, os modelos propostos ainda apresentam um bom desempenho.

5 CONSIDERAÇÕES FINAIS

Este estudo atingiu seu objetivo principal ao analisar as variáveis disponíveis e construir um modelo para classificação binária de afastamentos de trabalhadores na construção civil. Utilizando dados do Ministério do Trabalho e Emprego, aplicou-se uma abordagem estruturada para preparar as informações e realizar previsões sobre afastamentos. O desempenho dos modelos foi avaliado e destacou o potencial do aprendizado de máquina nesse contexto. Contudo, a ausência de dados sobre a duração dos afastamentos limitou a possibilidade de estimativas mais precisas, restringindo a análise à predição binária.

Nesse sentido, é sugerido que estudos futuros explorem a criação ou combinação de variáveis e até mesmo a consulta externa de outras bases de dados para determinar a data exata do retorno do afastamento e desenvolver modelos focados em regressão capazes de estimar o número exato de dias de afastamento, sempre que os dados estiverem disponíveis. A inclusão dessas variáveis pode aprimorar a precisão das previsões e fornecer *insights* valiosos sobre as situações de risco, favorecendo análises mais eficazes no setor da construção civil. Adicionalmente, uma abordagem híbrida, que combine a predição de fatalidade com a estimativa da duração do afastamento em casos não fatais, poderia oferecer uma compreensão mais completa das consequências dos acidentes, ampliando o alcance da análise.

É fundamental uma revisão ou integração das variáveis da CAT, de forma a fornecer informações relevantes à predição de acidentes na construção civil. Visto que a falta de informações explícitas sobre dados pós-acidente também restringiu a análise preditiva. O banco de dados se restringe a eventos já ocorridos, limitando uma abordagem mais preditiva e preventiva. Um incentivo para novas pesquisas pode ser superar essas barreiras ao explorar variáveis adicionais, como fatores individuais, ambientais e organizacionais, contribuindo para análises mais robustas e eficazes na prevenção de acidentes de trabalho. Além disso, pode-se investir em técnicas mais avançadas de aprendizado de máquina, utilizando redes neurais e aprendizado profundo para obter melhores relações entre os dados, de maneira mais complexa e robusta.

REFERÊNCIAS

BAKER, Henrietta; HALLOWELL, Matthew R.; TIXIER, Antoine J.-P. AI-based prediction of independent construction safety outcomes from universal attributes. **Automation in Construction**, [s.l.], v. 118, 2020, p. 103146. ISSN 0926-5805. doi: https://doi.org/10.1016/j.autcon.2020.103146.

BORGES, Nathália de Faria; VILAÇA, Isabela Pessanha; LAURINDO, Quézia Manuela Gonçalves. Acidentes do trabalho e cultura de segurança no setor da construção civil. **Revista Perspectivas Online:**

Exatas & Engenharia, v. 11, n. 33, p. 19-33, out. 2021. ISSN 2236-885X (Online). doi: https://doi.org/10.25242/885X113320212353.

BRASIL. Ministério da Previdência Social. **Anuário Estatístico da Previdência Social – 2023.** Brasília, DF, 2024. Disponível em: https://www.gov.br/previdencia/pt-br/noticias/2024/dezembro/anuario-estatistico-da-previdencia-social-2023-ja-esta-disponivel-para-consulta. Acesso em: 05 jan. 2025.

BRASIL. Ministério do Trabalho e Emprego. **Relatório de Transparência Salarial de Mulheres e Homens.** Brasília, DF, 2024. Disponível em: https://www.gov.br/trabalho-e-emprego/pt-br/noticias-e-conteudo/2024/Fevereiro/mte-e-mmulheres-tiram-duvidas-sobre-o-relatorio-de-transparencia-salarial/copy_of_RelatriodeIgualdadesalarialdeMulhereseHomens_09022024.pptx. Acesso em: 10 fev. 2025.

CHOI, Jongko; GU, Bonsung; CHIN, Sangyoon; LEE, Jong-Seok. Machine learning predictive model based on national data for fatal accidents of construction workers. **Automation in Construction,** [s.l.], v. 110, 2020. ISSN 0926-5805. doi: https://doi.org/10.1016/j.autcon.2019.102974.

EWALD, Fiona Katharina; BOTHMANN, Ludwig; WRIGHT, Marvin N.; BISCHL, Bernd; CASALICCHIO, Giuseppe; KÖNIG, Gunnar. A guide to feature importance methods for scientific inference. In: LONGO, L.; LAPUSCHKIN, S.; SEIFERT, C. (eds.). Explainable Artificial Intelligence. xAI 2024. **Communications in Computer and Information Science,** v. 2154. Cham: Springer, 2024. ISSN 1865-0929. doi: https://doi.org/10.48550/arXiv.2404.12862.

FONTANEDA, Ignacio; CAMINO LÓPEZ, Miguel A.; GONZÁLEZ ALCÁNTARA, Oscar J.; GREINER, Birgit A. The “Weekday Effect”: A Decrease in Occupational Accidents from Monday to Friday—An Extension of the “Monday Effect”. BioMed Research International, v. 2024, p. 1-12, 2024. ISSN 2314-6133. doi: https://doi.org/10.1155/2024/4792081.

GOMIDES, Luciana de Melo; ABREU, Mery Natali Silva; ASSUNÇÃO, Ada Ávila. Desigualdades ocupacionais e diferenças de gênero: acidentes de trabalho, Brasil, 2019. **Revista de Saúde Pública,** v. 58, p. 13, 2024. ISSN 1518-8787. doi: http://dx.doi.org/10.11606/s1518-8787.2024058005342.

JAFARI, Parinaz; MOHAMED, Emad; PEREIRA, Estacio; KANG, Shih-Chung; ABOURIZK, Simaan. Leading Safety Indicators: Application of Machine Learning for Safety Performance Measurement. **ISARC - International Symposium on Automation and Robotics in Construction,** 2019. ISSN 2413-5844. doi: https://doi.org/10.22260/ISARC2019/0067.

KANG, Kyung-Su; KOO, Choongwan; RYU, Han-Guk. An interpretable machine learning approach for evaluating the feature importance affecting lost workdays at construction sites. **Journal of Building Engineering,** [s.l.], v. 53, 2022, p. 104534. ISSN 2352-7102. doi: https://doi.org/10.1016/j.jobe.2022.104534.

KANG, Kyungsu; RYU, Hanguk. Predicting types of occupational accidents at construction sites in Korea using random forest model. **Safety Science,** [s.l.], v. 120, 2019, p. 226-236. ISSN 0925-7535. doi: https://doi.org/10.1016/j.ssci.2019.06.034.

KIM, Ji-Myong; ADHIKARI, Manik Das; BAE, Junseo; YUMB, Sang-Guk. A deep neural network algorithm-based approach for predicting recovery period of accidents according to construction scale. **Heliyon,** v. 10, n. 11, e32215, 2024. doi: https://doi.org/10.1016/j.heliyon.2024.e32215.

KOC, Kerim; EKMEKCİOĞLU, Ömer; GURGUN, Asli Pelin. Determining susceptible body parts of construction workers due to occupational injuries using inclusive modelling. **Safety Science,** v. 164, 2023. ISSN 0925-7535. doi: https://doi.org/10.1016/j.ssci.2023.106157.

KOC, Kerim; EKMEKCİOĞLU, Ömer; GURGUN, Asli Pelin. Developing a National Data-Driven Construction Safety Management Framework with Interpretable Fatal Accident Prediction. **Journal of Construction Engineering and Management,** v. 149, n. 4, 2023, p. 04023010. ISSN: 0733-9364. doi: https://doi.org/10.1061/JCEMD4.COENG-12848

KOC, Kerim; EKMEKCIOĞLU, Ömer; GURGUN, Asli Pelin. Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. **Automation in Construction,** [s.l.], v. 131, 2021, p. 103896. ISSN 0926-5805. doi: https://doi.org/10.1016/j.autcon.2021.103896.

KOC, Kerim; EKMEKCIOĞLU, Ömer; GURGUN, Asli Pelin. Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. Engineering, Construction and Architectural Management, v. 30, n. 10, p. 3792–3820, 2023. ISSN 0969-9988. doi: https://doi.org/10.1108/ECAM-04-2022-0305.

LEE, Jae Yun; YOON, Young Geun; OH, Tae Keun; PARK, Seunghee; RYU, Sang Il. A Study on Data Pre-Processing and Accident Prediction Modelling for Occupational Accident Analysis in the Construction Industry. **Applied Sciences,** v. 10, n. 21, p. 7949, 2020. ISSN 2076-3417. doi: https://doi.org/10.3390/app10217949.

MANU, Patrick A.; ANKRAH, Nii A.; PROVERBS, David G.; SURESH, Subashini. Investigating the multi-causal and complex nature of the accident causal influence of construction project features. **Accident Analysis and Prevention,** v. 48, p. 126-133, jan. 2012. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2011.05.008.

POH, Clive Q.X.; UBEYNARAYANA, Chalani Udhyami; GOH, Yang Miang. Safety leading indicators for construction sites: A machine learning approach. **Automation in Construction,** v. 93, p. 375-386, 2018. ISSN 0926-5805. doi: https://doi.org/10.1016/j.autcon.2018.03.022.

SMARTLAB. **Relatório de Análise de Acidentes de Trabalho.** São Paulo, 2024. Disponível em: https://smartlabbr.org/sst/localidade/0?dimensao=covid. Acesso em: 10 fev. 2025.

TIXIER, Antoine J.-P.; HALLOWELL, Matthew R.; TIXIER, Henrietta Baker. Safer Together: Machine Learning Models Trained on Shared Accident Datasets Predict Construction Injuries Better than Company-Specific Models. **arXiv preprint arXiv:2301.03567,** 2023. doi: https://doi.org/10.48550/arXiv.2301.03567.

TOPTANCÍ, Şura; ERGINEL, Nihal; ACAR, Ilgın. Predicting the severity of occupational accidents in the construction industry using standard and regularized logistic regression models. **Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi,** v. 12, n. 3, p. 778-798, 2023. doi: https://doi.org/10.28948/ngumuh.1212385.

TÖZER, Kemal Dirgen; GÜRCANLI, Gürkan Emre; YARKINER, Zalihe. Analysis of workday losses due to falls from scaffoldings in the construction industry. **Journal of Construction Engineering, Management & Innovation,** [s.l.], v. 5, n. 1, p. 15-27, 2022. ISSN 2630-5771. doi: https://doi.org/10.31462/jcemi.2022.01015027.

ZHU, Rongchen; HU, Xiaofeng; HOU, Jiaqi; LI, Xin. Application of machine learning techniques for predicting the consequences of construction accidents in China. **Process Safety and Environmental Protection,** [s.l.], v. 145, 2021, p. 293-302. ISSN 0957-5820. doi: https://doi.org/10.1016/j.psep.2020.08.006.