Universidade de São Paulo
Faculdade de Saúde Pública

FERNANDO TIMOTEO FERNANDES

**Machine learning em saúde e segurança do trabalhador:
perspectivas, desafios e aplicações**

São Paulo
2021

FERNANDO TIMOTEO FERNANDES

**Machine learning em saúde e segurança do trabalhador:
perspectivas, desafios e aplicações**

Versão Revisada

Tese apresentada ao programa de pós-graduação em Saúde Pública da Faculdade de Saúde Pública da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de Concentração: Saúde Pública

Orientador: Prof. Dr. Alexandre Dias Porto Chiavegatto Filho

São Paulo

2021

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo da Publicação

Ficha elaborada pelo Sistema de Geração Automática a partir de dados fornecidos pelo(a)
autor(a)

Bibliotecária da FSP/USP: Maria do Carmo Alvarez - CRB-8/4359

Fernandes, Fernando Timoteo
Machine learning em saúde e segurança do trabalhador:
perspectivas, desafios e aplicações / Fernando Timoteo
Fernandes; orientador Alexandre Dias Porto Chiavegatto
Filho. -- São Paulo, 2021.
117 p.

Tese (Doutorado) -- Faculdade de Saúde Pública da
Universidade de São Paulo, 2021.

1. Aprendizado de Máquina. 2. Saúde do Trabalhador. I.
Dias Porto Chiavegatto Filho, Alexandre, orient. II. Título.

Nome: FERNANDES, Fernando Timoteo

Título: Machine learning em saúde e segurança do trabalhador: perspectivas, desafios e aplicações

Tese apresentada à Faculdade de Saúde Pública da
Universidade de São Paulo para obtenção do título
de Doutor em Ciências.

Aprovado em:

Banca Examinadora

Prof. Dr.	_____
Instituição:	_____
Julgamento:	_____

Prof. Dr.	_____
Instituição:	_____
Julgamento:	_____

Prof. Dr.	_____
Instituição:	_____
Julgamento:	_____

DEDICATÓRIA

Aos meus filhos, Alice e Arthur.

AGRADECIMENTOS

A Deus por guiar minhas escolhas durante minha vida e por proporcionar o privilégio de ter uma família unida que me apoiou durante toda minha jornada.

Aos meus pais, Maria Lúcia e Manoel por sempre me incentivarem nos estudos e pelos sacrifícios para propiciar condições para meu desenvolvimento.

À minha esposa, Olívia, pelo companheirismo, pelo apoio constante e incondicional.

Ao meu amigo Ricardo, por me apresentar à área da saúde pública, me incentivando e abrindo portas tanto na saúde pública quanto na saúde e segurança do trabalhador.

Ao meu amigo Jefferson, pelas longas conversas e discussões acaloradas informais e pelo apoio técnico na saúde dos professores.

Aos meus amigos pós-graduandos do Labdaps/USP, pela excelência técnica e aprendizagem mútua, em especial aos parceiros de pesquisa como Tiago, Cristiane, André e Thales. Me sinto muito honrado em ter participado deste grupo.

Ao meu orientador, Alexandre Dias Porto Chiavegatto Filho, pela competência, gentileza e dedicação ao qual conduz seus orientandos e pelo meu progresso como pesquisador.

Aos professores da Faculdade de Saúde Pública, pela excelência e qualidade das aulas lecionadas que contribuíram muito em minha formação acadêmica.

Aos membros da secretaria de pós-graduação da Faculdade de Saúde Pública da USP, pela atenção e ajuda durante o percurso do doutorado.

FERNANDES, F. T. **Machine learning em saúde e segurança do trabalhador: perspectivas, desafios e aplicações.** 2021.117f. Tese (Doutorado) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2021.

RESUMO

Algoritmos de *machine learning* têm impactado a área da saúde nos últimos anos. Muita dessa popularidade deve-se aos ganhos de performance preditiva em comparação aos modelos estatísticos tradicionais, já que estes algoritmos conseguem capturar relações não-lineares e complexas, além de permitirem o uso de diferentes tipos de dados. Esta pesquisa objetiva descrever as diferentes técnicas recentes de *machine learning* e como elas podem ser aplicadas na saúde e segurança do trabalhador (SST). Os resultados da tese estão organizados em três artigos científicos. No primeiro artigo, foi realizada uma revisão bibliográfica para entender o panorama de uso de *machine learning* na saúde pública e em SST. Foram identificadas e categorizadas aplicações de aprendizagem supervisionada e não-supervisionada, e os principais problemas de pesquisa correspondentes. No segundo artigo, foram aplicados algoritmos de aprendizagem supervisionada para predição de absenteísmo por doença e doença relacionada ao trabalho em professores da rede pública municipal do Estado de São Paulo entre 2014 a 2018 (n=174.294), usando como fonte de dados a Relação Anual de Informações Sociais (RAIS). Cinco algoritmos foram comparados de acordo com o valor da área abaixo da curva (AUC). Todos os algoritmos obtiveram AUROC superior a 0,76. O melhor algoritmo (redes neurais artificiais) obteve AUROC de 0,79, com acurácia de 71,52%, sensibilidade de 72,86% e especificidade de 70,52%. Foi possível realizar predições que forneceram estimativas de risco no ensino infantil, para subsidiar a prevenção de afastamento por morbidades em professores, utilizando dados públicos e anônimos. No terceiro e último artigo, foram desenvolvidos modelos preditivos para identificar, antecipadamente, trabalhadores com risco de diagnóstico positivo para doença pulmonar obstrutiva crônica (DPOC). O estudo utilizou dados da coorte prospectiva do UK Biobank, de indivíduos acompanhados desde 2006, filtrando aqueles que preencheram o questionário de histórico ocupacional (n=120.289). Desses, 1731 (1,4%) foram diagnosticados com DPOC. Ao todo, foram selecionadas 26 variáveis, entre dados demográficos, exames laboratoriais, hábitos e sintomas, para a construção de modelos generalistas para predição de DPOC. Além disso, foi selecionado um subconjunto de participantes (n=7.628) com histórico ocupacional na indústria da construção civil e na mineração com possível exposição a poeira de sílica, para desenvolver modelos especialistas. Desses, 237 (3,11%) tiveram diagnóstico de DPOC. O modelo generalista obteve AUROC de 0.845, e o modelo especialista obteve AUROC de 0.841. As cinco principais variáveis preditoras foram idade, tosse crônica, tabagismo, histórico de asma e expectoração. Os resultados mostram que é possível prever risco individual de diagnóstico de DPOC na população geral e nos trabalhadores expostos a poeiras minerais utilizando variáveis comumente coletadas na atenção primária. Nesta tese, mostramos a viabilidade de uso de modelos preditivos na saúde do trabalhador tanto para prognóstico quanto para diagnóstico, com boa performance preditiva. Espera-se que este estudo possa contribuir para uma maior adoção de modelos preditivos por pesquisadores em SST, permitindo identificar antecipadamente trabalhadores expostos a riscos ambientais de forma a auxiliar o início de medidas preventivas que inibam ou minimizem os riscos.

Palavras-chave: Aprendizado de Máquina; Prognóstico; Diagnóstico; Saúde do Trabalhador.

FERNANDES, F. T. **Machine learning in public health and occupational safety and health: perspectives, techniques and applications.** 2021.117f. Thesis (Ph. D.) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2021.

ABSTRACT

Machine learning algorithms have gained prominence in the health area in recent years. Much of this popularity is due to predictive performance gains when compared to traditional statistical models, as these algorithms are able to capture non-linear relationships and to handle different types of data. This research aims to describe the different machine learning techniques and how these techniques can be applied in occupational safety and health (OSH). The results are organized into three scientific articles. In the first manuscript, a literature review was carried out to understand the panorama of machine learning use in public health and OSH. Supervised and unsupervised learning algorithms were identified and categorized, and main research problems were listed. In the second article, supervised learning algorithms were developed to predict absenteeism due to illness and work related illness in teachers from all public municipal schools in the State of São Paulo between 2014 and 2018 (n=174.294) available from the *Relação Anual de Informações Sociais* (RAIS). Five algorithms were compared according to the value of the area under the receiver operating characteristic curve (AUROC). All algorithms obtained AUROC greater than 0.76. The best algorithm (artificial neural networks) obtained an AUROC of 0.79, with an accuracy of 71.52%, sensitivity of 72.86% and specificity of 70.52%. It was possible to make assertive predictions, which provide estimates of risk, providing subsidies for preventing sick leave in teachers using public and anonymous data. In the third and last article, predictive models were developed to identify workers at risk of a positive diagnosis for chronic obstructive pulmonary disease (COPD). The study used data from the UK Biobank prospective cohort from individuals followed since 2006, filtering those who completed the occupational history questionnaire (n=120.294). Of these, 1731 (1.4%) had a positive diagnosis of COPD. In all, 26 variables were selected, including demographic data, laboratory tests, habits and symptoms, for the development of generalist models for the prediction of COPD. In addition, a subset of individuals (n=7628) with an occupational background in the construction and mining industry, with possible exposure to mineral dusts was selected to develop specialized models. Of these, 237 (3.11%) were diagnosed with COPD. The generalist model obtained AUROC of 0.845, and the specialist model, an AUC of 0.841. The five main predictive variables were age, chronic cough, smoking, earlier diagnosis of asthma and chronic sputum. The results show that it is possible to predict individual risk of COPD diagnosis in the general population and in workers exposed to silica dust using variables commonly collected in primary care. In this research, we showed the feasibility of using predictive models in worker health for both prognosis and diagnosis, with good predictive performance. We believe that this study can contribute to a greater adoption of predictive models by OSH researchers, allowing the early identification of workers exposed to risks and the adoption of preventive measures that inhibit or minimize risks.

Keywords: Machine learning; Prognosis; Diagnostic; Occupational Health.

LISTA DE SIGLAS

AUC	<i>Area Under the Curve</i>
AUROC	<i>Area Under the Receiver Operating Characteristic Curve</i>
BVS	Biblioteca Virtual em Saúde
CBO	Classificação Brasileira de Ocupações
CNAE	Classificação Nacional de Atividades Econômicas
DORT	Distúrbios Osteomusculares Relacionados ao Trabalho
DPOC	Doença Pulmonar Obstrutiva Crônica
EQM	Erro Quadrático Médio
IBGE	Instituto Brasileiro de Geografia e Estatística
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LER	Lesões por esforço repetitivo
MDA	<i>Mean Absolute Distance</i>
PCA	Principal Component Analysis
VPP	Valor Predito Positivo
VPN	Valor Predito Negativo
RAIS	Relação Anual de Informações Sociais
RBSO	Revista Brasileira de Saúde Ocupacional
RMSE	<i>Root Mean Square Error</i>
ROC	<i>Receiver Operating Characteristic</i>
ROSE	<i>Random Over-Sampling Examples</i>
SGBD	Sistema Gerenciador de Banco de Dados
SMOTE	<i>Synthetic minority over-sampling technique</i>
SOM	<i>Self-Organizing Map</i>
SST	Saúde e Segurança do Trabalhador
SVM	<i>Support Vector Machines</i>
XGBoost	eXtreme Gradient Boosting

SUMÁRIO

1	INTRODUÇÃO	12
1.1	<i>Machine learning</i> para prevenção e prognóstico	13
1.2	<i>Machine learning</i> para diagnóstico	14
1.3	Desafios e limitações	15
2	MACHINE LEARNING: VISÃO GERAL, PROBLEMAS E MÉTRICAS.	16
2.1	Tipos de aprendizado de máquina e pré-processamento.....	16
2.1.1	Aprendizagem supervisionada.....	16
2.1.2	Aprendizagem não-supervisionada	18
2.1.3	Pré-processamento	20
2.2	Sobreajuste, sub-ajuste e algoritmos de ML.....	21
2.2.1	Reamostragem, Validação Cruzada e <i>Ensemble</i>	22
2.2.2	Balanceamento de classes	24
2.2.3	Algoritmos de machine learning	25
2.3	Métricas para avaliação de modelos preditivos	31
2.3.1	Métricas para modelos de classificação	31
2.3.2	Métricas para modelos de regressão.....	33
3	OBJETIVOS.....	35
3.1	OBJETIVO GERAL	35
3.2	OBJETIVOS ESPECÍFICOS	35
4	MÉTODOS.....	35
4.1	MÉTODO DO ARTIGO 1.....	35
4.1.1	Delineamento do estudo	35
4.1.2	Revisão e análise dos artigos	36
4.2	MÉTODO DO ARTIGO 2.....	36
4.2.1	Fonte de dados e delineamento do estudo	36
4.2.2	População de estudo	37
4.2.3	Variáveis do estudo	38
4.2.4	Processamento e análise dos dados	38
4.2.5	Aspectos éticos	39
4.3	MÉTODO DO ARTIGO 3.....	39
4.3.1	Fonte de dados e delineamento do estudo	39
4.3.2	População de Estudo	40
4.3.3	Variáveis do estudo	41

4.3.4	Processamento e análise dos dados	42
4.3.5	Aspectos Éticos	42
5	RESULTADOS E DISCUSSÃO	43
5.1	ARTIGO 1 – Publicado na Revista Brasileira de Saúde Ocupacional.....	43
5.1.1	Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho	43
5.2	ARTIGO 2 – Publicado na Revista de Saúde Pública.....	61
5.2.1	Predição de absenteísmo docente na rede pública com machine learning	61
5.3	ARTIGO 3 – Ainda não enviado para publicação conforme as regras do PPG-SP.....	79
5.3.1	Machine Learning para Predição de Doença Pulmonar Obstrutiva Crônica (DPOC) em adultos	79
6	CONCLUSÕES E CONSIDERAÇÕES FINAIS.....	100
7	REFERÊNCIAS BIBLIOGRÁFICAS.....	103
	ANEXOS.....	107
	Anexo A - Aprovação da Comissão de Ética em Pesquisa para o estudo de absenteísmo docente.	107
	Anexo B - Aprovação da Comissão de Ética em Pesquisa para o estudo de larga escala de coorte prospectiva do UK Biobank	109
	Anexo C - Aprovação da concessão de material. <i>Material Transfer Agreement</i>	117

1 INTRODUÇÃO

A relação doença e trabalho tem sido discutida há séculos, com destaque para a publicação da obra de Ramazzini¹, considerado pai da medicina do trabalho¹, no ano de 1700, em que foram descritas as doenças comuns entre grupos de trabalhadores como, por exemplo, os mineiros e os pedreiros. No estudo, são elencados os processos de trabalho de diversas ocupações, materiais utilizados e os agravos causados aos trabalhadores devido a longas exposições a riscos ambientais.

No estudo de Mendes e Dias², é possível acompanhar desde o surgimento da medicina do trabalho, inicialmente voltada para evitar o absenteísmo por doenças ou incapacitações permanentes, até a consolidação da área de saúde e segurança do trabalhador (SST), com a identificação de fatores de risco ambientais, intervenções nos locais de trabalho e valorização das condições físicas e psicológicas dos trabalhadores, assim como o surgimento de regulamentações para a proteção dos trabalhadores.

No Brasil, os riscos ambientais aos quais os trabalhadores estão expostos, podem ser categorizados em riscos físicos, químicos, biológicos, ergonômicos e de acidentes³. Alguns exemplos de riscos físicos são: excesso de ruído, vibrações, radiações ionizantes, radiações não-ionizantes, temperaturas muito altas ou muito baixas, quedas, entre outros. Em relação aos riscos químicos, podemos citar a exposição a gases como o benzeno e a inalação de poeiras minerais, entre outros. Os riscos biológicos incluem a infecção por micro-organismos, como vírus, bactérias e toxinas. Os riscos ergonômicos incluem esforço físico intenso, postura inadequada, jornadas de trabalho prolongadas e situações causadoras de estresse físico ou psicológico. Alguns exemplos de risco de acidentes são: iluminação inadequada, eletricidade, máquinas sem proteção, incêndio ou explosão.

O diagnóstico antecipado e o prognóstico de evolução de adoecimentos, podem melhorar a qualidade de vida dos trabalhadores. Neste sentido, o diagnóstico preciso, permite adotar medidas preventivas para evitar novos adoecimentos.

Os algoritmos de *machine learning* têm ganhado destaque na área da saúde tanto para diagnóstico de doenças^{4,5}, quanto para prognóstico na evolução das condições dos pacientes⁶. Sua popularidade tem aumentado devido aos ganhos de performance preditiva em comparação com modelos tradicionais, à diversidade de ferramentas para análise de dados e à capacidade de lidar com diferentes tipos de dados e de grandes volumes.⁷ As

bases desses algoritmos têm em comum o aprendizado estatístico e a teoria da probabilidade.⁸⁻¹⁰

Na área da SST, têm sido utilizados algoritmos de *machine learning* em estudos internacionais para a detecção de distúrbios relacionados ao trabalho¹¹, acidentes de trabalho¹², diagnóstico de doenças por imagem^{13,14}, e demais doenças relacionadas à exposição ocupacional, desde doenças ocasionadas por exposição prolongada a agentes físicos e químicos^{15,16}, a transtornos mentais¹⁷, apresentando, de modo geral, performance preditiva superior em comparação a técnicas convencionais.¹⁸

Observa-se um aumento de publicações que aplicam *machine learning* nas principais revistas científicas na área da saúde pública, em parte devido à maior disponibilidade de algoritmos¹⁹⁻²², ferramentas²³⁻²⁵, técnicas²⁶, interpretabilidade²⁷⁻²⁹ e confiança nos resultados obtidos, assim como a redução de custos computacionais. No entanto, ainda são escassos os estudos que citam o uso de algoritmos de *machine learning* na área de SST no Brasil.

Esta pesquisa apresenta um panorama das experiências de uso de algoritmos de *machine learning* na área de SST no Brasil e no mundo e implementa algoritmos para a predição de risco de absenteísmo por doença e diagnóstico antecipado de doença utilizando dados públicos e anônimos. A seguir são apresentadas algumas dessas experiências, para prognóstico e para diagnóstico na área da saúde pública e SST.

1.1 *Machine learning* para prevenção e prognóstico

Atualmente, já é possível encontrar na literatura estudos que relatam o uso de *machine learning* para prognóstico na área da saúde pública, seja na estimativa de tempo de sobrevivência de pacientes com câncer³⁰, evolução³¹ e vigilância de doenças contagiosas³², como a COVID-19, Influenza e outros vírus, medicina de precisão³³, como resposta a diferentes doses e medicamentos e uso de medicamentos direcionados às características do paciente, incluindo análise do genoma, efeitos do uso de medicamentos³⁴ em pacientes com sintomas de depressão e eventos de longo prazo³⁵, como o prognóstico de agravamento ou melhoria em pacientes internados.

Na área de SST, há estudos sobre o efeito da exposição contínua a agentes físicos³⁶⁻³⁸, incidência de transtornos mentais relacionados ao estresse ocupacional^{39,40}, lesões por esforço repetitivo (LER) e Distúrbios Osteomusculares Relacionados ao Trabalho (DORT)⁴¹, disfonia¹¹ e demais morbidades, mas ainda há poucos estudos em SST quando comparada com as demais áreas da saúde.

As técnicas utilizadas vão desde algoritmos que analisam dados não estruturados como textos, imagens e sons, até a análise de dados estruturados como tabelas em bases administrativas ou prontuários eletrônicos. Os algoritmos geram estimativas de risco para problemas de classificação em que o desfecho é categórico ou estimativas numéricas para problemas de regressão.

Reconhecer o risco de desenvolver uma morbidade dado o ambiente ao qual o trabalhador está inserido, pode ajudar a prevenir sua ocorrência, diminuir os casos ou minimizar os riscos ocupacionais⁴², adotando medidas preventivas⁴³. Além disso, muitos trabalhadores desconhecem os riscos ambientais em sua organização e, portanto, não sabem como se proteger. Ao conhecer os riscos, representados por probabilidades geradas por meio de modelos preditivos, pode-se adotar uma cultura de segurança⁴⁴ empresarial no intuito de evitá-los ou minimizá-los.

Nesta tese, abordaremos o uso de *machine learning*, utilizando dados públicos e anônimos para prever o absenteísmo por doença em profissionais com alta carga de estresse ocupacional.

1.2 *Machine learning* para diagnóstico

O uso de algoritmos de *machine learning* para diagnóstico é bastante vasto na saúde pública, seja no diagnóstico de doenças por imagem⁴⁵, com algoritmos de redes neurais artificiais¹⁴, no diagnóstico de doenças usando dados de exames laboratoriais⁴⁶ ou na reclassificação de diagnósticos por meio de análise de bases de dados administrativas⁴⁷ e relatos textuais.⁴⁸

Na área de SST, podemos citar o uso de algoritmos de *machine learning* na identificação de grupos com alto risco de diagnóstico positivo para doenças comuns e doenças raras relacionadas ao trabalho, como doenças respiratórias¹³, cânceres

relacionados ao trabalho⁴⁹, reabilitação após ferimentos⁵⁰, diagnóstico de transtornos mentais⁵¹, entre outros.

Nesta tese, será abordado o uso de algoritmos de *machine learning* em dados públicos e anônimos para prever o risco de diagnóstico positivo para doença pulmonar obstrutiva crônica.

1.3 Desafios e limitações

Apesar de se observar um aumento no uso de algoritmos de *machine learning* na área da saúde pública e em SST, é preciso considerar as características inerentes dos dados analisados.

Ao lidar com desfechos raros, como por exemplo, cânceres ocupacionais, é comum a presença de conjuntos de dados muito desbalanceados²⁶, em que há uma proporção muito maior de não-casos (sadios) do que de casos (doentes). Tal desbalanceamento, se não tratado, pode ocasionar problemas em que o algoritmo identifica muito bem os não-casos, porém, não consegue identificar os casos adequadamente, apresentando um baixo poder preditivo positivo, gerando um classificador ineficiente.

Além disso, após o desenvolvimento do algoritmo, deve-se considerar a aplicação prática no cotidiano, considerando o ambiente em que os profissionais atuam e selecionando somente as variáveis passíveis de serem coletadas e que realmente contribuam para a tomada de decisão a partir do modelo ajustado⁵². Uma outra questão é a generalização do algoritmo para as populações estudadas, em que um algoritmo pode ser extremamente eficiente para uma população específica, mas pode necessitar de novo treinamento e ajustes para outras populações. Isso não inviabiliza o uso do algoritmo, mas requer cautela ao aplicá-lo de forma generalista.

Portanto, é necessário conhecer esses desafios e limitações de maneira a tentar minimizá-los no desenvolvimento de modelos preditivos. A seguir, será realizada uma breve descrição dos tipos de aprendizagem de máquina, problemas e técnicas de balanceamento.

2 **MACHINE LEARNING: VISÃO GERAL, PROBLEMAS E MÉTRICAS.**

2.1 Tipos de aprendizado de máquina e pré-processamento

Em relação aos tipos de aprendizado de máquina, os tipos mais comuns são aprendizado supervisionado e não-supervisionado.⁵³ No aprendizado supervisionado, o desfecho de um conjunto de dados é conhecido e rotulado por um especialista no domínio. Há algoritmos para problemas de classificação (quando o desfecho é uma variável categórica) e outros para problemas de regressão (quando o desfecho é uma variável quantitativa contínua).⁵⁴ No aprendizado não-supervisionado, não existe uma variável resposta e o objetivo consiste, por exemplo, em agrupar populações parecidas (*clustering*) ou reduzir a dimensionalidade de um conjunto de dados, em que conjuntos com muitas variáveis independentes são reduzidos a dimensões menores para facilitar a análise, sem perder as características dos dados.

É importante mencionar que existem outras formas de aprendizagem como a aprendizagem semi-supervisionada e o aprendizado por reforço. Na aprendizagem semi-supervisionada, uma pequena parte dos dados é rotulada e combinada com uma grande parte de dados não rotulados⁵⁵; no aprendizado por reforço, o algoritmo aprende o que fazer de maneira a maximizar uma recompensa numérica.⁵⁶ Nesse caso, o algoritmo não recebe uma tarefa específica e deve descobrir por tentativa e erro quais ações geram a maior recompensa.

Na sequência, os dois tipos mais comuns de aprendizagem de máquina são apresentados em mais detalhes.

2.1.1 Aprendizagem supervisionada

No processo de aprendizagem supervisionada um especialista classifica (rotula) as observações de acordo com seu conhecimento no domínio. Os dados rotulados são utilizados para treinamento do modelo. Nos testes do modelo, retira-se o rótulo e o algoritmo prediz a observação de acordo com os padrões aprendidos durante a fase de treino. Os dados classificados pelo algoritmo são comparados com os dados de fato rotulados para a avaliação da performance preditiva.

A Figura 1 ilustra o processo de rotulagem e predição realizada em uma observação não rotulada para algoritmos de aprendizagem supervisionada.

Figura 1 - Processo de rotulagem e predição em novos dados

Fonte: Adaptado de Silva et al.⁵⁴

Conforme ilustrado na Figura 1, é possível observar um conjunto de dados representados por vetores $\rightarrow_x$, onde cada variável preditora x_i descreve uma característica do conjunto de dados, como por exemplo, cor, tamanho ou forma. A rotulagem é o processo de associar um valor por um especialista para as observações de um conjunto de dados. No exemplo da Figura 1, dado um vetor $\rightarrow_{x1}$, com variáveis independentes x_{11} , x_{12} , x_{13} , o algoritmo realiza a classificação da observação na classe estimada (y_1).

Para a construção de um modelo de aprendizagem supervisionado, após a rotulagem e definição de um conjunto de dados inicial, separa-se uma parte da amostra para realizar o treinamento (ex: 70%) e ajuste dos hiperparâmetros do modelo e outra parte da amostra (ex: 30%) para testes de desempenho do modelo.

Como exemplo de aprendizagem supervisionada, o estudo realizado por Correa et al.⁵⁷, utilizou uma série de imagens de manchas de pele, em que um especialista classifica algumas dessas imagens como tumores benignos ou malignos. O algoritmo é treinado com esses dados rotulados para realizar a classificação, utilizando características como assimetria e coloração. Nesse estudo foi utilizado o algoritmo de aprendizado supervisionado denominado máquina de vetor de suporte (em inglês, *Support Vector Machines*, SVM). O estudo obteve uma acurácia superior a 90% na classificação de tumores malignos.

O estudo de Goh e Ubeynarayana⁵⁸ utilizou o algoritmo SVM e outros algoritmos^{10,19} como *Random Forest*, K-NN, Árvores de Decisão, *Naïve Bayes* e

regressão logística, para classificar relatos de acidentes na área da construção civil em tipos de acidentes específicos, aplicando técnicas de mineração textual de dados.

A Figura 2 a seguir ilustra como o SVM classificaria em duas classes linearmente separáveis, como no estudo realizado por Correa et al.⁵⁷. O algoritmo tem o objetivo de separar os dados maximizando a margem entre as classes.

Figura 2 - Separação de margens entre classes com o uso do algoritmo SVM. Os círculos claros representam uma classe e os círculos escuros outra classe. As linhas tracejadas representam as diferentes opções para separar as classes.

Fonte: Adaptado de Han et al.⁵⁹

2.1.2 Aprendizagem não-supervisionada

A aprendizagem não-supervisionada, por sua vez, é frequentemente aplicada para problemas de agrupamento (também chamado de clusterização), ou para a redução de dimensionalidade. Nos casos de agrupamentos, os dados são analisados por suas similaridades ou suas dissimilaridades, e agrupados por meio de medidas de distância como, por exemplo, a distância euclidiana, que é a distância entre os pontos de um vetor $P = (p_1, p_2, \dots, p_n)$ e outro vetor $Q = (q_1, q_2, \dots, q_n)$, cuja fórmula é apresentada a seguir:

$$\sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

Uma forma de agrupamento não supervisionado é o agrupamento por partição, em que cada ponto deve pertencer ao menos a uma partição k , ou cluster, sendo que o desenvolvedor do modelo define inicialmente o número de partições. O algoritmo k -médias (do inglês, *k-means*), é um exemplo desse tipo de agrupamento e inicia selecionando aleatoriamente k centroides do conjunto de dados, e a cada iteração agrupa os pontos mais próximos ao centróide por meio da menor distância euclidiana. Os centroides são atualizados pela média dos pontos de cada cluster e o algoritmo termina (converge) quando não há mudança significativa dos centroides. A Figura 3 ilustra o uso do algoritmo *k-means* para diferentes configurações de centroides (k).

Figura 3 - 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 ($k=4$).

Fonte: Elaborado pelo autor.

É importante mencionar que existem outras técnicas de comparação entre vetores com outras medidas de similaridade como, por exemplo, distância de cossenos (*cosine similarity*), similaridade de Jaccard, entre outras medidas.^{60,61}

Como exemplo, o estudo realizado por Olson et al.⁶² utilizou o algoritmo *k-means* para agrupamento de perfis. Nesse estudo, 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.³⁹ 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 desses dois estudos foi identificar grupos parecidos de profissionais que se beneficiariam de tratamentos ou intervenções.

Outra técnica de aprendizagem não supervisionada bastante utilizada é a técnica denominada análise de componentes principais (do inglês, *Principal Component Analysis* - PCA), que objetiva a redução de dimensionalidade de conjuntos de dados. No PCA, são criadas novas variáveis, denominadas componentes principais, sendo que os primeiros componentes gerados visam capturar a maior variabilidade da combinação linear dos preditores¹⁰, conforme a fórmula a seguir:

$$PC_i = (a_{j1}x \text{ Preditor } 1) + (a_{j2}x \text{ Preditor } 2) + \dots + (a_{jn}x \text{ Preditor } N)$$

Onde i corresponde ao número do componente principal e j corresponde ao peso atribuído ao preditor para o componente atual.

2.1.3 Pré-processamento

Antes de executar os modelos de *machine learning*, é necessário transformar as variáveis de maneira que os algoritmos consigam processá-las da melhor forma possível. As variáveis quantitativas normalmente são padronizadas para evitar problemas de escala usando, por exemplo, o escore-z (*z-score*). As variáveis qualitativas com mais de duas categorias podem ser transformadas em variáveis *dummy*, em que cada categoria é transformada em uma nova variável binária com valores 0 ou 1, deixando-se uma das variáveis como referência. Quando todas as categorias são mantidas, o processo é conhecido como *one-hot encoding*.

Além disso, deve-se realizar a limpeza dos dados, removendo valores duplicados, bem como valores ruidosos ou aberrantes, assim como levar em consideração os valores ausentes (*missing*), podendo-se optar por excluir as observações correspondentes, ou realizar a imputação pela média, mediana, moda ou demais métricas, dependendo do tipo da variável. Para conjuntos com dados desbalanceados ou amostras pequenas, é importante considerar uma estratégia de imputação para evitar a perda de observações.

2.2 Sobreajuste, sub-ajuste e algoritmos de ML

Algoritmos de *machine learning* foram desenvolvidos para resolver problemas complexos, combinando diferentes técnicas para melhorar seu desempenho. O objetivo é obter a melhor estimativa utilizando um modelo matemático que consiga capturar a variabilidade dos dados e que consiga permanecer eficiente em novos conjuntos de dados. Há modelos que utilizam combinações de outros modelos na busca do melhor resultado possível. Essas combinações de múltiplos modelos preditivos são conhecidas como *ensemble*.

No entanto, assim como modelos de regressão e classificação tradicionais, os algoritmos de *machine learning* podem apresentar problemas como sub-ajuste (*underfitting*), quando o modelo não consegue se ajustar à variabilidade dos dados existentes e gera estimativas ruins, e sobreajuste (*overfitting*), quando o modelo se ajusta muito bem para a amostra utilizada no treinamento, mas não consegue bons resultados para novos dados.

Esses problemas podem ocorrer devido a diversos fatores, tais como tamanho amostral insuficiente, poucas variáveis preditoras, pré-processamento inadequado dos dados e alta colinearidade entre as variáveis preditoras.

Além disso, vieses (*bias*) podem ser introduzidos durante a etapa de construção dos modelos, no ajuste de parâmetros que melhor representam os dados utilizados. Por exemplo, problemas com alto viés (*high bias*) podem surgir quando uma função matemática simplista, que envolve poucos preditores, é utilizada e acaba subestimando os resultados, ocasionando o *underfitting*. Uma das formas de melhorar o modelo com essas características, seria introduzir novas variáveis preditoras ou adicionar variáveis polinomiais a partir das variáveis existentes. Caso o modelo tenha um número adequado de preditores, mas possua uma amostra insuficiente, pode ocorrer outro problema conhecido como alta variância (*high variance*) em que o modelo se ajusta muito bem à amostra de treinamento mas é ruim para lidar com novos dados. Nessa situação, acontece o sobre-ajuste (*overfitting*), em que o modelo se torna vulnerável a qualquer variabilidade dos dados como, por exemplo, a introdução de novas observações, e se torna ineficiente para novas previsões. O desafio ao construir modelos de *machine learning* é encontrar um equilíbrio entre viés e variância, conhecido como *Bias-Variance Tradeoff*.

Algumas técnicas podem ser aplicadas para melhorar a performance preditiva dos modelos como quando são utilizadas amostras pequenas e os modelos desenvolvidos possuem alta variância. O uso de penalizações e a introdução de novos dados sintéticos podem auxiliar na diminuição da alta variância do modelo. Esse tipo de abordagem é discutido na próxima seção.

2.2.1 Reamostragem, Validação Cruzada e *Ensemble*

Uma forma de melhorar o desempenho dos algoritmos e evitar problemas como *overfitting* ou *underfitting* é aplicar técnicas de reamostragem (*resampling*) e validação cruzada (*cross-validation*).

Existem técnicas de validação cruzada como, por exemplo, a técnica de *Leave-One-Out Cross Validation (LOOCV)* e a técnica *Leave-Group-Out Cross Validation (LGOCV)* que permitem definir, respectivamente, o número (n) ou a proporção (%) de dados destinados para validação. Por exemplo, na técnica *LOOCV*, separa-se uma observação para validação e o restante dos dados ($n-1$) é usado para treinamento do modelo. Repete-se o processo n vezes. O modelo tem sua performance aferida usando a média de todas as iterações. A seguir, é apresentada a fórmula para estimar o erro quadrático médio após a validação cruzada usando *LOOCV*:

$$CV_n = \frac{1}{n} \sum_{i=1}^n MSE_i$$

A técnica *LOOCV* é bastante eficiente, mas pode ser computacionalmente custosa, caso o número de repetições (n) seja muito alto.

Uma alternativa à técnica *LOOCV*, é usar a validação cruzada *K-Fold*, em que o conjunto de dados é dividido em k subconjuntos de aproximadamente mesmo tamanho. O primeiro subconjunto (*fold*) é usado para validação e os demais subconjuntos ($k-1$) são usados para treinamento do modelo. O processo é repetido k vezes, sendo que a cada iteração, um grupo diferente de observações é usado para validação.⁹ Os resultados são combinados para estimativa de performance. A Figura 4 a seguir ilustra o processo de validação cruzada *K-Fold* usando 5 subconjuntos ($K=5$):

Figura 4 – Exemplo de treinamento usando validação cruzada *5-Fold* (n=10, 5 grupos com oito observações para treinamento e duas observações para validação).

Fonte: Elaborado pelo autor.

A seguir, é apresentada a fórmula para estimar o erro quadrático médio por meio da validação *k-Fold*, usando-se a média de todas as iterações:

$$CV_k = \frac{1}{k} \sum_{i=1}^k MSE_i$$

Existem ainda outras técnicas de validação cruzada, como o *repeated k-fold*, em que o processo *k-fold* é executado *n* vezes. Por exemplo, cinco repetições de um *10-fold*, geram 50 iterações.

As técnicas de validação cruzada permitem melhorar o desempenho dos algoritmos e torná-los mais flexíveis quando lidam com novos dados. No entanto, é preciso ainda avaliar os demais fatores que possam gerar problemas de ajuste, como avaliar o tamanho da amostra. Existem técnicas de *resampling* capazes de lidar com amostras menores e que podem minimizar problemas como *overfitting* ou *underfitting*.

A técnica de *bootstrap* é uma técnica de *resampling* que, a partir do conjunto de dados original gera *n* amostras, a partir da seleção aleatória de observações do conjunto de dados original, sendo que as observações podem vir a ser selecionadas mais de uma vez (com reposição). As observações não utilizadas na reamostragem são utilizadas para predição. Portanto, são gerados *n* modelos em que o resultado final é dado pela média da predição destes modelos. A Figura 5, a seguir, ilustra como funciona o *Bootstrap*.

Figura 5 - Reamostragem com *Bootstrap*

Fonte: Elaborado pelo autor

Uma das técnicas mais conhecidas que usa essa abordagem é o *Bootstrap Aggregation* ou *Bagging*.

A partir dos dados originais, são gerados n conjuntos de dados com a técnica de *bootstrap* aos quais modelos preditivos são ajustados utilizando um algoritmo de *machine learning* simples (Ex: árvores de decisão). As predições dos modelos são então combinadas. Para problemas de classificação, é utilizado o voto da maioria (*majority vote*), em que a categoria predita com maior ocorrência nos modelos gerados é a vencedora. Para problemas de regressão, é utilizada a média das predições. A técnica de *bagging* pode ser utilizada para reduzir alta variância, pois utiliza a média ou a maioria de votos de diferentes modelos, tornando-o mais flexível. É comum a sua aplicação em modelos como árvores de decisão que sofrem deste tipo de problema⁶, resultando em *ensembles* chamados de *Bagging trees*.

Além das técnicas de *resampling* também podem ser aplicadas técnicas como o balanceamento de carga para que o algoritmo não fique condicionado à classe com maior frequência. Na próxima seção, apresentamos opções de balanceamento de classes em problemas de classificação, assim como métricas de comparação de algoritmos de classificação.

2.2.2 Balanceamento de classes

Quando as classes estão desbalanceadas, os algoritmos de *machine learning* acertarão mais a categoria majoritária, mas terão um desempenho ruim na predição da

categoria minoritária. Técnicas de balanceamento, como *oversampling* e *undersampling*, podem ser utilizadas para minimizar este comportamento. Na técnica de *oversampling* são gerados dados sintéticos das classes minoritárias, resultando em um conjunto de dados mais balanceado. No entanto, pode-se perder parte da variabilidade dos dados, pois dados da classe minoritária são replicados. Já na técnica de *undersampling*, algumas observações da classe majoritária são excluídas ao acaso do conjunto de dados, sem geração de dados sintéticos, de modo que a proporção dos dados seja equânime. Também há uma outra forma que realiza tanto *oversampling* para geração de novos dados sintéticos da classe minoritária, e *undersampling* para diminuição das observações da classe majoritária.

Alguns softwares como o R e o Python já possuem pacotes para o balanceamento de classes como os pacotes ROSE (*Random Over-Sampling Examples*)⁶³ e SMOTE (*Synthetic minority over-sampling technique*).⁶⁴

A seguir, são apresentados alguns dos algoritmos de *machine learning* utilizados nos modelos desenvolvidos nesta pesquisa.

2.2.3 Algoritmos de machine learning

2.2.3.1 Árvores de Decisão

Modelos baseados em árvore geram regras condicionais aninhadas (do tipo se então, ou *if...else*) para particionar o conjunto de dados em setores menores.¹⁰ Em outras palavras, dado um conjunto de regras (*if...else*) é realizada uma decisão para definição do desfecho. As árvores de decisão podem ser aplicadas tanto para regressão quanto para classificação. Resultam em modelos facilmente interpretáveis devido à sua simplicidade. No entanto, podem apresentar problemas de sobreajuste e alta variância, em que o modelo se ajusta muito bem aos dados de treino mas é ruim ao lidar com novos dados, pois pequenas variações nos dados têm um impacto grande na predição do modelo.

A árvore de decisão gera regras de acordo com as variáveis preditoras utilizadas, como no exemplo hipotético da Figura 6.

Figura 6 – Regras hipotéticas geradas para classificar um tumor em benigno (N) ou maligno (S).

Fonte: Elaborado pelo autor.

Conforme observado na Figura 6, de acordo com os valores das variáveis utilizadas, identifica-se a ocorrência do desfecho como S (Sim) ou N (Não).

Os algoritmos de árvore são utilizados em diversos algoritmos do tipo *ensemble* como o *Bagging Trees* descritos na seção 2.2.1 e no *Random Forest* e *eXtreme Gradient Boosting (XGBoost)*, descritos a seguir.

2.2.3.2 Random Forest

Random Forest é um algoritmo do tipo *ensemble* bastante popular que surgiu no início da década de 2000¹⁹. Este algoritmo combina resultados de classificadores fracos (como as árvores de decisão) para chegar a um resultado final. Assim como no *bagging trees*, são utilizadas amostras *bootstrap* e, adicionalmente, para cada divisão (*split*) de árvores a partir do topo, são selecionados aleatoriamente diferentes preditores. Ao alcançar um limite de divisões e árvores definidas pelo especialista, utiliza-se a média de todas as previsões das árvores ajustadas para gerar o valor predito pelo modelo para problemas de regressão, ou o voto majoritário para problemas de classificação.

O algoritmo lida apenas com variáveis numéricas, portanto, antes de utilizá-lo, é necessário realizar o pré-processamento do conjunto de dados para a transformação de variáveis categóricas.

A Figura 7, a seguir, ilustra um exemplo do processo de treinamento do algoritmo.

Figura 7 – Processo de Treinamento do algoritmo *Random Forest*. O algoritmo treina n árvores em amostras *bootstrap*. Os nós cinzas representam as divisões da árvore. Os nós amarelos representam uma categoria e os nós azuis outra categoria.

Fonte: Elaborado pelo autor.

Na Figura 7 é possível observar que a partir da amostra separada para treinamento, o algoritmo cria n árvores, e para cada árvore, utiliza uma amostra *bootstrap* dos dados de treino. Então cada árvore é treinada na amostra *bootstrap* com diferentes variáveis predictoras em cada nó para minimizar o sobreajuste.

A Figura 8, a seguir, ilustra como o algoritmo classificaria um novo dado após realizada a etapa de treinamento.

Figura 8 – Classificação por voto majoritário. Dado um vetor $\vec{x}$ simulando novos dados, o algoritmo utiliza o resultado combinado das n árvores para definir a categoria predita.

Fonte: Elaborado pelo autor.

Na Figura 8, é possível observar a classificação realizada por cada árvore em que a classificação final do algoritmo é dada pelo voto majoritário.

A seguir, abordaremos um outro algoritmo, também baseado em árvore, que utiliza a técnica de *Boosting* para realizar sua predição.

2.2.3.3 eXtreme Gradient Boosting

O algoritmo eXtreme Gradient Boosting ou *XGBoost* também é um algoritmo *ensemble* baseado em árvore, capaz de lidar com um grande volume de observações utilizando processamento paralelo.²⁰ A proposta do algoritmo é treinar um modelo e usar os resíduos deste modelo (diferença entre valor predito e valor observado) para treinar os modelos subsequentes (*Gradient Boosting*). Também usa classificadores fracos, como algoritmos de árvore de decisão (*Gradient Boosting trees*), para produzir um classificador robusto que combina os resultados de dois ou mais destes classificadores para obter resultados superiores a um único classificador.⁶⁵ A cada nova iteração do algoritmo, os pesos são reajustados de maneira que as observações erradas ganham pesos maiores e as observações corretas, pesos menores. O *XGBoost* pode utilizar computação paralela em *hardware* convencional, o que o torna rápido e bastante popular.

A Figura 9, a seguir, mostra o processo de treinamento usando *gradiente boosting trees*, processo usado no algoritmo do *XGBoost*.

Figura 9 - Treinamento usando *Gradient Boosting Trees* para regressão. Passo 1) Inicia-se treinando a primeira árvore com os dados da amostra de treino. Passo 2) as demais árvores são treinadas com os resíduos da anterior e a cada iteração. 3) soma-se à predição inicial as predições das demais árvores para estimar o valor final.

Fonte: Elaborado pelo autor.

Além do processo ilustrado na Figura 9, o algoritmo conta com vários hiperparâmetros para prevenir *overfitting*, como a taxa de aprendizado (*eta* ou *learning rate*), que possibilita configurar, a cada passo, os pesos para a predição final do modelo, regularização (*lambda*), que permite reduzir a sensibilidade a dados extremos, poda da árvore (*pruning*, parâmetro: *gamma*), profundidade da árvore (*max_depth*), entre outros.

A seguir, abordaremos outros algoritmos que não são baseados em árvores, mas que também foram utilizados nos artigos desta pesquisa.

2.2.3.4 Support Vector Machines

O algoritmo *Support Vector Machines* (SVM) surgiu na década de 1960. Com o objetivo de classificar grupos distintos, o SVM visa maximizar a distância, chamada de margem, entre as diferentes categorias. Para tanto, usam-se observações limítrofes,

chamadas de vetores de suporte para estabelecimento das margens. Quanto maior a distância entre as categorias, melhor a classificação realizada. Um exemplo desse algoritmo foi ilustrado anteriormente na Figura 2 na seção 2.1.1.

2.2.3.5 Redes Neurais Artificiais

As redes neurais artificiais são inspiradas na biologia e baseiam-se em representações de um cérebro humano, onde o algoritmo tenta de certa forma mimetizar o comportamento de um neurônio. Cada neurônio tem uma ou mais entradas (dendritos) e, após o processamento dessas entradas, devolve os valores processados em uma ou mais saídas, simulando o terminal de um axônio. Uma rede neural pode ter um ou mais neurônios encadeados e múltiplas camadas ocultas de processamento (nesse caso, conhecidas como *deep learning*) para a realização de funções mais complexas. O processo de classificação ocorre após funções de ativação, que são funções matemáticas acionadas em geral quando a soma ponderada chega a um limiar estabelecido.⁶⁶ A Figura 10 a seguir ilustra um modelo de rede neural artificial.

Figura 6 - Comparação de uma rede neural artificial em relação à representação de neurônios biológicos. a) Representação de um neurônio humano. b) Representação matemática de um neurônio artificial. c) Representação da estrutura de uma rede de neurônios humanos interligados. d) Representação da estrutura de uma rede neural artificial com mais de uma camada oculta.

Fonte: Adaptado de Gonçalves et al. (2013).⁶⁷

Para realizar a classificação, cada ligação de entrada (x_i) recebe um peso numérico (w_i) que determina a força do sinal, podendo ser compreendido como a representatividade

de cada atributo descritivo do conjunto de dados (x_i) (vide Figura 10-b). Utiliza-se, então, uma soma ponderada, com o auxílio da função de entrada, e em seguida, é aplicada uma função de ativação para processar o valor obtido pela soma ponderada. Na função de ativação, pode-se utilizar diferentes modelos estatísticos de acordo com o problema em análise. Para problemas de classificação simples é comum o uso da função logística. Os pesos são atualizados a cada execução (iteração) do algoritmo, normalmente calculando-se a diferença entre o resultado esperado, que já foi rotulado, e o resultado obtido. Esse processo é denominado treinamento da rede neural. A quantidade de neurônios e de camadas pode variar e representam os hiperparâmetros do algoritmo. Estruturas com mais de uma camada oculta, são conhecidas como *deep learning*.

A seguir são apresentadas métricas para comparação do desempenho dos modelos.

2.3 Métricas para avaliação de modelos preditivos

2.3.1 Métricas para modelos de classificação

Os modelos de classificação resultam em probabilidades individuais ou categoria predita. Para estes modelos, as métricas utilizadas são: acurácia (proporção de acertos), sensibilidade (a razão de verdadeiros positivos preditos corretamente sobre os verdadeiros positivos), especificidade (a razão de verdadeiros negativos preditos corretamente sobre verdadeiros negativos), precisão (razão de verdadeiros positivos sobre verdadeiros positivos e falsos positivos), também chamado de valor predito positivo (VPP), e a relação entre precisão e sensibilidade, chamado de F-Score. A métrica F-Score é a média harmônica entre precisão (VPP) e sensibilidade, variando de 0 a 1, em que 1 representa precisão e sensibilidade perfeitas. Esta métrica pode auxiliar na escolha do melhor classificador, pois resume a relação entre duas variáveis.

A análise dos valores preditos e observados pode ser feita visualmente utilizando uma tabela chamada de matriz de confusão.

A Tabela 1 a seguir apresenta a matriz de confusão entre valores preditos e observados na classificação de uma doença hipotética.

Tabela 1 - Matriz de Confusão

Predição	Realidade	
	Doentes	Não Doentes
Doentes	Verdadeiros Positivos (VP)	Falsos Positivos (FP)
Não Doentes	Falsos Negativos (FN)	Verdadeiros Negativos (VN)

A partir da Tabela 1, pode-se calcular a acurácia, sensibilidade, especificidade, precisão e escore F1 do modelo:

- **Acurácia:** $\frac{\text{Verdadeiros positivos} + \text{Verdadeiros Negativos}}{\text{Total}}$
- **Sensibilidade:** $\frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}}$
- **Especificidade:** $\frac{\text{Verdadeiros Negativos}}{\text{Falsos Positivos} + \text{Verdadeiros Negativos}}$
- **Precisão:** $\frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}}$
- **F1 Score:** $2 * \frac{\text{precisão} * \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}}$

Em situações de doenças raras como cânceres ocupacionais em grandes populações, seria comum encontrar muito mais observações de não doentes do que de doentes. Neste cenário, identificar os doentes é mais importante do que identificar os não doentes. O algoritmo poderia alcançar boa acurácia, porém teria uma baixa sensibilidade na identificação de doentes. Nessas situações, é importante analisar não apenas a acurácia, mas também o equilíbrio entre sensibilidade e especificidade, assim como a precisão do modelo, também conhecida como valor predito positivo (VPP).

A avaliação de performance de modelos preditivos se torna ainda mais complexa quando é necessário comparar todas essas métricas em vários modelos distintos. Uma estratégia que permite visualizar e comparar os resultados destas métricas consiste em utilizar o valor da área abaixo da curva ROC (*Receiver Operating Characteristics*).⁶⁸ A curva ROC é originária da área de engenharia e surgiu para analisar a performance na detecção de sinais elétricos, popularizando-se na avaliação de modelos classificatórios. A Figura 11 a seguir ilustra uma curva ROC.

Figura 11 - Exemplo da relação entre taxa de falso positivo (FPR) e taxa de verdadeiro positivo (TPR) pela curva ROC

Fonte: Elaborado pelo autor

A curva ROC mostra a relação entre sensibilidade e especificidade para a escolha do melhor ponto de corte (*cutoff point*) para a probabilidade predita pelo modelo ajustado ao problema analisado. Em algumas situações é melhor ter maior sensibilidade do que especificidade, como no caso de diagnóstico de doenças em que é mais importante acertar mais os casos de adoecimento do que os casos negativos. Quanto mais à esquerda e para cima estiver a curva, melhor será a discriminação do modelo ajustado.

Para facilitar a tarefa de comparação da performance de diferentes modelos preditivos, é comum utilizar a área abaixo da curva (AUC), em que um único valor é usado para a seleção do melhor modelo.

A seguir são apresentadas as métricas para problemas de regressão, em que se tem como variável resposta uma variável contínua.

2.3.2 Métricas para modelos de regressão

Para os modelos de regressão, busca-se minimizar a diferença entre o valor predito ($\hat{y}_i$) e valor observado (y_i). Quanto mais próximo estiver o valor predito do observado, melhor ajustado estará o modelo. Uma forma comum de calcular o erro de um modelo é por meio do erro quadrático médio (EQM)⁹, dado por:

$$EQM = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

A raiz quadrada do erro quadrático médio (do inglês, *Root Mean Square Error - RMSE*) é frequentemente utilizada na comparação de modelos. A Figura 12 a seguir ilustra como é calculado o erro para um modelo hipotético:

Figura 12 – Exemplo de cálculo usando a métrica da raiz quadrada do erro quadrático médio.

Fonte: Elaborado pelo autor.

Quanto menor o valor do erro, mais próximas dos dados observados estarão as previsões e, portanto, diz-se que o modelo está melhor ajustado.

A Figura 13 a seguir ilustra a comparação entre dois modelos ajustados a dados fictícios.

Figura 13 – Comparação entre dois modelos hipotéticos e suas proximidades com o valor observado. O modelo A representa um modelo linear. O modelo B representa um modelo quadrático mais próximo aos dados observados e, portanto, com menor EQM.

Fonte: Elaborado pelo autor.

3 OBJETIVOS

3.1 OBJETIVO GERAL

Utilizar algoritmos de *machine learning* para criar modelos preditivos capazes de identificar desfechos de saúde em trabalhadores.

3.2 OBJETIVOS ESPECÍFICOS

1. Descrever o panorama de uso de algoritmos de *machine learning* usados para predição na área de saúde pública e SST por meio de revisão bibliográfica.
2. Desenvolver e avaliar modelos preditivos de absenteísmo por doença ou doença relacionado ao trabalho em professores expostos a condições com alto grau de estresse.
3. Desenvolver e avaliar modelos preditivos de DPOC em trabalhadores supostamente expostos a agentes químicos.

4 MÉTODOS

Para responder aos objetivos desta tese, foram desenvolvidos três artigos científicos. O primeiro artigo realiza uma revisão bibliográfica para identificar a aplicação de algoritmos de *machine learning* na Saúde Pública e em SST. O segundo artigo utiliza dados públicos e anônimos para desenvolver um modelo preditivo capaz de identificar professores com alto risco de afastamento por doenças ou doenças relacionadas ao trabalho. O último artigo utiliza dados do UK Biobank para desenvolver modelos preditivos de DPOC em trabalhadores supostamente expostos a poeiras minerais.

A seguir, são apresentados os métodos específicos empregados em cada um dos artigos realizados.

4.1 MÉTODO DO ARTIGO 1

4.1.1 Delineamento do estudo

Para atender ao objetivo específico de descrever o panorama do uso de *machine learning* na saúde pública e em SST, foi realizada uma revisão bibliográfica em um dos principais repositórios de artigos científicos, a Biblioteca Virtual em Saúde (BVS). Foram pesquisados os termos em inglês *machine learning* e *data mining*. Esse último termo foi

adicionado à pesquisa por referir-se a uma área correlata que inclui o uso de algoritmos de *machine learning*. A consulta foi restrita apenas aos títulos dos artigos para facilitar a identificação dos temas na área da saúde no período entre 2016 e 2018, resultando em 3458 artigos.

4.1.2 Revisão e análise dos artigos

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. Os artigos que geraram dúvida em relação ao uso na saúde humana foram analisados individualmente. Após a remoção de duplicidades e artigos não relacionados, foram identificados 3283 artigos, sendo 623 em 2016, 888 em 2017 e 1772 em 2018. Portanto, observou-se um crescimento constante no uso de algoritmos de *machine learning* nas áreas da saúde pública e SST.

A maioria dos artigos observados utilizaram algoritmos de *machine learning* para resolver problemas de classificação ou de regressão com algoritmos de aprendizado supervisionado, ou para agrupamento de perfis com o uso de aprendizagem não-supervisionada.

O artigo apresenta um resumo dos principais algoritmos utilizados, tanto para aprendizagem supervisionada, quanto para aprendizagem não-supervisionada, bem como os diferentes objetivos de pesquisa.

Este manuscrito foi publicado na Revista Brasileira de Saúde Ocupacional (RBSO).

4.2 MÉTODO DO ARTIGO 2

Este artigo refere-se ao objetivo específico de desenvolver um modelo de prognóstico para predição de absenteísmo docente por doença ou doença relacionada ao trabalho.

4.2.1 Fonte de dados e delineamento do estudo

Foi utilizada a base de dados anônima da Relação Anual de Informações Sociais (RAIS) para realizar um corte transversal dos registros individualizados (microdados)⁶⁹ no período entre 2014 e 2018, de vínculos empregatícios de professores que atuam na

educação infantil da rede pública municipal de todos os municípios do Estado de São Paulo, que informaram dados à RAIS.

A RAIS mantém informações fornecidas anualmente por empresas públicas e privadas sobre os vínculos empregatícios formais de seus funcionários, tendo informações sobre contratações e demissões realizadas no ano corrente, assim como informações sobre jornada de trabalho e afastamentos por doença ou doença relacionada ao trabalho.

4.2.2 População de estudo

Foram incluídos professores da rede pública municipal de todos os municípios do Estado de São Paulo que informaram dados à RAIS, referentes a docentes com nível superior e vínculo empregatício ativo até 31/12 do ano corrente, atuando no ensino pré-escolar, classificados na RAIS com a Classificação Brasileira de Ocupações (CBO) 231105 (Professores de Ensino Superior no Ensino Infantil, quatro a seis anos).

Como forma de identificação do desfecho, foram utilizados os campos de causa de afastamento presentes na RAIS, filtrando-se pelos códigos 30 (doença de trabalho) e 40 (doença). Não foram consideradas aposentadorias decorrentes de doenças. Para a identificação do tipo de estabelecimento (federal, estadual, municipal ou privado), foi utilizada a informação da natureza jurídica do estabelecimento (códigos 1031,1066,1120,1155,1180,1244), filtrando somente pelos estabelecimentos municipais. Para a identificação da atividade do estabelecimento, foram utilizados os códigos 841160 (administração pública em geral) e a divisão 85 (educação) da Classificação Nacional de Atividades Econômica versão 2.0 (CNAE 2.0) fornecidos na RAIS.

Do total de registros, 99,7% (n=174.346) estavam cadastrados no CNAE 841160, sendo que o restante, 0,03% (n=52), foi excluído por pertencer a um CNAE não relacionado ao ensino infantil pré-escolar. Também foram excluídos registros em que o município de exercício do trabalhador não correspondia à unidade federativa de São Paulo, resultando no total de 174.294 registros. Destes, 75.680 (43,42%) tiveram ao menos um afastamento por doença ou doença relacionada ao trabalho. Nas situações em que o município de exercício do trabalhador não foi informado, foi adotado o município do estabelecimento.

Para capturar parte das condições de trabalho a que esses profissionais estavam expostos, foi vinculada a informação da média de alunos por turma na etapa pré-escolar

por município, fornecida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)⁷⁰, considerando que algumas doenças comuns aos docentes estão associadas à acústica do ambiente e à complexidade relacionada à atuação do docente⁷¹, e adicionado o número de habitantes por município, fornecido pela Fundação Seade.⁷² Os dados do INEP e da Fundação SEADE foram vinculados à amostra selecionada da RAIS por meio do código do município segundo o Instituto Brasileiro de Geografia e Estatística (IBGE).

4.2.3 Variáveis do estudo

Ao todo, foram selecionadas onze variáveis preditoras: sexo, idade, pós-graduação, tamanho do estabelecimento (porte da escola de acordo com o número de funcionários), tipo de vínculo empregatício, primeiro emprego, tempo no emprego, horas contratadas, média anual da quantidade de salários mínimos, média de alunos por turma no município e número de habitantes no município de atuação do trabalhador. A variável resposta foi representada pela combinação dos afastamentos por doença ou por doença relacionada ao trabalho, sendo considerada positiva na presença de pelo menos um afastamento, e negativa caso contrário.

As variáveis média de alunos por turma e população, apesar de se referirem ao município e não ao indivíduo, foram utilizadas como um *proxy* para capturar parte das condições de trabalho em que os docentes atuaram, dada a limitação de se utilizar dados anônimos e não haver a identificação da escola específica em que o trabalhador atuava.

4.2.4 Processamento e análise dos dados

A Figura 14 a seguir apresenta todo o processo, desde a carga dos dados, até a análise da performance dos modelos preditivos.

Figura 7 - Processamento e análise dos dados

Fonte: Elaborado pelo autor

Os dados passaram por uma etapa inicial de pré-processamento, para tratar as variáveis ausentes e transformar os dados para que pudessem ser utilizados na construção dos diferentes tipos de modelos preditivos de *machine learning*¹⁰. Dentre as variáveis selecionadas, todas estavam totalmente preenchidas. Variáveis com mais de duas categorias foram representadas por variáveis *dummy*, em que para cada categoria, é gerada uma nova variável com valores 0 ou 1. As variáveis contínuas foram padronizadas por meio de escore-z (z-score). Foi testada a correlação entre as variáveis numéricas e foi constatada correlação muito forte entre o número de habitantes no município e a média de alunos (0,93), assim optou-se por dicotomizar a variável de número de habitantes utilizando a definição de 500 mil habitantes, para identificação de atuação em uma cidade grande⁷³, de maneira a capturar parte das dificuldades de atuação destes trabalhadores inerentes a diferentes tamanhos de cidades.

Para a fase de pré-processamento e carga de dados, foi utilizado o sistema gerenciador de banco de dados (SGBD), MS SQL Server. Para a análise de dados e construção dos modelos preditivos foi utilizado o software R.

4.2.5 Aspectos éticos

O projeto foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo, sob o nº 4.031.362, CAAE 30786620.5.0000.5421.

Este manuscrito foi publicado na Revista de Saúde Pública (RSP). (ANEXO II).

4.3 MÉTODO DO ARTIGO 3

Este artigo refere-se ao objetivo específico de desenvolver um modelo de diagnóstico para a predição de DPOC em trabalhadores supostamente expostos à poeiras minerais.

4.3.1 Fonte de dados e delineamento do estudo

Foi utilizada a base de dados do UK Biobank. A base de dados do UK Biobank é uma grande coorte prospectiva com dados socioeconômicos e biomédicos de mais de 500.000 participantes. A base de dados possui informações coletadas por meio de questionários, avaliações presenciais e exames laboratoriais, incluindo histórico ocupacional dos participantes. Os participantes foram recrutados no período entre 2006 e

2010⁷⁴ no Reino Unido com idades entre 40 a 69 anos. Foram considerados os dados da primeira entrevista, assim como os primeiros exames laboratoriais (*baseline*).

4.3.2 População de Estudo

Para esta análise, foram selecionados os participantes que preencheram o questionário de histórico ocupacional, totalizando 120.289 participantes. O desfecho considerado foi diagnóstico de DPOC, auto-relatado ou diagnosticado por internação hospitalar, ou enfisema pulmonar diagnosticado por médicos durante o acompanhamento médico da coorte. Sabe-se que o termo enfisema já foi utilizado como sinônimo de DPOC e hoje é entendido como um componente fisiopatológico da doença.⁷⁵ Foram considerados os casos diagnósticos a partir de 40 anos ou mais, tanto para DPOC quanto para enfisema. Dessa forma, os modelos desenvolvidos com essa população são mais genéricos e aplicam-se à população geral com ou sem exposições específicas.

Em seguida, foi aplicado um filtro para criação de um subconjunto de dados contendo participantes que tiveram algum histórico ocupacional de atividades vinculadas à construção civil, mineração e pavimentação, utilizando os 4 primeiros dígitos da classificação padronizada de ocupações (SOC 2000⁷⁶) do campo de código de ocupação do UK Biobank (*Jog Coding*; 22601). Essas ocupações foram selecionadas devido ao conhecimento de sua maior exposição a agentes químicos, como poeiras minerais, como a sílica, podendo apresentar consequentemente uma maior prevalência de DPOC.⁷⁷⁻⁷⁹ A seleção resultou em 7.628 observações para serem utilizadas para treinamento e testes de modelos especializados.

A Figura 15, a seguir, ilustra o processo de seleção da população de estudo e dos desfechos.

Figura 8- Processo de Seleção de população. a) Seleção de população para construção de modelos genéricos. b) Seleção de população para construção de modelos especializados.

Fonte: Elaborado pelo autor.

4.3.3 Variáveis do estudo

Inicialmente foram selecionadas 40 variáveis, como variáveis sociodemográficas, hemograma, medidas antropométricas, comportamentais (tabagismo) e sintomas respiratórios, independentemente do histórico ocupacional. Em seguida foi utilizada a técnica de seleção de variáveis (*feature selection*) conhecida como Boruta⁸⁰ em todo o conjunto de dados, com as 40 variáveis iniciais. Essa técnica visa identificar as variáveis mais relevantes para o desfecho, podendo melhorar a performance do modelo desenvolvido. A técnica de seleção de variáveis identificou 22 variáveis relevantes: idade, IMC, hematócritos, contagem de leucócitos, Volume Corpuscular Médio (VCM), fosfatase alcalina, contagem de monócitos, contagem de neutrófilos, frequência de pulso, proteína C reativa, creatinina, HDL, LDL, uréia, fumante/ex-fumante, tosse crônica, expectoração, dispneia, histórico de asma, contagem de linfócitos, albumina e triglicérides. Além das 22 variáveis, foram adicionadas mais 4 variáveis para complementar dados de hemograma, comorbidades, exposição ao tabaco e histórico ocupacional relacionadas à indústria da construção: eosinófilos, exposição ao tabaco no trabalho, exposição ao tabaco em casa e histórico de trabalho na construção civil, totalizando 26 variáveis.

4.3.4 Processamento e análise dos dados

Para evitar que os modelos desenvolvidos priorizassem a classe majoritária devido à característica de alto desbalanceamento dos dados em relação aos casos e não casos de DPOC, foi utilizada a técnica de subamostragem aleatória (*random undersampling*) nos dados de treinamento, em que as observações da classe majoritária são selecionadas aleatoriamente para exclusão, de maneira a melhorar a proporção de casos e não casos.²⁶

As variáveis foram pré-processadas para que pudessem ser utilizadas nos modelos preditivos. As variáveis com mais de uma categoria foram transformadas em *dummy* e as variáveis contínuas foram padronizadas pelo z-score. Também foi feita a análise de correlação entre as variáveis numéricas. Variáveis com correlação superior a 0.90 (ex: hematócrito e hemoglobina), foram manualmente descartadas, mantendo apenas uma das variáveis (ex: hematócrito).

Todas as análises foram realizadas utilizando a linguagem de programação Python com a biblioteca *scikit-learn*.

4.3.5 Aspectos Éticos

O projeto do UK Biobank recebeu aprovação do Comitê de Ética em Pesquisa Multicêntrico do Noroeste (*North West Multi-Centre Research Ethics Committee*) sob o número 16/NW/0274 renovando a referência anterior 11/NW/0382 e obteve consentimento escrito de todos os participantes anteriormente ao início do estudo.

Este manuscrito ainda não foi submetido a nenhum periódico seguindo as normas do Programa de Pós-Graduação em Saúde Pública da Faculdade de Saúde Pública da USP para defesa de teses.

5 RESULTADOS E DISCUSSÃO

5.1 ARTIGO 1 – Publicado na Revista Brasileira de Saúde Ocupacional.

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

Página de Título

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

Fernando Timoteo Fernandes^{1,2}
ORCID: <https://orcid.org/0000-0003-2548-9685>

Alexandre Dias Porto Chiavegatto Filho²
ORCID: <http://orcid.org/0000-0003-3251-9600>

¹ Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho. São Paulo, SP, Brasil.

² Universidade de São Paulo, Faculdade de Saúde Pública. São Paulo, SP, Brasil.

Contato:

Fernando Timoteo Fernandes

E-mail:

fernando.fernandes@fundacentro.gov.br

RESUMO

Introdução: a variedade, volume e velocidade de geração de dados (*big data*) possibilita 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 Segurança e Saúde 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 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 new and more complex analyzes.

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 on OSH. **Results:** literature shows that data mining with machine learning algorithms for predictive purposes on OSH and public health tend to 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 possible analyze large volume of morbidity and mortality data. These techniques are not meant to replace human from the equation, but rather to assist in decision-making processes, to serve as a tool for the statistical analysis of OSH data and to discover new information that may support public prevention policies.

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

1 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), é ao mesmo tempo um desafio e uma oportunidade para pesquisas na área da Segurança e Saúde no Trabalho (SST). Essa variedade e grande quantidade de dados é conhecida como *big data* e tem levado a uma mudança nas formas tradicionais de análise de dados¹.

Sistemas gerenciadores de bancos de dados, como o MySQL* e *softwares* estatísticos como SAS, SPSS e outros já lidam com grandes volumes de dados há anos. No entanto, dificuldades de escalabilidade e processamento podem surgir ao atualizar ou adicionar novos conjuntos de dados², de diferentes tipos e fontes e em tempo real, como nos casos de dados 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 então 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 dispendir 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 de todo o 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 um treinamento prévio realizado por um especialista³. Para facilitar o entendimento das técnicas empregadas nos estudos demonstrados 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 (Prefácio, p.VIII) . Tais agentes buscam executar suas ações de maneira a maximizar suas chances de sucesso para algum objetivo. 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 (1950), em que se discutia de forma informal conceitos de inteligência e aprendizado de máquina⁴ sobre o que poderia ser considerada uma máquina inteligente. A inteligência pode estar associada à tomada de decisão racional em que o agente busca alcançar o melhor resultado, ou na impossibilidade, o melhor resultado esperado de forma autônoma³.

Nesse cenário de automatização e busca de melhores resultados, surgem processos analíticos que auxiliam na tomada de decisão como, por exemplo, 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*, 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 previsões ou classificações dos dados disponíveis.⁶ Há duas grandes técnicas de *machine learning*, a aprendizagem supervisionada e a aprendizagem não-supervisionada, que também serão abordadas ao longo deste manuscrito.

Atualmente, o estudo de algoritmos de *machine learning* tem ganhado destaque devido à sua 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}, previsão de riscos⁹ e biomedicina¹⁰. Tais técnicas também têm sido empregadas como ferramentas complementares em estudos epidemiológicos.¹¹

Neste ensaio serão discutidos exemplos de uso de *data mining* e *machine learning* na área da saúde e segurança no trabalho (SST) e seus resultados, de forma a explorar sua aplicação nessa área 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* de aprendizagem não-supervisionada e aprendizagem supervisionada. Em seguida, serão apresentados exemplos de estudos que utilizam combinações de diferentes técnicas de *machine learning* para a previsão de riscos ocupacionais e outras perspectivas de análises.

2 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, em que 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 então para a etapa de seleção dos dados, em que são filtrados os dados relevantes para a pesquisa. Em seguida os dados são transformados e consolidados de acordo com os propósitos da mineração. Realize-se então o *data mining*, em que são aplicadas técnicas para descoberta de padrões nas bases de dados, por meio de algoritmos computacionais⁵. Nesta fase podem ser utilizados algoritmos de *machine learning* para tarefas de previsão, associação ou agrupamento dos dados⁵ (p.13). Após a etapa de *data mining*, inicia-se a etapa de análise dos padrões encontrados e finalmente a etapa de apresentação dos resultados e do conhecimento adquirido.

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

Fonte: Elaborado pelos autores.

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.

O uso de *data mining*, combinado com algoritmos de *machine learning*, pode vir a auxiliar o especialista da saúde em momentos críticos que demandem decisões rápidas, ou quando há uma deficiência dos recursos apresentados, seja uma imagem 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 profissional⁷. 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 em que 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 errata, artigos e periódicos não relacionados à saúde humana como, por exemplo, os referentes ao uso na agricultura e pecuária. Artigos que geraram dúvida em relação ao uso na saúde humana foram analisados individualmente. Os dados na Tabela 1 mostram uma tendência de crescimento no uso de *data mining* e *machine learning* na área da saúde.

Tabela 1– Artigos citando uso de *data mining* e/ou *machine learning* identificados na BVS

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

Em relação ao uso de algoritmos de *machine learning*, a maioria dos problemas de análise podem ser inseridos em duas principais categorias¹⁴ (p.26): 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.

2.1 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 agrupamento dos dados não-estruturados de forma a encontrar padrões e informações relevantes para cada conjunto de textos.

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*. Então, 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 de *stopwords*. Em seguida, os termos são reduzidos aos seus radicais, removendo prefixos e sufixos, para evitar variações para a mesma palavra, como por exemplo, nas palavras “interessante” e “interessantíssimo”, que passam a mesma ideia, apenas com intensidades diferentes. A redução do termo ao seu radical é conhecida como *stemming*. Enfim, contabiliza-se a presença dos termos em forma de uma representação vetorial binária ou uma representação de frequências do termo.⁵

A seguir, apresentamos os conceitos dos dois principais tipos de aprendizagem de máquina e alguns estudos que utilizam estas técnicas na área de saúde pública e SST.

2.2 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 suas 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 k , 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 e inicia selecionando aleatoriamente k centroides do conjunto de dados, e a cada iteração, 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* e o algoritmo termina (converge) quando não há mudança significativa dos centroides. A **Figura 2 (A)** ilustra um gráfico de dispersão de um conjunto de dados não categorizado e as **Figura 2 (B)**, **Figura 2 (C)** e **Figura 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).

Fonte: Elaborado pelos autores

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

O algoritmo *k-means* utiliza medidas de distância, como, por exemplo, a distância quadrática euclidiana, para agrupar os pontos em cada centróide. 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}

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, 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, denominados componentes principais, sendo que os primeiros componentes gerados, visam capturar a maior variabilidade da combinação linear dos preditores²⁰, conforme

Equação 2, onde i corresponde ao número do componente principal e j corresponde ao peso atribuído ao preditor para o componente atual.

Equação 1 - Cálculo de Componentes Principais

$$PC_i = (a_{j1}x \text{ Preditor } 1) + (a_{j2}x \text{ Preditor } 2) + \dots + (a_{jn}x \text{ Preditor } N)$$

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 onde não há padrões conhecidos. A diferença é que com o uso dos SOMs, é possível visualizar n dimensões em um gráfico bidimensional. No estudo, 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 em que 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 escolhidas, assim como as probabilidades de mortalidade em 8 anos, de acordo com os perfis identificados pelo SOM.

2.3 Aprendizagem Supervisionada

A aprendizagem supervisionada baseia-se em dados preparados para treinamento em que já 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 ajuste do modelo e outra parte da amostra deste mesmo conjunto de dados 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 identificar os preditores que mais contribuíram para aumento do risco de hipertensão. Dos indivíduos participantes do estudo, separou-se uma parte das observações para o treinamento do modelo e outra parte para testar sua acurácia. O desempenho do modelo foi aferido com base nos resultados de teste do modelo, apresentando performance preditiva superior 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 um tumor benigno ou maligno. Para realizar a classificação com o algoritmo de aprendizado supervisionado denominado máquina de vetor de suporte (em inglês, *Support Vector Machines*, SVM), foram utilizadas características como assimetria, coloração, entre outras. A imagem então é vetorizada, ou seja, os pixels da imagem são transformados em uma sequência de números $\rightarrow x_i$ em que x representa uma única observação e i sua identificação no conjunto

de imagens. Conforme o padrão da imagem se enquadre 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.

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

Fonte: Elaborado pelos autores

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

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.

Um 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.

Um outro estudo realizado por Torres et al.²⁷ utilizou o algoritmo *superlearner* para prever 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.

3 Exemplos de Aplicação de Text Mining e Machine Learning em SST

Um estudo realizado por Marucci-Wellman et al.²⁸ realizou mineração textual em 15.000 registros de narrativas de solicitações de seguro por acidente ou lesão, extraídas 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*, 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, 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* Weka³⁰ para aplicar *data mining* em classificação de páginas *web* de Saúde. O *software* já possui alguns 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 (*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, como 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.

Um 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 um medicamento, utilizando relatos em redes sociais, revelando importantes preocupações em relação aos efeitos colaterais e ao custo do medicamento.

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, como por exemplo, pela

análise dos efeitos do trabalho em turnos de trabalho 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, em que é possível aplicar diferentes algoritmos para fases distintas e processamento. O trabalho de Christen³⁶ utiliza este conceito para realizar *linkage* de bases de dados, em que 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.

4 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
	SVM	Classificação de melanomas	Correa et al. ⁷
		Estimar internações hospitalares	Lucini et al. ⁴¹
	<i>Naïve Bayes</i>	Classificação de declarações textuais de morbidades relacionadas ao trabalho.	Marucci-Wellman et al. ²⁸
	<i>Random Forests</i>	Predição de gravidez de risco com base em fatores de risco da mãe	Pan et al. ²⁴
		Deteção de diabetes não diagnosticada.	Oliveira et al. ²⁵
Não Supervisionada	SOM	Identificação de padrões de qualidade do ar em séries históricas.	Pearce et al. ²¹
			Akay et al. ³⁴

<i>K-means</i>	Analisar sentimentos de pacientes que passaram por tratamentos de câncer baseado em dados de fóruns especializados.	Gao et al. ²²
	Análise exploratória e identificação de subgrupos populacionais	Olson et al. ¹⁸
	Identificação de grupos de caminhoneiros de acordo com comportamentos e hábitos alimentares.	Lee et al. ¹⁹
	Identificação de grupos expostos a diferentes níveis da síndrome de <i>Burnout</i> em equipes médicas.	

O Quadro 1 exhibe 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 adequa 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 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.	Gao et al. ²²
	Identificação de perfis similares com uso de K-Means.	Olson et al. ¹⁸
		Lee et al. ⁴²
Predição de desfechos como adoecimento ou mortalidade.	Predição de diabetes com uso de redes neurais, K-NN e <i>Random Forest</i> .	Oliveira et al. ²⁵
	Predição de hipertensão e riscos associados com mineração textual e árvore de decisão.	Ramezankhani et al. ²³
	Predição de óbitos infantis e fatores de risco com uso de árvores de decisão.	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.	Allen et al. ⁴³
	Análise da qualidade do ar em séries históricas com o uso de SOM.	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 <i>Naïve Bayes</i> .	Marucci-Wellman et al. ²⁸

	Mineração textual e uso de SVM para predição de internações hospitalares.	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. ⁴⁴

5 Considerações Finais

Conforme relatado ao longo deste ensaio, é possível observar diferentes aplicações de *data mining* e algoritmos de *machine learning* na área de saúde pública e 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 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 há uma perspectiva positiva na adoção desta 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 de Governo, o surgimento de novas fontes de dados, como observatórios de SST⁴⁵ e dados de 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 o conteúdo publicado.

Referências

1. Chiavegatto Filho ADP. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. *Epidemiol e Serviços Saúde*. 2015 Jun;24(2):325–32.
2. Madden S. From Databases to Big Data. *IEEE Comput Soc*. 2012;0–2.
3. Russel S, Norvig P. *Inteligência artificial*. Elsevier, editor. 2013.
4. Turing AM. Turing. Computing machinery and intelligence. *Mind* [Internet]. 1950;59(236):433–60. Available from: papers2://publication/uuid/E74CAAC6-F3DD-47E7-AEA6-5FB511730877
5. Silva L, Peres S, Boscarioli C. *Introdução à Mineração de Dados com aplicações em R*. Elsevier, editor. 2016.
6. Mitchell TM. Machine Learning and Data Mining. *Commun Acm*. 1999;42(11).
7. Correa DNL, Paniagua LRB, Noguera JLV, Pinto-Roa DP, Toledo LAS. Computerized Diagnosis of Melanocytic Lesions Based on the ABCD Method. *Proc - 2015 41st Lat Am Comput Conf CLEI 2015*. 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. *Gestão & Produção*. 2006 May;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. *Groza T, editor. 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. Fayyad 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, CA, itd: Morgan Kaufmann. 2012. 745 p.
14. Gareth J, Witten D, Hastie T, Tibishirani R. *An Introduction to Statistical Learning with Applications in R*. 1st ed. Springer; 2013. 426 p.
15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. Springer; 2016. 764 p.
16. Elmagarmid AK, Member S. 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(March):28–34.
18. Olson R, Thompson S V, 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. *Inq (United States)*. 2016;53(2).
20. Kuhn M, Johnson K. *Applied Predictive Modeling*. Springer. Springer New York, New York, NY; 2013. 1–595 p.
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 Heal*. 2014;13:56.
22. Gao S, Mutter S, Casey A, Mäkinen V-P. Numero: a statistical framework to define multivariable subgroups in complex population-based datasets. *Int J Epidemiol*. 2018;(June):1–6.
 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).
 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. *Sao Paulo Med J*. 2017;135(3):234–46.
 26. Vianna R, Moro C, Moysés S, 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;(July):1–11.
 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 da, 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 Heal Informatics*. 2009;1(1):1–6.
 30. Weka 3: Data Mining Software in Java [Internet]. [cited 2018 Jan 22]. Available from: www.cs.waikato.ac.nz/ml/weka/
 31. NIOSH. Industry and Occupation Coding [Internet]. [cited 2018 Mar 4]. Available from: <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. *Int World Wide Web Conf Comm*. 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;19(1):210–8. Available from: <http://ieeexplore.ieee.org/document/6851846/>
 35. Carvalho F, Guedes GP. Night Sleep Deprivation: Computational Analysis of Language Effects. *Proc 23rd Brazillian Symp Multimed Web* [Internet]. 2017;221–4. Available from: <http://doi.acm.org/10.1145/3126858.3131595>
 36. Christen P. A two-step classification approach to unsupervised record linkage. In: 6th Australasian Data Mining Conference (AusDM'07). 2007. p. 111–9.
 37. Kim T-W, Koh D-H, Park C-Y. Decision tree of occupational lung cancer using

- classification and regression analysis. *Saf Health Work*. 2010 Dec;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;171–80.
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 Sep;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 Aug 29;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;
42. 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. *Inq (United States)*. 2016;53(2).
43. 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):1–10.
44. 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).
45. Observatório Digital de Saúde e Segurança no Trabalho - Smartlab de Trabalho Decente MPT - OIT. 2017 [Internet]. 2017 [cited 2017 Jun 23]. Available from: <https://observatoriosst.mpt.mp.br/>

5.2 ARTIGO 2 – Publicado na Revista de Saúde Pública.

5.2.1 Predição de absenteísmo docente na rede pública com machine learning

Página de Título

Predição de absenteísmo docente na rede pública com *machine learning*

Prediction of absenteeism in public schools teachers with machine learning

Fernando Timoteo Fernandes^{1,2}

ORCID: <https://orcid.org/0000-0003-2548-9685>

Alexandre Dias Porto Chiavegatto Filho³

ORCID: <http://orcid.org/0000-0003-3251-9600>

¹ Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho. São Paulo, SP, Brasil.

² Fundacentro. São Paulo, SP, Brasil

³ Universidade de São Paulo, Faculdade de Saúde Pública. São Paulo, SP, Brasil.

Contato:

Fernando Timoteo Fernandes

E-mail:

fernando.fernandes@fundacentro.gov.br

RESUMO

OBJETIVO: Predizer o risco de ausência laboral decorrente de morbidades dos docentes que atuam na educação infantil na rede pública municipal, com o uso de algoritmos de *machine learning*. **MÉTODOS:** Trata-se de um estudo transversal utilizando dados secundários, públicos e anônimos da Relação Anual de Informações Sociais, selecionando professores da educação infantil que atuaram na rede pública municipal do estado de São Paulo entre 2014 e 2018 (n = 174.294). Foram também vinculados dados da média de alunos por turma e número de habitantes no município. Os dados foram separados em treinamento e teste, utilizando os registros de 2014 a 2016 (n = 103.357) para treinar cinco modelos preditivos e os dados de 2017 a 2018 (n = 70.937) para testar seus desempenhos em dados novos. A performance preditiva dos algoritmos foi avaliada por meio do valor da área abaixo da curva ROC (AUROC). **RESULTADOS:** Todos os cinco algoritmos testados apresentaram área abaixo da curva acima de 0,76. O algoritmo com melhor performance preditiva (redes neurais artificiais) obteve 0,79 de área abaixo da curva, com acurácia de 71,52%, sensibilidade de 72,86%, especificidade de 70,52% e kappa de 0,427 nos dados de teste. **CONCLUSÃO:** É possível predizer casos de afastamentos por morbidade em docentes da rede pública com *machine learning* usando dados públicos. O melhor algoritmo apresentou melhor resultado da área abaixo da curva quando comparado ao modelo de referência (regressão logística). Os algoritmos podem contribuir para predições mais assertivas na área da saúde pública e da saúde do trabalhador, permitindo acompanhar e ajudar a prevenir afastamentos por morbidade desses trabalhadores.

DESCRIPTORIOS: Absenteísmo. Fatores de Risco. Aprendizado de Máquina Supervisionado. Professores Escolares. Educação Infantil.

ABSTRACT

OBJECTIVE: To predict the risk of absence from work due to morbidities of teachers working in early childhood education in the municipal public schools, using machine learning algorithms.

METHODS: This is a cross-sectional study using secondary, public and anonymous data from the *Relação Anual de Informações Sociais*, selecting early childhood education teachers who worked in the municipal public schools of the state of São Paulo between 2014 and 2018 (n = 174,294). Data on the average number of students per class and number of inhabitants in the municipality were also linked. The data were separated into training and testing, using records from 2014 to 2016 (n = 103,357) to train five predictive models, and data from 2017 to 2018 (n = 70,937) to test their performance in new data. The predictive performance of the algorithms was evaluated using the value of the area under the ROC curve (AUROC). **RESULTS:** All five algorithms tested showed an area under the curve above 0.76. The algorithm with the best predictive performance (artificial neural networks) achieved 0.79 of area under the curve, with accuracy of 71.52%, sensitivity of 72.86%, specificity of 70.52%, and kappa of 0.427 in the test data. **CONCLUSION:** It is possible to predict cases of sickness absence in teachers of public schools with machine learning using public data. The best algorithm showed a better result of the area under the curve when compared with the reference model (logistic regression). The algorithms can contribute to more assertive predictions in the public health and worker health areas, allowing to monitor and help prevent the absence of these workers due to morbidity.

DESCRIPTORS: Absenteeism. Risk Factors. Supervised Machine Learning. School Teachers. Early Childhood Education.

1 Introdução

Profissionais da área da educação básica apresentam altos índices de afastamento por doença relacionada ao trabalho, situando-se entre as primeiras posições no Brasil.¹ Entre as suas principais doenças estão os distúrbios de voz, doenças do aparelho respiratório, sistema osteomuscular e transtornos mentais e comportamentais^{2,3}, com o crescimento recente de notificações relacionadas à saúde mental, como no caso do Estado de São Paulo.⁴ As condições de trabalho a que estão expostos, com longas jornadas, turmas numerosas⁵ e falta de reconhecimento, acabam agravando o problema, com consequências diretas na vida particular dos docentes.⁶

No entanto, ainda há poucos indicadores que analisam as condições dos docentes em cada etapa do ensino básico, desde a educação infantil até o ensino médio, áreas que apresentam distinções e especificidades no contexto de trabalho⁷. Estudos que buscam gerar indicadores normalmente utilizam valores agregados das diferentes etapas do ensino básico¹, ou são baseados em inquéritos em regiões específicas que dependem da ação do respondente, podendo gerar altos custos operacionais.^{8,9}

O uso de indicadores agregados para a análise de doenças pode omitir situações críticas de morbidade e não corresponder à realidade das diferentes condições de trabalho em que os docentes estão expostos. O problema se agrava ao analisar professores da rede municipal, em que os dados de afastamento somente são acessíveis por meio de secretarias de saúde ou por departamentos de perícias médicas de cada município.²

Uma forma de analisar a situação de morbidade por etapa de ensino, e a relação de trabalho com o adoecimento de professores, é buscar fontes de dados oficiais públicas, que possuam registros individualizados e atualizados com ampla cobertura geográfica.

Nesse cenário, a partir de 2011, o acesso aos dados provenientes do governo federal foi ampliado, mediante a publicação da Lei 12.527/2011, conhecida como Lei de Acesso à Informação (LAI). Essa legislação permite que sejam solicitadas informações a qualquer órgão público e prevê que as entidades

e órgãos públicos se antecipem e publiquem seus dados e informações na internet em formatos abertos e não-proprietários, conhecidos como dados abertos⁵⁹. Um exemplo é a Relação Anual de Informações Sociais (RAIS), instituída pelo Decreto nº 76.900 de 23/12/1975, atualmente sob responsabilidade do Ministério da Economia. Os dados são divulgados anualmente de forma pública e anônima, contendo registros individualizados em todo o território nacional.¹¹

A RAIS tem como principal finalidade subsidiar a geração de estatísticas do trabalho e do mercado formal (privado e estatutário) para as entidades governamentais. No entanto, também são disponibilizados os tipos de afastamento como, por exemplo, doença ou doença relacionada ao trabalho, sem informações da doença específica. O preenchimento é obrigatório a todas as empresas públicas e privadas desde 1977.¹²

Nos dados de afastamento de professores na RAIS, verifica-se que em média mais de 40% dos professores da educação infantil se afastaram do trabalho ao menos uma vez ao ano devido a morbidades no período entre 2014 e 2018, superando as demais etapas do ensino básico e apontando para a necessidade de analisar esse grupo específico.

O presente estudo propõe construir modelos preditivos capazes de estimar risco de ausência laboral dos docentes da educação infantil que atuam na rede municipal, considerando todos os municípios do Estado de São Paulo, utilizando inteligência artificial e algoritmos de aprendizagem supervisionada (*machine learning*) conhecidos pelo alto desempenho, principalmente na área da saúde.¹³ Espera-se que o melhor modelo proposto possa ser aplicado para estimar o risco de absenteísmo por doenças em professores, utilizando dados projetados dos trabalhadores (ex: tempo de ocupação, carga horária, renda) e do ambiente de trabalho (ex: alunos por turma), de forma a subsidiar políticas públicas direcionadas de prevenção de ausência laboral, evitando novos adoecimentos, melhorando a qualidade de vida desses profissionais e, consequentemente, a qualidade do ensino na rede pública.

2 Materiais e Métodos

Trata-se de um estudo transversal para prever risco de afastamento por morbidade (doença ou doença relacionada ao trabalho) em professores que atuam na educação infantil, no período pré-escolar de toda a rede pública municipal do Estado de São Paulo pelo uso de dados secundários, públicos e anônimos.

2.1 Fontes de Dados

Foram utilizados dados individualizados em seu menor nível de desagregação, conhecidos também como microdados¹¹, de vínculos empregatícios formais fornecidos pela RAIS no período entre 2014 e 2018.

A RAIS mantém informações fornecidas anualmente pelas empresas públicas e privadas sobre os vínculos empregatícios formais de seus funcionários, tendo informações como contratações e demissões realizadas no ano corrente, assim como informações sobre jornada de trabalho e afastamentos por doença ou doença relacionada ao trabalho.

Foram incluídos professores da rede pública municipal de todos os municípios do Estado de São Paulo que informaram dados à RAIS, em que os docentes possuíam nível superior e vínculo empregatício ativo até 31/12 do ano corrente, atuando no ensino pré-escolar, classificados na RAIS com a CBO 231105 (Professores de Ensino Superior no Ensino Infantil, quatro a seis anos).

Como forma de identificação do desfecho, foram utilizados os campos de causa de afastamento presentes na RAIS, filtrando-se pelos códigos 30 (doença de trabalho) e 40 (doença). Não foram consideradas aposentadorias decorrentes de doenças. Para a identificação do tipo de estabelecimento (federal, estadual, municipal ou privado), foi utilizada a informação da natureza jurídica do estabelecimento (códigos 1031,1066,1120,1155,1180,1244). Para identificação da atividade do estabelecimento, foram utilizados os códigos 841160 (administração pública em geral) e a divisão 85 (educação) da classificação nacional de atividades econômica versão 2.0 (CNAE 2.0) fornecidos na RAIS.

Cerca de 99,7% (n=174.346) dos registros estavam cadastrados no CNAE 841160, sendo que o restante 0,03% (n=52) pertenciam a um CNAE não relacionado ao ensino infantil pré-escolar, sendo excluídos da seleção. Também foram excluídos registros em que o município de exercício do trabalhador não correspondia à unidade federativa de São Paulo, resultando no total de 174.294 registros. Nas situações em que o município de exercício do trabalhador não foi informado, foi adotado o município do estabelecimento.

Para capturar parte das condições de trabalho em que esses profissionais estavam expostos, foi vinculada a informação da média de alunos por turma na etapa pré-escolar, fornecida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)¹⁴, considerando que algumas doenças comuns aos docentes estão associadas à acústica do ambiente e à complexidade na atuação do docente³, e adicionado o número de habitantes por município, fornecido pela Fundação Seade.¹⁵ Os dados do INEP e da Fundação SEADE foram vinculados à amostra selecionada da RAIS por meio do código IBGE do município.

Ao todo, foram selecionadas 11 variáveis preditoras: sexo, idade, pós-graduação, tamanho do estabelecimento, tipo de vínculo empregatício, primeiro emprego, tempo no emprego, horas contratadas, média anual da quantidade de salários mínimos, média de alunos por turma e número de habitantes no município de atuação do trabalhador.

2.2 Preparação dos Dados

Os dados passaram por uma etapa inicial de pré-processamento, para tratar as variáveis ausentes e transformar os dados para que pudessem ser utilizados na construção dos diferentes tipos de modelos preditivos de *machine learning*.¹⁶ Dentre as variáveis selecionadas, todas estavam totalmente preenchidas, sem valores ausentes (*missing*). Variáveis com mais de duas categorias foram representadas por um conjunto de variáveis chamadas *dummy*, em que para cada categoria, é gerada uma nova variável com valores de 0 ou 1. As variáveis contínuas foram padronizadas por meio de *escore-z* (*z-score*). Foi testada a correlação entre as variáveis numéricas e foi constatada alta correlação entre o número de habitantes no município e a média de alunos

(0,93), assim optou-se por dicotomizar a variável de número de habitantes utilizando a definição de 500 mil habitantes para identificação de atuação em uma cidade grande.¹⁷

Para a fase de pré-processamento e carga de dados, foi utilizado o sistema gerenciador de banco de dados (SGBD), MS SQL Server. Para a análise de dados e construção dos modelos preditivos foi utilizado o software R.

2.3 Construção dos modelos preditivos

Ao todo, foram desenvolvidos cinco algoritmos de aprendizado supervisionado de *machine learning*: regressão logística, árvores de decisão¹⁶, *random forest*¹⁸, *xgboost*¹⁹ e redes neurais artificiais.²⁰ Algoritmos de *machine learning* estão sujeitos a problemas como o sub-ajuste (*underfitting*), quando o modelo não consegue se ajustar à variabilidade inerente dos dados gerando estimativas ruins, e sobreajuste (*overfitting*) quando o modelo se ajusta muito bem para a amostra utilizada no treinamento, mas não obtém bons resultados para novos dados. Uma forma de melhorar o desempenho dos algoritmos e evitar problemas como *overfitting* ou *underfitting* é aplicar técnicas de reamostragem (*resampling*) e validação cruzada (*cross-validation*) para a seleção de hiperparâmetros.¹⁶

Na validação cruzada, o conjunto de dados é separado em duas partes. A primeira parte é destinada ao treinamento do algoritmo e a segunda parte é utilizada para ajustes dos hiperparâmetros do modelo, simulando novos dados e selecionando os hiperparâmetros que otimizem a métrica de performance escolhida. Essa fase de treinamento permite que o algoritmo obtenha um melhor desempenho preditivo quando se depara com novos conjuntos de dados. Uma das técnicas de validação cruzada mais conhecidas é a *k-fold*, onde *k* significa o número de subconjuntos dos dados de testes que serão utilizados para ajustes dos hiperparâmetros do modelo durante a fase de treinamento.²¹

Para a construção dos modelos, separou-se uma parte dos dados (60%) referente ao período entre 2014 e 2016 (n=103.357) para treinamento, incluindo validação cruzada com a técnica *k-fold* utilizando 10 partições para definição dos hiperparâmetros. A outra parte dos registros (40%), referentes ao período entre

2017 e 2018 (n=70.937), foi utilizada para a avaliação de desempenho dos modelos, omitindo-se o resultado sobre afastamento por doença e aplicando o modelo treinado para estimar o desfecho em dados futuros. Para avaliar o desempenho dos modelos, foram analisadas medidas como acurácia, sensibilidade, especificidade, valor preditivo positivo (PPV) e valor preditivo negativo (NPV). Para selecionar o melhor modelo, foi utilizado o valor da área abaixo da curva ROC (AUROC)²². Foi utilizado o coeficiente kappa de Cohen para avaliar a concordância entre os valores previstos e observados.

Este estudo segue as orientações de descrição de modelos preditivos multivariados para diagnóstico ou prognóstico (*Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – TRIPOD*).²³

2.4 Critérios de Confidencialidade

O projeto foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo, sob o nº 4.031.362, CAAE 30786620.5.0000.5421.

3 Resultados

3.1 População de Estudo

Foi realizada inicialmente a análise descritiva dos dados de treinamento (2014 a 2016) e teste (2017 a 2018). Os dados de treinamento compreendem 103.357 registros de vínculos empregatícios de professores, havendo 45.419 (43,94%) registros de afastamentos por doença ou doença relacionada ao trabalho. Os dados de teste correspondem a 70.937 registros e 30.261 casos de afastamento por doença (42,65%). A Tabela 1 mostra os resultados dessa análise.

Tabela 2 - Análise descritiva dos dados de treinamento (n=103.357) e teste (n=70.937).

Variável	Categoria	Treinamento		Teste	
		n	%	n	%
Sexo	0 – Masculino	22.120	21,40	16.055	22,63
	1 – Feminino	81.237	78,60	54.882	77,37
Pós-Graduado(a)	0 – Não	100.904	97,63	68.412	96,44
	1 – Sim	2.453	2,37	2.525	3,56
Idade	-	43,74 (9,73)		43,97(9,70)	
Empresa Grande*	0 – Não	19	0,02	19	0,03
	1 – Sim	103.338	99,98	70.918	99,97
Tipo de Vínculo	0 – Estatutário	96.682	93,54	65.210	91,93
	1 – CLT	5.867	5,68	4.493	6,33
	2 – Temporário	808	0,78	1.234	1,74
Primeiro Emprego	0 – Não	100.105	96,85	65.236	91,96
	1 – Sim	3.252	3,15	5.701	8,04
Tempo Emprego (em meses)	-	120,49 (92,31)		123,29 (93,84)	
Horas Contratadas	-	33,28 (6,60)		33,28 (6,33)	
Salários Mínimos	-	5,03 (3,04)		5,04 (3,04)	
Cidade Grande**	0 – Até 500.000 hab.	30.866	29,86	22.212	31,31
	1 – Acima de 500 mil	72.491	70,14	48.725	68,69
Alunos por turma	-	27,42 (5,75)		26,74 (5,50)	

*Acima de 100 funcionários

** Classificação de cidades - OECD ¹⁷ – pequena; média; grande (acima de 500 mil) e metrópole

Em seguida, foi realizada a análise das variáveis independentes para verificar se há relação significativa com o desfecho, utilizando o teste de associação pelo qui-quadrado de Pearson para as variáveis categóricas. A Tabela 2 mostra os resultados, em que é possível observar que a maioria das variáveis apresenta associação significativa com o desfecho.

Tabela 3 - Análise de associação com o desfecho da amostra de treinamento

Variável	Categoria	Professores afastados por doença			teste χ^2 *
		Sim	Não	Total	
		n (%)	n (%)	n (%)	
Dados do Trabalhador					
Sexo	0 – Masculino	9.930 (44,89)	12.190 (55,11)	22.120(100,0)	<0,001
	1 – Feminino	35.489(43,69)	45.748 (56,31)	81.237 (100,0)	
Pós-Graduado(a)	0 – Não	45.075 (44,67)	55.829 (55,33)	100.904 (100,0)	<0,0001
	1 – Sim	344 (14,02)	2.109 (85,98)	2.453 (100,0)	
Idade	43,74 (9,73)				
Dados da Empresa					
Empresa Grande	0 – Não	11 (57,89)	8 (42,11)	19 (100,00)	0,220
	1 – Sim	45.408 (43,94)	57.930(56,06)	103.338 (100,00)	
Dados do Vínculo					
Tipo de Vínculo	0 – CLT	746 (12,72)	5.121 (87,28)	5.867 (100,0)	<0,0001
	1 – Estatutário	44.500 (46,03)	52.182 (53,97)	96.682 (100,0)	
	2 – Temporário	173 (21,41)	635 (78,59)	808 (100,0)	
Primeiro Emprego	0 – Não	44.616 (44,57)	55.489 (55,43)	100.105 (100,0)	<0,0001
	1 – Sim	803 (24,69)	2.449 (75,31)	3.252 (100,0)	
Tempo Emprego (em meses)	120,49 (92,31)				
Horas Contratadas	33,28 (6,60)				
Salários Mínimos	5,03 (3,04)				
Dados Ambientais					
Cidade Grande	0 – Até 500.000	3.422 (11,09)	27.444 (88,91)	30.866 (100,0)	<0,0001
	1 – Acima de 500 mil habitantes	41.997 (57,93)	30.494 (42,07)	72.491 (100,0)	
Alunos por turma	27,04 (5,60)				

*Teste de associação pelo χ^2 : H_0 : não existe associação; H_a : existe associação

3.2 Análise dos modelos preditivos

A Figura 1, mostra as curvas ROC de cada modelo desenvolvido em que foram incluídas 11 variáveis (tipo de vínculo empregatício, número de horas contratadas, tempo de emprego em meses, média da quantidade de salários mínimos do ano, sexo, tamanho do estabelecimento, possuir pós-graduação,

idade, primeiro emprego, média de alunos por turma pré-escolar no município) e número de habitantes.

Figura 1 – Comparação dos modelos com base nos resultados de predição utilizando dados de testes (2017 e 2018)

A Tabela 3 mostra os resultados do desempenho de cada algoritmo. O melhor modelo foi o de redes neurais artificiais, que obteve 0,79 de área abaixo da curva, com acurácia de 71,52%, sensibilidade de 72,86%, especificidade de 70,52%, PPV de 64,77% e NPV de 77,74%. O modelo também apresentou a melhor coeficiente kappa de 0.4266.

Tabela 4 - Resultados dos modelos desenvolvidos utilizando dados de teste.

Algoritmo	AUC	Sensibilidade	Especificidade	Acurácia	PPV	NPV	Kappa
Reg. Logística	0.7792 [0.7759-0.7826]	0.7754	0.6568	0.7074	0.6270	0.7972	0.4195
Árvore de Decisão	0.7756 [0.7722-0.7790]	0.7008	0.7243	0.7142	0.6541	0.7649	0.4212
<i>Random Forest</i>	0.7670 [0.7634-0.7703]	0.7715	0.6571	0.7059	0.6260	0.7944	0.4134
<i>XGBoost</i>	0.7843 [0.7810-0.7876]	0.7235	0.6970	0.7083	0.6398	0.6970	0.4136
Redes Neurais Artificiais	0.7902[0.7867-0.7934]	0.7286	0.7052	0.7152	0.6477	0.7774	0.4266

Os hiperparâmetros dos modelos de machine learning que obtiveram os maiores valores de AUROC foram os seguintes: árvore de decisão (cp-complexity-parameter=0.000385), random forest (mtry- variáveis randomicamente selecionadas a cada divisão=7), xgboost (nrounds=100, max_depth=4, gamma=0.6, colsample_bytree=0.8, min_child_weight=2, subsample=1), redes neurais artificiais (size=19, decay=0.2).

Cada algoritmo utiliza um conjunto de regras diferentes para se ajustar aos dados com o menor erro possível, sendo que algumas variáveis se tornam mais importantes que outras para a capacidade preditiva do algoritmo. A importância das variáveis refere-se a quanto a variável contribuiu para realizar previsões acuradas em cada modelo. Para os modelos lineares utiliza-se de estatística t tradicional para contabilizar a importância das variáveis. Nos demais modelos, a importância depende da implementação de cada algoritmo.

Por exemplo, no caso de árvores de decisão, se uma variável é utilizada com maior frequência nas divisões de folhas contribuindo para maior pureza do nó, ela terá uma importância maior para o modelo final, significando que sua remoção prejudicará a performance geral do modelo. Nos modelos que utilizam múltiplas árvores, como o random forest, a importância das variáveis é calculada a cada divisão das folhas de uma árvore e somada com o resultado das demais árvores.²¹ Nos modelos de redes neurais, para este estudo foi utilizado o método “Weights”²⁴ em que se utiliza uma combinação dos valores de saída dos neurônios das camadas ocultas e os valores dos pesos de cada conexão para

calcular a maior contribuição e representatividade de cada variável de entrada no resultado da classificação. Neste estudo, foi utilizado o pacote caret⁷⁰ do R que já implementa o cálculo de importância de variáveis de acordo com cada modelo.²⁶

Apesar de não ter interpretação causal²⁷, as variáveis podem dar informações relevantes sobre a sua relação com o desfecho. A Figura 2 demonstra como cada variável contribuiu para os modelos desenvolvidos.

Figura 2 – Importância das variáveis por modelo preditivo

As variáveis que mais contribuíram para o melhor modelo (redes neurais artificiais) foram: quantidade de salários mínimos, cidade grande, média de alunos por turma, tempo no emprego e empresa grande, respectivamente.

4 Discussão

A RAIS pode fornecer indícios da situação de morbidade dos trabalhadores por meio do campo de causa de afastamento. A ampla cobertura e o registro individualizado, facilitam a seleção e análise de dados de uma população de trabalhadores formais com a mesma ocupação em um determinado Estado, o que permite acompanhar ao longo dos anos a evolução de afastamentos por doença e predizer quais profissionais apresentam maior risco de adoecimento nos próximos anos. Para o melhor algoritmo (redes neurais artificiais), obteve-se 0,79 da área abaixo da curva ROC, sensibilidade de 72,86% e 71,52% de acurácia, usando apenas 11 variáveis preditoras.

Os resultados mostram a viabilidade da utilização de fontes anonimizadas para a predição de morbidade de docentes do ensino pré-escolar com algoritmos de *machine learning*. As variáveis que mais contribuíram para o melhor modelo podem ser exploradas em estudos futuros, para verificar se há relação causal. Trabalhar em grandes cidades, renda média do trabalhador e média de alunos por turma podem ter impactos diretos na saúde dos professores. No caso da média de alunos, há evidências na literatura que corroboram este raciocínio, como problemas na acústica do ambiente e elevado ruído devido a maior agitação em sala de aula.^{3,28}

Algoritmos de *machine learning* têm sido utilizados em diferentes contextos na saúde do trabalhador para auxiliar na predição de morbidades relacionadas às atividades destes profissionais.^{29,30} Na área da educação, não foram encontrados estudos que utilizaram *machine learning* ou demais modelos preditivos para risco de absenteísmo docente como principal desfecho, havendo alguns estudos que analisaram a associação de variáveis preditoras com o desfecho^{31,32}, não tendo como a predição do desfecho o principal objetivo. A predição de absenteísmo docente pode identificar professores que precisam de maior atenção, permitindo adotar medidas para evitar afastamentos por doenças relacionadas à ocupação^{3,28}, diminuindo a ausência ao trabalho e consequentemente aprimorando a qualidade do ensino.³³

O estudo possui algumas limitações. Por se tratar de dados anônimos, não foi possível verificar se os docentes atuam em mais de uma escola, podendo

o mesmo possuir múltiplos vínculos empregatícios municipais. Também não foi possível identificar se houve um afastamento prévio por doença pré-existente que possa estar relacionado aos afastamentos futuros. Por fim, foi analisado um número baixo de variáveis preditoras devido à disponibilidade da RAIS. Espera-se que no futuro, com uma coleta mais extensa realizada por meio de questionários, sejam possíveis resultados ainda mais robustos por parte dos algoritmos.

Em conclusão, somente com dados públicos e anônimos, foi possível obter boa sensibilidade (73%) e especificidade (71%) na identificação de risco de absenteísmo por doença, independente da causa da morbidade. A RAIS apesar de não ter sido criada para fins da saúde, mostrou-se viável para análise de absenteísmo por morbidades, principalmente por conter variáveis das relações de trabalho, podendo ser explorada em outras categorias profissionais. Os modelos desenvolvidos podem futuramente ser aplicados nas demais etapas de ensino e em outras categorias ocupacionais.

Algoritmos de *machine learning* podem contribuir com predições mais assertivas na análise de absenteísmo de professores do ensino infantil da rede pública municipal, podendo ser usados na elaboração de indicadores de morbidade ou para subsidiar políticas públicas prevencionistas direcionadas a essa categoria profissional.

Referências

1. DIEESE. Anuário do Sistema Público de Emprego, Trabalho e Renda. 2016. 134 p.
2. Gasparini SM, Barreto SM, Assunção AÁ. O professor, as condições de trabalho e os efeitos sobre sua saúde. *Educ e Pesqui*. 2005;31(2):189–99.
3. Medeiros AM de, Vieira M de T. Ausência ao trabalho por distúrbio vocal de professores da Educação Básica no Brasil. *Cad Saude Publica*. 2019;35:1–12.
4. G1. Número de professores afastados por transtornos em SP quase dobra em 2016 e vai a 50 mil [Internet]. 2017. Available from: <https://g1.globo.com/sp/sao-paulo/noticia/numero-de-professores-afastados-por-transtornos-em-sp-quase-dobra-em-2016-e-vai-a-50-mil.ghtml>
5. Rodríguez-Loureiro L, Artazcoz L, López-Ruiz M, Assunção AÁ, Benavides FG. Joint effect of paid working hours and multiple job holding on work absence due to health problems among basic education teachers in Brazil: the Educatel Study. *Cad Saude Publica*. 2019;35(suppl 1):1–12.
6. Silva J, Fischer F. Invasão multiforme da vida pelo trabalho entre professores de educação básica e repercussões sobre a saúde. *Rev Saude Publica*. 2020;1–8.
7. Assunção AÁ, Oliveira DA. Intensificação do trabalho e saúde dos professores. *Educ e Soc*. 2009;30(107):349–72.
8. Porto L, Oliveira N, Carvalho F, Araújo T. Construção de um índice de morbidade para professoras da educação básica. *Rev Baiana Saúde Pública*. 2008;15.
9. Maia EG, Claro RM, Assunção AÁ. Múltiplas exposições ao risco de faltar ao trabalho nas escolas da Educação Básica no Brasil. *Cad Saude Publica*. 2019;35(suppl 1):1–13.
10. BRASIL. Dados Abertos [Internet]. 2018 [cited 2018 Oct 1]. Available from: <http://dados.gov.br/noticia/primeira-lei-de-acesso-no-mundo-que-preve-dados-abertos>
11. BRASIL. Microdados RAIS e CAGED [Internet]. 2019 [cited 2020 Feb 7]. Available from: <http://pdet.mte.gov.br/microdados-rais-e-caged>
12. BRASIL. Decreto No 76.900, de 23 de dezembro de 1975 [Internet]. 1975 [cited 2020 Feb 28]. Available from: http://www.planalto.gov.br/ccivil_03/decreto/Antigos/D76900.htm
13. Fernandes FT, Chiavegatto Filho ADP. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. *Rev Bras Saúde Ocup*. 2019;44:1–12.
14. INEP. Indicadores Educacionais [Internet]. 2019 [cited 2019 Sep 18]. Available from: <http://portal.inep.gov.br/web/guest/indicadores-educacionais>
15. SEADE. Informações dos Municípios Paulistas [Internet]. 2019 [cited 2019 Feb 18]. Available from: <http://www.imp.seade.gov.br/frontend/#/>
16. Kuhn M, Johnson K. *Applied Predictive Modeling*. Springer. Springer New York, New York, NY; 2013. 1–595 p.
17. OECD. Urban population by city size [Internet]. 2018. Available from: <https://data.oecd.org/popregion/urban-population-by-city-size.htm>

18. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
19. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016. p. 785–94.
20. Bishop C. *Neural Networks for Pattern Recognition.* Oxford University Press; 1995.
21. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* 2nd ed. Springer; 2016. 764 p.
22. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 1997;
23. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1–73.
24. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Modell.* 2003;160(3):249–64.
25. Kuhn M. A Short Introduction to the caret Package. 2017 [cited 2017 Sep 22]; Available from: <https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf>
26. 15 Variable Importance | The caret Package [Internet]. [cited 2020 Aug 11]. Available from: <https://topepo.github.io/caret/variable-importance.html>
27. Santos HG dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018.
28. Rezende BA, De Medeiros AM, Da Silva AM, Assunção AA. Fatores associados à percepção de ruído ocupacional intenso pelos professores da educação básica no Brasil. *Rev Bras Epidemiol.* 2019;22:1–13.
29. 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 Aug 29;88(6):779–87.
30. 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. *Inq (United States).* 2016;53(2).
31. Ferris G, Bergin T, Wayne S. Personal Characteristics, Job Performance, and Absenteeism of Public School Teachers. *J Appl Soc Psychol.* 1988;552–63.
32. Rosenblatt Z, Shirom A. Predicting teacher absenteeism by personal background factors. *J Educ Adm.* 2005;43(2):209–25.
33. Miller RT, Murnane RJ, Willett JB. Do teacher absences impact student achievement? Longitudinal evidence from one urban school district. *Educ Eval Policy Anal.* 2008;30(2):181–200.

5.3 ARTIGO 3 – Ainda não enviado para publicação conforme as regras do PPG-SP.

5.3.1 Machine Learning para Predição de Doença Pulmonar Obstrutiva Crônica (DPOC) em adultos

Predição de Doença Pulmonar Obstrutiva Crônica (DPOC) em adultos com *machine learning*

Prediction of Chronic Obstructive Pulmonary Disease (COPD) in adults with machine learning

Fernando Timoteo Fernandes^{1,2}
ORCID: <https://orcid.org/0000-0003-2548-9685>

Thales Pardini Fagundes^{1,3}
ORCID: <https://orcid.org/0000-0002-3302-9913>

Alexandre Dias Porto Chiavegatto Filho¹
ORCID: <http://orcid.org/0000-0003-3251-9600>

¹ Universidade de São Paulo, Faculdade de Saúde Pública. São Paulo, SP, Brasil.

² Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho. São Paulo, SP, Brasil.

³ Hospital das Clínicas de Ribeirão Preto. Ribeirão Preto, SP, Brasil

Contato:

Fernando Timoteo Fernandes

E-mail:

fernando.fernandes@fundacentro.gov.br

RESUMO

Introdução: A doença pulmonar obstrutiva crônica (DPOC) é uma das principais morbidades em adultos acima de 40 anos. Parte da sua incidência na população é atribuída à exposição a fatores de risco no trabalho e no ambiente, incluindo o tabagismo e poeiras minerais. **Objetivos:** Identificar pacientes com maior risco de ter diagnóstico positivo de DPOC usando variáveis coletadas rotineiramente na atenção primária. **Métodos:** Foram analisados 120.289 participantes da base do UK Biobank acompanhados a partir de 2006. Desses, 1.731 (1,4%) tiveram diagnóstico positivo de DPOC. Ao todo, foram selecionadas 26 variáveis, entre dados demográficos, exames laboratoriais, hábitos e sintomas, para o ajuste de um modelo preditivo generalista de DPOC com algoritmos de *machine learning* (redes neurais artificiais, *extra trees*, *random forests*, *catboost*, *extreme gradient boosting and light gbm*). Além disso, foi selecionado um subconjunto de 7.628 participantes com histórico laboral na indústria da construção e da mineração para treinar um modelo especializado. Entre esses, 237 (3,11%) tiveram o diagnóstico positivo da doença no período. Todos os dados foram divididos aleatoriamente: 70% foram alocados para treinamento dos algoritmos e 30% para testes de performance do modelo ajustado. **Resultados:** Ambos os modelos apresentaram boa performance preditiva. O modelo generalista apresentou AUROC de 0.845, sensibilidade de 0.751 e especificidade de 0.782. No modelo especialista, observou-se AUROC de 0.841, sensibilidade de 0.817 e especificidade de 0.761. As cinco principais variáveis preditoras foram idade, tosse crônica, tabagismo, histórico de asma e expectoração. **Conclusão:** Os resultados mostram que é possível prever risco individual de diagnóstico de DPOC na população geral e em trabalhadores supostamente expostos a poeiras minerais usando variáveis comumente coletadas na atenção primária.

Keywords: Supervised Machine Learning; Chronic Obstructive Pulmonary Disease; COPD;

ABSTRACT

Introduction: Chronic obstructive pulmonary disease (COPD) is one of the most common diseases in adults over 40 years of age and is attributable to exposure to environmental and work factors, including tobacco and mineral dusts. **Objectives:** The aim of this study is to identify patients at higher risk of having a positive diagnosis of COPD through variables collected routinely in primary care. **Methods:** This research has been conducted using UK Biobank Resource. We analyzed data from 120,289 participants from the UK Biobank, recruited between 2006 and 2010 who answered the occupational history questionnaire. Of the total, 1731 (1.4%) had a positive diagnosis of COPD. Twenty-six variables were selected, including demographic, clinical and laboratory tests to build a generalist machine learning model capable of predicting COPD. In addition, a subset of 7,628 participants who work in construction and mining industry were selected to train a specialized model. Of these, 237 (3.11%) were diagnosed with the disease. Dataset were randomly divided: 70% were allocated for training the models and 30% for performance tests. **Results:** The generalist model presented AUROC of 0.845, sensitivity of 0.751 and specificity of 0.782. In the specialized model, we observed AUROC of 0.841, sensitivity of 0.817, specificity of 0.761. Top five main predictive variables for COPD were age, chronic cough, smoking, earlier diagnosis of asthma and chronic sputum. **Conclusion:** Machine learning algorithms can predict the risk of been diagnosed with COPD both in the general population and in workers supposedly exposed to mineral dust with satisfactory performance through data easily obtained in primary care.

Keywords: Supervised Machine Learning; Chronic Obstructive Pulmonary Disease; COPD.

1 INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is ranked in the top positions of diseases that most affect adults aged 50 and over when considering the years of life lost or the time of life living with the disease (Disability Adjusted Life Years – DALY).¹ In the United Kingdom, it is estimated that approximately 3 million people live with the disease.²

About 15% to 19% of COPD cases are attributable to occupational exposure.^{3,4} Cigarette smoke⁵, air pollution^{6,7}, diesel exhaust⁸ and silica dust⁹ are examples of agents found in work environment associated with the development of the disease.

Some studies investigated occupations with significant exposure to physical agents¹⁰ and evaluated the risk of developing COPD and other related diseases (e.g., silicosis¹¹). The results showed that construction and mining workers are more affected by respiratory diseases compared to those in occupations that do not have exposure to chemical agents.^{12,13,14}

Presence of COPD also increases the risk of developing lung cancer and anticipating the diagnosis can prevent disease worsening.¹⁵ Spirometry exams can diagnosis COPD, using forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC), considering values under 0,70 of FEV1/FVC ratio being COPD positive. Although spirometry being the gold standard for diagnosing COPD, in some situations it may not be properly used, especially for junior doctors, due to lack of training or familiarity with the equipment, or even knowing if there is any equipment available in the health facility.¹⁶

Early diagnosis of COPD in primary care or occupational health clinical centers could improve patients' quality of life, considering that the decline in forced expiratory volume in 1 second (FEV1) occur more quickly in the first stages of the disease¹⁷, allowing interventional measures to interrupt or control disease progression.

Models of artificial intelligence that predicts COPD in primary care could help health professionals ordering tests such as spirometry and diffusing capacity of the lung for carbon monoxide (DLCO) to confirm the presence of the disease. In addition, it is expected that the models can be adapted to low and middle-income countries (LMICs), such as Brazil, in which the estimated prevalence of the disease is about 17% among adults over 40 years.¹⁸ In other words, artificial intelligence have the potential to aid

doctors in figuring out the pre-test probability of COPD and to prioritize patients who should undergo spirometry.¹⁹

The aim of this study is to build machine learning models capable of finding the pre-test probability of the diagnosis of COPD, using variables routinely collected in primary care and in specialized worker health centers. Two approaches were considered for predictive model adjustment, one more generalist, aiming to be used in primary care, and the other more specialized, focused on patients who have been supposedly exposed to mineral dust.

2 METHODS

2.1 Population and data source

The UK Biobank database is a large prospective cohort with socioeconomic and biomedical data and has more than 500,000 participants randomly selected in the UK from 2006 to 2010²⁰ aged 40 to 69 years. The database has information collected through questionnaires, face-to-face assessments, and laboratory tests, including the occupational history of the participants.

For this study, individuals who completed the occupational history questionnaire were selected, totaling 120,289 participants. The considered outcome was COPD self-reported or hospital diagnosed^{5,21}. Since the term emphysema has already been used as a synonym for COPD, and is now understood as histological find of the disease²², a specific variable for the diagnosis of “emphysema” was also considered as an outcome of COPD. Diagnosis of emphysema or COPD were considered only when diagnosed with 40 years or older. Models developed with this population apply to the general population with or without specific exposures to environmental agents.

Then, a new filter was applied to create a subset of data holding participants who had some occupational history related to construction, mining and paving activities, using the first four digits of the standardized classification of occupations (SOC 2000²³) in UK Biobank occupation code field (Job Coding; 22601), resulting in 7,628 observations to be used for training and test the most specialized models. The selected occupation codes are available in supplementary table 1.

Figure 1 illustrates the population selection process, the outcomes and model development.

Figure 1. Population selection. a) Population selection for the development of generic models.
b) Population selection for the development of specialized modes.

2.2 Variable selection and predictive model development

Initially, 40 variables were chosen, including sociodemographic, hemogram, anthropometric measures, behavior (tobacco smoking) and respiratory symptoms variables. Then, the feature selection technique known as Boruta²⁴ was applied on the entire dataset including all 40 variables. This technique aims to show the most relevant variables to an outcome and may improve the performance of the developed models. The application of the variable selection technique found 22 relevant variables: age, BMI, hematocrit, leukocyte count, mean corpuscular volume (MCV), alkaline phosphatase (ALP), monocyte count, neutrophil count, pulse rate, C-reactive protein, creatinine, HDL, LDL, urea, ever smoker, chronic cough, sputum, dyspnea, history of asthma, lymphocyte count, albumin and triglycerides. In addition to these variables, another four variables were included: eosinophils count, exposure to tobacco at work, exposure to tobacco at home and work history in civil construction, totaling 26 variables. The models were developed using the twenty six selected variables.

Prior to model development, the selected predictors were preprocessed. Those with more than one category were transformed into dummy variables and continuous variables were standardized by its z-score. Correlation analysis between numerical variables was also carried out for those with a correlation greater than 0.90 (e.g., hematocrit and hemoglobin) and only one of them remained (e.g., hematocrit). To assess the predictive performance of the adjusted models on test data, metrics such as accuracy, sensitivity (also called recall), specificity, positive predicted value (PPV or precision), negative predicted value (NPV) and the F1 score were used. The latter metric is the harmonic mean between precision and sensitivity. A higher area under the ROC (AUROC) curve was the main criteria for the selection of the best adjusted model.

Six popular machine learning algorithms were selected: artificial neural networks²⁵, extra trees²⁶, random forests²⁷, catboost²⁸, extreme gradient boosting²⁹ and light gradient boosting machine³⁰. The models were trained with 70% of the data and evaluated with 30% of the remaining data to measure predictive performance. The cross-validation technique with 10 partitions (K-Fold) and the Bayesian optimization technique were used to adjust the hyperparameters (HyperOpt³¹) on the training data.

To avoid the algorithm prioritize the majority class due to the imbalanced datasets, we applied random under sampling in the training data, in which the observations of the majority class were randomly selected for exclusion, which approximates the proportion of cases and not cases.³¹

After model development, the variable contribution for each model was evaluated using the SHAP³³ technique. All analyzes were conducted using the python programming language with scikit-learn library.

3 RESULTS

3.1 Descriptive statistics

Table 1 shows the descriptive analysis of the data used for the generalist model. Sex variable although not used in the model was considered to show demographic characteristics.

Table 1 – Descriptive analysis comparing COPD cases and non-cases.

		Negative COPD (n=118.558)	Positive COPD (n=1.731)	Total (120.289)	
Variable		Mean (SD)	Mean (SD)	Mean (SD)	p-value
Demographics					
Sex					
	male, n (%)	52472 (44.26%)	955 (55.17%)	53427 (44.42%)	p<0.0001 ^a
	female, n (%)	66086 (55.74%)	776 (44.83%)	66862 (55.58%)	
Age (years)		55.993 (7.666)	60.756 (5.942)	56.062 (7.665)	p<0.0001 ^b
Age Group					
	40-49	27509 (23.20%)	95 (5.49%)	27604 (22.95%)	p<0.0001 ^a
	50-59	45055 (38.00%)	503 (29.06%)	45558 (37.87%)	
	60-69	45650 (38.50%)	1118 (64.59%)	46768 (38.88%)	
	70+	344 (0.29%)	15 (0.87%)	359 (0.30%)	
Vital Signs and anthropometrics measures					
Pulse rate (bpm)		68.821 (11.334)	72.438 (12.76)	68.873 (11.364)	p<0.0001 ^b
BMI		26.746 (4.534)	28.086 (5.37)	26.766 (4.549)	p<0.0001 ^b
Tobacco exposure					
Ever Smoker					
	yes, n (%)	69.505 (58.63%)	1446 (83.54%)	70.951 (58.68%)	p<0.0001 ^a
	no, n (%)	49.053 (41.37%)	285 (16.46%)	49.338 (41.02%)	
Exposure to tobacco at home					
	yes, n (%)	9536 (8.04%)	157 (9.07%)	9.693 (8.06%)	p<0.0001 ^a
	no, n (%)	109022 (91.96%)	1574 (90.93%)	110.596 (91.94%)	
Exposure to tobacco at work					
	yes, n (%)	66.861 (56.40%)	1275 (73.66%)	68.136 (56.64%)	p<0.0001 ^a
	no, n (%)	51.697 (43.60%)	456 (26.34%)	52.153 (43.36%)	
Comorbidities					
Diabetes					
	yes, n (%)	3807 (3.21%)	137 (7.91%)	3.944 (3.28%)	p<0.0001 ^a
	no, n (%)	114.751 (96.79%)	1594 (92.09%)	116.345 (96.72%)	
Cardiovascular diseases					
	yes, n (%)	28.861 (24.34%)	731 (42.23%)	29.592 (24.60%)	p<0.0001 ^a
	no, n (%)	89.697 (75.66%)	1000 (57.77%)	90.697 (75.40%)	

Previous asthma diagnosis					
yes, n (%)	14.626 (12.34%)	593 (34.26%)	15.219 (12.65%)	p<0.0001 ^a	
no, n (%)	103.932 (87.66%)	1138 (65.74%)	105.070 (87.35%)		
Symptoms					
Chronic cough					
yes, n (%)	15.460 (13.04%)	856 (49.45%)	16.316 (13.56%)	p<0.0001 ^a	
no, n (%)	103098 (96.96%)	875 (50.55%)	103.973 (86.44%)		
Chronic sputum					
yes, n (%)	9.640 (8.13%)	703 (40.61%)	10.343 (8.60%)	p<0.0001 ^a	
no, n (%)	108.918 (91.87%)	1.028 (59.39%)	109.946 (91.40%)		
Dyspnea					
yes, n (%)	2.881 (2.43%)	266 (15.37%)	3.147 (2.62%)	p<0.0001 ^a	
no, n (%)	115.677 (97.57%)	1465 (84.63%)	117.142 (97.38%)		
Count of Blood Cells					
Leukocytes (10^9 cells/L)	6.67 (2.041)	7.56 (1.979)	6.682 (2.043)	p<0.0001 ^b	
Neutrophils (10^9 cells/L)	4.075 (1.315)	4.769 (1.605)	4.085 (1.322)	p<0.0001 ^b	
Basophils (10^9 cells/L)	0.033 (0.049)	0.043 (0.08)	0.033 (0.050)	p<0.0001 ^b	
Eosinophils (10^9 cells/L)	0.17 (0.136)	0.208 (0.166)	0.171 (0.137)	p<0.0001 ^b	
Hematocrit (%)	41.021 (3.462)	41.969 (3.591)	41.035 (3.466)	p<0.0001 ^b	
Monocyte count (10^9 cells/L)	0.465 (0.197)	0.531 (0.228)	0.466 (0.198)	p<0.0001 ^b	
Lymphocyte count (10^9 cells/L)	1.921 (1.264)	2.005 (0.752)	1.922 (1.258)	p<0.0001 ^b	
MCV (fL)	91.161 (4.282)	92.234 (4.797)	91.176 (4.292)	p<0.0001 ^b	
MCHC (g/dL)	34.531 (1.043)	34.412 (0.982)	34.529 (1.042)	p<0.0001 ^b	
Albumin (g/L)	45.375 (2.580)	44.718 (2.746)	45.365 (2.583)	p<0.0001 ^b	
C-Reactive Protein (mg/L)	2.22 (3.874)	4.02 (6.476)	2.246 (3.929)	p<0.0001 ^b	
Creatinine (umol/L)	71.846 (15.614)	75.395 (28.733)	71.897 (15.882)	p<0.0001 ^b	
Urea (mmol/L)	5.333 (1.285)	5.653 (1.903)	5.337 (1.296)	p<0.0001 ^b	
Alkaline Phosphatase (U/L)	80.982 (23.174)	87.656 (27.246)	81.077 (23.78)	p<0.0001 ^b	
HDL (mmol/L)	1.484 (0.382)	1.427 (0.408)	1.483 (0.382)	p<0.0001 ^b	
LDL (mmol/L)	3.582 (0.845)	3.396 (0.904)	3.58 (0.846)	p<0.0001 ^b	

Tryglicerydes (mmol/L)	1.668 (0.97)	1.812 (1.061)	1.67 (0.971)	p<0.0001 ^b
Occupational exposure				
Work history in construction industry				
yes, n (%)	7.391 (6.23%)	237 (13.69%)	7.628 (6.34%)	p<0.0001 ^a
no, n (%)	111.167 (93.77%)	1.494 (86.31%)	112.661 (93.66%)	

^a chi-square association test

^b Student's t test for independent samples

As shown in Table 1, the variables used to develop the models were related to the outcome with statistical significance according to Pearson's chi-square test for categorical variables and the Student's t test for continuous variables.

3.2 Predictive Performance of Algorithms

The predictive performance of generalist and specialist algorithms was analyzed. For generalist models, the entire database with occupational history questionnaire fulfilled was used (n = 120,289) with COPD outcome (n = 1731, 1.43%). For the specialized models, that consider only workers supposedly exposed to mineral dusts, the filtered database (n = 7,628) with COPD outcome (n = 237, 3.11%) was used.

The correspondent performance of the five algorithms for each approach were compared, using the 95% confidence interval of the AUROC curve.

Table 2 shows the performance of each adjusted model according to test set as results.

Table 2 – Predictive performance of adjusted models

Strategy	AUROC [95% C.I.]	Sensitivity	Specificity	PPV	NPV	F1
Generalist						
<i>Extra trees</i>	0.838 [0.82;0.86]	0.713	0.803	0.050	0.995	0.094
<i>Random Forest</i>	0.844 [0.83;0.86]	0.755	0.789	0.050	0.995	0.093
<i>XGBoost^a</i>	0.845 [0.83;0.86]	0.751	0.782	0.048	0.995	0.090
<i>Lightgbm</i>	0.837 [0.82;0.85]	0.840	0.671	0.036	0.997	0.069
<i>Catboost</i>	0.842 [0.83;0.86]	0.742	0.785	0.048	0.995	0.090
<i>MLP</i>	0.828 [0.81;0.84]	0.669	0.835	0.056	0.994	0.103
Specialist						
<i>Extra trees^b</i>	0.841 [0.80;0.89]	0.817	0.761	0.099	0.992	0.176
<i>Random Forest</i>	0.845 [0.80;0.89]	0.817	0.740	0.091	0.992	0.165
<i>XGBoost</i>	0.840 [0.79;0.89]	0.817	0.699	0.080	0.992	0.146
<i>Lightgbm</i>	0.829 [0.78;0.88]	0.958	0.340	0.044	0.996	0.085
<i>Catboost</i>	0.766 [0.70;0.83]	0.732	0.729	0.080	0.988	0.144
<i>MLP</i>	0.719 [0.66;0.78]	0.901	0.364	0.043	0.991	0.083

^a Best generalist model according to AUC^b Best specialized model according to AUC

As seen in Table 2, all models presented high predictive performance with AUROC around 0.80. The best generalist model was the one that employed the XGboost algorithm, with an AUROC of 0.845 [0.83; 0.86], sensitivity of 0.751 and specificity of 0.782. The best specialist model was the one that used the Extra Trees algorithm that reached AUROC curve of 0.841 [0.80; 0.89], sensitivity of 0.817 and specificity of 0.761. Hyperparameters of the best models are presented in supplementary table 2. A comparison of the AUROC curve of the models is available in Supplementary Figure 1.

The variables with the best ranking according to SHAP technique are shown in Figure 2. SHAP aims to explain the prediction of an instance by computing the contribution of each variable to the prediction. SHAP comes from game theory and tell

how each feature contribute to the prediction, considering all other players (variables). The more contribution to the prediction outcome, greater the "payout".

Figure 2 shows the top 10 variables according to SHAP for generalist and specialized models.

Figure 2 – Variable importance according to SHAP. The x-axis shows the SHAP value of each variable for each observation and the impact towards predicting an outcome. The y-axis shows variable importance by rank. Colors are based on variable values, where high values are red and low values are blue. a) Top 10 variables for the best generalist model (Random Forest). b) Top 10 variables for the best specialist model (Extra trees).

In Figure 2, age, chronic cough, never smoker, absence of asthma history and absence of chronic sputum were the five variables that most contributed to the generalist model (Figure 2a). For the specialist model, the five main variables were chronic cough, chronic sputum, absence of asthma history, ever smoker and age.

Figure 3 shows the probability density curves for both models.

Figure 3 – Risk probability density curves. a) Density curves to predict COPD risk in the generalist model. B) Density curves to predict COPD in the specialist model.

The results in Figure 3 show that although there is an intersection between the curves, the distinction between classes is clear without full overlapping, supporting the good performance.

Figure 4 shows risk group clustering among participants with COPD according to the predictions of the developed models on test data.

Figure 4 – Observed COPD according to predicted risk groups on test set. Low risk refers to first to sixtieth decile. Medium risk refers to seventieth to eightieth decile. High risk refers to ninetieth to twentieth decile.

Results from Figure 4 shows that both models correctly grouped more than 85% of observed cases from medium to high in the test datasets.

4 Discussion

When analyzing the data of more than 120,289 people from the UK Biobank database, it was found that 1,731 (1.4%) people were diagnosed at some point in life with COPD or pulmonary emphysema. This proportion increases when we select workers with supposedly some exposure to recognized risk agents as mineral dusts such as silica⁹, and in 7,628 participants, 237 (3.11%) people developed the disease, showing a higher prevalence. The present study showed that the use of machine learning models to predict the probability of diagnosing COPD using common variables in primary care (pre-test) is possible and has satisfactory performance. Patients predicted with higher risks, may be followed for further COPD exams such as spirometry.

The presence of COPD can compromise the patient's quality of life for itself, in addition to being associated with more serious conditions, such as lung cancer³⁴ and death.^{35,36} It is known that the identification of the disease is still a challenge. The definition of early COPD is heterogeneous, but there is consensus among specialists about the possible benefit of intervening before the appearance of clinical manifestations.³⁷

The results of this study show that it is feasible to determine a pre-test probability for COPD diagnosis in patients who are received in the primary care, using routinely collected variables and taking the patient's health history. Both the generalist model (AUROC: 0.845) and the specialized model (AUROC: 0.841) had good performance, with the area under the curve between both models overlapping, showing similar decision-making capacity. The specialist model achieves a higher precision performance (10%) compared to the generalist model (4.8%), which can be explained due to the higher prevalence of the disease in the specialist model.

The identification of people at higher risks of being diagnosed with COPD can help in the early diagnosis of the disease, especially in places lacking basic equipment, such as spirometry equipment, as in developing countries.³⁸ This study increases the chances of finding adults at risk for COPD and who need confirmatory spirometry, in comparison with other studies.¹⁹

Age, chronic cough, smoking, history of asthma and chronic sputum were important predictors for the adjusted models and they are also risk factors and clinical

manifestations known to be associated with COPD.^{39,40} In addition to these variables, it is interesting to note that some blood markers, such as C-reactive protein, white blood cell count, neutrophil count, mean corpuscular volume also played an important role in the decision making process of the algorithms, being listed among the top 10 variables for both models, which makes sense considering the inflammatory and hypoxemic nature of the disease.^{41,42}

This study has a few limitations, such as the high data imbalance, which may cause majority class prioritization⁴³ during model development. To minimize this problem, we applied random undersampling technique in the majority class during training process. Some occupations were already been explored in relation to the significant association with the development of COPD in the UK Biobank database.¹² However, the authors analyzed the association of each occupation with COPD.

Finally, we conclude that the early prediction of COPD can be performed with good predictive performance using machine learning algorithms. The use of these algorithms can help the medical team in decision making in situations where availability of test such as spirometry are limited.

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10258):1204–22.
2. NICE. Chronic obstructive pulmonary disease in adults [Internet]. NICE Guideline. 2011 [cited 2021 Apr 27]. Available from: www.nice.org.uk/guidance/qs10
3. Blanc PD, Torén K. Occupation in chronic obstructive pulmonary disease and chronic bronchitis: an update. *Int J Tuberc Lung Dis*. 2007 Mar;11(3):251–7.
4. Hnizdo E, Sullivan PA, Bang KM, Wagner G. Association between chronic obstructive pulmonary disease and employment by industry and occupation in the US population: A study of data from the Third National Health and Nutrition Examination Survey. *Am J Epidemiol*. 2002;156(8):738–46.
5. Agustí A, Celli B. Natural history of COPD: gaps and opportunities. *ERJ Open Res* [Internet]. 2017;3(4):00117–2017. Available from: <http://dx.doi.org/10.1183/23120541.00117-2017>
6. Sarkar C, Zhang B, Ni M, Kumari S, Bauermeister S, Gallacher J, et al. Environmental correlates of chronic obstructive pulmonary disease in 96 779 participants from the UK Biobank: a cross-sectional, observational study. *Lancet Planet Heal* [Internet]. 2019;3(11):e478–90. Available from: [http://dx.doi.org/10.1016/S2542-5196\(19\)30214-1](http://dx.doi.org/10.1016/S2542-5196(19)30214-1)
7. Zhao J, Li M, Wang Z, Chen J, Zhao J, Xu Y, et al. Role of PM2.5 in the development and progression of COPD and its mechanisms. *Respir Res*. 2019;20(1):1–13.
8. Cullinan P. Occupation and chronic obstructive pulmonary disease (COPD). *Br Med Bull*. 2012;104(1):143–61.
9. Hnizdo E, Vallyathan V. Chronic obstructive pulmonary disease due to occupational exposure to silica dust: A review of epidemiological and pathological evidence. *Occup Environ Med*. 2003;60(4):237–43.
10. Brown TP, Rushton L. Mortality in the UK industrial silica sand industry: 2. A retrospective cohort study. *Occup Environ Med*. 2005;62(7):446–52.
11. Hoy RF, Glass DC, Dimitriadis C, Hansen J, Hore-Lacy F, Sim MR. Identification of early-stage silicosis through health screening of stone benchtop industry workers in Victoria, Australia. *Occup Environ Med*. 2020;1–7.
12. De Matteis S, Jarvis D, Hutchings S, Darnton A, Fishwick D, Sadhra S, et al. Occupations associated with COPD risk in the large population-based UK Biobank cohort study. *Occup Environ Med*. 2016;73(6):378–84.
13. Cassidy A, Mannetje A, Van Tongeren M, Field JK, Zaridze D, Szeszenia-Dabrowska N, et al. Occupational exposure to crystalline silica and risk of lung cancer: A multicenter case-control study in Europe. *Epidemiology*. 2007;18(1):36–43.
14. Stocks SJ, Turner S, McNamee R, Carder M, Hussey L, Agius RM. Occupation and work-related ill-health in UK construction workers. *Occup Med (Chic Ill)*. 2011;61(6):407–15.
15. Sekine Y, Katsura H, Koh E, Hiroshima K, Fujisawa T. Early detection of COPD is important for lung cancer surveillance. *Eur Respir J*. 2012;39(5):1230–40.
16. Roberts, N. J., Smith, S. F. & Partridge, M. R. Why is spirometry underused in the diagnosis of the breathless patient: a qualitative study. *BMC Pulm. Med*. 11, 37 (2011).

17. Decramer M, Cooper CB. Treatment of COPD: The sooner the better? *Thorax*. 2010;65(9):837–41.
18. Cruz MM, Pereira M. Epidemiology of chronic obstructive pulmonary disease in Brazil: A systematic review and meta-analysis. *Cienc e Saude Coletiva*. 2020;25(11):4547–57.
19. Hill K, Hodder R, Blouin M, Heels-Ansdell D, Guyatt G, Goldstein R. Identifying adults at risk of COPD who need confirmatory spirometry in primary care: Do symptom-based questions help? *Can Fam Physician*. 2011;57(2):51–7.
20. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015 Mar;12(3):e1001779.
21. Ito K, Barnes PJ. COPD as a disease of accelerated lung aging. *Chest* [Internet]. 2009;135(1):173–80. Available from: <http://dx.doi.org/10.1378/chest.08-1419>
22. Barnes PJ. Chronic Obstructive Pulmonary Disease. *N Engl J Med* [Internet]. 2000;343(4):269–80. Available from: <https://doi.org/10.1056/NEJM200007273430407>
23. Office for National Statistics. SOC 2000 [Internet]. 2000 [cited 2021 Apr 27]. Available from: <https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassification/soc/socarchive>
24. Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw*. 2010;36(11):1–13.
25. Bishop C. *Neural Networks for Pattern Recognition*. Oxford University Press; 1995.
26. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;(November 2005):3–42.
27. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
28. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *CoRR* [Internet]. 2018;abs/1810.1. Available from: <http://arxiv.org/abs/1810.11363>
29. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. p. 785–94.
30. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 30, 3146–3154 (2017).
31. Bergstra J, Yamins D, Cox D. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. *Proc 12th Python Sci Conf*. 2013;(Scipy):13–9.
32. He H, Ma Y. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons; 2013. 216 p.
33. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;2017-Decem(Section 2):4766–75.
34. Brenner DR, McLaughlin JR, Hung RJ. Previous lung diseases and lung cancer risk: A systematic review and meta-analysis. *PLoS One*. 2011;6(3):1–10.
35. Hartl S, Lopez-Campos JL, Pozo-Rodriguez F, Castro-Acosta A, Studnicka M, Kaiser B, et al. Risk of death and readmission of hospital-admitted COPD exacerbations: European COPD Audit. *Eur Respir J* [Internet]. 2016;47(1):113–21. Available from: <https://erj.ersjournals.com/content/47/1/113>
36. Mannino DM, Kiriz VA. Changing the burden of COPD mortality. *Int J Chron Obstruct Pulmon Dis*. 2006;1(3):219–33.

37. Di Marco F, Balbo P, de Blasio F, Cardaci V, Crimi N, Girbino G, et al. Early management of COPD: where are we now and where do we go from here? A Delphi consensus project. *Int J Chron Obstruct Pulmon Dis* [Internet]. 2019 Feb 4;14:353–60. Available from: <https://pubmed.ncbi.nlm.nih.gov/30787604>
38. Menezes AMB, Perez-Padilla R, Jardim JRB, Muiño A, Lopez MV, Valdivia G, et al. Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *Lancet* (London, England). 2005 Nov;366(9500):1875–81.
39. Rabe KF, Watz H. Chronic obstructive pulmonary disease. *Lancet* [Internet]. 2017;389(10082):1931–40. Available from: [http://dx.doi.org/10.1016/S0140-6736\(17\)31222-9](http://dx.doi.org/10.1016/S0140-6736(17)31222-9)
40. Postma DS, Rabe KF. The Asthma-COPD Overlap Syndrome. *N Engl J Med*. 2015 Sep;373(13):1241–9.
41. Silva DR, Gazzana MB, Knorst MM. C-reactive protein levels in stable COPD patients: a case-control study. *Int J Chron Obstruct Pulmon Dis* [Internet]. 2015 Aug 31;10:1719–25. Available from: <https://pubmed.ncbi.nlm.nih.gov/26357470>
42. Aksu F, Capan N, Aksu K, Ofluoğlu R, Canbakan S, Yavuz B, et al. C-reactive protein levels are raised in stable Chronic obstructive pulmonary disease patients independent of smoking behavior and biomass exposure. *J Thorac Dis* [Internet]. 2013 Aug;5(4):414–21. Available from: <https://pubmed.ncbi.nlm.nih.gov/23991296>
43. Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: A review. *Int J Pattern Recognit Artif Intell*. 2009;23(4):687–719.

Supplementary Table 1 – Selected occupations.

SOC Code	Occupation
112 – Production Managers	
1122	Managers in construction
1123	Managers in mining and energy
312 - Draughtspersons And Building Inspectors	
3123	Building inspectors
521 - Metal Forming, Welding And Related Trades	
5211	Smiths and forge workers
5212	Moulders, core makers, die casters
5213	Sheet metal workers
5214	Metal plate workers, shipwrights, riveters
5215	Welding trades
5216	Pipe fitters
522 - Metal Machining, Fitting And Instrument Making Trades	
5221	Metal machining setters and setter-operators
5222	Tool makers, tool fitters and markers-out
5223	Metal working production and maintenance fitters
524 - Electrical Trades	
5241	Electricians, electrical fitters
531 – Construction Trades	
5311	Steel erectors
5312	Bricklayers, masons
5313	Roofers, roof tilers and slaters
5315	Carpenters and joiners
5316	Glaziers, window fabricators and fitters
5319	Construction trades n.e.c.
532 – Building Trades	
5321	Plasterers
5322	Floorers and wall tilers
5323	Painters and decorators
549 - Skilled Trades n. e. c.	
5491	Glass and ceramics makers, decorators and finishers
5492	Potter, ceramic caster, ceramic artist, ceramic lithographer
812 - Plant And Machine Operatives	
8122	Coal mine operatives
8123	Quarry workers and related operatives
814 – Construction Operatives	
8141	Scaffolders, staggers, riggers
8142	Road construction operatives
8143	Rail construction and maintenance operatives
8149	Construction operatives n.e.c.
912 – Elementary Construction Occupations	
9121	Labourers in building and woodworking trades
9129	Labourers in other construction trades n.e.c.

* 4 digit codes not listed were disregarded for this study.

Supplementary Table 2 – Best algorithms's hiperparameters according to AUC.

Strategy	Hiperparameters
Generalist	<i>colsample_bylevel</i> : 1, <i>colsample_bynode</i> : 1,
	<i>colsample_bytree</i> : 0.55, <i>gamma</i> : 0.65,
	<i>importance_type</i> : 'gain', <i>learning_rate</i> : 0.075,
	<i>max_delta_step</i> : 0, <i>max_depth</i> : 6,
<i>XGBoost</i>	<i>min_child_weight</i> : 3.0, <i>n_estimators</i> : 75,
	<i>reg_alpha</i> : 1.25, <i>reg_lambda</i> : 1,
	<i>scale_pos_weight</i> : 1.25, <i>subsample</i> : 0.55,
Specialist	<i>bootstrap</i> : True, <i>max_depth</i> : 40,
	<i>max_features</i> : sqrt,
<i>Extra trees</i>	<i>min_samples_leaf</i> : 4, <i>min_samples_split</i> : 10,
	<i>n_estimators</i> : 175

Supplementary Figure 1. Area under the curve comparison between best models.

6 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Nesta tese, foram utilizadas técnicas de aprendizagem supervisionada de *machine learning* em problemas de classificação de diferentes áreas da saúde ocupacional, tais como a saúde dos professores e nos trabalhadores que atuam diretamente ou indiretamente na construção civil. No primeiro artigo foi feita uma revisão bibliográfica para conhecer o estado da arte dos algoritmos utilizados nas áreas de saúde pública e de SST com o intuito de identificar objetivos de uso e tendências, percebendo-se uma ampla aceitação das técnicas de *machine learning* pela comunidade científica da área da saúde, dado o incremento anual de número de estudos que utilizaram tais algoritmos. Há relatos de uso tanto de aprendizagem supervisionada quanto de aprendizagem não-supervisionada, conforme descritos no capítulo 2 desta tese. Nota-se uma evolução na descrição dos resultados dos achados dos artigos e das métricas reportadas, com o uso de novas métricas como o F1 score, mais destaque à precisão e às calibrações dos modelos. O artigo apresenta um quadro destinado a elencar estudos que utilizaram algoritmos de aprendizagem supervisionada e não-supervisionada, assim como um quadro com os principais temas e objetivos de estudo.

No segundo artigo, foram desenvolvidos modelos supervisionados para prever o afastamento por doenças e doenças relacionadas ao trabalho em professores da rede pública municipal do ensino infantil. O melhor modelo mostrou que é possível prever o absenteísmo docente ocasionado por doenças com boa performance preditiva usando apenas dados públicos. O estudo tem limitações como a não possibilidade de acompanhar a trajetória dos docentes e não ser possível contabilizar o número de vínculos empregatícios simultâneos, devido à anonimidade dos dados. No entanto, observa-se um grande potencial para novos estudos exploratórios em bases públicas e anônimas para diferentes ocupações formais. No caso dos docentes, é possível estimar o risco de

ausência laboral em uma determinada etapa de ensino, o que pode melhorar as estatísticas atuais que apresentam dados agregados para todo o ensino básico.⁸¹

No terceiro artigo, foi abordado o uso de algoritmos de *machine learning* para identificar pacientes com risco elevado de doença pulmonar obstrutiva crônica (DPOC). A proposta deste modelo é a de que ele seja utilizado como um pré-teste a partir de dados coletados rotineiramente na atenção primária, podendo representar uma triagem inicial para a realização subsequente de exames confirmatórios. Foram comparados modelos generalistas, para todas as ocupações, com modelos filtrados apenas para ocupações que pudessem ter sido expostas a agentes químicos, que por sua vez, podem aumentar os riscos de DPOC. Ambos os modelos tiveram boa performance preditiva, com destaque para a área abaixo da curva superior no modelo generalista e a melhor precisão no modelo especialista. Esses modelos, apenas com dados coletados na atenção primária, podem servir de triagem para diagnóstico antecipado de DPOC, doença comum a trabalhadores expostos por diversos anos à poeira de sílica.

Dentre as técnicas utilizadas nos dois artigos, além das métricas analisadas para seleção de melhor performance, deu-se um enfoque também na interpretabilidade dos algoritmos para avaliar quais foram as principais variáveis preditoras para cada algoritmo. No primeiro artigo, utilizou-se a importância de variáveis e, no segundo artigo foi utilizado o Shapley Additive Values (SHAP). Apesar de não ter interpretação causal, algumas variáveis se destacam e podem direcionar novos estudos para a coleta de variáveis semelhantes que melhorem a performance preditiva dos modelos. Também cabe citar a técnica de seleção de variáveis Boruta, que pode facilitar a seleção de múltiplos preditores em conjuntos de dados com múltiplas variáveis.

A proposta de aplicação em temas distintos, mostra que de algoritmos de *machine learning* apresentam bons resultados na área de saúde pública e de SST. Espera-se que

este estudo possa contribuir para uma maior adoção de modelos preditivos pelos pesquisadores em SST, permitindo identificar antecipadamente trabalhadores expostos a riscos de desfechos de saúde adversos devido à sua ocupação e sugerir a adoção de medidas preventivas que inibam ou minimizam esses riscos.

7 REFERÊNCIAS BIBLIOGRÁFICAS

1. Ramazzini, B. *As doenças dos trabalhadores*. (Fundacentro, 2016).
2. Mendes, R. & Dias, E. C. Da medicina do trabalho à saúde do trabalhador. *Rev. Saúde Pública* 25, 341–349 (1991).
3. BRASIL. MINISTÉRIO DO TRABALHO E EMPREGO. Portaria GM/SSSTb nº25, de 29/12/1994. Norma Regulamentadora N° 9-Programa de Prevenção de Riscos Ambientais. Diário Oficial da República Federativa do Brasil, 1994.
4. Cruz, J. A. & Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2, 59–77 (2007).
5. Correa, D. N. L., Paniagua, L. R. B., Noguera, J. L. V., Pinto-Roa, D. P. & Toledo, L. A. S. Computerized Diagnosis of Melanocytic Lesions Based on the ABCD Method. *Proc. - 2015 41st Lat. Am. Comput. Conf. CLEI 2015* 19, 1–22 (2015).
6. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17 (2015).
7. Maity, N. G. & Das, S. Machine learning for improved diagnosis and prognosis in healthcare. in *2017 IEEE Aerospace Conference* 1–9 (2017). doi:10.1109/AERO.2017.7943950
8. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer, 2016). doi:10.1007/b94608
9. Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. (Springer, 2013).
10. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer New York, New York, NY, 2013). doi:10.1007/978-1-4614-6849-3
11. Gilbert, N., Assar, R. & Martinez, R. Creation of classification model using machine learning; to detect dysphonia work-related. *Occup. Environ. Med.* 76, A80–A80 (2019).
12. Sarkar, S., Vinay, S., Raj, R., Maiti, J. & Mitra, P. Application of optimized machine learning techniques for prediction of occupational accidents. *Comput. Oper. Res.* 106, 210–224 (2019).
13. Liu, H. *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* 9, 366 (2009).
14. Wang, X. *et al.* Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. *Occup. Environ. Med.* 77, 597–602 (2020).
15. Chen, Z. *et al.* Metabolomics screening of serum biomarkers for occupational exposure of titanium dioxide nanoparticles. *Nanotoxicology* 0, 1–18 (2021).
16. Aquilina, N. J., Delgado-Saborit, J. M., Bugelli, S., Ginies, J. P. & Harrison, R. M. Comparison of Machine Learning Approaches with a General Linear Model To Predict Personal Exposure to Benzene. *Environ. Sci. Technol.* 52, 11215–11222 (2018).
17. Shatte, A. B. R., Hutchinson, D. M. & Teague, S. J. Machine learning in mental health: a scoping review of methods and applications. *Psychol. Med.* 49, 1426–1448 (2019).
18. Id, A. B., Thielmann, A. & Weltermann, B. Chronic stress in practice assistants : An analytic approach comparing four machine learning classifiers with a standard logistic regression model. *PLoS One* 1–15 (2021). doi:10.1371/journal.pone.0250842
19. Breiman, L. Random forests. *Mach. Learn.* 45, 5–32 (2001).

20. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016). doi:10.1145/2939672.2939785
21. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *CoRR* **abs/1810.1**, (2018).
22. Bishop, C. *Neural Networks for Pattern Recognition*. (Oxford University Press, 1995).
23. R Studio. (2021). Available at: www.rstudio.com. (Accessed: 10th May 2021)
24. Jupyter Project. (2021). Available at: <https://jupyter.org/>. (Accessed: 10th May 2021)
25. Google. Google Colab. (2021). Available at: <https://colab.research.google.com/>. (Accessed: 10th May 2021)
26. He, H. & Ma, Y. *Imbalanced learning: foundations, algorithms, and applications*. (John Wiley & Sons, 2013).
27. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. (2021). Available at: <https://christophm.github.io/interpretable-ml-book/>. (Accessed: 10th May 2021)
28. Adadi, A. & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).
29. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 4766–4775 (2017).
30. Gensheimer, M. F. *et al.* Automated model versus treating physician for predicting survival time of patients with metastatic cancer. *J. Am. Med. Informatics Assoc.* (2020). doi:10.1093/jamia/ocaa290
31. Fernandes, F. T. *et al.* A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. *Sci. Rep.* **11**, 3343 (2021).
32. Allen, C., Tsou, M. H., Aslam, A., Nagel, A. & Gawron, J. M. Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS One* **11**, 1–10 (2016).
33. Plant, D. & Barton, A. Machine learning in precision medicine: lessons to learn. *Nat. Rev. Rheumatol.* **17**, 5–6 (2021).
34. Rajpurkar, P. *et al.* Evaluation of a Machine Learning Model Based on Pretreatment Symptoms and Electroencephalographic Features to Predict Outcomes of Antidepressant Treatment in Adults With Depression: A Prespecified Secondary Analysis of a Randomized Clinical Trial. *JAMA Netw. open* **3**, e206653 (2020).
35. Heo, J. *et al.* Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke* **50**, 1263–1265 (2019).
36. Farhadian, M., Aliabadi, M. & Darvishi, E. Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods. *Indian J. Occup. Environ. Med.* **19**, 84–9 (2015).
37. 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* **88**, 779–787 (2015).
38. Zhao, Y. *et al.* Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. *Ear Hear.* **40**, 690–699 (2019).
39. Lee, Y. C., Huang, S. C., Huang, C. H. & Wu, H. H. A new approach to identify high burnout medical staffs by kernel K-means cluster analysis in a regional teaching hospital in Taiwan. *Inq. (United States)* **53**, (2016).

40. Librenza-Garcia, D. et al. Prediction of depression cases, incidence, and chronicity in a large occupational cohort using machine learning techniques: an analysis of the ELSA-Brasil study. *Psychol. Med.* 1–9 (2020). doi:DOI: 10.1017/S0033291720001579
41. Sasikumar, V. & Binoosh, S. C. A. B. A model for predicting the risk of musculoskeletal disorders among computer professionals. *Int. J. Occup. Saf. Ergon.* **26**, 384–396 (2020).
42. de Oliveira, J. R. G. A importância da ginástica laboral na prevenção de doenças ocupacionais. *Rev. Educ. Física* **76**, (2017).
43. Mendes, R. Importância das pequenas empresas industriais no problema de acidentes do trabalho em São Paulo. *Rev. Saúde Pública* **10**, 315–325 (1976).
44. Chaga, D. A importância da cultura de segurança na prevenção dos acidentes de trabalho. (2016). Available at: <https://repositorium.sdum.uminho.pt/handle/1822/20167>. (Accessed: 15th May 2021)
45. William, W., Ware, A., Basaza-Ejiri, A. H. & Obungoloch, J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Comput. Methods Programs Biomed.* **164**, 15–22 (2018).
46. Batista, A. F. de M., Miraglia, J. L., Donato, H. R. & Chiavegatto Filho, A. D. P. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv* (2020).
47. Marucci-Wellman, H. R., Corns, H. L. & Lehto, M. R. Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. *Accid. Anal. Prev.* **98**, 359–371 (2017).
48. Vallmuur, K. et al. Harnessing information from injury narratives in the ‘big data’ era: understanding and applying machine learning for injury surveillance. *Inj. Prev.* (2016). doi:10.1136/injuryprev-2015-041813
49. Kim, T.-W., Koh, D.-H. & Park, C.-Y. Decision tree of occupational lung cancer using classification and regression analysis. *Saf. Health Work* **1**, 140–8 (2010).
50. Fong, J., Ocampo, R., Gross, D. P. & Tavakoli, M. Intelligent Robotics Incorporating Machine Learning Algorithms for Improving Functional Capacity Evaluation and Occupational Rehabilitation. *J. Occup. Rehabil.* **30**, 362–370 (2020).
51. Zhernova, P., Bodyanskiy, Y., Yatsenko, B. & Zavgorodnii, I. Detection and Prevention of Professional Burnout Using Machine Learning Methods. in *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)* 218–221 (2020). doi:10.1109/TCSET49122.2020.235426
52. Foster, K. R., Koprowski, R. & Skufca, J. D. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *Biomed. Eng. Online* **13**, 1–9 (2014).
53. Kuhn, M. A Short Introduction to the caret Package. (2017).
54. Silva, L., Peres, S. & Boscarioli, C. *Introdução à Mineração de Dados com aplicações em R*. (2016).
55. Goldberg, X. *Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning* **6**, (Morgan, 2009).
56. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (The MIT Press, 2018).
57. Correa, D. N. L., Paniagua, L. R. B., Noguera, J. L. V., Pinto-Roa, D. P. & Toledo, L. A. S. Computerized Diagnosis of Melanocytic Lesions Based on the ABCD Method. *Proc. - 2015 41st Lat. Am. Comput. Conf. CLEI 2015* **19**, 1–22 (2015).
58. Goh, Y. M. & Ubeynarayana, C. U. Construction accident narrative classification: An evaluation of text mining techniques. *Accid. Anal. Prev.* **108**, 122–130 (2017).
59. Han, J., Kamber, M. & Pei, J. *Data Mining: Concepts and Techniques*. San Francisco, CA, itd: Morgan Kaufmann (2012). doi:10.1016/B978-0-12-381479-1.00001-0

60. Elmagarmid, A. K. & Member, S. Duplicate Record Detection : A Survey. *IEEE Trans. Knowl. Data Eng.* **19**, 1–16 (2007).
61. Al-Anazi, S., Almahmoud, H. & Al-Turaiki, I. Finding Similar Documents Using Different Clustering Techniques. *Procedia Comput. Sci.* **82**, 28–34 (2016).
62. Olson, R. *et al.* Sleep, Dietary, and Exercise Behavioral Clusters Among Truck Drivers With Obesity: Implications for Interventions. *J. Occup. Environ. Med.* **58**, 314–321 (2016).
63. Lunardon, N., Menardi, G. & Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **6**, 79–89 (2014).
64. Chawla, N., Bowyer, K., O. Hall, L. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* (2002). doi:10.1613/jair.953
65. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann. Stat.* **28**, 337–407 (2002).
66. Russel, S. & Norvig, P. *Inteligência artificial*. (2013).
67. Goncalves, V., Maria, K. & Silva, A. B. F. da. Applications of Artificial Neural Networks in Chemical Problems. in *Artificial Neural Networks - Architectures and Applications* (InTech, 2013). doi:10.5772/51275
68. Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognit.* (1997). doi:10.1515/bchm.1997.378.7.697
69. BRASIL. Microdados RAIS e CAGED. (2019). Available at: <http://pdet.mte.gov.br/microdados-rais-e-caged>. (Accessed: 7th February 2020)
70. INEP. Indicadores Educacionais. (2019). Available at: <http://portal.inep.gov.br/web/guest/indicadores-educacionais>. (Accessed: 18th September 2019)
71. Medeiros, A. M. de & Vieira, M. de T. Ausência ao trabalho por distúrbio vocal de professores da Educação Básica no Brasil. *Cad. Saúde Pública* **35**, 1–12 (2019).
72. SEADE. Informações dos Municípios Paulistas. (2019). Available at: <http://www.imp.seade.gov.br/frontend/#/>. (Accessed: 18th February 2019)
73. OECD. Urban population by city size. (2018). Available at: <https://data.oecd.org/popregion/urban-population-by-city-size.htm>.
74. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
75. Barnes, P. J. Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **343**, 269–280 (2000).
76. Office for National Statistics. SOC 2000. (2000). Available at: <https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/socarchive>. (Accessed: 27th April 2021)
77. De Matteis, S. *et al.* Occupations associated with COPD risk in the large population-based UK Biobank cohort study. *Occup. Environ. Med.* **73**, 378–384 (2016).
78. Cassidy, A. *et al.* Occupational exposure to crystalline silica and risk of lung cancer: A multicenter case-control study in Europe. *Epidemiology* **18**, 36–43 (2007).
79. Stocks, S. J. *et al.* Occupation and work-related ill-health in UK construction workers. *Occup. Med. (Chic. Ill)*. **61**, 407–415 (2011).
80. Kursu, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
81. DIEESE. Anuário do Sistema Público de Emprego, Trabalho e Renda: mercado de trabalho. 2016. São Paulo; 2016.

ANEXOS

Anexo A - Aprovação da Comissão de Ética em Pesquisa para o estudo de absenteísmo docente.

USP - FACULDADE DE SAÚDE PÚBLICA DA UNIVERSIDADE DE SÃO PAULO - FSP/USP
---	---

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Estimando a morbidade de professores do ensino básico: Uso de aprendizado de máquina com dados públicos e anônimos.

Pesquisador: FERNANDO TIMOTEO FERNANDES

Área Temática:

Versão: 1

CAAE: 30788620.5.0000.5421

Instituição Proponente: Faculdade de Saúde Pública da Universidade de São Paulo - FSP/USP

Patrocinador Principal: Faculdade de Saúde Pública da Universidade de São Paulo - FSP/USP

DADOS DO PARECER

Número do Parecer: 4.031.362

Apresentação do Projeto:

Trata-se de projeto de pesquisa para estimar a morbidade de professores do ensino básico com uso de aprendizado de máquina com dados públicos e anônimos. O pesquisador justifica que os profissionais da educação básica no Brasil possuem altos índices de afastamento por doenças e que a predição do risco individual de afastamento pode auxiliar na elaboração de medidas preventivas, evitando novos adoecimentos e melhorando a qualidade de vida desses profissionais.

Objetivo da Pesquisa:

Estimar o risco de ausência laboral decorrente de morbidades dos docentes da educação básica infantil na rede pública municipal do estado de SP utilizando algoritmos de machine learning com dados públicos.

Avaliação dos Riscos e Benefícios:

Os riscos são mínimos porque serão utilizados apenas dados secundários de domínio público e anônimo. O benefício da pesquisa é desenvolver um modelo preditivo que estime com boa performance as probabilidades de ausência por adoecimento de trabalhadores, baseado somente nos vínculos empregatícios.

Comentários e Considerações sobre a Pesquisa:

Análise de dados abertos, secundários e anônimos de n=174.294 professores da educação infantil

Endereço: Av. Doutor Arnaldo, 715		CEP: 01.246-904
Bairro: Cerqueira Cesar		
UF: SP	Município: SAO PAULO	
Telefone: (11)3061-7779	Fax: (11)3061-7779	E-mail: coep@fsp.usp.br

USP - FACULDADE DE SAÚDE
PÚBLICA DA UNIVERSIDADE
DE SÃO PAULO - FSP/USP

Continuação do Parecer: 4.031.362

que atuaram na rede pública municipal em todos os 645 municípios do Estado de São Paulo entre 2014 e 2018 disponibilizados pela Relação Anual de Informações Sociais (RAIS).

Considerações sobre os Termos de apresentação obrigatória:

Foram apresentados os seguintes termos de apresentação obrigatória: projeto de pesquisa detalhado; dispensa de TCLE tendo em vista que o mesmo utilizará somente dados abertos secundários de domínio público e anônimos. A justificativa para dispensa de TCLE foi escrita de forma adequada; folha de rosto FSP/USP assinada.

Recomendações:

Aprovação do projeto de pesquisa.

Conclusões ou Pendências e Lista de Inadequações:

Projeto sem óbices.

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1295632.pdf	15/04/2020 18:42:14		Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	SOLICITACAO_DISPENSA_TCLE_ASSINADO.pdf	15/04/2020 18:41:49	FERNANDO TIMOTEO FERNANDES	Aceito
Projeto Detalhado / Brochura Investigador	Projeto_FSP_Doutorado_Fernando_Professores_PlatBrasil.pdf	15/04/2020 18:41:33	FERNANDO TIMOTEO FERNANDES	Aceito
Folha de Rosto	FolhaRosto_Assinada.pdf	15/04/2020 17:45:29	FERNANDO TIMOTEO FERNANDES	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Endereço: Av. Doutor Arnaldo, 715
Bairro: Cerqueira Cesar CEP: 01.246-904
UF: SP Município: SÃO PAULO
Telefone: (11)3061-7779 Fax: (11)3061-7779 E-mail: coep@fsp.usp.br

Página 02 de 03

USP - FACULDADE DE SAÚDE
PÚBLICA DA UNIVERSIDADE
DE SÃO PAULO - FSP/USP

Continuação do Parecer: 4.031.362

SÃO PAULO, 16 de Maio de 2020

Assinado por:
José Leopoldo Ferreira Antunes
(Coordenador(a))

Anexo B - Aprovação da Comissão de Ética em Pesquisa para o estudo de larga escala de coorte prospectiva do UK Biobank

Referência:

UK Biobank. Disponível em https://www.ukbiobank.ac.uk/media/nigbkoox/favourable-ethical-opinion-and-rtb-approval-16-nw_-0274-200778-may-2016.pdf. (Accessed: 26th May 2021)

Health Research Authority

North West - Haydock Research Ethics Committee

3rd Floor - Barlow House
4 Minshull Street
Manchester
M1 3DZ

Telephone: 0207 104 8012

13 May 2016

Dr Tim Peakman
UK Biobank Limited
1-4 Spectrum Way
Adswold
Stockport
Cheshire
SK3 0SA

Dear Dr Peakman

Title of the Research Tissue Bank:	UK Biobank: a large scale prospective epidemiological resource
REC reference:	16/NW/0274
Designated Individual:	Dr Tim Peakman
IRAS project ID:	200778

The Research Ethics Committee reviewed the above application at the meeting held on 10 May 2016. Thank you for attending with Mr Jonathan Sellors and Ms Nicola Doherty to discuss the application.

We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this favourable opinion letter. The expectation is that this information will be published for all studies that receive an ethical opinion but should you wish to provide a substitute contact point, wish to make a request to defer, or require further information, please contact the REC Manager Ms Rachel Katzenellenbogen, nrescommittee.northwest-haydock@nhs.net. Under very limited circumstances (e.g. for student research which has received an unfavourable opinion), it may be possible to grant an exemption to the publication of the study.

Favourable opinion

The members of the Committee present gave a favourable ethical opinion of the above research tissue bank on the basis described in the application form and supporting documentation, subject to the conditions specified below.

The Committee has also confirmed that the favourable ethical opinion applies to all research

projects conducted in the UK using tissue or data supplied by the tissue bank, provided that the release of the tissue or data complies with the attached conditions. It will not be necessary for these researchers to make project-based applications for ethical approval. They will be deemed to have ethical approval from this committee. You should provide the researcher with a copy of this letter as confirmation of this. The Committee should be notified of all projects receiving tissue and data from the tissue bank by means of an annual report.

This application was for the renewal of a Research Tissue Bank application. The previous REC Reference number for this application was 11/NW/0382.

Conditions of the favourable opinion

The favourable opinion is subject to the following conditions being met prior to the start of the study.

Research governance

Under the Research Governance Framework (RGF), there is no requirement for NHS research permission for the establishment of research tissue banks in the NHS. Applications to NHS R&D offices through IRAS are not required as all NHS organisations are expected to have included management review in the process of establishing the research tissue bank.

Research permission is also not required by collaborators at tissue collection centres (TCCs) who provide tissue or data under the terms of a supply agreement between the organisation and the research tissue bank. TCCs are not research sites for the purposes of the RGF.

Research tissue bank managers are advised to provide R&D offices at all TCCs with a copy of the REC application for information, together with a copy of the favourable opinion letter when available. All TCCs should be listed in Part C of the REC application.

NHS researchers undertaking specific research projects using tissue or data supplied by the research tissue bank must apply for permission to R&D offices at all organisations where the research is conducted, whether or not the research tissue bank has ethical approval.

Site-specific assessment (SSA) is not a requirement for ethical review of research tissue banks.

Summary of discussion at the meeting

Social or scientific value: scientific design and conduct of the study

The Committee were pleased to see that UK Biobank was constantly re-evaluating itself with regards to new technology and data collection. This meant that new tests were undertaken and new data and tissue collected allowing UK Biobank to grow and develop as a resource.

The Committee were very pleased that this resource was open access and also that researchers had to register to use it. The Committee noted that while UK Biobank owned the resource they had no preferential access. The Committee very happy to note that all research results had to be sent back to UK Biobank as part of a transparency agenda.

The Committee noted that the data was being used by a broad range of researchers and asked how use would be maximised in the future.

You explained that originally UK Biobank had been designed to be used in case control studies. However, you had been able to demonstrate that centralised generation of large datasets had advantages of cost, standardisation and a lack of gaps. This meant that it was being used in more than just case control studies.

You said that genotyping was being done on all participants and that they were currently measuring 34 biomarkers with the data available to all. You said that you were currently working up a proposal to measure 40 markers of infectious disease and were also looking at developing strategies to look at proteins and metabolites. You said it was important to maximise the tissue so that, for example, you wouldn't use tissue simply to measure glucose, but if you could run tests that delivered a lot of data, including glucose, then the data would be gathered in a good way.

You said that the data was linked to various registers, including deaths, cancer and hospital visit. 30% of English participants had primary care information and this was a lot higher for Welsh and Scottish participants. This meant that you would be able to create a plausible calendar as to when the data would be mature for more common conditions and then you would put out a call for researchers.

The Committee noted that one of the criteria for accessing the biobank was that the research be "in the public interest". The Committee asked if any applications had been turned down because they had not been in the public interest.

You said that no applications had been turned down because they were not in the public interest. In fact, only 2 or 3 requests for samples had been turned down and that was because they had either requested too much or actually did not need to turn to a biobank to do their research.

The Committee agreed that it had been an exemplary submission and had led to an interesting and informative discussion. The Committee looked forward to the publication regarding imaging and the reporting of findings and hoped the researchers would advise them of when it was published and how it could be accessed.

Care and protection of research participants: respect for potential and enrolled participants' welfare and dignity

The Committee agreed that the systems in place to avoid identifying participants were robust. Always growing and considering and developing.

The Committee noted that UK Biobank was regularly in touch with participants via newsletter and held an Annual General Meeting. The Committee agreed this was very important if participants were to stay motivated and interested as without this no new data or tissue could be added.

The Committee noted that the imaging Participant Information Sheet and consent form said that GPs would be contacted if anything clinically significant was discovered. The Committee noted that UK Biobank had had a policy of not feeding back findings and wondered if this policy had now changed. The Committee also agreed that they needed to know more about how significant

clinical findings were determined. For example, carotid arteries narrowed as people aged, so would all narrowing be reported or just ones with a certain percentage of narrowing.

The Committee asked what the current position was regarding feeding back clinically significant findings.

You said that the position had not changed, although it was reconsidered on a regular basis. When participants came for their baseline visit in 2007-2010 if something was spotted during the visit, then it was fed back. However, assay or other research findings were not fed back.

With regard to imaging, which could lead to acute findings such as cancer, you explained that you had spent 5-6 years working out the best protocol for that. The end result was that if the radiographer observed something that concerned them it was flagged and a radiologist would assess it. If the radiologist determined that it was significant then it was reported to the GP.

You explained that during the imaging pilot you had run two protocols, the one that is in current use, and a second one that involved a radiologist screening all of the images. After follow up it became clear that this was hugely problematic, not because of cost or expediency, but because it had led to 200 false positives. At the extreme end there had been a lung section and a removal of ovaries for people with false positives. Scaling this up to 100,000 people meant there could be 20,000 false positives.

You said that you had spoken to participants and to imaging projects and it had been agreed that while the radiographers might miss things, the best protocol was to have radiologists only look at images flagged by radiographers. You also said that you would be publishing the results of this research shortly.

You said that, in short, the feedback policy was that anything of clinical significance discovered during data acquisition would be feedback but any other findings would not be.

The Committee agreed that this was acceptable, especially as it was all made very clear to participants in information sheets.

The Committee asked why radiologists were diagnosing so many false positives.

You said that the images were research scans which, despite what many participants had thought, were not more detailed than ones taken for clinical purposes. Additionally, the radiologists did not have any of the other information they would have in normal diagnosis.

The Committee agreed that the level of commitment required from participants was high and the Committee agreed they would like to know how many participants had withdrawn and how many had simply been lost to contact. However, the Committee was impressed with the way UK Biobank kept participants informed of new developments and asked how many participants had been lost to contact or withdrawal.

You said that just over 1,000 participants had withdrawn with about 600 of them having requested all tissue and data be removed from the resource. The Committee said that

while annual communications always sparked some withdrawals, the benefits of the communication far outweighed that problem.

You said that most communication was by email, including web based questionnaires. However, it was easier to keep in touch with people by post because if they moved you could usually find their new address. Also it was impossible to know how many emails were opened and read, so no one knew who actually read the newsletter.

You explained that response rates to questionnaires had actually gone up over time and that there had been a 50% response rate to the request for participants willing to wear an accelerometer. In fact, you had managed to recruit 100,000 participants to do that.

You said that you were now also starting to use mobile technology to contact participants.

Other ethical issues were raised and resolved in preliminary discussion before your attendance at the meeting.

Duration of ethical opinion

The favourable opinion is given for a period of five years from the date of this letter and provided that you comply with the standard conditions of ethical approval for Research Tissue Banks set out in the attached document. You are advised to study the conditions carefully. The opinion may be renewed for a further period of up to five years on receipt of a fresh application. It is suggested that the fresh application is made 3-6 months before the 5 years expires, to ensure continuous approval for the research tissue bank.

Research Tissue Bank Renewals

The Research Tissue Bank has been renewed for a further five years from the end of the previous five year period. The previous five year period ran from 17 June 2011 to 17 June 2016. This Research Tissue Bank may be renewed for further periods of five years at a time by following the process described in the above paragraph.

Approved documents

The documents reviewed and approved at the meeting were:

<i>Document</i>	<i>Version</i>	<i>Date</i>
Human Tissue Authority licence [HTA Licence 12002 & 12624]		26 July 2010
IRAS Checklist XML [Checklist_27042016]		27 April 2016
Other [Table 1: Comparison of the sample collection for the baseline assessment and imaging pilot]	1.0	02 March 2016
Other [Table 2: Progress with key cohort-wide linkages Q1-2 2016]	1.0	02 March 2016
Other [UK Biobank Ethics & Governance Framework]	3.0	01 October 2007
Other [Figure 1: Submitted Access Applications by areas of interest]	1.0	02 March 2016
Other [Table 3: Biochemistry assays being performed in all 500,000 participants]	1.0	02 March 2016
Other [Participant Withdrawal Form]	1.1	10 February 2012

Other [UK Biobank Newsletter June 2015]	1.0	22 June 2015
Other [Data Dictionary Showcase Sept15]	Sept 2015	24 March 2016
Other [Curriculum vitae - Timothy Peakman]	March 2016	24 March 2016
Other [RTB Report March 2016]	1.0	24 March 2016
Other [Appendix: Occupational Questionnaire]	1.0	27 August 2014
Other [Appendix: Occupational Questionnaire Invitation Text]	1.0	27 August 2014
Other [Appendix: Occupational Questionnaire Reminder Invitation Text]	1.0	27 August 2014
Other [Revised Imaging Invitation Email]	1.0	06 October 2014
Other [Imaging Reminder Text & SMS]	1.0	18 November 2014
Other [Feedback in the UK Biobank Imaging pilot study]	Jan 2014	29 January 2014
Other [Invitation letter for deliberative group interviews]	1.0	06 October 2014
Other [Imaging 2nd Invite email HTML]	0.1	01 January 2016
Other [Imaging 2nd Invite email PLAIN]	0.2	01 January 2016
Other [Imaging Participant pre-screening questionnaire]	1.3	27 October 2015
Other [Imaging Exit Survey]	0.1	01 January 2016
Other [Invite email reminder 6-month questionnaire HTML]	0.1	01 October 2015
Other [Invite email reminder 6-month questionnaire PLAIN]	0.1	01 October 2015
Other [Invite email reminder 6-week questionnaire HTML]	0.1	01 October 2015
Other [Invite email reminder 6-week questionnaire PLAIN]	0.1	01 October 2015
Other [Invite email reminder understanding consent questionnaire HTML]	0.1	01 October 2015
Other [Invite email reminder understanding consent questionnaire PLAIN]	0.1	01 October 2015
Other [Appendix 1: Mental Health Questionnaire]	1.2	23 March 2016
Other [Appendix 2: Rationale and tools used in Mental Health Questionnaire]	1.1	04 March 2016
Other [Appendix 3: Invitation email Mental Health Questionnaire]	1.2	11 March 2016
Other [Appendix 4: Reminder email Mental Health questionnaire]	1.2	11 March 2016
Other [Appendix 5: Reminder partial responder email Mental Health questionnaire]	1.1	11 March 2016
Other [Appendix 6: Last chance email Mental Health questionnaire]	1.0	11 March 2016
Other [Repeat Assessment email invitation]	1.0	09 August 2012
Other [Repeat Assessment invite letter]	1.0	26 March 2012
Other [Repeat Assessment confirmation letter]	1.0	11 July 2012
Other [Confirmation of imaging appointment letter]	1.0	08 April 2016
Other [Activity Monitor Information Letter]	26/03/2012	26 March 2012
Other [Activity Monitor Invitation Letter]	26/03/2012	26 March 2012
Other [Activity Monitor Return Reminder]	26/03/2012	26 March 2012
Other [UK Biobank Assessment form]	20061124	24 November 2006
Other [Diet Questionnaire]	1.0	11 April 2016
Other [UK Biobank Participant Invite letter]	1.0	11 April 2016
Other [Touch-screen questionnaire]	1.0	11 April 2016

Other [Touch-screen questionnaire addendum]	1.0	11 April 2016
Other [Cognitive Function tests]	1.0	26 March 2013
Other [Cognitive Function Web Questionnaire email invitation]	1.0	26 March 2013
Other [Cognitive Function Web Questionnaire email reminder]	1.0	26 March 2013
Other [Cognitive Function Web Questionnaire email reminder partial responder]	1.0	26 March 2013
Other [UK Biobank Protocol]	21/03/2007	21 March 2007
Other [UK Biobank Protocol addendum 1]	09/04/2009	09 April 2009
Other [UK Biobank Protocol addendum 2]	02/07/2009	02 July 2009
Other [Text Message to request email address]	1.0	20 April 2016
Other [UK Biobank TIME study invitation]	2.2	15 April 2016
Other [Imaging Questionnaire to assess participant understanding of consent]	January 2014	01 January 2014
Other [Imaging Participant Questionnaire sent at 6 weeks to assess IF]	January 2014	01 January 2014
Other [Imaging Participant Questionnaire sent at 6 months to assess impact of IF]	January 2014	01 January 2014
Other [Imaging Questionnaire sent to participants who did not receive IF feedback]	January 2014	01 January 2014
Other [Imaging Letters notifying participant and participant's GP of potentially serious incidental finding]	1.0	01 January 2014
Other [Imaging GP questionnaire sent at 6 months to assess the later impact of feedback of IF]	1.0	01 April 2015
Participant consent form [UK Biobank Consent form]	20061124	24 November 2006
Participant consent form [Consent Form for the imaging assessment: UK Biobank]	Jan 2014	29 January 2014
Participant information sheet (PIS) [Participant Information Leaflet]	21/04/2010	21 April 2010
Participant information sheet (PIS) [Biobank Imaging Information Leaflet]	Dec 2015	01 December 2015
Participant information sheet (PIS) [Repeat Assessment Participant Information Leaflet]	26/03/2012	26 March 2012
Participant information sheet (PIS) [Further Information Leaflet]	001	08 April 2016
Participant information sheet (PIS) [Biobank Imaging Information Leaflet including ECG monitoring]	2.0	26 November 2014
Protocol for management of the tissue bank [UK Biobank Access Procedures]	1.0	01 November 2011
REC Application Form [RTB_Form_24032016]		24 March 2016

Licence from the Human Tissue Authority

Thank you for providing a copy of the above licence.

Membership of the Committee

The members of the Ethics Committee who were present at the meeting are listed on the attached sheet. Dr Tim Sprosen helped set up UK Biobank and was a member of the Scientific Steering Committee. It was agreed that Dr Sprosen would leave the room during the discussion and take no part in the discussion or decision making. Dr Valerie Siddall, Alternate Vice-Chair, would chair

that portion of the meeting.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

The attached standard conditions give detailed guidance on reporting requirements for research tissue banks with a favourable opinion, including:

- Notifying substantial amendments
- Submitting Annual Progress reports

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:

<http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/>

HRA Training

We are pleased to welcome researchers and R&D staff at our training days – see details at <http://www.hra.nhs.uk/hra-training/>

16/NW/0274	Please quote this number on all correspondence
------------	--

Yours sincerely

Dr Tim S Sprosen
Chair

E-mail: nrescommittee.northwest-haydock@nhs.net

Anexo C - Aprovação da concessão de material. *Material Transfer Agreement.*

DocuSign Envelope ID: 1C9F4D0B-59E1-4FEF-92A2-1692AAA99C9A

Annex II: Material Transfer Agreement with the Applicant for data and/or samples

Dear Professor Filho

UK Biobank is pleased to approve your Application Reference Number 30140 to use the UK Biobank Resource. Execution of this Material Transfer Agreement (MTA) and payment of the Access Charges are the final steps before access is granted. UK Biobank's approval of this Application is valid for 90 days, after which the Applicant Principal Investigator (PI) will need to re-apply for access. The content of UK Biobank's standard MTA, and the conditions contained within it, are non-negotiable.

Parties

This is an agreement between UK Biobank Limited on the one hand and the Applicant Institution (University of Sao Paulo) on the other hand. The Applicant PI is not a party to the MTA, however, UK Biobank requires that the Applicant PI acknowledges that the provisions of this MTA have been "read and understood" by the Applicant PI so that they are fully aware of their Institution's obligations to both UK Biobank and to UK Biobank's participants.

The Applicant Institution will be responsible for the conduct of any and all of the Applicant Researchers involved in this Research Project.

Structure of agreement

The MTA will become effective on receipt by UK Biobank of:

- (i) A copy of this MTA Agreement executed by an authorised signatory of the Applicant Institution and confirmed as "read and understood" by the Applicant PI; and
- (ii) Cleared funds covering the Access Charges from the Applicant Institution.

UK Biobank will then promptly send a dated confirmatory email.

Provision of samples and/or data

Annex A summarises the data and/or samples that UK Biobank will make available to the Applicant in accordance with their approved Application Reference Number 30140. The timeframe and methodology by which the data and/or samples will be dispatched is also set out in Annex A.

Payment

The Access Charges which are payable are set out in Annex B. This also serves as an invoice on which VAT will be included (as appropriate). The derivation of these Access Charges is also set out in Annex B.

This payment should be submitted in cleared funds to Barclays Bank PLC, Account name: UK Biobank Limited, Account number: 33069427 and Sort code: 20-24-09.

Standard terms and schedules

This Agreement incorporates the attached terms and conditions (including any documents and/or materials that are referred to in them), the Annexes and where applicable the contents of the Preliminary and Main Application Forms with Reference Number 30140.

Yours faithfully

Accepted and agreed

For and on behalf of UK Biobank/Effective Date
(Jonathan Sellers / Company Solicitor)

26/2/2020

For and on behalf of Applicant Institution
(Please sign and print your name and position)

Oswaldo Tanaka
Dean School of Public Health
Principal Investigator

(Please sign and print your name and position)

Alexandre Chiavegatto Filho

Alexandre
Chiavegatto Filho
Associate Professor