RBSO Revista Brasileira de Saúde Ocupacional
ISSN: 2317-6369 (online)
http://dx.doi.org/10.1590/2317-6369000019418

Tema livre/Ensaio

Fernando Timoteo Fernandes[a,b]
https://orcid.org/0000-0003-2548-9685

Alexandre Dias Porto Chiavegatto Filho[b]
http://orcid.org/0000-0003-3251-9600

[a] Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho (Fundacentro). São Paulo, SP, Brasil.
[b] Universidade de São Paulo, Faculdade de Saúde Pública. São Paulo, SP, Brasil.

Contato:
Fernando Timoteo Fernandes
E-mail:
fernando.fernandes@fundacentro.gov.br

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

Os autores informam que o trabalho não foi baseado em dissertação ou tese e não foi apresentado em evento científico.

Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho

Data mining and machine learning perspectives for occupational safety and health

Resumo

Introdução: a variedade, volume e velocidade de geração de dados (*big data*) possibilitam novas e mais complexas análises. *Objetivo:* discutir e apresentar técnicas de mineração de dados (*data mining*) e de aprendizado de máquina (*machine learning*) para auxiliar pesquisadores de Saúde e Segurança no Trabalho (SST) na escolha da técnica adequada para lidar com *big data*. *Métodos:* revisão bibliográfica com foco em data mining e no uso de análises preditivas com *machine learning* e suas aplicações para auxiliar diagnósticos e predição de riscos em SST. *Resultados:* a literatura indica que aplicações de *data mining* com algoritmos de *machine learning* para análises preditivas em saúde pública e em SST apresentam melhor desempenho em comparação com análises tradicionais. São sugeridas técnicas de acordo com o tipo de pesquisa almejada. *Discussão: data mining* tem se tornado uma alternativa cada vez mais comum para lidar com bancos de dados de saúde pública, possibilitando analisar grandes volumes de dados de morbidade e mortalidade. Tais técnicas não visam substituir o fator humano, mas auxiliar em processos de tomada de decisão, servir de ferramenta para a análise estatística e gerar conhecimento para subsidiar ações que possam melhorar a qualidade de vida do trabalhador.

Palavras-chave: mineração de dados; aprendizado de máquina; saúde do trabalhador.

Abstract

Introduction: *variety, volume and data generation speed allow for new and more complex analyses.* Objective: *to discuss and present data mining and machine learning techniques to aid occupational safety and health (OSH) researchers to choose the suitable technique when dealing with large volumes of data.* Methods: *literature review to discuss data mining and machine learning predictive applications for aiding diagnosis and risk prevention in OSH.* Results: *literature shows that data mining with machine learning algorithms for predictive purposes in OSH and public health present better performance when compared to traditional analysis. According to the research purpose, different techniques are recommended.* Discussion: *data mining has become a common alternative when dealing with large databases in public health, making it possible to analyze large volume of morbidity and mortality data. These techniques are not meant to replace the human factor, but rather to assist in decision-making processes, to work as a tool for the statistical analysis of OSH data and to build up knowledge to subsidize actions that may improve worker's quality of life.*

Keywords: *data mining; machine learning; occupational safety and health.*

Recebido: 19/03/2018
Revisado: 30/08/2018
Aprovado: 07/12/2018

Introdução

A preparação, o processamento e a análise de grandes volumes de dados de origens distintas e de conjuntos de dados de instituições privadas ou órgãos governamentais, como dados estruturados tabulares do Sistema de Informação sobre Mortalidade (SIM), Sistema de Informações de Nascidos Vivos (Sinasc), Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH-SUS), Relação Anual de Informações Sociais (RAIS), Sistema de Informação de Agravos de Notificação (Sinan) e dados não estruturados, como imagens e textos (incluindo descrições em prontuários eletrônicos e registros tradicionais em papel), são, ao mesmo tempo, um desafio e uma oportunidade para pesquisas na área da Saúde e Segurança no Trabalho (SST). Esta variedade, combinada com a grande quantidade de dados, é conhecida como *big data* e tem levado a uma mudança nas formas tradicionais de análise de dados[1].

Sistemas gerenciadores de bancos de dados, como o MySQL, e *softwares* estatísticos, como SAS, SPSS e outros, lidam com grandes volumes de dados há anos. No entanto, dificuldades de escalabilidade e processamento podem surgir ao atualizar ou adicionar conjuntos de dados[2] de diferentes tipos e fontes em tempo real, como nos casos provenientes das redes sociais.

Outro problema surge ao lidar com dados não estruturados, como textos e imagens, quando é necessário, por exemplo, solicitar a um especialista que classifique uma imagem ou um registro administrativo textual em uma determinada categoria para que possam ser desenvolvidos modelos preditivos a partir de um conjunto de dados classificados. Além disso, a cada nova atualização dos dados, é preciso gerar um novo modelo preditivo e dispender mais tempo do especialista.

Para tornar o processo de análise menos trabalhoso, pode-se empregar técnicas que facilitem o trabalho do especialista de forma a minimizar o tempo dispendido em tarefas de classificação e agrupamento. Essas técnicas exigem uma breve preparação do conjunto de dados para que uma sequência finita de instruções programadas, denominada algoritmos, realize a classificação do conjunto de maneira automática. A partir dessa classificação, o especialista pode aferir a acurácia do algoritmo e utilizar o mesmo modelo criado para a análise de novos dados. Dentre essas técnicas, podemos citar a mineração de dados e a aprendizagem de máquina, que empregam conceitos de inteligência artificial para tomar decisões baseadas em treinamentos prévios realizados por especialistas[3]. Para facilitar o entendimento das técnicas empregadas nos estudos discutidos ao longo deste ensaio, definiremos brevemente esses conceitos de modo a fornecer uma visão geral.

A inteligência artificial pode ser definida pelo "estudo de agentes que recebem percepções do ambiente e executam ações"[3] (p. VIII). Tais agentes buscam executar suas ações de maneira a maximizar as chances de sucesso para seus objetivos. O campo da inteligência artificial discute, por exemplo, a capacidade de agentes físicos (máquinas) ou lógicos (programas de computador) tomarem decisões com base em dados captados por meio de sensores ou alimentados por meio de intervenção humana. O tema é antigo e já era discutido muito antes da clássica publicação de Turing[4], que discutia de forma informal conceitos de inteligência, aprendizado de máquina e sobre o que poderia ser considerada uma máquina inteligente. A inteligência pode estar associada à tomada de decisão racional em um processo no qual o agente busca alcançar o melhor resultado ou, na impossibilidade deste, o melhor resultado esperado de forma autônoma[3].

Neste cenário de automatização e busca de melhores resultados, surgem processos analíticos que auxiliam na tomada de decisão, como o processo de mineração de dados. A mineração de dados (do inglês, *data mining*) pode ser definida como o "processo automático ou semiautomático de explorar analiticamente grandes bases de dados"[5] (p. 10). O processo de *data mining* busca descobrir padrões e novas informações a partir de um determinado conjunto de dados. Esse processo será apresentado ao longo deste ensaio.

Relacionado aos campos de estudo da inteligência artificial e *data mining*, existe outro conceito que muitas vezes se confunde com o próprio *data mining* e que é denominado aprendizado de máquina (do inglês, *machine learning*). No *data mining*, são utilizadas técnicas para descobrir propriedades de um conjunto de dados existente e possíveis correlações de diferentes atributos desse conjunto, podendo ser utilizados algoritmos de *machine learning* para construir modelos que realizam predições ou classificações dos dados disponíveis[6]. Há duas grandes técnicas de *machine learning*: a aprendizagem supervisionada e a aprendizagem não supervisionada, as quais também serão abordadas ao longo deste manuscrito.

Atualmente, o estudo de algoritmos de *machine learning* tem ganhado destaque devido à alta performance preditiva em análises de grandes volumes de dados. Na área da saúde, é cada vez mais frequente o uso de *data mining* e *machine learning* no auxílio ao processo de diagnóstico[7,8], predição de riscos[9] e biomedicina[10]. Tais técnicas também têm sido

empregadas como ferramentas complementares em estudos epidemiológicos[11].

Neste ensaio, serão discutidos exemplos e resultados de uso de *data mining* e *machine learning* na área da saúde e segurança no trabalho (SST), de forma a explorar sua aplicação neste ramo e facilitar o processo de escolha dessas técnicas.

Iniciaremos discutindo o processo de *data mining* como um todo, seguido por relatos de uso de algoritmos de *machine learning* e de aprendizagens não supervisionada e supervisionada. Em seguida, serão apresentados exemplos de estudos que utilizam combinações de diferentes técnicas de *machine learning* para a predição de riscos ocupacionais e outras perspectivas de análises.

Visão geral de *data mining* e *machine learning*

A mineração de dados está inserida no processo de descoberta de conhecimento em bases de dados (em inglês, *knowledge discovery in databases* [KDD])[12]. Segundo Han et al.[13], o processo de KDD envolve uma sequência de etapas, como a limpeza de dados, na qual são tratados valores ruidosos e *outliers* (valor atípico ou aberrante). Após a limpeza dos dados, pode-se realizar a integração de novos conjuntos de dados. Segue-se a etapa de seleção, na qual somente os dados relevantes são filtrados para a pesquisa. Em seguida, os dados são transformados e consolidados de acordo com os propósitos da mineração. Realiza-se, então, o *data mining*, no qual são aplicadas técnicas para descoberta de padrões nas bases de dados, por meio de algoritmos computacionais[5]. Nesta fase, podem ser utilizados algoritmos de *machine learning* para tarefas de predição, associação ou agrupamento dos dados[5]. Após a etapa de *data mining*, inicia-se a etapa de análise dos padrões encontrados e, finalmente, a etapa de apresentação dos resultados e a descoberta de conhecimento.

A **Figura 1** ilustra as etapas do processo de KDD no qual está inserida a etapa de *data mining*.

O uso de *data mining*, combinado com algoritmos de *machine learning*, pode auxiliar o especialista da saúde em momentos críticos que demandem decisões rápidas quando há uma deficiência dos recursos apresentados, por exemplo imagens de baixa resolução, ou por condições precárias de trabalho que podem levar a um diagnóstico impreciso por fatores externos, como falta de tempo e elevado nível de estresse do profissional[7]. Com o uso de algoritmos de *machine learning*, podem ser criados modelos de predição ou agrupamento para identificação prévia de riscos ou diagnóstico de doenças a partir de determinados sintomas de um paciente, auxiliando o profissional da saúde na tomada de decisão, principalmente no caso de doenças raras com as quais o profissional tem pouca experiência.

Para analisar como essas técnicas estão sendo utilizadas na área da saúde, foram pesquisados estudos disponíveis na Biblioteca Virtual em Saúde (BVS), utilizando como chaves de busca os termos "*data mining*" e "*machine learning*", somente em inglês (**Tabela 1**). A consulta foi restrita apenas aos títulos dos artigos para facilitar a identificação das áreas em que estão sendo aplicadas essas técnicas. Foram excluídos artigos em duplicidade, artigos de correção ou erratas, artigos e periódicos não relacionados à saúde humana, como artigos nas áreas de agricultura e pecuária. Artigos que geraram dúvida em relação ao uso na saúde humana foram analisados individualmente. Os dados na **Tabela 1** mostram a tendência de crescimento no uso de *data mining* e *machine learning* na área da saúde.

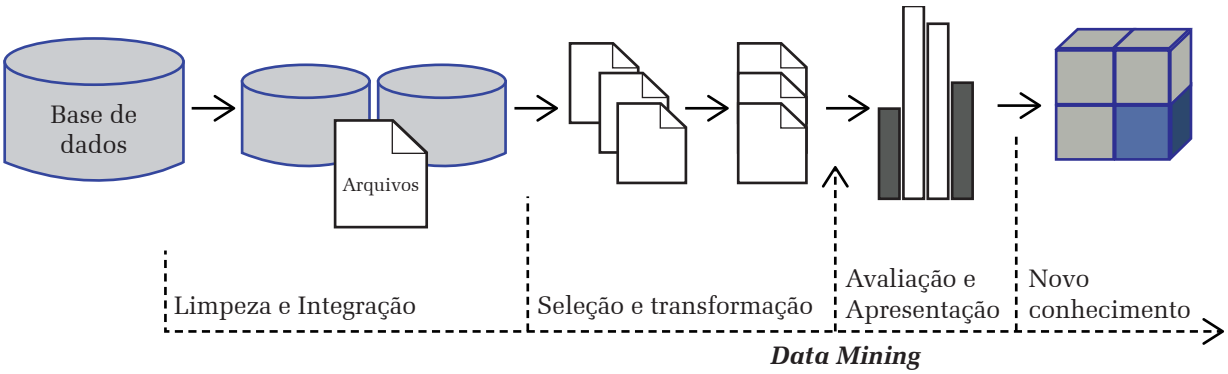

Figura 1 *Data mining* como parte do processo de descoberta de conhecimento em base de dados (KDD). A etapa de *data mining* acontece após as etapas de seleção e transformação de dados para tentar descobrir padrões nos dados

Tabela 1 Artigos citando uso de *data mining* e/ou *machine learning* identificados na Biblioteca Virtual em Saúde

Ano	*Número de artigos*	*Duplicidade*	*Não relacionados*	*Total*
2016	678	3	52	623
2017	938	5	45	888
2018	1842	2	68	1772
Total	**3458**	**10**	**165**	**3283**

Em relação ao uso de algoritmos de *machine learning,* a maioria dos problemas de análise pode ser inserido em duas categorias principais[14]: aprendizado supervisionado, em que o desfecho de um conjunto de dados é conhecido, ou seja, existe um valor da variável resposta a ser predito; e o aprendizado não supervisionado, em que não existe uma variável resposta específica, por exemplo no caso de identificar populações parecidas de acordo com suas similaridades ou reduzir a dimensionalidade de um conjunto de dados.

Para as análises supervisionadas, deve-se utilizar, de acordo com o tipo de variável resposta a ser predita, técnicas de classificação ou de regressão. Nas tarefas de classificação, dado um conjunto de dados em que a variável a ser predita é categórica, estima-se a categoria de um novo exemplar por meio da análise de seus atributos e das categorias existentes. Nas tarefas de regressão, o resultado da análise de um novo exemplar em um conjunto de dados é uma variável contínua[5].

A seguir, serão abordados alguns conceitos de uma vertente da mineração de dados bastante comum na análise de textos, a mineração de dados textuais (do inglês, *text mining*) e, nas próximas seções, serão abordadas formas de aprendizagem não supervisionada e supervisionada, assim como exemplos de aplicações em SST.

Mineração textual em registros administrativos

Para encontrar informações de qualidade em um conjunto de dados textuais, a mineração textual (em inglês, *text mining*) utiliza conceitos de *machine learning* e de outras disciplinas, como a computação linguística e a estatística[13]. As tarefas incluem a categorização e o agrupamento de dados não estruturados de forma a encontrar padrões e informações relevantes para cada conjunto.

Entre suas aplicações, a mineração textual pode ser utilizada para a análise de registros administrativos que contenham campos textuais que complementam ou descrevem aquele registro. Esses textos são chamados de dados não estruturados, pois não estão previamente organizados, como no caso de dados estruturados contidos, por exemplo, em tabelas, que contém uma coluna para cada informação. Assim como os dados estruturados, os dados não estruturados precisam ser pré-processados para que possam ser aplicadas análises posteriores.

A mineração textual consiste em analisar um conjunto de documentos denominado *corpus*. Para cada documento é gerada uma lista de termos, também chamados de *tokens,* em que são eliminados os termos considerados irrelevantes para a análise, como artigos, preposições, caracteres especiais e até mesmo termos definidos pelo usuário em uma lista especial denominada *stopwords*. Em seguida, os termos são reduzidos aos seus radicais, removendo prefixos e sufixos para evitar variações da mesma palavra, como em "interessante" e "interessantíssimo", que passam a mesma ideia com intensidades diferentes. A redução do termo ao seu radical é conhecida como *stemming*. Enfim, contabiliza-se a presença dos termos em forma de representação vetorial binária ou representação de frequências do termo[5].

A seguir, apresentamos os conceitos dos dois principais tipos de aprendizagem de máquina e alguns estudos que utilizam essas técnicas nas áreas de saúde pública e de SST.

Aprendizagem não supervisionada

Na aprendizagem não supervisionada, os algoritmos buscam padrões em registros com características similares, comparando os valores de seus atributos.

A aprendizagem não supervisionada é frequentemente aplicada para problemas de agrupamento (também chamado de *clusterização*) ou para redução de dimensionalidade de conjuntos de dados multivariados. Nos casos de agrupamento, os dados são analisados por suas similaridades ou dissimilaridades e agrupados por meio de medidas de distância, como a distância quadrática euclidiana. A distância quadrática euclidiana mede a distância entre duas observações (x, w) e todas as

variáveis de cada observação (p), cuja fórmula é apresentada na **Equação 1**:

Equação 1 Distância quadrática euclidiana em um espaço *n*-dimensional

$$distância(x_i, w_i) = \sum_{j=1}^{p} (x_{ij} - w_{ij})^2$$

Fonte: adaptado de Hastie et al.[15]

Uma forma de agrupamento não supervisionado é o agrupamento por partição, em que cada ponto deve pertencer a pelo menos uma partição ou *cluster*, sendo que o usuário define inicialmente o número de partições. O algoritmo k-médias (do inglês, *k-means*) é um exemplo desse tipo de agrupamento. Ele inicia selecionando aleatoriamente centroides do conjunto de dados e, a cada interação, agrupa os pontos mais próximos ao centroide por meio da menor distância quadrática euclidiana. Os centroides são também atualizados pela média dos pontos de cada *cluster*. O algoritmo termina (converge) quando não há mudança significativa dos centroides. A **Figura 2 (A)** ilustra o gráfico de dispersão de um conjunto de dados não categorizado. As **Figuras 2 (B)**, **2 (C)** e **2 (D)** ilustram a aplicação do algoritmo k-*means* sobre um conjunto de dados a partir da definição de 2 , 3 e 4 *clusters* (k).

O algoritmo *k-means* utiliza medidas de distância como a distância quadrática euclidiana para agrupar os pontos em cada centroide. Cabe citar que existem outras técnicas de comparação entre vetores com outras medidas de similaridade, como distância de cossenos (*cosine similarity*) e similaridade de Jaccard, entre outras[16,17].

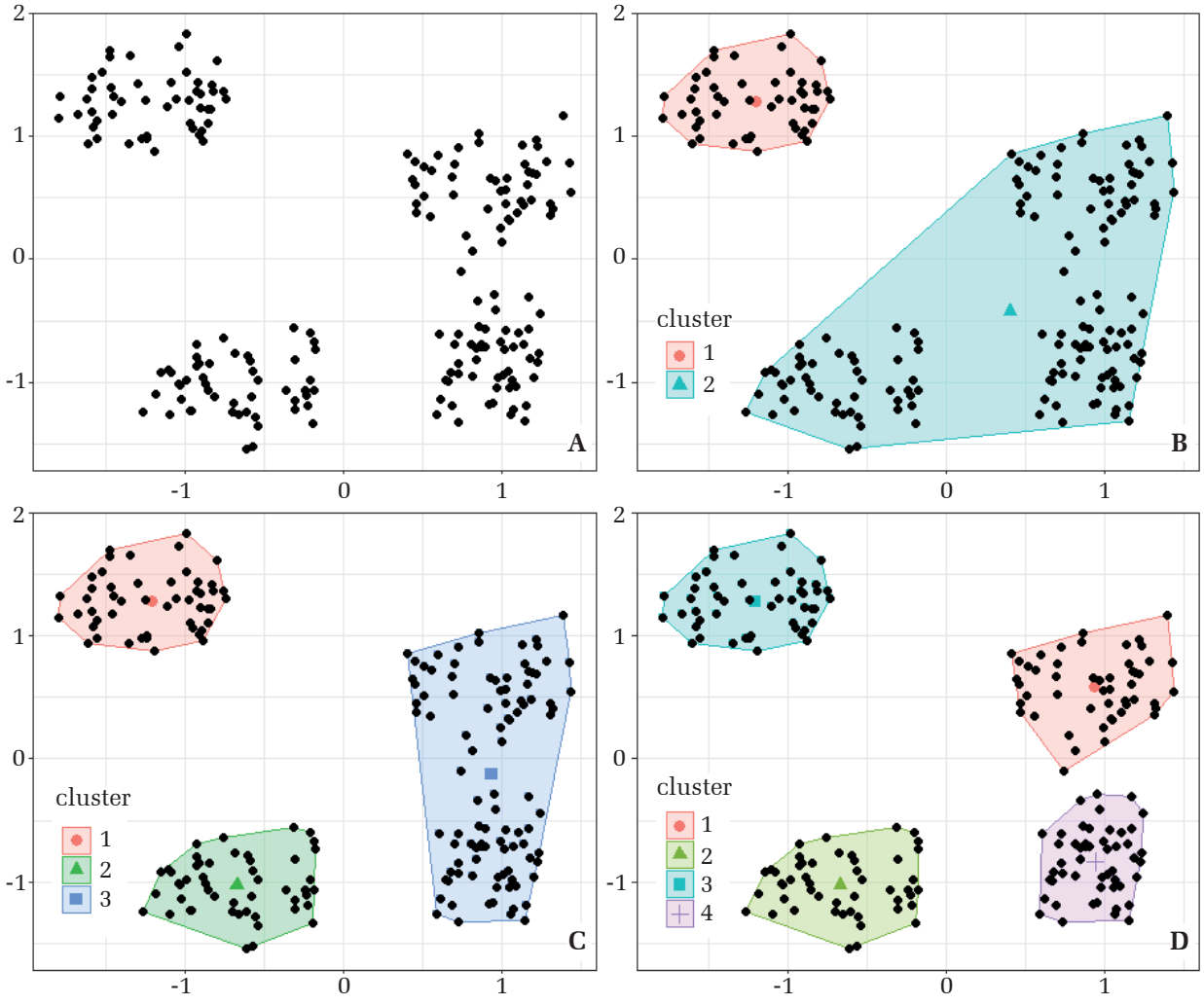

Figura 2 (A) Gráfico de dispersão de dados não categorizados. (B) Agrupamento utilizando o algoritmo *K-means* com 2 *clusters* (k=2). (C) Agrupamento utilizando 3 *clusters* (k=3). (D). Agrupamento utilizando 4 *clusters* (k=4)

Alguns estudos na área da saúde, como o estudo realizado por Olson et al.[18], utilizaram o algoritmo *k-means* para agrupamento de perfis. Nesse estudo[18], foram agrupados perfis de caminhoneiros de acordo com atributos comuns, como duração do sono, prática de exercícios e hábitos alimentares. Outro estudo, realizado por Lee et al.[19], utilizou uma variação do algoritmo *k-means* para agrupar perfis de equipes médicas de acordo com um questionário que avaliou a incidência e grau da síndrome de *burnout*. O objetivo nesses dois estudos foi identificar grupos parecidos de profissionais que se beneficiariam de tratamentos ou intervenções similares.

Outra técnica de aprendizagem não supervisionada bastante utilizada é a técnica denominada análise de componentes principais, ou PCA, para redução de dimensionalidade em conjuntos de dados multivariados. O PCA cria novas variáveis, denominadas componentes principais, sendo que os primeiros componentes gerados visam capturar a maior variabilidade da combinação linear dos preditores[20], conforme a **Equação 2**, em que *i* corresponde ao número do componente principal e *j* corresponde ao peso atribuído ao preditor para o componente atual.

Equação 2 Cálculo de componentes principais

$$PC_i = \left(a_{j1} x\ Preditor\ 1\right) + \left(a_{j2} x\ Preditor\ 2\right) + \ldots$$
$$\ldots + \left(a_{jn} x\ Preditor\ N\right)$$

Fonte: Kuhn et al.[20]

O estudo realizado por Pearce et al.[21] utilizou a técnica de agrupamento não supervisionado denominada mapa auto-organizável (*self-organizing map* [SOM]) ou algoritmo de Kohonen, que também é capaz de organizar dados complexos em grupos em que não há padrões conhecidos. A diferença é que, com o uso dos SOMs, é possível visualizar *n* dimensões em um gráfico bidimensional. No estudo de Pearce et al.[21], o SOM foi utilizado para analisar a qualidade de ar com base em séries históricas.

O trabalho realizado por Gao et al.[22] utiliza uma variação do SOM para realizar análises exploratórias nas quais o algoritmo cria subgrupos de populações com características similares, usando como base de dados os registros médicos de doentes renais. O algoritmo gera um mapa bidimensional representando cinco dimensões escolhidas e exibe a intensidade de concentração das variáveis selecionadas, assim como as probabilidades de mortalidade em oito anos, de acordo com os perfis identificados pelo SOM.

Aprendizagem supervisionada

A aprendizagem supervisionada se baseia em dados preparados para treinamento quando se sabe o desfecho de cada registro do conjunto de dados, em geral previamente rotulados por um especialista.

Para a construção de um modelo de aprendizagem supervisionada, após a definição de um subconjunto da base de dados inicial para a análise, separa-se uma parte da amostra para realizar o treinamento e o ajuste do modelo, e outra parte para testar o desempenho do modelo.

Por exemplo, o estudo realizado por Ramezankhani et al.[23] utiliza o algoritmo de árvore de decisão para a predição do desenvolvimento de hipertensão em uma coorte de adultos e para identificar os preditores que mais contribuíram para o aumento do risco de hipertensão. Dos indivíduos participantes do estudo, separou-se uma parte das observações para o treinamento do modelo e outra para testar sua acurácia. O desempenho do modelo foi aferido com base nos resultados de teste do modelo, apresentando performance preditiva superior se comparado com modelos estatísticos tradicionais.

O estudo realizado por Correa et al.[7] utilizou uma série de imagens de manchas de pele, sendo que, para cada imagem, há uma classificação associada a tumores benignos ou malignos. Para realizar a classificação foi utilizado o algoritmo de aprendizado supervisionado denominado máquina de vetor de suporte (em inglês, *support vector machines* [SVM]), e usadas características como assimetria, coloração, entre outras. A imagem então é vetorizada, ou seja, os *pixels* da imagem são transformados em uma sequência de números $\overrightarrow{xi}$ em que *x* representa uma única observação e *i* sua identificação no conjunto de imagens. Conforme o padrão da imagem se enquadra nas características de tumores malignos, o desfecho para aquela observação é predito como câncer (y_i) ou não. O estudo obteve uma acurácia superior a 90% na classificação de tumores malignos.

A **Figura 3** ilustra como o SVM classificaria duas classes linearmente separáveis, como no estudo realizado por Correa et al.[7]. O algoritmo tem o objetivo de separar os dados por classe, tentando maximizar a distância entre as classes.

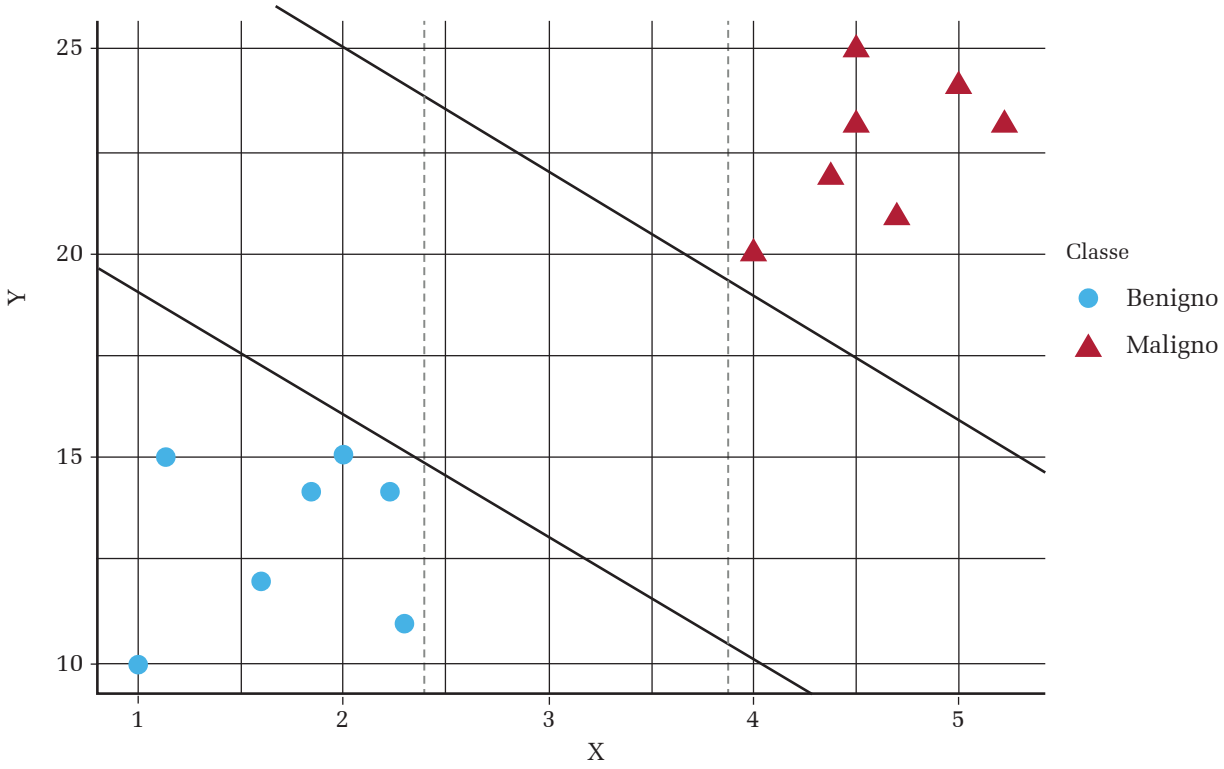

Figura 3 Dados linearmente separáveis utilizando o algoritmo SVM para predição do tipo de tumor

Nota: as linhas tracejadas representam opções para separar os dados em duas classes, utilizando hiperplanos. O eixo X representa uma variável de entrada (p.ex.: níveis de coloração do tumor) e o eixo Y representa outra variável de entrada (p.ex.: extensão da lesão). As linhas contínuas indicam a maior margem possível para separar os dois tipos de classes.

Outro estudo, realizado por Pan et al.[24], comparou a performance de algoritmos de aprendizagem supervisionada, como *Naïve Bayes* e *random forest*, para predição de partos considerados de risco, com base em dados históricos e em fatores de risco das mães, como estresse, condições socioeconômicas, má nutrição e idade, conseguindo uma performance preditiva 36% maior em comparação com análises estatísticas tradicionais.

Um estudo recente, realizado por Oliveira et al.[25], usou algoritmos supervisionados como K-NN, redes neurais artificiais, *Naïve Bayes* e *random forest* para a detecção de diabetes não diagnosticada utilizando dados do Estudo Longitudinal da Saúde do Adulto (Elsa-Brasil). O estudo comparou o desempenho desses algoritmos em relação à regressão logística, obtendo melhor desempenho com o algoritmo de redes neurais artificiais.

O estudo realizado por Vianna et al.[26] utilizou mineração textual em conjunto com algoritmos de aprendizagem supervisionada, como árvores de decisão, para analisar dados do Sinasc e do SIM para identificar fatores associados à mortalidade infantil, verificando que fatores como gravidez na adolescência estariam associados a maiores taxas de mortalidade.

Outro estudo realizado por Torres et al.[27] utilizou o algoritmo *superlearner* para predizer depressão em adultos de acordo com critérios sociais. O *superlearner* agrega um conjunto de algoritmos supervisionados, como *gradient boosted trees*, e seleciona o peso desses algoritmos de acordo com o desempenho individual alcançado. Esse tipo de algoritmo que engloba ou utiliza outros algoritmos é conhecido na área como *ensemble*[27].

Exemplos de aplicação de *text mining* e *machine learning* em SST

O estudo realizado por Marucci-Wellman et al.[28] realizou mineração textual em 15 mil registros de narrativas de solicitações de seguro por acidente ou lesão extraídos da base de dados de uma grande seguradora. As solicitações foram analisadas por meio de um algoritmo de aprendizagem supervisionada, utilizando como classificador o algoritmo *Naïve Bayes* e classificando cada registro com um código de acidentalidade interno de dois dígitos. O estudo obteve uma acurácia de 87% nas classificações realizadas em comparação com

o padrão-ouro. Do total de registros, 68% puderam ser automaticamente classificados, deixando apenas 32% dos registros para classificação manual e diminuindo a carga de trabalho dos analistas. Estudos brasileiros também utilizaram *text mining* para classificação de textos, como o trabalho realizado por Falcão et al.[29], que utilizou o *software* Weka[30] para aplicar *data mining* em classificação de páginas *web* de saúde. O *software* possui algoritmos de aprendizagem supervisionada, como *k*-vizinhos mais próximos (do inglês, *k-nearest neighbors)* e redes neurais artificiais, à disposição do pesquisador.

Outra ferramenta de *data mining* especializada em SST é a ferramenta gratuita disponibilizada pelo instituto de saúde ocupacional dos Estados Unidos, o National Institute for Occupational Safety and Health (NIOSH), para análise de narrativas. O *software* implementa tarefas de *data mining*, como *stemming*, entre outras tarefas, para realizar a codificação automatizada de ocupação e tipo de atividade[31,32]. A ferramenta está disponível para a língua inglesa e utiliza o padrão de codificação de ocupações internacional (SOC).

Outra aplicação é a análise de sentimentos ou análise de opiniões, que emprega técnicas de *text mining* junto com conceitos de outras áreas, por exemplo a psicologia, para consolidar opiniões sobre um assunto para grupos de pessoas. O estudo realizado por Araújo et al.[33] apresentou uma ferramenta própria para análise de sentimentos com recursos de aprendizado de máquina para a análise de dados de redes sociais.

O estudo realizado por Akay et al.[34] aplicou *text mining* e mapas auto-organizáveis para avaliar o sentimento de pacientes com câncer em relação ao uso de medicamentos, utilizando relatos em redes sociais e revelando importantes preocupações em relação aos efeitos colaterais e ao custo dos medicamentos.

Outras áreas também podem se beneficiar da análise de sentimentos de usuários a partir de relatos em redes sociais e sua relação com o trabalho, por exemplo pela análise dos efeitos do trabalho em turnos noturnos, conforme reportado por Carvalho et al.[35].

Cabe, ainda, citar o uso de combinações de técnicas supervisionadas e não supervisionadas para resolução de problemas mais complexos, para os quais é possível aplicar diferentes algoritmos para fases distintas e processamento. O trabalho de Christen[36] utiliza esse conceito para realizar *linkage* de bases de dados. Na fase inicial, são escolhidos exemplos de treinamento com o uso do algoritmo não supervisionado *k-means* e, em seguida, são utilizados outros algoritmos supervisionados, como SVM e K-NN para classificar os pares.

Resumo das técnicas de *data mining* e *machine learning*

Para resumir alguns dos algoritmos de *machine learning* utilizados com *data mining* citados ao longo deste ensaio e para citar outros algoritmos disponíveis, o **Quadro 1** categoriza, para cada tipo de aprendizagem, estudos recentes que utilizaram os diferentes tipos de análise na área de saúde pública e de saúde e segurança no trabalho.

Quadro 1 Algoritmos de *machine learning* por tipo de aprendizagem e exemplos de algoritmos aplicáveis na área da saúde e em SST

Tipo de aprendizagem	*Algoritmo*	*Aplicação*	*Referência*
Supervisionada	Árvores de decisão	Identificação de probabilidades de desenvolvimento de câncer de pulmão relacionado ao trabalho.	Kim et al.[37]
		Identificação de transmissão de sífilis da mãe para o filho.	Nakamura et al.[38]
		Identificação de características relacionadas à mortalidade infantil.	Vianna et al.[26]
	Redes neurais	Identificação de grupos de trabalhadores com altos riscos de pneumoconiose.	Liu et al.[39]
		Predição de quantidade de perda auditiva de trabalhadores de fábricas de aço.	Aliabadi et al.[40]

(Continua)

Quadro 1 Continuação…

	SVM	Classificação de melanomas. Estimar internações hospitalares.	Correa et al.[7] Lucini et al.[41]
	Naïve Bayes	Classificação de declarações textuais de morbidades relacionadas ao trabalho.	Marucci-Wellman et al.[28]
	Random Forests	Predição de gravidez de risco com base em fatores de risco da mãe. Detecção de diabetes não diagnosticada.	Pan et al.[24] Oliveira et al.[25]
Não supervisionada	SOM	Identificação de padrões de qualidade do ar em séries históricas. Analisar sentimentos de pacientes que passaram por tratamentos de câncer baseado em dados de fóruns especializados.	Pearce et al.[21] Akay et al.[34]
	K-means	Análise exploratória e identificação de subgrupos populacionais. Identificação de grupos de caminhoneiros de acordo com comportamentos e hábitos alimentares. Identificação de grupos expostos a diferentes níveis da síndrome de *burnout* em equipes médicas.	Gao et al.[22] Olson et al.[18] Lee et al.[19]

O **Quadro 1** exibe uma parte dos algoritmos disponíveis de *machine learning* para uso com *data mining* e análise preditiva, com o objetivo de exemplificar alguns estudos recentes na área e auxiliar o especialista na escolha do modelo que melhor se adeque à sua necessidade. Outras possibilidades de algoritmos populares incluem *gradient boosted trees*, regressões penalizadas de lasso e *ridge*, entre outros[24].

Com os resultados obtidos na pesquisa realizada na Biblioteca Virtual em Saúde e nos demais estudos citados neste ensaio, elaborou-se uma relação de finalidades, aplicações e estudos exemplificativos (**Quadro 2**).

Quadro 2 Finalidades, aplicações e estudos exemplificativos do uso de *data mining* e *machine learning*

Finalidade	*Aplicações*	*Estudos exemplificativos*
Complementar a análise exploratória e classificar perfis similares.	Identificação de perfis similares e frequências com uso de SOM adaptado. Identificação de perfis similares com uso de *k-means*.	Gao et al.[22] Olson et al.[18] Lee et al.[19]
Predição de desfechos como adoecimento ou mortalidade.	Predição de diabetes com uso de redes neurais, *k-nn* e *Random Forest*. Predição de hipertensão e riscos associados com mineração textual e árvore de decisão. Predição de óbitos infantis e fatores de risco com uso de árvores de decisão.	Oliveira et al. [25] Ramezankhani et al.[23] Vianna et al.[26]
Monitoramento e vigilância – epidemias e fatores ambientais.	Mineração de dados textuais de redes sociais e uso de SVM para identificar casos reais de influenza. Análise da qualidade do ar em séries históricas com o uso de SOM.	Allen et al.[42] Pearce et al.[21]
Suporte para decisão.	Classificação de melanomas por meio de análise de imagens e uso de SVM para categorização do desfecho.	Correa et al.[7]
Análise de registros administrativos.	Mineração textual em registros administrativos e classificação de doenças com uso de *Naïve Bayes*. Mineração textual e uso de SVM para predição de internações hospitalares.	Marucci-Wellman et al.[28] Lucini et al.[41]
Análise de sentimentos.	Mineração textual em uma rede social para analisar a opinião pública em relação à vacinação contra o vírus HPV. Uso do algoritmo SVM para classificar as opiniões como contrárias ou a favor da vacinação.	Du et al.[43]

Considerações finais

Conforme relatado ao longo deste ensaio, é possível observar diferentes aplicações de *data mining* e de algoritmos de *machine learning* na área de saúde pública e de saúde e segurança no trabalho. Na área de SST, foram abordados alguns estudos que realizaram predição de riscos ocupacionais, utilizando casos de morbidades em trabalhadores e estudos que permitem o agrupamento de dados volumosos com maior facilidade. Dessa forma, as técnicas apresentadas de aprendizagem não supervisionada podem auxiliar, por exemplo, na análise exploratória e na compreensão inicial de um conjunto dos dados a partir de seu agrupamento, enquanto os algoritmos de aprendizagem supervisionada podem ser utilizados para construir modelos preditivos pela aplicação de algoritmos, como SVM, redes neurais artificiais, entre outros. Portanto, entendemos que os algoritmos de *machine learning* facilitam a tomada de decisão ao lidar com grandes volumes de dados e que há uma perspectiva positiva na adoção dessa técnica para uso em novas pesquisas. Entretanto, importantes desafios permanecem, principalmente em relação à qualidade do preenchimento dos dados utilizados para as análises realizadas.

O campo de SST é bastante vasto e, conforme discutido neste ensaio, há um potencial muito grande para aplicação de *data mining* e *machine learning* devido ao grande volume de dados, a variedade e a velocidade de geração de novos dados, exigindo novas ferramentas, técnicas e algoritmos para lidar com *big data*. Essas técnicas já são utilizadas em diversas áreas da saúde pública. Os algoritmos podem ser utilizados para uma análise exploratória do conjunto de dados ou para desenvolver modelos preditivos que sejam capazes de estimar riscos de adoecimento de trabalhadores com base em históricos prévios de doenças. Com a disponibilização de microdados dos sistemas de saúde de várias esferas do governo e o surgimento de novas fontes de dados, como observatórios de SST[44] e redes sociais, concluímos que os algoritmos de *machine learning* em conjunto com o processo de *data mining* se tornam ferramentas poderosas para facilitar o processamento, a visualização dos dados e, consequentemente, a criação de modelos preditivos na área da saúde e de saúde e segurança no trabalho.

Contribuições de autoria

Fernandes TF contribuiu com o projeto do trabalho, com a análise e interpretação dos dados e com a elaboração do manuscrito. Chiavegatto Filho ADP contribuiu no planejamento, interpretação dos resultados e revisão do manuscrito. Ambos participaram na revisão crítica do manuscrito e na aprovação da sua versão final publicada e assumem integral responsabilidade pelo trabalho e pelo conteúdo publicado.

Referências

1. Chiavegatto Filho ADP. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol Serv Saúde. 2015;24(2):325-32.
2. Madden S. From databases to big data. IEEE Comput Soc. 2012;4-6.
3. Russel S, Norvig P. Inteligência artificial. Rio de Janeiro: Elsevier; 2013.
4. Turing AM. Computing machinery and intelligence. Mind 49 [Internet]. 1950 [citado em 9 out. 2019];59(236):433-60. Disponível em: https://www.csee.umbc.edu/courses/471/papers/turing.pdf
5. Silva L, Peres S, Boscarioli C. Introdução à mineração de dados com aplicações em R. Rio de Janeiro: Elsevier; 2016.
6. Mitchell TM. Machine learning and data mining. Commun ACM. 1999;42(11):30-6.
7. Correa DNL, Paniagua LRB, Noguera JLV, Pinto-Roa DP, Toledo LAS. Computerized diagnosis of melanocytic lesions based on the abcd method. Proceedings of the 41st Lat Am Comput Conf CLEI; 2015; Arequipa. Arequipa: CLEI; 2015;19(2):1-22.
8. Steiner MTA, Soma NY, Shimizu T, Nievola JC, Steiner Neto PJ. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. Gest Prod. 2006;13(2):325-37.
9. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One. 2017;12(4):e0174708.
10. Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol. 2004;25(8):690-5.
11. Marucci-Wellman HR, Corns HL, Lehto MR. Classifying injury narratives of large administrative databases for surveillance: a practical approach combining machine learning ensembles and human review. Accid Anal Prev. 2017;98:359-71.

12. Feyyad UM. Data mining and knowledge discovery: making sense out of data. IEEE Expert Intell Syst Their Appl. 1996;11(5):20-5.

13. Han J, Kamber M, Pei J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann; 2012.

14. James G, Witten D, Hastie T, Tibishirani R. An introduction to statistical learning with applications in R. Amsterdam: Springer; 2013.

15. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Amsterdam: Springer; 2016.

16. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: a survey. IEEE Trans Knowl Data Eng. 2007;19(1):1-16.

17. Al-Anazi S, Almahmoud H, Al-Turaiki I. Finding similar documents using different clustering techniques. Procedia Comput Sci. 2016;82:28-34.

18. Olson R, Thompson SV, Wipfli B, Hanson G, Elliot DL, Anger WK, et al. Sleep, dietary, and exercise behavioral clusters among truck drivers with obesity: implications for interventions. J Occup Environ Med. 2016;58(3):314-21.

19. Lee YC, Huang SC, Huang CH, Wu HH. A new approach to identify high burnout medical staffs by kernel K-means cluster analysis in a regional teaching hospital in Taiwan. Inquiry. 2016;53(2).

20. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.

21. Pearce JL, Waller LA, Chang HH, Klein M, Mulholland JA, Sarnat JA, et al. Using self-organizing maps to develop ambient air quality classifications: a time series example. Environ Health. 2014;13(1):56.

22. Gao S, Mutter S, Casey A, Mäkinen VP. Numero: a statistical framework to define multivariable subgroups in complex population-based datasets. Int J Epidemiol. 2019;48(2):369-74.

23. Ramezankhani A, Kabir A, Pournik O, Azizi F, Hadaegh F. Classification-based data mining for identification of risk patterns associated with hypertension in Middle Eastern population: a 12-year longitudinal study. Medicine (Baltimore). 2016;95(35):e4143.

24. Pan I, Nolan LB, Brown RR, Khan R, van der Boor P, Harris DG, et al. Machine learning for social services: a study of prenatal case management in Illinois. Am J Public Health. 2017;107(6):938-44.

25. Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo Á, Barreto SM, et al. Comparação de algoritmos de aprendizagem de máquina para construir um modelo preditivo para detecção de diabetes não diagnosticada – ELSA-Brasil: estudo de acurácia. São Paulo Med J. 2017;135(3):234-46.

26. Vianna RCXF, Moro CMCB, Moysés SJ, Carvalho D, Nievola J. Mineração de dados e características da mortalidade infantil. Cad Saúde Pública. 2010;26(3):535-42.

27. Torres JM, Rudolph KE, Sofrygin O, Glymour MM, Wong R. Longitudinal associations between having an adult child migrant and depressive symptoms among older adults in the Mexican Health and Aging Study. Int J Epidemiol. 2018;47(5):1432-42.

28. Marucci-Wellman HR, Lehto MR, Corns HL. A practical tool for public health surveillance: semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accid Anal Prev. 2015;84:165-76.

29. Falcão AEJ, Mancini F, Costa TM, Hummel AD, Texeira FO, Sigulem D, et al. InDeCS: Método automatizado de classificação de páginas Web de Saúde usando mineração de texto e Descritores em Ciências da Saúde (DeCS). J Health Informat. 2009;1(1):1-6.

30. Weka. Weka 3: Data Mining Software in Java [Internet]. 13 jul 2008 [citado em 22 jan 2018]. Disponível em: www.cs.waikato.ac.nz/ml/weka/

31. National Institute for Occupational Safety and Health. Industry and occupation coding [Internet]. 2018 [citado em 4 mar 2018]. Disponível em: https://www.cdc.gov/niosh/topics/coding/overview.html#input

32. Freeman MB, Pollack LA, Rees JR, Johnson CJ, Rycroft RK, Rousseau DL, et al. Capture and coding of industry and occupation measures: findings from eight national program of cancer registries states. Am J Ind Med. 2017;60(8):689-95.

33. Araújo M, Gonçalves P, Cha M, Benevenuto F. iFeel: a web system that compares and combines sentiment analysis methods. Proceedings of the Int World Wide Web Conf Comm; 2014; Seoul. Seoul: IW3CW; 2014:75-8.

34. Akay A, Dragomir A, Erlandsson B-E. Network-based modeling and intelligent data mining of social media for improving care. IEEE J Biomed Heal Informatics [Internet]. 2015 [citado em 9 out 2019];19(1):210-8. Disponível em: http://ieeexplore.ieee.org/document/6851846/

35. Carvalho F, Guedes GP. Night sleep deprivation: computational analysis of language effects. Proceedings 23rd Brazillian Symp Multimed Web [Internet]. 2017 [citado em 9 out 2019];221-4. Disponível em: http://doi.acm.org/10.1145/3126858.3131595

36. Christen P. A two-step classification approach to unsupervised record linkage. Proceedings of the 6th Australasian Data Mining Conference; 2007; Gold Coast. Gold Coast: AusDM; 2007:111-9.

37. Kim TW, Koh DH, Park CY. Decision tree of occupational lung cancer using classification and regression analysis. Saf Health Work. 2010;1(2):140-8.

38. Nakamura CY, Otero SD, Carvalho DR. Mineração de dados no enfrentamento da transmissão vertical da sífilis. J Heal Informatics. 2016;8:171-9.

39. Liu H, Tang Z, Yang Y, Weng D, Sun G, Duan Z, et al. Identification and classification of high risk groups for Coal Workers' Pneumoconiosis using

an artificial neural network based on occupational histories: a retrospective cohort study. BMC Public Health. 2009;9:366.

40. Aliabadi M, Farhadian M, Darvishi E. Prediction of hearing loss among the noise-exposed workers in a steel factory using artificial intelligence approach. Int Arch Occup Environ Health. 2015;88(6):779-87.

41. Lucini FR, Fogliatto FS, Silveira GJC, Neyeloff J, Anzanello MJ, Kuchenbecker RDS, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. Int J Med Inform. 2017;100:1-8.

42. Allen C, Tsou MH, Aslam A, Nagel A, Gawron JM. Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. PLoS One. 2016;11(7):e0157734.

43. Du J, Xu J, Song HY, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. BMC Med Inform Decis Mak. 2017;17(Suppl 2):69.

44. SmartLab. Observatório Digital de Segurança e Saúde no Trabalho [Internet]. 2017 [citado em 23 jun 2017]. Disponível em: https://observatoriosst.mpt.mp.br/