RBSO Revista Brasileira de Saúde Ocupacional
ISSN: 2317-6369 (online)
https://doi.org/10.1590/2317-6369/31620pt2023v48e4

Artigo de Pesquisa

Uso de técnicas de aprendizado de máquina para classificação de fatores que influenciam a ocorrência de dermatites ocupacionais

Using machine learning techniques to classify factors that influence the occurrence of occupational dermatitis

Ana Caroline Francisco da Rosa[a]
https://orcid.org/0000-0002-2225-8227

Edwin Vladimir Cardoza Galdamez[a]
https://orcid.org/0000-0002-1763-9332

Rodrigo Clemente Thom de Souza[b]
https://orcid.org/0000-0003-2435-8528

Maria das Graças Mota Melo[c]
https://orcid.org/0000-0003-2874-6771

Ana Luiza Castro Fernandes Villarinho[c]
https://orcid.org/0000-0002-4049-6114

Gislaine Camila Lapasini Leal[a]
https://orcid.org/0000-0001-8599-0776

[a]Universidade Estadual de Maringá (UEM), Programa de Pós-Graduação em Engenharia de Produção. Maringá, PR, Brasil.

[b]Universidade Federal do Paraná (UFPR), Campus Avançado de Jandaia do Sul. Jandaia do Sul, PR, Brasil.

[c]Fundação Oswaldo Cruz (Fiocruz), Escola Nacional de Saúde Pública Sergio Arouca (ENSP), Centro de Estudos da Saúde do Trabalhador e Ecologia Humana (CESTEH). Rio de Janeiro, RJ, Brasil.

Contato:
Edwin Vladimir Cardoza Galdamez
E-mail:
evcgaldamez@uem.br

Os autores declaram que o trabalho não foi subvencionado e que não há conflitos de interesses.

Os autores informam que o trabalho não foi apresentado em evento científico.

Resumo

Introdução: realizar a predição de doenças relacionadas ao trabalho é um desafio às organizações e ao poder público. Com as técnicas de aprendizado de máquina (AM), é possível identificar fatores determinantes para a ocorrência de uma doença ocupacional, visando direcionar ações mais efetivas à proteção dos trabalhadores. *Objetivo*: predizer, a partir da comparação de técnicas de AM, os fatores com maior influência para a ocorrência de dermatite ocupacional. *Métodos*: desenvolveu-se um código em linguagem R e uma análise descritiva dos dados e identificaram-se os fatores de influência de acordo com a técnica de AM que demonstrou melhor desempenho. O banco de dados foi disponibilizado pelo Serviço de Dermatologia Ocupacional da Fundação Oswaldo Cruz e contém informações de trabalhadores que apresentaram alterações cutâneas sugestivas de dermatite ocupacional no período de 2000-2014. *Resultados*: as técnicas com melhor desempenho foram: *neural network*, *random forest*, *support vector machine* e *naive Bayes*. As variáveis sexo, escolaridade e profissão foram as mais adequadas para os modelos de previsão de dermatite ocupacional. *Conclusão*: as técnicas de AM possibilitam predizer os fatores que influenciam a segurança e a saúde dos trabalhadores, os parâmetros que subsidiam a implantação de procedimentos e as políticas mais efetivas para prevenir a dermatite ocupacional.

Palavras-chave: doenças ocupacionais; dermatite ocupacional; predição; aprendizado de máquina; saúde do trabalhador.

Abstract

Introduction*: to predict work related diseases is a challenge for organizations and the governmental authorities. By means of machine learning (ML) techniques it is possible to identify factors that determine the occurrence of an occupational disease, aiming at taking more effective actions to protect workers.* Objective*: to predict, by comparing ML techniques, the factors which highly influence the occurrence of occupational dermatitis.* Methods*: we developed a code in R language and a descriptive analysis of the data and identified the influence factors according to the ML technique that presented the best performance. The database was made available by the Occupational Dermatology Service of Oswaldo Cruz Foundation and assembles information of the workers who experienced cutaneous alterations suggestive of occupational dermatitis between 2000-2014.* Results*: the techniques which presented the best performance were: neural network, random forest, support vector machine, and naive Bayes. Sex, schooling, and profession were the most adequate variables for the occupational dermatitis prediction models.* Conclusion*: ML techniques allowed to predict the factors that influence the workers' safety and health, as well as the parameters that subsidize the procedures implementation, and the most effective policies to prevent occupational dermatitis.*

Keywords: *occupational diseases; dermatitis, occupational; forecasting; machine learning; occupational health.*

Introdução

As atividades de gestão e prevenção de acidentes em Segurança e Saúde no Trabalho (SST) direcionam-se para a prevenção de acidentes de trabalho e doenças ocupacionais, assim como para a mitigação de tais eventos. Para que ocorra essa prevenção, os profissionais de SST atuam no desenvolvimento de ferramentas, mecanismos e estratégias com o propósito de reduzir a ocorrência de eventos que possam mostrar-se danosos às pessoas e à propriedade, pela inserção de barreiras e padronização dos procedimentos operacionais[1]. Com o intuito de mitigar os acidentes de trabalho e as doenças ocupacionais, a área de SST emprega seus esforços na implementação do programa de gerenciamento de riscos, manutenção e calibração dos equipamentos de proteção coletiva, fornecimento dos equipamentos de proteção individual, promoção de treinamentos e capacitações, além de análise e correção dos desvios identificados.

A área de SST deve acompanhar o desenvolvimento tecnológico da indústria, de forma a alterar os métodos e a operacionalização para a implementação de soluções efetivas e que garantem a saúde dos trabalhadores[2]. Os países ditos desenvolvidos, impulsionados pelos avanços tecnológicos e pelas mudanças na legislação, aprimoraram suas atividades de gestão em SST para ações cada vez mais proativas, antecipando-se aos perigos e os controlando ativamente, realidade ainda não alcançada por grande parte dos países em desenvolvimento[3]. Tarefas direcionadas à proteção da saúde dos trabalhadores carecem de gerenciamento, similarmente às demais funções do negócio[4], sendo assim, suas métricas necessitam de acompanhamento contínuo tanto em termos descritivos (para a análise dos eventos registrados) quanto em termos preditivos (para antecipar possíveis cenários futuros).

O uso de técnicas de aprendizado de máquina (AM) para a predição de eventos em saúde vem crescendo em função da quantidade de dados que as organizações produzem, dos novos algoritmos e ferramentas desenvolvidos pela academia e, também, por conta da digitalização dos serviços de saúde (diagnósticos digitais, resultados de exames armazenados em bancos de dados etc.)[5]. O AM pode ser entendido como a intersecção entre a ciência da computação, a engenharia e a estatística que busca o entendimento do processo no qual as informações estão inseridas para a identificação de padrões, construindo, assim, aproximações úteis e possibilitando a elaboração de previsões[6]. Para realizar tais previsões, supõe-se que o futuro não seja suficientemente distinto do passado (quando os dados foram coletados), assegurando sua confiabilidade[7].

Na literatura, observa-se um crescimento da pesquisa sobre o aprendizado de máquina e SST, conforme trabalho realizado por Zhao et al.[8], que buscou predizer, por meio de quatro técnicas de mineração, as perdas auditivas de trabalhadores expostos a ruídos industriais não gaussianos. Outra pesquisa que usou uma técnica de AM (*clustering*) para entender melhor padrões relacionados com a saúde ocupacional foi a de Saâdaoui et al.[9]. Na pesquisa, os autores modelaram um conjunto de dados composto por observações quantitativas e qualitativas de 813 trabalhadores, com o propósito de estimar o impacto de fatores como dor, estresse, lesões por esforço repetitivo, índice de massa, trabalho noturno etc., no estado de saúde dos trabalhadores (distúrbios na saúde).

O AM também vem sendo aplicado para determinar a probabilidade de ocorrer acidentes de trabalho com mineradores. A pesquisa realizada por Palei e Das[10] concluiu que a largura de galerias, o tipo de proteção utilizada nos telhados e a profundidade dos espaços confinados contribuem para a gravidade dos acidentes de trabalho.

O objetivo deste estudo é predizer, a partir da comparação de técnicas de AM, os fatores com maior influência para a ocorrência de dermatite ocupacional. Trata-se, portanto, de uma análise conduzida a partir da comparação de técnicas de AM supervisionadas e aplicadas em um banco de dados de trabalhadores que buscaram atendimento especializado na Fundação Oswaldo Cruz (Fiocruz) porque apresentaram alterações sugestivas de doenças ocupacionais relativas à pele. Os fatores considerados para a construção do modelo foram: sexo, idade, etnia, escolaridade, profissão, atopia (se a doença era preexistente no paciente ou na família), síndrome da pele excitada (SPE) e se se tratava de doença ocupacional (se o diagnóstico foi oriundo de uma doença do trabalho ou não).

Aprendizado supervisionado e técnicas de aprendizagem

O aprendizado supervisionado é caracterizado a partir de um conjunto de dados que possui uma variável-alvo pré-definida, e os registros são categorizados em relação a essa variável[11]. Existem algumas etapas executadas pelo aprendizado supervisionado após a seleção do algoritmo, que são o treinamento do conjunto de dados e a validação ou o teste para posterior avaliação do modelo[12]. Dessa forma, na fase de treino a partir do conjunto de dados, acontece a formulação dos hiperparâmetros; enquanto, na de validação, o algoritmo treinado é avaliado por meio dos dados de validação, sendo possível ajustar os parâmetros, caso o desempenho não seja suficientemente satisfatório durante a validação[13].

A tarefa de classificação consiste em uma das mais usadas no âmbito da SST, pois objetiva identificar a classe que determinado indivíduo integra[8,14]. Assim sendo, o modelo realiza a análise do conjunto de registros fornecidos – conjunto esse em que cada registro comporta a indicação da classe pertencente – e aprende com esses registros para que, na fase subsequente, classifique automaticamente um novo registro[15]. No setor da construção civil, a classificação foi empregada por Kang e Ryu[16] para analisar e ordenar, segundo a importância e a gravidade, os fatores que compõem um acidente de trabalho. Por outro lado, Rubaiyat et al.[17] utilizaram-se da classificação para ordenar os trabalhadores que exerciam suas funções seguramente, a partir das imagens do circuito de monitoramento.

A tarefa de regressão é similar ao processo de classificação, com a diferença de que, na regressão, os valores a que se recorre são contínuos e, na classificação, têm-se grupos discretos, sendo possível determinar o valor de uma variável a partir das informações das demais[18]. Dessa forma, conforme Wuest et al.[7], com a regressão é possível ordenar os indivíduos em qualquer valor real entre 0 e 1 e, na classificação, os registros são alocados em classes binárias, para elaborar um modelo de previsão de acidentes para instrutores e turistas durante a operação de um resort, valendo-se de dados do clima, do sistema *Radio Frequency IDentification* (RFID) e das informações administrativas[19].

Algoritmos de classificação e regressão

Os algoritmos ou técnicas de AM constituem "o caminho" para a resolução dos problemas que se fundamenta em formulação matemática e, também, no processo de recuperação e armazenamento das informações[12].

A técnica de *linear discriminant analysis* (LDA) visa encontrar a combinação linear dos recursos ou separar as classes de objetos ao modelar as diferenças entre as classes de dados, caracterizando-se como um algoritmo de classificação[7]. O algoritmo de *support vector machine* (SVM), por sua vez, enquadra-se como uma técnica de classificação e regressão que busca encontrar grupos nos dados por meio das curvas de limites, que são determinadas pela identificação dos principais pontos nos dados nomeados de vetores de suporte. Os vetores de suporte são aqueles em que, por alguma métrica e com alguma separação máxima entre eles, desenha-se uma curva de limite[12,20].

Outras duas técnicas de classificação e de regressão são a *K-nearest neighbors* (KNN) e a *classification and regression tree* (CART). A KNN constitui um modelo que memoriza a observação do treinamento para classificar os dados de teste não vistos, encontrando, assim, agrupamentos nos dados. Também é denominada de aprendizado preguiçoso, uma vez que não aprende nada durante a fase de treinamento, começando a funcionar apenas durante a fase de teste para comparar as observações do teste com as observações de treinamento mais próximas[11,18]. Por outro lado, a técnica CART cria uma árvore binária usando particionamento recursivo binário, caracterizando-se como não linear, pois divide os dados em subconjuntos com base em variáveis independentes disponíveis[11,21,22].

A técnica de regressão logística (LRG) refere-se a um modelo estatístico de uma ampla classe de modelos lineares generalizados (MLGs). O vetor Y nesse tipo de técnica é representado por *n* respostas observadas e é explicado por parâmetros sistemáticos (*p*) – variáveis explicativas – configurados via preditor linear e função de ligação. As variáveis por flexibilização da família exponencial podem assumir diversas distribuições, dentre elas a normal, de Poisson, binomial, gama etc. A LRG permite obter um resultado discreto a partir de um conjunto de variáveis que podem ser contínuas, discretas, dicotômicas ou qualquer mistura delas. A representação gráfica dessa técnica é definida pela curva logística e apresenta os melhores resultados quando as variáveis independentes são contínuas[12,20,23].

O algoritmo de *naive Bayes* (NB) opera sob a suposição de independência condicional de variáveis, apresentando facilidade para modelar sem que necessariamente existam esquemas complicados de estimativa de parâmetros iterativos. Para classificar um evento dessa forma, o NB utiliza a frequência de informações[12,20].

AdaBoost (ADA) é a técnica de *ensamble* mais conhecida e consiste em uma técnica de classificação. Um algoritmo *ensamble* realiza a combinação de distintas técnicas para abrandar suas limitações e, assim, produzir classificadores poderosos. A ideia principal do ADA é focar em instâncias que foram classificadas incorretamente ao realizar o treinamento. Dessa forma, o nível de foco é determinado por um peso atribuído a cada instância no conjunto de treinamento e, na primeira iteração, o mesmo peso é atribuído a todas as instâncias para que, nas iterações subsequentes, os pesos das classificações incorretas sejam aumentados, enquanto o das corretas diminuídos. Sendo assim, as deficiências são identificadas por pontos nos dados de alto peso, e as perdas de função são classificadas como perdas exponenciais[24,25].

O *random forest* (RF) é um algoritmo de classificação que realiza o cálculo da média das árvores de decisão que o compõe, o que minimiza o

componente de variação do modelo, aproximando-o de um modelo ideal. Assim, concentra-se na amostragem de observações e das variáveis de dados de treinamento com o objetivo de desenvolver árvores de decisão independentes e obter votação majoritária para então realizar a classificação[12,26].

A técnica C5.0 tem a função de seleção de atributos relevantes e de reforço. A função de reforço destina-se a gerar vários modelos (ensaios), ao invés de apenas um. O parâmetro *trial* controla o número de vezes em que os ensaios serão gerados. Para cada construção do modelo, é dada mais atenção às regras de classificação com maiores taxas de erros, tentando melhorá-las no próximo ensaio[11,27].

Outras técnicas de classificação são a *mixture discriminant analysis* (MDA) e a *patient rule induction method* (PRIM). A MDA pode ser vista como uma extensão do LDA, na qual as distribuições de classes são modeladas como misturas de distribuições de subclasses, com cada subclasse sendo representada por uma distribuição gaussiana[28,29]. O algoritmo PRIM, por sua vez, pesquisa um conjunto de sub-regiões do espaço de variáveis de entrada, dentro do qual o desempenho da resposta é consideravelmente melhor do que o de todo o domínio de entrada[30,31].

A *neural network* (NN) é constituída por um conjunto de algoritmos projetados para reconhecer padrões, interpretar os dados recebidos por um tipo de percepção de máquina e rotulá-los ou agrupá-los ao reconhecer seus padrões, simulando o comportamento do cérebro humano. Uma NN simples compõe-se de uma camada de entrada, uma camada de saída e, entre elas, uma camada oculta. As camadas, por sua vez, conectam-se por meio de nós e as conexões formam uma rede de nós interconectados nomeada de rede neural. De forma simplista, os dados são inseridos na rede neural por meio de uma camada de entrada, que se comunica com as camadas ocultas, e o processamento acontece nessas camadas através de um sistema de conexões ponderadas. Os nós existentes nas camadas ocultas realizam a combinação dos dados de entrada com um conjunto de coeficientes e atribuem distintos pesos às entradas, que são somadas. Na sequência, a soma passa pela função de ativação de um nó responsável por determinar a extensão que um resultado deve progredir na rede para influenciar o resultado final. Para concluir, as camadas ocultas se ligam à camada de saída[7,12,32].

Métodos

Para a pesquisa, foi utilizado um banco de dados, referente ao período de 2000 a 2014, de um Serviço de Dermatologia especializado em doenças do trabalho, localizado no Rio de Janeiro e pertencente ao acervo do Centro de Estudos da Saúde do Trabalhador e Ecologia Humana da Escola Nacional de Saúde Pública Sérgio Arouca da Fundação Oswaldo Cruz (CESTEH/ENSP/Fiocruz). Esse banco de dados se originou de prontuários médicos que contêm informações sobre o perfil epidemiológico de 616 trabalhadores. A pesquisa também foi submetida para a avaliação do comitê de ética da Universidade Estadual de Maringá (UEM), que concedeu parecer favorável, em 6 de janeiro de 2020, conforme processo nº CAAE 25497719.0.0000.0104.

O software utilizado foi o R, versão 3.6.1, juntamente com o ambiente para desenvolvimento integrado R Studio Desktop, versão 1.2.5019, visto que se trata de uma ferramenta livre que possui sua própria linguagem de programação.

Para a etapa de preparação dos dados, foram excluídos os casos em que não foi possível determinar se a dermatite foi ou não originada do trabalho, que totalizaram 56 indivíduos, e foi verificado se, para as demais classes de dados, os campos haviam sido preenchidos corretamente, para posterior aplicação das etapas do AM. Não foram verificados *missing values* na base de dados e todas as variáveis, com exceção da idade, foram convertidas para variáveis categóricas.

Uma vez estabelecida a classificação do que constitui e do que não constitui uma dermatite ocupacional, foi verificado que as saídas são conhecidas por meio da configuração do enquadramento quanto ao uso do aprendizado supervisionado. As técnicas foram escolhidas considerando a disponibilidade de pacotes previamente existentes na ferramenta, cujos acrônimos são: CART, LDA, SVM, KNN, PRIM, RF, LRG, MDA, C5.0, NB, NN e ADA. Os parâmetros requeridos para cada técnica são distintos, pois cada uma busca resolver o mesmo problema por "caminhos" diferentes. Portanto, para todas as técnicas, fez-se uso dos parâmetros padrão (default).

Para particionamento do banco de dados, 80% das informações foram empregadas para composição da base de treino e os 20% restantes foram utilizados para teste, por ser uma proporção usual em trabalhos científicos[7]. Em relação ao controle do cenário de teste, foi empregado o método de validação cruzada com dez repetições, munido da acuracidade para escolha do melhor modelo. Dessa forma, todas as técnicas passaram pelos mesmos critérios de controle. Os critérios para avaliar o melhor classificador são dados no **Quadro 1**. Neste caso, será escolhido o classificador que apresentar os melhores resultados na maioria das medidas propostas.

Quadro 1 Métricas utilizadas para problemas de duas classes

Medida	*Fórmula*
Erro	$E = \frac{fn + fp}{fp + vp + fn + vn}$
Acuracidade	$A = \frac{vn + vp}{fp + vp + fn + vn} = 1 - E$
Precisão	$P = \frac{vp}{vp + fp}$
Sensitividade	$S = \frac{vp}{vp + fn}$
Especificidade	$Es = \frac{vn}{fp + vn}$
F_1 *Score*	$F1 = \frac{2 * P * R}{P + R}$
Detecção	$D = \frac{vp}{fp + vp + fn + vn}$
Prevalência	$Pr = \frac{fp + vp}{fp + vp + fn + vn}$

Fonte: Adaptado de Callahan e Shah[13].

As medidas constantes do **Quadro 1** originam-se da matriz de confusão que avalia a assertividade das técnicas de mineração no tocante à classificação original do conjunto de dados. Dessa forma, se a técnica de mineração classificou como positiva (classe de interesse), e no conjunto de dados original a classe também era positiva, estabelecem-se os verdadeiros positivos (*vp*), o que também ocorre para os negativos, constituindo, assim, os verdadeiros negativos (*vn*)[33]. Se a classe de interesse for classificada como negativa pela técnica de mineração, tem-se um falso negativo (*fn*), e se a classe que não era de interesse for classificada como positiva, ocorre um falso positivo (*fp*)[34].

Há, ainda, uma outra medida de avaliação da qualidade do ajuste do modelo nomeada de índice kappa, que avalia a reprodutibilidade entre dois conjuntos de dados. Um índice kappa apresenta concordância quase perfeita quando os valores variam de 0,80 a 1; concordância substantiva, de 0,60 a 0,79; e concordância moderada, de 0,40 a 0,59. Caso a concordância esteja entre 0,20 e 0,39, diz-se que ela é leve; de 0 a 0,19, concordância pobre; e, caso o índice kappa seja menor que 0, estabelece-se que há ausência de concordância[34].

Vários pacotes foram utilizados para realizar a comparação das técnicas de AM. O primeiro pacote utilizado foi o readxl, que realiza a importação de planilhas[35]. O pacote com a maior quantidade de funções utilizadas foi o caret, que abrange desde a partição do conjunto de dados, preparação e controle do cenário de teste até o treino e a construção da matriz de confusão[36]. Acerca dos algoritmos de AM, foi utilizada a versão mais recente disponível para cada pacote em junho de 2020. Além disso, os algoritmos não carecem de importação, pois se trata de dependências do caret[d*], porém, para localização dos parâmetros, necessitam de apresentação: *nnet*, *LiblineaR*, *rpart*, *MASS*, *kernlab*, *randomForest*, *mda*, *fastAdaboost*, *C50*, *plyr*, *klaR*, *supervisedPRIM*[36]. A análise gráfica dos dados foi realizada a partir dos resultados gerados pelo pacote *ggplot2*[37,38].

Em seguida, procedeu-se ao método de avaliação dos algoritmos. A metodologia iniciou-se com a chamada dos pacotes que seriam utilizados, tendo como segunda etapa a importação do arquivo no qual se localizavam os dados. O conjunto de dados já havia sido previamente tratado, razão pela qual não foram inclusas etapas de transformação. Como terceira etapa, estabeleceu-se o particionamento do conjunto de dados em treino e teste com base na variável de resposta (ocupacional). Constituindo a quarta etapa, definiu-se o elemento a ser considerado como *vp* no referido conjunto de dados. Na quinta etapa, ocorreu a preparação do cenário de teste indicando que a variável de resposta possui apenas duas soluções possíveis, que seria utilizado o método de validação cruzada com dez repetições e que foi permitido o processamento paralelo. Na sexta etapa, houve a definição dos parâmetros dos modelos, que deve ser repetida para

d Para obter informações sobre os pacotes disponíveis para serem utilizados com o caret, acessar: https://topepo.github.io/caret/available-models.html.

cada técnica estudada. Para que a ferramenta avaliasse o algoritmo, foi necessário identificar o método (nome da função no pacote de origem). Na sétima fase, ocorreu a comparação dos resultados das referidas técnicas de AM. Foi necessário, portanto, referenciar cada técnica e, posteriormente, gerar gráficos e tabelas comparativos para que o tomador de decisão escolha qual(is) a(s) melhor(es) técnica(s) para descrição do conjunto de dados. Escolhida(s) a(s) melhor(es) técnica(s), estabeleceu-se a última etapa, caracterizada pela classificação e avaliação dos fatores de influência na ocorrência da lesão ocupacional.

Resultados e discussão

Doenças ocupacionais são entendidas como perdas para a saúde do trabalhador em virtude do exercício profissional, enquanto os acidentes do trabalho se caracterizam como acontecimentos não planejados capazes de causar danos às pessoas e à propriedade[39]. No ano de 2020, o Brasil registrou 445.814 acidentes de trabalho, sendo 313.575 (70,3%) acidentes típicos, 59.520 (13,4%) acidentes de trajeto e, 30.599 (6,9%), em decorrência de doença do trabalho ou de origem ignorada[40].

Como consequência dessas perdas acarretadas por acidentes de trabalho e doenças ocupacionais, os sistemas de saúde em todo o mundo dedicam parte considerável de seus recursos para o tratamento de lesões que poderiam ser evitadas com ações preventivas das organizações[1]. Dessa forma, o estudo da prevenção de lesões ocupacionais não é de interesse exclusivo das empresas privadas, mas também da sociedade, uma vez que a existência de um trabalhador lesionado implica retardo no avanço dos padrões de vida e, também, no aumento de custos previdenciários[3]. O emprego de AM é uma das alternativas para reduzir proativamente a frequência de doenças e acidentes ocupacionais[5].

Para proporcionar um melhor entendimento a respeito das particularidades do conjunto de dados, é apresentada uma breve caracterização a seguir. Das oito variáveis em estudo, somente a idade se configura como contínua, enquanto as demais são compostas por classes (discretas). Dos 559 trabalhadores, apenas 288 desenvolveram uma doença no ambiente de trabalho, enquanto os 271 remanescentes apresentaram lesões de origem não ocupacional, e somente 40 pessoas foram diagnosticadas com SPE. Quanto ao sexo, a população em questão é composta de 223 homens e 336 mulheres. No que se refere à etnia, 196 pessoas declararam-se brancas, 165 pardas, 111 negras e 87 indivíduos não declararam. A categoria profissão, por sua vez, compôs-se de 27 atividades laborativas, sendo as que apresentaram maior representatividade as profissões relacionadas à limpeza, com 137 trabalhadores, seguidas por pedreiro ou servente, com 54 representantes. Tratando-se da atopia, 309 pessoas não apresentaram histórico de antecedentes no que concerne às doenças de pele, 223 declararam moléstias anteriores na pele, em si ou na família, e 28 trabalhadores não relataram ocorrência prévia do tipo de lesão em questão. Com respeito à escolaridade, dos doze possíveis níveis, apenas quatro concentraram grande parte dos trabalhadores, a saber: nível fundamental incompleto, com 148 trabalhadores; nível médio completo, com 124; grau acadêmico não declarado, com 83; e nível fundamental incompleto, com 79.

Em relação à comparação das técnicas de AM, os resultados coletados referentes às métricas estão dispostos na **Tabela 1**.

Ao analisar a **Tabela 1**, é possível verificar que todas as técnicas apresentaram acuracidade entre 55% e 69,4%, sensitividade entre 49,1% e 80,7% e especificidade entre 50% e 66,7%. O erro médio entre as variáveis foi de 37,5%, e a maior prevalência encontrada foi de 64%, enquanto o índice kappa figurou na grande maioria das técnicas com concordância leve.

Sobre a precisão, a média dos resultados encontrados foi de 0,623, mostrando que somente essa quantia foi corretamente classificada como verdadeira e que a média da sensitividade encontrada se situou em 0,68. O F_1 *Score* apresentou como medida máxima 0,730, para NN, e mínima de 0,549, para C5.0, que é um indicativo de que, com ajustes adicionais dos hiperparâmetros em detrimento da adoção dos parâmetros default, as técnicas apresentam o potencial de alcançar um desempenho superior.

A **Figura 1** apresenta um painel comparativo das técnicas de AM empregadas em relação à curva *receiver operating characteristic* (ROC), ordenadas segundo a mediana da acuracidade.

Como é possível observar na **Figura 1**, cada técnica apresenta um comportamento característico que varia segundo o percentual de acertos nas predições. As técnicas PRIM e KNN foram as únicas nas quais a curva ROC se mostrou, em alguns pontos, inferior à curva diagonal. As melhores técnicas em termos de *area under the curve* (AUC) foram NB, com 71,11%, e RF e NN, com um índice de 70,6% e 70,1%, respectivamente.

A especificidade apresentou grande variabilidade e quantidade superiores de *outliers* em comparação à sensitividade.

Quanto aos fatores de influência no desenvolvimento de uma doença do trabalho relacionada à pele no conjunto de dados estudado, não foi possível determinar a melhor técnica com os parâmetros default. Por se tratar de um caso relacionado a doenças, escolheu-se a técnica que apresentou os maiores índices de sensitividade

(capacidade de acertar o resultado positivo), que foi a NN. De acordo com essa técnica, os fatores de maior influência no desenvolvimento de uma doença ocupacional foram sexo (100%), escolaridade (91,99%), profissão (28,17%), etnia (11,85%), SPE (11,33%) e idade (7,43%). Somente a variável atopia não apresentou importância no desenvolvimento de doenças ocupacionais, conforme apresentado na **Figura 2**.

Tabela 1 Comparação entre técnicas de AM por meio de métricas

Parâmetro	*NB*	*RF*	*NN*	*SVM*	*MDA*	*CART*	*C5.0*	*ADA*	*LRG*	*LDA*	*PRIM*	*KNN*
Acuracidade	0,667	0,667	**0,694**	0,64	0,658	0,676	0,586	0,613	0,55	0,55	0,604	0,595
Kappa	0,331	0,329	**0,383**	0,275	0,316	0,349	0,175	0,225	0,097	0,097	0,201	0,185
Sensitividade	0,737	0,772	**0,807**	0,737	0,649	0,737	0,491	0,614	0,597	0,597	0,737	0,684
Especificidade	0,593	0,556	**0,574**	0,537	0,667	0,611	0,685	0,611	0,5	0,5	0,463	0,5
Precisão	0,656	0,647	**0,667**	0,627	0,673	0,667	0,622	0,625	0,557	0,557	0,592	0,591
F_1 *Score*	0,694	0,704	**0,73**	0,677	0,661	0,7	0,549	0,62	0,576	0,576	0,656	0,634
Detecção	0,378	0,396	**0,414**	0,378	0,333	0,378	0,252	0,315	0,306	0,306	0,378	0,351
Prevalência	0,577	0,613	**0,622**	0,604	0,496	0,568	0,405	0,505	0,55	0,55	0,64	0,595
Erro	0,333	0,333	**0,306**	0,36	0,342	0,324	0,414	0,387	0,451	0,451	0,396	0,405
AUC	0,711	0,706	**0,701**	0,688	0,683	0,665	0,659	0,646	0,617	0,616	0,591	0,581

NB: *naive Bayes*; RF: *random forest*; NN: *neural network*; SVM: *support vector machine*; MDA: *mixture discriminant analysis*; CART: *classification and regression tree*; ADA: *AdaBoost*; LRG: regressão logística; LDA: *linear discriminant analysis*; PRIM: *patient rule induction method*; KNN: *K-nearest neighbors*; AUC: *area under the curve*.

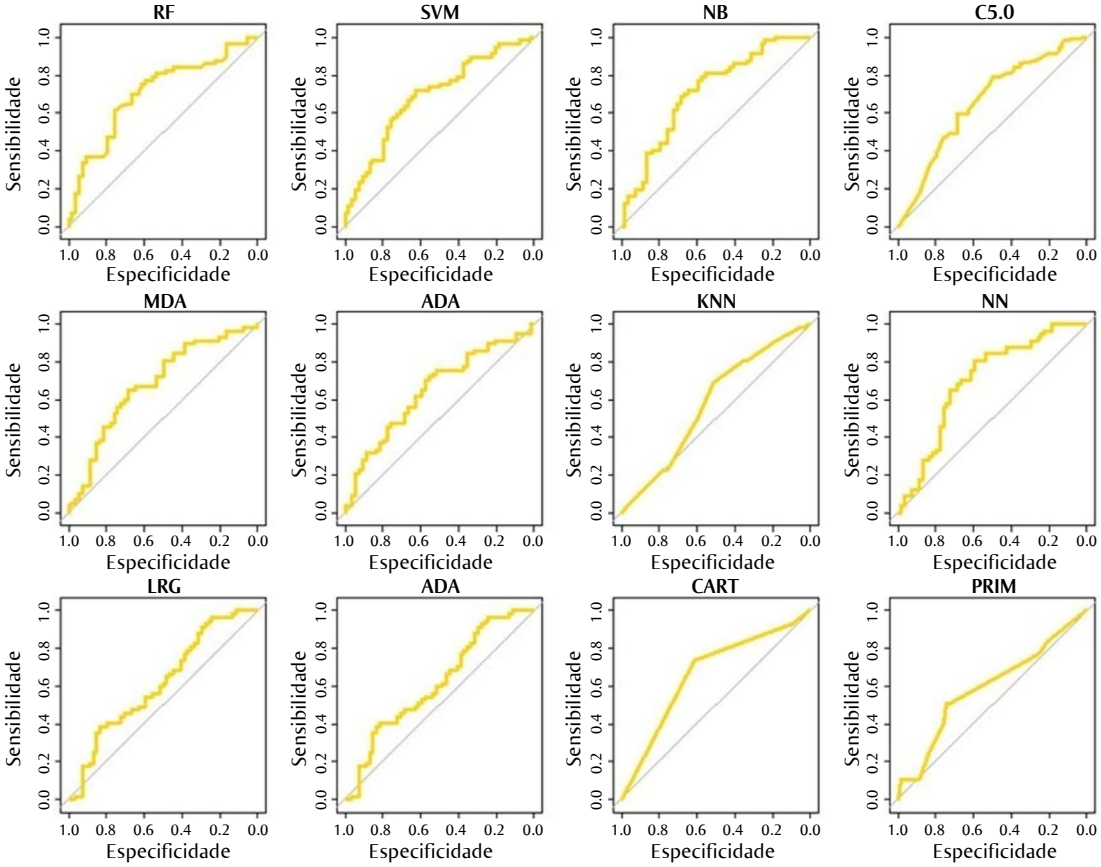

Figura 1 Comparação entre as curvas ROC

RF: *random forest*; SVM: *support vector machine*; NB: *naive Bayes*; MDA: *mixture discriminant analysis*; ADA: *AdaBoost*; KNN: *K-nearest neighbors*; NN: *neural network*; LRG: *regressão logística*; LDA: *linear discriminant analysis*; CART: *classification and regression tree*; PRIM: *patient rule induction method*.

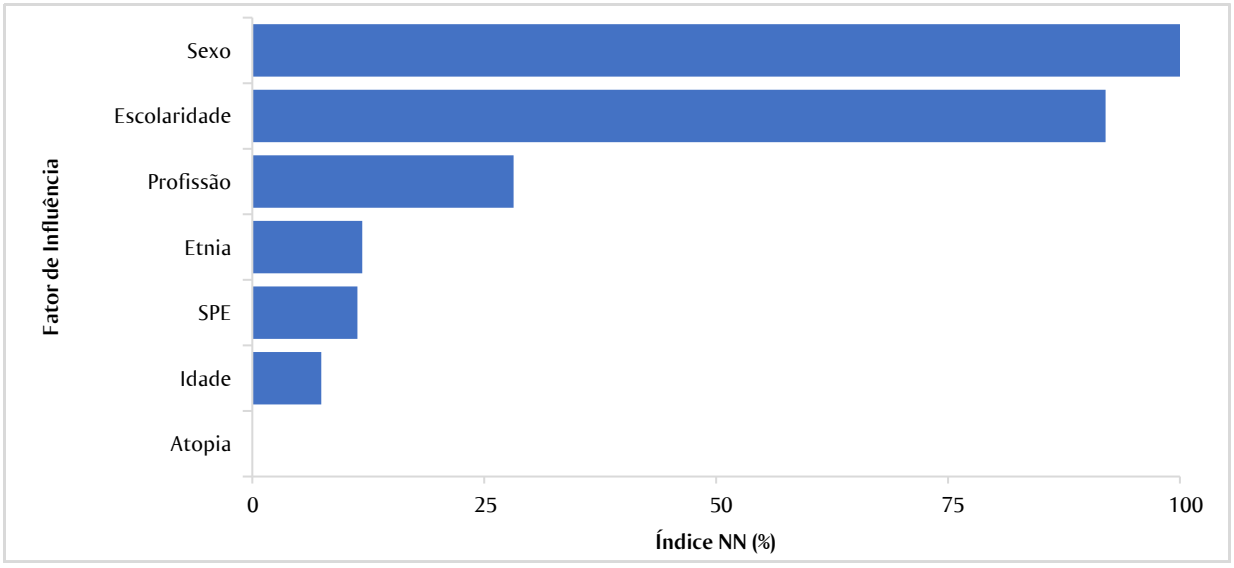

Figura 2 Importância das variáveis para a *neural network* (NN)
SPE: síndrome da pele excitada.

Os fatores determinantes para a ocorrência de uma doença de pele relacionada ao trabalho foram sexo e escolaridade. Tratando-se do fator sexo, para cada homem que não desenvolveu uma doença relacionada ao trabalho, tem-se 1,72 que desenvolveram. No tocante às mulheres, a relação revelou-se inversamente proporcional, de forma que, a cada mulher que apresentou uma doença de pele sem relação com o trabalho, 0,78 mulheres apresentaram doenças relacionadas ao trabalho. Isso é um indicativo de que os homens buscam atendimento com maior frequência do que as mulheres, quando relacionado ao trabalho. Quanto à escolaridade, por sua vez, quanto menor o nível de instrução, maior a chance de incidência de uma doença do trabalho.

A principal limitação da pesquisa está relacionada ao número de observações reunidas no banco de dados. Além disso, utilizar outras variáveis clínicas, como os exames laboratoriais, pode influenciar a aprendizagem dos algoritmos e, desta maneira, oferecer soluções mais assertivas (classificação) sobre as dermatites ocupacionais.

O AM aplicado à SST pode ser, e vem sendo, empregado para predizer doenças e/ou acidentes ocupacionais, conforme explorado por este trabalho e pelos estudos conduzidos por Zhao et al.[8], Saâdaoui et al.[9] e Palei e Das[10].

Conclusão

As técnicas de aprendizado de máquina (AM) apresentam potencial para auxiliar as ações preditivas de SST, pois consideram os eventos anteriores para formular modelos preditivos que subsidiam a tomada de decisão dos profissionais que atuam nos locais de trabalho com o objetivo de reduzir acidentes de trabalho e doenças ocupacionais.

Após a implementação do script no banco de dados utilizando apenas os parâmetros default de cada algoritmo, foi possível realizar a comparação das técnicas de aprendizado de máquina CART, LDA, SVM, KNN, PRIM, RF, LRG, MDA, C5.0, NB, NN e ADA, fazendo uso de métricas. As técnicas que apresentaram maior potencial para descrever com níveis superiores o banco de dados em questão foram: RF, SVM, NN e NB. Com o intuito de exemplificar a forma como o AM classifica as variáveis de influência, apresentou-se o resultado da classificação de NN, cujos principais influenciadores em ordem decrescente foram: sexo, escolaridade e profissão.

Fundamentado em um conjunto de dados de trabalhadores, foi possível identificar, a partir de um modelo de AM, os fatores (sexo, escolaridade e profissão) que prevalecem no surgimento da dermatite ocupacional. É importante destacar que as variáveis empregadas para a elaboração dos modelos preditivos consistiram em sua grande maioria em classes do tipo discretas, variáveis que comumente são armazenadas pelas organizações e instituições especializadas em saúde ocupacional e que podem ser utilizadas para validar outras hipóteses de pesquisas referentes à dermatite ocupacional, contribuindo para a construção de um conhecimento científico que subsidie políticas de prevenção na SST.

Para pesquisas futuras, recomenda-se utilizar técnicas de AM de busca automática de parâmetros para determinação dos hiperparâmetros – o que confere maior acuracidade –, tais como: *GridSearch*, *RandomSearch* e *Bayesian Optimization Model Tuning*. Além disso, recomenda-se avaliar, para cada técnica de AM, as variáveis que exercem maior influência no desenvolvimento de uma doença relacionada ao trabalho. Em termos práticos, após a avaliação da criticidade segundo o algoritmo para cada variável, essa informação deve ser confrontada com os dados reais para validação dos gestores e especialistas visando posterior implementação.

Referências

1. Provan DJ, Rae AJ, Dekker SWA. An ethnography of the safety professional's dilemma: safety work or the safety of work? Saf Sci. 2019;117:276-89.
2. Badri A, Gbodossou A, Nadeau S. Occupational health and safety risks: towards the integration into project management. Saf Sci. 2012;50(2):190-8.
3. Badri A, Boudreau-Trudel B, Souissi AS. Occupational health and safety in the industry 4.0 era: a cause for major concern? Saf Sci. 2018;109:403-11.
4. Neely A. The performance measurement revolution: why now and what next? International Journal of Operations & Production Management. 1999;19(2):205-28.
5. Fernandes FT, Chiavegatto Filho ADP. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. Rev Bras Saude Ocup. 2019;44:e13.
6. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255-60.
7. Wuest T, Weimer D, Irgens C, Thoben KD. Machine learning in manufacturing: advantages, challenges, and applications. Prod Manuf Res. 2016;4(1):23-45.
8. Zhao Y, Li J, Zhang M, Lu Y, Xie H, Tian Y, et al. Machine learning models for the hearing impairment prediction in workers exposed to complex industrial noise: a pilot study. Ear Hear. 2019;40(3):690-9.
9. Saâdaoui, F, Bertrand PR, Boudet G, Rouffiac K, Dutheil F, Chamoux A. A dimensionally reduced clustering methodology for heterogeneous occupational medicine data mining. IEEE Trans Nanobioscience. 2015;14(7):707-15.
10. Palei SK, Das SK. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: an approach. Saf Sci. 2009;47(1):88-96.
11. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst. 2008;14:1-37.
12. Mehta P, Bukov M, Wang CH, Day AGR, Richardson C, Fisher CK, et al. A high-bias, low-variance introduction to machine learning for physicists. Phys Rep. 2019;810:1-124.
13. Callahan A, Shah NH. Machine learning in healthcare. In: Sheikh A, Cresswell KM, Wright A, Bates DW, editors. Key advances in clinical informatics: transforming health care through health information technology. London: Elsevier; 2017. p. 279-91.
14. Sarkar S, Verma A, Maiti J. Prediction of occupational incidents using proactive and reactive data: a data mining approach. In: Maiti J, Ray PK, editors. Industrial safety management: 21st century perspectives of Asia. Singapore: Springer; 2018. p. 65-79.
15. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-9.
16. Kang K, Ryu H. Predicting types of occupational accidents at construction sites in Korea using random forest model. Saf Sci. 2019;120:226-36.
17. Rubaiyat AHM, Toma TT, Kalantari-Khandani M, Rahman SA, Chen L, Ye Y, et al. Automatic detection of helmet uses for construction safety. Proceedings of the International Conference on Web Intelligence Workshops; 2016; Omaha. Piscataway: IEEE; 2016. p. 135-42.
18. Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J. 2014;18(2):50-7.
19. Bohanec M, Delibašić B. Data-mining and expert models for predicting injury risk in ski resorts. Proceedings of the First International Conference on Decision Support System Technology; 2015; Belgrade. Cham: Springer; 2015. p. 46-60.
20. Nanda G, Grattan KM, Chu MT, Davis LK, Lehto MR. Bayesian decision support for coding occupational injury data. J Safety Res. 2016;57:71-82.
21. Shin DP, Park YJ, Seo J, Lee DE. Association rules mined from construction accident data. KSCE Journal of Civil Engineering. 2018;22:1027-39.
22. Cheng CW, Yao HQ, Wu TC. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. J Loss Prev Process Ind. 2013;26(6):1269-78.
23. Nelder JA, Wedderburn WM. Generalized linear models. J R Stat Soc Ser A Stat Soc. 1972;135(3):370-84.
24. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Proceedings of the Second European Conference on Computational Learning Theory; 1995; Barcelona. Berlin: Springer; 1995. p. 23-37.

25. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4):e1249.
26. Xu B, Huang JZ, Williams G, Wang Q, Ye Y. Classifying very high-dimensional data with random forests built from small subspaces. International Journal of Data Warehousing and Mining. 2012;8(2):44-63.
27. Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. J Med Syst. 2017;41(4):69.
28. Clemmensen L, Hastie T, Witten D, Ersbøll B. Sparse discriminant analysis. Technometrics. 2011;53(4):406-13.
29. Halbe Z, Aladjem M. Model-based mixture discriminant analysis – an experimental study. Pattern Recognit. 2005;38(3):437-40.
30. Kwak DS, Kim KJ, Lee MS. Multistage PRIM: patient rule induction method for optimisation of a multistage manufacturing process. Int J Prod Res. 2010;48(12):3461-73.
31. Nannings B, Abu-Hanna A, de Jonge E. Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. Int J Med Inform. 2008;77(4):272-9.
32. Parodi P. Computational intelligence with applications to general insurance: a review: I – The role of statistical learning. Annals of Actuarial Science. 2012;6(2):307-43.
33. Guns R, Lioma C, Larsen B. The tipping point: F-score as a function of the number of retrieved items. Inf Process Manag. 2012;48(6):1171-80.
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
35. Wickham H, Bryan J. Package 'readxl': read Excel files [Internet]. [local desconhecido]: CRAN; 2023 [citado em 15 jul 2020]. Disponível em: https://cran.r-project.org/web/packages/readxl/readxl.pdf
36. Kuhn M. The caret Package [Internet]. [local desconhecido]: Max Kuhn; 2019 [citado em 15 mar 2020]. Disponível em: https://topepo.github.io/caret/index.html
37. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. Package 'ggplot2': create elegant data visualisations using the grammar of graphics [Internet]. [local desconhecido]: CRAN; 2023 [citado em 15 jul 2020]. Disponível em: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
38. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. Package 'pROC': display and analyze ROC curves [Internet]. [local desconhecido]: CRAN; 2022 [citado em 15 jul 2020]. Disponível em: https://cran.r-project.org/web/packages/pROC/pROC.pdf
39. Brasil. Lei nº 8.213, de 24 de julho de 1991: dispõe sobre os Planos de Benefícios da Previdência Social e dá outras providências. Diário Oficial da União [Internet]. 25 jul 1991 [citado em 20 abr 2020];1:14809. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8213cons.htm
40. INSS – Comunicação de Acidente de Trabalho – CAT [Internet]. Brasília (DF): Ministério do Trabalho e Previdência; [citado em 3 abr 2023]. Disponível em: https://dados.gov.br/dados/conjuntos-dados/inss-comunicacao-de-acidente-de-trabalho-cat1

Contribuições de autoria

Rosa ACF, Galdamez EVC, Souza RCT, Melo MGM, Villarinho ALCF e Leal GCL contribuíram igualmente para a concepção do estudo, o levantamento, a análise e a interpretação dos dados, a elaboração, as revisões críticas do manuscrito e a aprovação da versão final publicada e assumem responsabilidade pública integral pelo trabalho realizado e o conteúdo aqui publicado.

Disponibilidade de dados

Os autores declaram que o conjunto de dados que dá suporte aos resultados deste estudo não está disponível publicamente por serem dados referentes ao prontuário de pacientes/trabalhadores atendidos em um serviço de saúde.

Recebido: 14/09/2020
Revisado: 24/06/2021
Aprovado: 25/06/2021

Editor-Chefe responsável:
Eduardo Algranti