UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

ANA CAROLINE FRANCISCO DA ROSA

Avaliação de Técnicas de Mineração de Dados em um Conjunto de Dados de Dermatite Ocupacional Obtido a partir de um Serviço Especializado

Maringá
2021

ANA CAROLINE FRANCISCO DA ROSA

Avaliação de Técnicas de Mineração de Dados em um Conjunto de Dados de Dermatite Ocupacional Obtido a partir de um Serviço Especializado

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia de Produção, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Engenharia de Produção.
Área de concentração: Engenharia de Produção

Orientador: Prof. Dr. Edwin Vladimir Cardoza Galdamez
Coorientador: Prof. Dr. Rodrigo Clemente Thom de Souza

Maringá
2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá, PR, Brasil)

R788a

Rosa, Ana Caroline Francisco da
Avaliação de técnicas de mineração de dados em um conjunto de dados de dermatite ocupacional obtido a partir de um serviço especializado / Ana Caroline Francisco da Rosa. -- Maringá, PR, 2021.
123 f. : il. color., figs., tabs.

Orientador: Prof. Dr. Edwin Vladimir Cardoza Galdamez.
Coorientador: Prof. Dr. Rodrigo Clemente Thom de Souza.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Engenharia de Produção, Programa de Pós-Graduação em Engenharia de Produção, 2021.

1. Mineração de Dados. 2. Dermatoses Ocupacionais. 3. Saúde e Segurança Ocupacional. 4. Machine learning. 5. Testes cutâneos. I. Cardoza Galdamez, Edwin Vladimir, orient. II. Souza, Rodrigo Clemente Thom de, coorient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Engenharia de Produção. Programa de Pós-Graduação em Engenharia de Produção. IV. Título.

CDD 23. ed. 005.117
Marinalva Aparecida Spolon Almeida - 9/1094

FOLHA DE APROVAÇÃO

ANA CAROLINE FRANCISCO DA ROSA

Avaliação de Técnicas de Mineração de Dados em um Conjunto de Dados de Dermatite Ocupacional Obtido a partir de um Serviço Especializado

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia de Produção, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Engenharia de Produção pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Prof. Dr. Edwin Vladimir Cardoza Galdamez - Orientador
Universidade Estadual de Maringá – PGP - UEM

Prof. Dr. Rafael Henrique Palma Lima
Universidade Tecnológica Federal do Paraná – UTFPR / PGP - UEM

Prof. Dr. Cassius Tadeu Scarpin
Universidade Federal do Paraná – UFPR

Prof. Dr. Rodrigo Clemente Thom de Souza - Coorientador
Universidade Federal do Paraná – UFPR / PGP – UEM

Aprovada em: 25 de fevereiro de 2021.
Local de Defesa: *Video call link: https://meet.google.com/nqv-uyrf-jgh*.

AGRADECIMENTOS

O Programa de Pós-graduação em Engenharia de Produção foi um divisor de águas na minha vida profissional e acadêmica. Ao redigir este texto, percebo o quanto amadureci, tornei-me uma pessoa mais confiante e com uma visão perspicaz sobre o contexto ao qual estou inserida. Quando ingressei no Programa, acreditei que nunca mais ia "colocar o pé na indústria" (o que soa até um tanto contraditório uma vez que o programa objetiva estudar os processos industriais) pois estava frustrada com minhas experiências profissionais e acreditei que a academia seria a solução para os meus "conflitos" de carreira. Ao cursar as disciplinas ainda no primeiro semestre, percebi que as atividades profissionais que havia exercido antes do mestrado tinham e muito ao que acrescentar na minha futura jornada acadêmica e passei a considerar meu retorno ao mercado de trabalho. No segundo semestre do mestrado, consegui um emprego numa transportadora para realizar análise de dados e fui apresentada à ferramenta incrível que é o Power BI onde aprendi não apenas sobre o setor de transportes, mas também sobre linguagem SQL e gestão de manutenção de frota. Além disso, tive um "intensivo" com o professor Edwin de como aliar o conhecimento acadêmico ao profissional. Como a vida é feita de escolhas, após reflexões sobre a evolução de minha vida profissional em março de 2020 decidi procurar outras oportunidades de trabalho onde em junho do mesmo ano fui aprovada no processo seletivo da Louis Dreyfus Company (LDC) como Supervisora de Turno Produção na unidade de Apucarana. Na LDC, tenho aprendido a "arte" que é fazer a gestão de mais de uma equipe de trabalho e também vivenciado o impacto de minhas decisões e o alinhamento com outras áreas exercem na rotina de trabalho de toda a unidade.

Conforme citado até o momento, muita coisa mudou e gostaria de agradecer e honrar a Deus e também algumas pessoas e instituições que muito me ampararam nessa jornada. A Deus por ter me ajudado a perseverar, a buscar nele a minha melhor versão e por ter colocado no meu caminho as pessoas certas pra me aconselhar, amparar e instruir. Ao meu esposo Bruno pela compreensão, paciência e apoio nas distintas situações que vivenciamos. Aos meus pais pelo suporte, respeito às minhas escolhas e em especial à minha mãe Vilma por desde que eu era pequena me encorajar a lutar pelo que quero e me ensinar literalmente a ser profissional. Aos professores Edwin, Rodrigo e Camila por terem me escolhido a ser orientanda de vocês, por terem confiado em mim para abordar um tema tão complexo e me instruírem a ser uma pesquisadora de excelência. À Fiocruz por conceder a base de dados e à CAPES pelo mantimento do Programa de Pós-graduação em Engenharia de Produção.

EPÍGRAFE

“The price of light is less than the cost of darkness”.
“O preço da luz é menor que o custo da escuridão”.

(ARTHUR C. NIELSEN)

Avaliação de Técnicas de Mineração de Dados em um Conjunto de Dados de Dermatite Ocupacional Obtido a partir de um Serviço Especializado

RESUMO

Dermatoses ocupacionais caracterizam-se pela alteração das mucosas, pele e seus anexos que são mantidas, agravadas, condicionadas ou causadas por agentes existentes na atividade laboral ou no ambiente de trabalho. Diante disso, conhecer os fatores que influenciam a ocorrência de uma dermatose ocupacional é importante para o estabelecimento de ações preventivas de proteção ao trabalhador. No Brasil, as principais fontes de dados para a investigação dos acidentes são os benefícios concedidos pela previdência, as informações do Sistema Único de Saúde além de instituições de referência como a Fiocruz (Fundação Oswaldo Cruz). Nos últimos anos, pesquisas têm sido desenvolvidas para aprimorar a assertividade no diagnóstico clínico, investigar de acidentes de trabalho, melhorar políticas públicas e uma de tais vertentes é o aprendizado de máquina. Técnicas de aprendizado de máquina visam aprender com as informações passadas para então encontrar padrões de comportamento que expliquem com considerável nível de confiabilidade os eventos futuros. A presente dissertação apresenta como objetivo de usar técnicas de *machine learning* para identificar padrões de comportamento entre as variáveis, determinar os fatores de influência e avaliar a relevância dos testes cutâneos da bateria padrão brasileira de testes de contato na incidência de dermatoses ocupacionais. Para alcançar tal objetivo, é proposto o uso da metodologia *KDD* (*Knowledge Discovery in Databases*) em um banco de dados do serviço especializado em dermatologia do trabalho da Fiocruz. Como resultados da pesquisa, foram comparadas as técnicas de mineração de dados *Random Forest, Catboost, Neural Network,* Regressão Logística, *Adaboost, Extreme Gradient Boosting* em termos das métricas Acuracidade, Sensitividade, Especificidade, Erro, *Recall*, Precisão, F_1 *Score*, Prevalência, índice *Kappa* e *Area Under the Curve* em dois cenários. As técnicas que apresentaram maior acuracidade no segundo cenário foram *Random Forest* seguido por *Catboost* e *Adaboost.* As variáveis de maior influência para *Random Forest* foram: mãos e antebraços, profissão, mês, diagnóstico de dermatite irritativa e escolaridade. Para avaliar a relevância dos testes cutâneos na incidência das dermatoses ocupacionais realizou-se os testes de Mann-Whitney e *t-student* e foi constatado que não há indícios de que realizar os testes cutâneos sejam relevantes para determinar a classificação de dermatoses ocupacionais.

Palavras-chave: Dermatoses Ocupacionais. Aprendizado de Máquina. Saúde e Segurança Ocupacional. Testes Cutâneos.

Evaluation of Data Mining Techniques in an Occupational Dermatitis Data Set Obtained from a Specialized Service

ABSTRACT

Occupational dermatoses are characterized by changes in mucous membranes, skin and their attachments that are maintained, aggravated, conditioned or caused by existing agents in work activity or in the work environment. Therefore, knowing the factors that influence the occurrence of an occupational dermatosis is important for the establishment of preventive actions to protect workers. In Brazil, the main sources of data for the investigation of accidents are the benefits granted by social security, information from the Unified Health System and reference institutions such as Fiocruz (Oswaldo Cruz Foundation). In recent years, research has been developed to improve assertiveness in clinical diagnosis, investigate work accidents, improve public policies and one such aspect is machine learning. Machine learning techniques aim to learn from past information and then find patterns of behavior that explain with considerable level of reliability future events. This dissertation presents as objective to use machine learning techniques to identify patterns of behavior among the variables, determine the factors of influence and evaluate the relevance of skin tests of the Brazilian standard battery of contact tests in the incidence of occupational dermatoses. To achieve this goal, it is proposed the use of the KDD (Knowledge Discovery in Databases) methodology in a database of the specialized service in dermatology of Fiocruz's work. As results of the research, we compared the techniques of data mining Random Forest, Catboost, Neural Network, Logistic Regression, Adaboost, Extreme Gradient Boosting in terms of accuracy, sensitivity, specificity, error, recall, precision, F1 score, prevalence, Kappa Area index and under the curve in two scenarios. The techniques that presented the highest accuracy in the second scenario were Random Forest followed by Catboost and Adaboost. The variables most influenced for Random Forest were: hands and forearms, profession, month, diagnosis of irritative dermatitis and schooling. To evaluate the relevance of skin tests in the incidence of occupational dermatoses, Mann-Whitney and t-student tests were performed and it was found that there is no evidence that performing skin tests is relevant to determine the classification of occupational dermatoses.

***Keywords*:** Occupational Dermatoses. Machine Learning. Occupational Health and Safety. Patch Tests.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADA	*AdaBoost*
AUC	*Area Under the Curve*
BNB	*Bernoulli Naïve Bayesian*
CART	*Classification and Regression Trees*
CAT	*Catboost*
CID	Classificação Internacional das Doenças
Cesteh	Centro de Estudos da Saúde do Trabalhador e Ecologia Humana
DRT	Dermatoses Relacionadas ao Trabalho
DT	*Decision Tree*
Ensp	Escola Nacional de Saúde Pública Sérgio Arouca
ETC	*Extra Tree Classifier*
FIOCRUZ	Fundação Oswaldo Cruz
GBM	*Gaussian Naïve Bayesian*
INSS	Instituto Nacional do Seguro Social
KDD	*Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbors*
LDA	*Linear Discriminant Analysis*
LRG	*Logistic Regression*
MDA	*Mixture Discriminant Analysis*
NB	*Naïve Bayes*
NN	*Neural Network*
OSHA	*Occupational Safety and Health Administration*
PAC	*Passive Agressive Classifier*
RF	*Random Forest*
RFID	*Radio-Frequency IDentification*
RNC	*Radius Neighbors Classifier*
ROC	*Receiver Operating Characteristic*
SINAN	Sistema de Informação de Agravos de Notificação
SSO	Saúde e Segurança Ocupacional
SVM	*Support Vector Machine*

SUMÁRIO

INTRODUÇÃO

1

1.1 CONTEXTUALIZAÇÃO

A saúde do trabalhador compõe a saúde pública e é de incumbência do governo intervir nas relações entre o trabalho e a saúde (BRASIL, 2021). São categorizados como trabalhadores todos os homens e mulheres que exercem algum tipo de atividade remunerada para o sustento próprio e/ou de sua família no domínio formal ou informal da economia (WHO, 2021a). As políticas públicas no que se referem à saúde do trabalhador, apresentam o propósito de promover a proteção do trabalhador por meio do desenvolvimento de ações de mitigação e prevenção dos riscos existentes no local de trabalho. Além de promoção da proteção do trabalhador, compete ao governo o fornecimento da assistência em relação ao diagnóstico, tratamento e reabilitação das moléstias ocasionadas pelo trabalho (TAMERS *et al.*, 2019).

Doenças ocupacionais são caracterizadas por serem adquiridas ou desencadeadas pelo exercício profissional por conta de condições específicas constantes nas atividades desempenhadas ou inertes ao local no qual o ofício é desenvolvido (BRASIL, 1991). Além de prejuízos à saúde do trabalhador, as doenças ocupacionais apresentam impactos econômicos e à sociedade que incluem: o empobrecimento dos trabalhadores e de suas famílias, redução da produtividade e capacidade de trabalho do trabalhador além do aumento de custos ao sistema de saúde e à previdência (CHEN *et al.*, 2020).

Em meio à vasta gama de doenças profissionais existentes, um grupo relativamente comum no âmbito mundial é o das dermatoses de contato ocupacionais que de acordo Keegel *et al.* (2009) a cada 10.000 horas de trabalho têm-se no mínimo 1,3 pessoas acometidas com uma variante dessa moléstia e até o máximo de 8,1 trabalhadores por país. Tal variação é estabelecida segundo o nível de desenvolvimento da nação e também quanto a acessibilidade aos serviços de saúde de forma que quão mais precário o acesso a recursos financeiros e aos

serviços de saúde maiores serão os índices de doenças do trabalho relacionadas à pele. Dermatoses ocupacionais são caracterizadas por inflamações reativas da pele em decorrência do contato direto com produtos químicos, podendo manifestar-se sobremaneira nas formas irritativa e alérgica segundo os compostos presentes no local de trabalho (LAMPEL; POWELL, 2019). Para averiguar se uma pessoa apresenta ou não aversão à determinado composto químico aplica-se testes cutâneos, testes esses que podem ser específicos para a profissão (quando há suspeita de doença ocupacional) ou padrões de testes para uma determinada população (GAWKRODGER, 2001).

Para a proposição de políticas de proteção à saúde do trabalhador, os governos se baseiam no perfil epidemiológico referentes aos dados coletados pelos sistemas de saúde e aos benefícios concedidos pela previdência com o intuito de descrever as características de suas populações (LEVTEROVA, 2014). Uma das formas de diferenciar as doenças comuns das doenças ocasionadas pelo trabalho é o uso de técnicas de *machine learning* que pautadas em formulação matemática são capazes de encontrar padrões nos dados, padrões esses que não seriam facilmente identificados apenas com a observação do conjunto de dados (FORSTING, 2017). Em relação os algoritmos de *machine learning* utilizados para a proposição de políticas envolvendo saúde pública conforme a revisão de literatura feita por Santos *et al.* (2019) encontrou-se 250 artigos publicados entre 2009 e 2018. Dentre os estudos localizados por Santos *et al.* (2019) destacaram-se os seguintes algoritmos de classificação: *Support Vector Machine* presente em 64 estudos, Árvores de Decisão em 57 artigos, *Random Forest* (RF) em 48, Regressão Logística (LRG) em 43 trabalhos e Naïve Bayes (NB) com 33 representantes. Novos algoritmos têm sido continuamente desenvolvidos com o propósito de investigar os fenômenos cientificamente como o algoritmo *Catboost* (CAT) criado em meados de 2018 (HANCOCK; KHOSHGOFTAAR, 2020).

Além das informações coletadas pelos sistemas de saúde e previdenciário, são de grande valor o conhecimento produzido por instituições de referência como a Fiocruz (Fundação Oswaldo Cruz). A Fiocruz possui o Centro de Estudos da Saúde do Trabalhador e Ecologia Humana (Cesteh), sediado na Escola Nacional de Saúde Pública Sérgio Arouca (Ensp) situado na cidade do Rio de Janeiro, que apresenta diversos serviços especializados à saúde e dentre eles está o Serviço de Dermatoses Relacionadas ao Trabalho (DRT) onde são tratados pacientes com sintomas sugestivos de dermatoses ocupacionais que foram previamente atendidos na rede pública de saúde. Para composição da presente dissertação, foi concedido o acesso a base de dados de trabalhadores atendidos no DRT entre os anos 2000 e 2014.

1.2 OBJETIVOS

O objetivo geral desta dissertação é usar técnicas de *machine learning* para identificar padrões de comportamento entre as variáveis, determinar os fatores de influência e avaliar a relevância dos testes cutâneos da Bateria Padrão Brasileira de Testes de Contato na incidência de dermatoses ocupacionais dos pacientes atendidos na Fiocruz no DRT. Para alcançar o objetivo principal são propostos os seguintes objetivos específicos:

- comparar os algoritmos de *machine learning*: Regressão Logística, *Adaboost* (ADA), *Neural Network* (NN), *Random Forest*, *Extreme Gradient Boosting* (XGB) e *Catboost*;
- utilizar as métricas Acuracidade, Sensitividade, Especificidade, Erro, *Recall*, Precisão, F_1 *Score*, Prevalência, índice *Kappa* e *Area Under the Curve* (AUC) para comparação dos algoritmos;
- confrontar a importância das variáveis de *Catboost* com a técnica que apresentou melhores resultados dentre as 5 técnicas restantes.

1.3 JUSTIFICATIVA

A saúde do trabalhador é um tema que carece ser melhor compreendido sobre distintas perspectivas uma vez que além do interesse governamental e da sociedade em relação à proteção de sua força de trabalho tem-se o viés das empresas tanto públicas quanto privadas na figura de empregador (YANAR; LAY; SMITH, 2019). Os empregadores dessa forma, precisam de informação para poderem traçar suas próprias políticas de prevenção aos acidentes de trabalho tendo assim planos de ação para a eliminação e/ou mitigação dos riscos em suas instalações.

As dermatoses ocupacionais são um tipo de doenças consideravelmente comuns aos quais a população brasileira está exposta e são poucos os trabalhadores com plena ciência dos riscos aos quais estão sujeitos. Para melhor compreender a relação entre as dermatoses e o trabalho bem como diferenciar as dermatoses cuja origem distingue-se do trabalho foi aplicado o processo *KDD* (*Knowledge Discovery in Databases*) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A metodologia *KDD* é iniciada com a seleção, pré-processamento e transformação do conjunto de dados (LUO, 2008). Para que fosse possível a replicabilidade dos resultados encontrados com a aplicação das técnicas de *machine learning* descreveu-se as etapas de pré-processamento e transformação das observações sob considerável nível de detalhamento. Além

de descrever o passo a passo do pré-processamento e transformação, foram apresentadas algumas relações entre as variáveis que não seriam possíveis de serem estabelecidas sem que a etapa anterior ocorresse.

Com a aplicação do *KDD* é possível eliminar o viés da opinião pessoal para o que o modelo computacional classifique como relevante com base nas premissas da técnica empregada os fatores de influência no evento em estudo (KARPATNE *et al.*, 2017). Existem vários algoritmos capazes de descrever um fenômeno. É preciso, portanto, compreender os fatores envolvidos e verificar o tipo de aprendizado que mais se enquadra para o atingimento do objetivo proposto para então analisar o tipo de tarefa e aplicar as técnicas de mineração (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Em relação ao aprendizado supervisionado onde as saídas são conhecidas é prudente fazer o uso de métricas para comparação dos algoritmos, de forma que a técnica que se sobressair sobre as demais pode ser eleita para descrição do evento (CALLAHAN; SHAH, 2017). Tais métricas são originadas da aplicação do modelo obtido no treinamento na amostra de dados que não compôs a população de treinamento de forma que de posse dos erros e acertos da classificação estabelece-se a matriz de confusão (GUNS; LIOMA *et al.*, 2012). Além de comparar as técnicas com as métricas, é possível a realização de testes estatísticos para avaliar os resultados produzidos durante a etapa de treinamento do modelo e ainda confrontar tais respostas com as métricas oriundas da matriz de confusão.

Ao considerar os envolvidos, o objeto de estudo e a metodologia proposta, não foram verificados até o momento do desenvolvimento da presente dissertação, estudos brasileiros que façam uso de técnicas de aprendizado de máquina no âmbito das dermatoses ocupacionais. Os trabalhos encontrados, por sua vez, apresentaram abordagem exclusivamente descritiva. Um exemplo de pesquisa que fez uso da estatística descritiva para as dermatoses do trabalho foi o de Lise *et al.* (2018) que buscaram descrever o perfil dos trabalhadores constantes no banco de dados do Sistema de Informação de Agravos de Notificação (SINAN) dentre os anos de 2007 e 2014 que totalizam 4.710 casos.

1.4 ESTRUTURA DO TEXTO

A presente dissertação está disposta em capítulos. No Capítulo 2, situa-se a revisão de literatura, onde é possível encontrar definições gerais sobre Saúde e Segurança Ocupacional (SSO) (Seção 2.1) seguido por uma breve descrição a respeito das revoluções industriais e seu relacionamento com SSO situado na Seção 2.2. A Seção 2.3, por sua vez, trata das definições do ponto de vista

clínico sobre dermatoses ocupacionais enquanto a Seção 2.4 dispõe relativamente às estratégias de prevenção para as dermatoses ocupacionais. Na Seção 2.5, está alocada informações sobre o processo *KDD*, enquanto na 2.6 descreve-se a respeito de *machine learning*. A Seção 2.7 apresenta as técnicas de classificação aplicadas neste estudo enquanto a Seção 2.8 dispõe sobre estudos de *machine learning* em SSO.

O Capítulo 3 apresenta a metodologia de condução da pesquisa, enquanto no Capítulo 4 são expostos os resultados encontrados bem como as discussões pertinentes. O Capítulo 4 está estruturado da seguinte estrutura: na Seção 4.1 situa-se a análise inicial da base de dados, na 4.2 localiza-se a comparação entre técnicas de *machine learning* e na 4.3 são apresentadas algumas perspectivas futuras para pesquisa.

No capítulo 5, para findar este estudo estabece-se as conclusões. Na sequência apresenta-se o Apêndice A o qual expõe o *script* em linguagem R seguido pelo Apêndice B que dispõe a correlação existente entre os compostos dos testes cutâneos. O Apêndice C, por sua vez, exibe as medidas de posição das técnicas de mineração enquanto o Apêndice D proporciona o resultado referente à importância das variáveis para as técnicas com as melhores métricas. Após os Apêndices apresenta-se os anexos, de forma que no Anexo A expõe os benefícios concedidos pelo INSS (Instituto Nacional do Seguro Social) em decorrência de dermatoses ocupacionais de dezembro de 2018 a novembro de 2020. O Anexo B, por sua vez, apresenta a tradução do questionário nórdico de doenças ocupacionais relativas à pele seguido pelo Anexo C que dispõe o parecer do Comitê de Ética em Pesquisa.

REFERENCIAL TEÓRICO

2

2.1 SAÚDE E SEGURANÇA OCUPACIONAL

A Saúde e Segurança Ocupacional refere-se às práticas de proteção da saúde e da segurança do trabalhador em seu ambiente de trabalho de forma a promover ações que minimizem a probabilidade de ocorrências de acidentes e doenças ocupacionais (COMBERTI; DEMICHELA; BALDISSONE, 2018).

Acidente de Trabalho é entendido por Moura *et al.* (2017) como acontecimentos não planejados capazes de provocar lesões às pessoas e/ou à propriedade durante a execução das atividades operacionais consideradas comuns. Acidentes relacionados ao trabalho não provocam unicamente perdas relativas ao capital humano, ocasionam também perdas econômicas e sociais que acarretam em ineficiência, retardam o avanço econômico do país ao causar prejuízos financeiros devido às paradas de produção, quebras de maquinário e danos à reputação do negócio que implicam ainda em perdas de competitividade (FERNÁNDEZ-MUÑIZ; MONTES-PEÓN; VÁZQUEZ-ORDÁS, 2012).

Doenças Ocupacionais são entendidas segundo a legislação brasileira, disposta na Lei 8.213 de 24 de julho de 1991, como perdas à saúde do trabalhador em virtude do exercício profissional, podendo ser classificada como doença ocupacional (quando relacionada à toda classe profissional). Pode ser classificada como doença do trabalho (quando oriunda do ambiente em que a atividade laboral é exercida).

A atuação dos profissionais de SSO têm se concentrado em duas vertentes principais: na prevenção de acidentes e doenças ocupacionais e também na mitigação de tais eventos (PROVAN; RAE; DEKKER, 2019; BREWER; HOLT; MALIK, 2018). Para que ocorra a prevenção das doenças e acidentes oriundos do trabalho, os profissionais de SSO atuam no desenvolvimento de ferramentas, mecanismos e estratégias. Tais ferramentas, mecanismos e estratégias possuem o propósito de reduzir a ocorrência de eventos que possam se mostrar danosos às pessoas e à propriedade com a inserção de barreiras e padronização dos

procedimentos operacionais para execução das atividades laborais (UHRENHOLDT MADSEN; HASLE; LIMBORG, 2019). Com o intuito de mitigar os acidentes e doenças ocupacionais, a área de saúde e segurança emprega seus esforços na implementação do programa de gerenciamento de riscos, na manutenção e calibração dos equipamentos de proteção coletivas, no fornecimento dos equipamentos de proteção individuais, promoção de treinamentos e capacitações além de análise e correção dos desvios identificados (COMBERTI *et al.*, 2015).

2.2 AS REVOLUÇÕES INDUSTRIAIS E SSO

A industrialização passou por transformações notáveis desde o evento da 1ª Revolução Industrial datada no início do século XVIII na Inglaterra (LESO; FONTANA; IAVICOLI, 2018). A primeira grande revolução da indústria foi marcada pela introdução aos processos produtivos de máquinas movidas a vapor conferindo padronização aos produtos, agilidade aos processos além da criação dos navios e trens também movidos por combustão. Em contrapartida a tal avanço tecnológico, riscos até o momento inexistentes como ruídos, operação com vasos de pressão, umidade elevada e extensivas jornadas de trabalho passaram então a fazer parte do dia-a-dia dos trabalhadores (BADRI; BOUDREAU-TRUDEL; SOUISSI, 2018).

A 2ª Revolução Industrial teve seu início no século XIX sendo caracterizada pela introdução da eletricidade aos processos de produção, possibilitando então a distribuição de energia a partir de uma central (LIU *et al.*, 2020). Além de mudanças na matriz energética industrial, a segunda revolução industrial trouxe consigo inovações na conservação e armazenamento dos alimentos, possibilitou o desenvolvimento dos aviões, democratizou o acesso à automóveis e transformou o conceito de comunicação com a invenção do rádio e telefone. Como consequência da implementação de tais inovações além do aumento do conforto às pessoas, a segunda revolução industrial proporcionou novos riscos ao trabalho tais como: choques e descargas elétricas de grande intensidade e manipulação de produtos químicos cujos perigos eram desconhecidos. O movimento sindical ganhou força com a segunda revolução onde foram desenvolvidas as primeiras legislações de relativas à limitação do horário de trabalho além da realização dos primeiros estudos de proteção à saúde do trabalhador (LESO; FONTANA; IAVICOLI, 2018).

Em relação à 3ª Revolução Industrial ocorrida no século XX, foi observado o desenvolvimento da eletrônica, linhas de montagem motorizadas e a produção se tornou mais automatizada, portanto, focada no desempenho (BADRI; BOUDREAU-TRUDEL; SOUISSI,

2018). Durante a terceira revolução industrial foi percebido um avanço em relação à SSO, podendo ser constatado inclusive a participação de empregadores e empregados na resolução dos problemas de segurança (LIU *et al.*, 2020).

No que diz respeito à 4ª Revolução Industrial ou Indústria 4.0, realidade vivenciada no momento, estabelece-se como uma revolução global na manufatura (DALENOGARE *et al.*, 2018). Tal revolução é marcada pela interconectividade, digitalização da manufatura, integração dos ativos físicos aos ecossistemas digitais em toda a cadeia de valor. Ao tratar a SSO, a 4ª Revolução Industrial caracteriza-se pela gestão integrada de SSO, por melhores práticas de trabalho, por supervisão e controle dos ambientes e equipamentos em tempo real (LESO; FONTANA; IAVICOLI, 2018). Além dos pontos citados até o momento, a Indústria 4.0 caracteriza-se pela existência de ferramentas e padrões sofisticados de gestão, equipamentos de operação mais segura capazes de proporcionar antecipação dos riscos ocupacionais, atuando, portanto, proativamente (BADRI; BOUDREAU-TRUDEL; SOUISSI, 2018).

Ao analisar o processo de industrialização como um todo, pode ser percebido que a industrialização proporcionou a criação de um novo mercado de trabalho que sob caráter emergencial conduziu ao aumento de condições de trabalho deploráveis nas quais, homens mulheres e crianças arriscavam a própria vida para sustentar-se (LIU *et al.*, 2020). A inexperiência da força de trabalho em conjunto com a ignorância dos empregadores ao que hoje conhecemos como SSO cobrou um preço tão alto, que sob pressão pública forçou os legisladores a intervir de forma que as organizações passaram a ser responsáveis pela eliminação dos perigos na fonte e/ou mitigação dos mesmos quando aplicável (BADRI; BOUDREAU-TRUDEL; SOUISSI, 2018).

Por mais que o mundo esteja vivenciando a quarta revolução industrial, tal afirmativa não significa que todas as empresas controlem proativamente os perigos constantes em suas instalações (LIU *et al.*, 2020). Dessa forma, ainda existem empresas cujos processos, formas de gestão e segurança mais se assemelham à segunda e terceira revoluções industriais o que implica numa maior exposição ao risco dos trabalhadores que ali desempenham suas atividades. Na maioria dos países industrializados e nos em desenvolvimento que possuem uma equipe de gestão madura conforme Badri, Boudreau-Trudel e Souissi (2018) é possível constatar que a segurança não é apenas uma palavra, mas um componente crucial ao sucesso financeiro, tal como controle de qualidade, produtividade e custos.

2.3 DOENÇAS DA PELE RELACIONADAS AO TRABALHO

A pele está em constante contato com o ambiente externo ao corpo humano, fator esse, que faz com que a mesma esteja suscetível a uma grande variedade de potenciais agressores. Em virtude de tal exposição, é por meio da pele que o organismo manifesta a princípio uma potencial lesão, relacionando-se com os riscos presentes no ambiente sejam do tipo físicos, químicos ou biológicos (MELO; VILLARINHO; LEITE, 2019).

Tratando-se das lesões relacionadas ao trabalho, o Ministério da Saúde do Brasil, através do Departamento de Ações Programáticas Estratégicas caracteriza a dermatose ocupacional da seguinte maneira: "é toda alteração das mucosas, pele e seus anexos que seja diretamente ou indiretamente causada, condicionada, mantida ou agravada por agentes presentes na atividade ocupacional ou no ambiente de trabalho" (BRASIL, 2006).

Existem alguns critérios a ser considerados para determinação do nexo causal entre a doença da pele manifestada e o trabalho que são: 1. Os primeiros sintomas devem aparecer após a designação de uma estação de trabalho que é conhecida por envolver um risco dermatológico (ainda que o trabalhador não tenha manifesto anteriormente os sintomas); 2. Os sintomas tendem a diminuir quando o trabalhador não está exercendo suas atividades (comumente no dia de folga ou após o trabalho) e piorar quando o trabalho é iniciado; 3. No ambiente de trabalho deve existir um agente etiológico (causal) (KEEFE *et al.*, 2020).

Há ainda alguns potenciais fatores que precisam ser considerados ao relacionar a dermatose com o trabalho: idade, sexo, etnia, temperatura, umidade, condições de trabalho, posições durante o trabalho, não observância em relação às normas de higiene, segurança e saúde ocupacional e, ainda, atividades não relacionadas com o trabalho desenvolvidas durante a vida pessoal (PACHECO, 2018). Dermatite ou eczema de contato é definida como uma inflamação eczematosa reativa da pele ocorrida após o contato direto com um produto químico, podendo ocasionalmente ser manifesta também, ao estabelecer contato com agentes biológicos ou físicos (HOLNESS, 2014).

O CID-10 (Classificação Internacional das Doenças) é uma ferramenta de classificação padrão para a gestão de saúde e epidemiologia, amplamente utilizado para fornecer um panorama geral da saúde de países e grupos de interesse (WHO, 2021b). Dessa forma, para classificar uma determinada moléstia e caracterizá-la de forma universalmente entendível, os estados-membros da Organização Mundial da Saúde adotaram o CID-10 como ferramenta de caracterização das doenças apresentadas (WHO, 2021b). Tratando-se das dermatoses ocupacionais, as mesmas estão codificadas no CID-10 nos grupos L23, L24 e L25. No grupo

L23, estão dispostas as dermatites alérgicas de contato, na categoria L24 as dermatites de contato irritativas e na L25 apresenta-se as dermatites de contato não especificadas (BAINS; NASH; FONACIER, 2019).

A dermatite de contato irritativa decorre dos efeitos tóxicos e pró-inflamatórios de substâncias alcalinas ou ácidas fracas que, não sendo capazes de causar queimadura e/ou necrose acarretam em irritação cutânea (LAMPEL; POWELL, 2019). Dessa forma, um eczema de contato é caracterizado por descamação, vermelhidão da pele e, em alguns casos por bolhas. Estes podem surgir horas após o contato com irritantes fortes ou semanas depois do contato contínuo com agentes irritantes fracos, onde o indivíduo acometido sente-se com queimação e dor (MILAM; COHEN, 2019). Alguns dos irritantes mais comuns a esse tipo de eczema são: água, sabões, graxas, ácidos e álcalis, fibra de vidro, poeira e detergentes (BEHROOZY; KEEGEL, 2014). O CID-10 particiona a L24, portanto, as dermatites de contato irritativas em dez subcategorias segundo o agente causador da lesão a saber: detergentes, óleos e gorduras, solventes, cosméticos, drogas em contato com a pele, outros produtos químicos, alimentos em contato com a pele, plantas e de causas não especificadas (WHO, 2021b).

A dermatite de contato do tipo alérgica pertencente ao grupo L23 segundo Pacheco (2018) e Lampel e Powell (2019), por sua vez, ocorre em pessoas com predisposição individual como resposta imunológica que se apresenta na forma de vermelhidão, formação de bolhas, exsudação (saída de líquido através dos poros com consistência viscosa), escamação, formação de crosta, elevação sólida e limitada da pele e, em alguns casos podendo ser manifesta na figura de liquenificação (processo de alteração da pele de causa psicossomática que toma aspecto semelhante a liquens).

O tempo de aparecimento das lesões após o contato varia de algumas horas até dias dependendo da concentração do agente e pré-disponibilidade do indivíduo (HOLNESS, 2014). As fontes causadoras de dermatites alérgicas consideradas mais comuns são cosméticos, sais metálicos, germicidas, plantas, aditivos da borracha, resinas plásticas, látex e medicamentos tópicos (MILAM; COHEN, 2019). A L23, apresenta como subclassificações das dermatites alérgicas de contato dez categorias para as moléstias que foram dispostas também segundo sua originação e, incluem: metais, adesivos, cosméticos, drogas em contato com a pele, corantes, outros produtos químicos, alimentos em contato com a pele, plantas, outros agentes e de causa não especificada (GONZÁLEZ-LÓPEZ *et al.*, 2018). A categoria L25 conforme anteriormente mencionado, faz referência as dermatites de contato não especificadas e apresenta como subclasses os mesmos agentes causadores existentes na L24 (WHO, 2021b).

Existe um consenso a respeito da subnotificação de doenças da pele relacionadas ao trabalho em virtude da ausência de diagnóstico devido à dificuldade de acesso a profissionais da saúde que conduz a um segundo problema que consiste na automedicação. Nas situações em que o paciente é atendido por um profissional da saúde constata-se um baixo preparo no que se refere ao estabelecimento do nexo causal entre a lesão e a atividade laboral (HOLNESS, 2014).

No Brasil, há ainda alguns fatores que tornam mais difícil a existência de informações relativas às dermatoses ocupacionais: escassez de serviços especializados, não-informação dos trabalhadores em relação aos perigos aos quais estão expostos, grande quantidade de trabalhadores que exercem atividades informais e a descentralização das informações do SUS (MELO; VILLARINHO; LEITE, 2019).

O estudo realizado por Lise *et al.* (2018) muniu-se dos dados do SINAN com o intuito verificar o panorama das doenças da pele de origem ocupacional no Brasil para os casos apresentados de 2007 a 2014. Para atingir seus objetivos, os autores fizeram uso de todo o banco de dados disponibilizado para consulta, abrangendo dessa forma, os 4.710 casos ocorridos durante todo o período considerado. Como resultados, as partes do corpo mais atingidas foram a cabeça e as mãos (27,54%), a faixa etária com maior incidência foi de 35 a 49 anos com 1.852 casos (39,3%). De toda a população estudada 15,4% dos indivíduos apresentaram incapacidade temporária, 1,8% desenvolveram invalidez parcial permanente e 0,2% dos casos mostraram-se com invalidez total permanente em decorrência de alguma variação de dermatites de contato.

De acordo com o conjunto de dados disponibilizado pelo INSS, de dezembro de 2018 a novembro de 2020, 2.217 segurados fizeram uso dos benefícios concedidos pelo INSS em virtude de dermatoses ocupacionais conforme disponível no Anexo A. As três categorias com maior incidência foram respectivamente dermatites alérgicas de contato com 547 casos (24,7%), dermatites de contato não especificada com 395 (17,82%) ocorrências e dermatite alérgica de contato de causa não especificada, acometendo o total de 188 (8,48%) trabalhadores. Os 2.217 casos custaram à previdência o montante de R$ 2.773.246,00 pagos a 1.191 (53,72%) mulheres e 1.026 (46,27%) homens residentes em todos os estados da nação conforme apresentado na Figura 2.1.

Figura 2.1 – Beneficiados por Estado

Fonte: Comitê de Dados Abertos do INSS (2020)

Como é possível estabelecer na Figura 2.1, o estado com maior incidência de benefícios em virtude de dermatoses ocupacionais foi Minas Gerais (405 – 18,27%), seguido por São Paulo (367 – 16,55%), Rio Grande do Sul (193 – 8,71%), Rio de Janeiro (140 – 6,31%) e Santa Catarina (147 – 6,63%). As unidades federativas com menor incidência foram Tocantins (10 – 0,45%), Amapá (7 – 0,32%), Roraima (7 – 0,32%) e Acre (6 – 0,27%), enquanto o Distrito Federal contou com 88 (3,97%) indivíduos acometidos com alguma variante de dermatose ocupacional. Em relação à idade, a categoria compreendida por pessoas de 18 a 27 anos totalizou 133 (5,99%) indivíduos, de 28 a 37 anos 391(17,64%) pessoas, trabalhadores entre 38 a 47 anos somaram o montante de 638 (28,77%) pessoas, de 48 a 57 foram no total 702 (31,66%) beneficiados e segurados com idade superior a 57 anos totalizou 353 (15,92%) trabalhadores (COMITÊ DE DADOS ABERTOS DO INSS, 2020).

2.4 PREVENÇÃO PARA AS DERMATOSES OCUPACIONAIS

As estratégias efetivas de prevenção para as doenças ocupacionais são essenciais para reduzir a incidência de doenças relacionadas ao trabalho bem como os altos custos econômicos, sociais e humanos em decorrência de tais moléstias (KEEFE *et al.*, 2020).

A prevenção primária objetiva prevenir as lesões e doenças do trabalho antes que aconteçam podendo envolver intervenções únicas ou combinações de distintos métodos bem como se concentrar num perigo ou doença específicos ou compreender vários perigos num dado ambiente de trabalho (WEISTENHÖFER *et al.*, 2011). De acordo com a revisão sistemática realizada por Keefe *et al.* (2020) são 5 as principais classes de prevenção primária: 1) legislação, 2) vigilância, 3) medidas de controle de exposição, 4) educação e treinamento e 5) abordagens multifacetadas resultadas da combinação de múltiplos métodos.

Se tratando das dermatoses ocupacionais, as atividades de prevenção primária com maior efetividade são a educação e treinamento combinados com vigilância (KEEFE *et al.*, 2020). Conforme o estudo realizado por Van Gils *et al.* (2011) foi evidenciado que os programas de educação e treinamento contribuem para a redução de dermatoses relacionadas ao trabalho e que intervenções ocupacionais como medidas administrativas e acompanhamento ativo da liderança induzem a mudanças comportamentais importantes.

O Programa Dinamarquês de Proteção à pele foi desenvolvido para possibilitar a comparação de populações no que se refere às dermatoses oriundas do trabalho e a tradução do questionário proposto situa-se no Anexo B (ANDRUP, 2021). Como é possível observar no Anexo B, o questionário apresenta 10 dimensões onde cada dimensão corresponde a uma letra do alfabeto (ex. A, C, etc.) que visam descrever com detalhamento o quadro clínico e epidemiológico dos indivíduos a saber: Histórico Ocupacional e Demográfico Geral (G), Histórico dos Sintomas Atópicos (A), Dermatites (D), Fatores Agravantes (F), Impacto à Qualidade de Vida e Consequências (C), Urticária de Contato (U), Sintomas na Pele (S), Testes de Pele (T), Exposição (E) e Saúde Geral (H). O estudo de Johansen *et al.*, (2012) se baseou no questionário nórdico para investigar como a educação e o treinamento são capazes de influenciar na carreira de cabelereiras para prevenir dermatoses ocupacionais uma vez que é sabido que diversas profissionais deixam a profissão após a ocorrência de lesões severas desta natureza. Foi constatado, com o uso de regressão logística, que profissionais treinados a respeito dos riscos da omissão do uso de EPIs e da gravidade das doenças da pele oriundas do trabalho tendem a não apresentar dermatoses ocupacionais.

Uma forma para encontrar o caminho para a efetiva prevenção visando o controle de exposição de dermatoses ocupacionais é a realização de exames clínicos também conhecidos como periódicos, e dentre eles podem ser recomendados os testes cutâneos. A bateria padrão brasileira de testes de contato foi desenvolvida nos anos de 1993 a 2000 para padronizar as baterias de testes cutâneos com o intuito de atender a realidade brasileira com base na resposta alérgica em decorrência de reação hipersensitiva sendo constituída por 30 compostos químicos (DERMATOLOGIA, 2000). No Quadro 2.1 são apresentados os compostos químicos que compõem a bateria padrão brasileira, a concentração de cada composto conforme a classificação da bateria padrão e o código a qual o composto é adotado para as análises apresentadas no Capítulo 4.

Quadro 2.1 – Bateria padrão brasileira de testes de contato

Substância	**Concentração**	**Código**	**Substância**	**Concentração**	**Código**
Antraquinona	2%	A	Neominina	20%	P
Bálsamo-do-peru	25%	B	Nitrofurazona	1%	Q
Benzocaína	5%	C	Parabenos	15%	R
Bricromato de Potássio	0,5%	D	Parafenilenodiamina	1%	S
Butilfenol-p-terciário	1%	E	Perfure-mix	7%	T
Carba-mix	3%	F	PPD-mix	0,4%	U
Cloreto de cobalto	1%	G	Promerazima	1%	V
Colofônia	20%	H	Propilenoglicol	10%	W
Etilenodiamina	1%	I	Quatemium	2%	X
Formaldeído	1%	J	Quinolina-mix	6%	Y
Hidroquinona	1%	K	Resina epóxi	1%	Z
Irgasan	1%	L	Sulfato de níquel	5%	AA
Kathon CG	0,5%	M	Terebintina	10%	AB
Lanolina	30%	N	Timerosal	0,05%	AC
Mercapto-mix	2%	O	Tiuram-mix	1%	AD

Fonte: Adaptado de Dermatologia (2000)

Os testes cutâneos são realizados com substâncias aplicadas no dorso por meio de contensores *Finn Chambers* com realização de leituras com 48 e 96 horas conforme o critério de leitura adotado pelo *International Contact Dermatitis Research Group* (LACHAPELLE; MAIBACH, 2012).

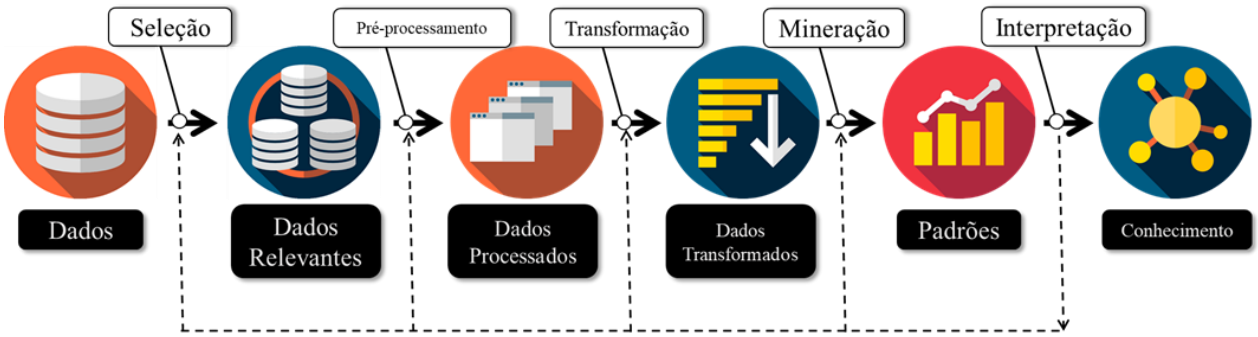
No estudo realizado por Reis, Oliveira e Festino (2012) num ambulatório de medicina do trabalho situado na cidade de Campinas no estado de São Paulo foram avaliados 180 trabalhadores atendidos entre 1999 e 2008 cujas profissões de maior incidência foram serviços de limpeza, pedreiro e metalúrgico. Constatou-se, então, que os compostos de maior relevância das doenças da pele relativas ao trabalho para a população em estudo foram sulfato de níquel, cloreto de cobalto, bicromato de potássio, carba-mix e timerosal.

2.5 PROCESSO *KDD*

A descoberta de conhecimento em base de dados, representada pelo acrônimo *KDD* estabelece-se como a área de pesquisa que lida com a descoberta de conhecimento em conjuntos de dados gerados a partir de processos experimentais e observacionais (KARPATNE *et al.*, 2017). Trata-se de um processo interativo e iterativo, que apresenta várias etapas, cujo objetivo é a identificação de padrões válidos e potencialmente úteis a partir de um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Para Pazmiño-Maji, García-Peñalvo e Conde-González (2017), o *KDD* é um processo de conversão de dados brutos em informações úteis que consiste em uma série de passos de transformação, passos esses que se iniciam com o pré-processamento dos dados e são finalizados na etapa de interpretação dos resultados conforme apresentado na Figura 2.2. A complexidade do processo *KDD* consiste na dificuldade de percepção e interpretação dos inúmeros fatos observados durante o processo e conjugação dinâmica de tais interpretações para a decisão das ações a serem realizadas em cada situação (LUO, 2008).

Figura 2.2 - Etapas do Processo *KDD*



Fonte: Adaptado Fayyad, Piatetsky-Shapiro e Smyth (1996)

Como é possível observar na Figura 2.2 o processo *KDD* requer vários passos de iteração e interação com o usuário. Antes, porém que esses passos sejam iniciados, é necessário que o usuário estabeleça os objetivos e metas, para que o conhecimento produzido na aplicação do *KDD* contribua com o propósito inicial da aplicação do processo (LUO, 2008). Em relação à interação, diz-se que a descoberta de conhecimento em base de dados é dependente do usuário para determinar e alterar parâmetros, refinar os padrões encontrados e interpretar os resultados obtidos. Tratando-se da iteração, esta é marcada pela existência de um conjunto de etapas a serem executadas que propiciam o retorno a etapas anteriormente executadas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Os dados podem ser apresentados em distintos formatos, armazenados e centralizados em um repositório de dados ou distribuídos em múltiplos locais. Para que seja possível o atingimento dos objetivos da mineração é importante fazer uso de boas práticas na seleção dos dados a serem utilizados no trabalho a ser desenvolvido. Nesta fase inicial tem-se como entrada, portanto, os dados oriundos das distintas fontes, o processo de seleção consiste na escolha das informações e na determinação do nível de granularidade, e, como saída é apresentado os dados relevantes à pesquisa (PAZMIÑO-MAJI; GARCÍA-PEÑALVO; CONDE-GONZÁLEZ, 2017).

O propósito do pré-processamento de dados é a remoção ou atenuação de possíveis ruídos presentes no conjunto de dados (LUO, 2008). Um exemplo para o processo de pré-processamento consiste na exclusão das informações que apresentam características excepcionais como *outliers* (valores atípicos) que são capazes de alterar o comportamento do sistema. A etapa de pré-processamento apresenta como entrada os dados selecionados e o processamento confere aos dados características propícias para a determinação de padrões (KARPATNE *et al.*, 2017).

A etapa de transformação objetiva transformar dados brutos em dados apropriados para a extração de padrões abrangendo a categorização, normalização, a transformação dos dados não-estruturados em estruturados, etc. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A mineração de dados refere-se às atividades de análise de dados, descoberta de problemas e oportunidades não facilmente identificáveis em seus relacionamentos, formação de modelos computacionais baseados nessas descobertas e, utiliza-se tais modelos para a previsão de comportamentos futuros (KARPATNE *et al.*, 2017).

O pós-processamento ou interpretação compreende o tratamento do conhecimento obtido na mineração de dados de forma que o mesmo seja avaliado viabilizando a apresentação de resultados úteis para posterior tomada de decisão (SZELKA; WRONA, 2016). Indicadores de desempenho ou métodos de teste de hipóteses podem ser aplicados durante esta fase para eliminar possíveis resultados apócrifos (LUO, 2008). O conhecimento estabelecido após o pós-processamento consiste na representação através de valores numéricos os resultados obtidos possibilitando a identificação de padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.6 *MACHINE LEARNING*

Machine learning pode ser entendido como a intersecção entre a ciência da computação, engenharia e matemática que busca o entendimento do processo no qual as informações estão inseridas para a identificação de padrões, construindo assim, aproximações úteis e a elaboração de previsões (JORDAN; MITCHELL, 2015). Para a realização de tais previsões, supõem-se que o futuro, não seja suficientemente distinto do passado (quando os dados foram coletados), de forma a assegurar sua confiabilidade (WUEST *et al.*, 2016).

A aplicação de métodos de aprendizado de máquina em bases de dados nomeia-se mineração de dados. Faz-se portanto, a analogia com o processo de extração de minérios, situação a qual a partir de um grande volume de terra e matéria-prima extraídos de uma mina é obtido uma pequena quantidade de material consideravelmente precioso após o processamento (YANAR; LAY; SMITH, 2019). A mineração em base de dados, portanto, possui como *inputs* um grande volume de informações que passam por um processamento e apresentam *outputs* modelos simples e de uso valioso, preferencialmente com alta precisão preditiva. De acordo com Wuest *et al.*, (2016), na manufatura, modelos de *machine learning* são utilizados por exemplo, para otimização, controle e solução de problemas, enquanto na medicina conforme Forsting (2017), o aprendizado é empregado comumente na elaboração do diagnóstico e determinação da epidemiologia dos pacientes.

Para fazer uso de técnicas de *machine learning*, não é necessário o uso exclusivo em dados estruturados (na forma de tabela), pode-se ainda realizar a aplicação em outros tipos de fontes de informação como voz, imagem e texto que são dados classificados como não-estruturados (BUCZAK; GUVEN, 2016).

O aprendizado de máquina não é de uso exclusivo do processo de descoberta de conhecimento em base de dados, sendo também largamente empregado na inteligência artificial (WU *et al.*, 2008). Para que um sistema seja considerado inteligente, o mesmo necessita de estar inserido num ambiente sujeito a mudanças, e, ainda conseguir aprender com tais transformações (ALANAZI; ABDULLAH; QURESHI, 2017). Uma vez que o sistema é capaz de aprender e adaptar-se a tais alterações, o projetista ou programador do sistema não precisará fornecer as soluções para todas as possíveis situações (BUCZAK; GUVEN, 2016).

2.6.1 Tipos de Aprendizado

O aprendizado de máquina se refere à programação de computadores para a otimização de um critério de desempenho fazendo uso de dados de exemplo ou experiências anteriores (KAKHKI; FREEMAN; MOSHER, 2019). Existem várias formas de aprendizado e as mais conhecidas são o aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado de reforço.

Em relação ao aprendizado supervisionado, de acordo com Wu *et al.* (2008), o conjunto de dados possui uma variável-alvo pré-definida e os registros são categorizados em relação a essa variável, indicando, portanto, a existência de um fornecimento de respostas. Existem algumas etapas que são executadas pelo aprendizado supervisionado após a seleção do algoritmo, conforme Mehta *et al.* (2019) que são o treinamento do conjunto de dados e a validação ou o teste para a avaliação do modelo. Dessa forma, na fase de treino a partir do conjunto de dados acontece a formulação dos hiperparâmetros e na de validação, o algoritmo treinado é avaliado utilizando os dados de validação, sendo possível ajustar os parâmetros caso o desempenho não esteja satisfatório o suficiente durante a validação (CALLAHAN; SHAH, 2017).

No aprendizado não supervisionado, não há o fornecimento de respostas de forma que o próprio algoritmo deve identificar os *clusters* ou grupos existentes no conjunto de dados (VALÊNCIO *et al.*, 2011). Há ainda uma outra aplicação para o aprendizado não supervisionado que consiste na redução da dimensão dos dados, convertendo-os em uma massa de informações relativamente menor, porém sem que ocorra a perda de representatividade do banco de dados original, o que possibilita maior eficiência e manutenção da qualidade do resultado (COMBERTI; BALDISSONE; DEMICHELA, 2015). O aprendizado não supervisionado, portanto, objetiva descobrir as classes desconhecidas de itens por agrupamento, identificando e aproximando os registros similares (redução de dimensão).

Tratando-se do aprendizado de reforço ou *reinforcement learning*, as informações de treinamento são fornecidas pelo ambiente, de forma que os dados de desempenho do sistema e suas respostas são fornecidas por um sinal numérico de reforço (MEHTA *et al.*, 2019). Além disso, o algoritmo precisa descobrir quais ações fornecem os melhores resultados (sinal numérico de reforço) por tentativa e erro ao invés de ser informado (DOLTSINIS; FERREIRA; LOHSE, 2012). O *reinforcement learning*, portanto, emula o processo de aprendizagem realizado pelos seres humanos com base na resposta ambiental sequencial.

2.6.2 Tarefas de Mineração

Em relação às tarefas realizadas pela mineração de dados, as principais categorias incluem: descrição, classificação, regressão, agrupamento, redução de dimensionalidade, predição e associação (MEHTA *et al.*, 2019).

A tarefa de descrição, conforme Saâdaoui *et al.* (2015), fornece caminhos para a interpretação dos resultados encontrados, sendo amplamente aplicada para descrever os padrões e tendências revelados pelos conjuntos de informações. Para Sidey-Gibbons e Sidey-Gibbons (2019), além de fornecer uma explicação para o comportamento dos dados, as atividades relacionadas às tarefas de descrição explicam também os relacionamentos entre as variáveis num ambiente complexo, aumentando, portanto, a compreensão a respeito de produtos, processos e pessoas.

A tarefa de agrupamento, por sua vez, objetiva identificar e aproximar os registros similares, de maneira que um grupo é uma coleção de registros que apresentam similaridade, porém distintos dos outros registros constantes nos demais agrupamentos (ALANAZI; ABDULLAH; QURESHI, 2017). Diferentemente da regressão e da classificação, Deo (2015) esclarece a respeito de que o agrupamento não necessita de categorização prévia.

A técnica de associação objetiva encontrar grupos de itens que apresentam a tendência de ocorrer juntos nas instâncias presentes na base de dados, identificando a relação entre atributos (WU *et al.*, 2008). A associação, dessa forma, busca explicar as associações fundamentais entre itens de maneira que a existência de determinados itens em uma operação implicará na presença de outros itens na mesma operação (BUCZAK; GUVEN, 2016).

A redução de dimensionalidade é uma importante tarefa na mineração de dados pois possibilita a eliminação de subconjuntos de atributos a partir do conjunto original de variáveis que por diversas vezes apresentam elevado número de dimensões (SAÂDAOUI *et al.*, 2015). Um grande número de dimensões eleva consideravelmente a complexidade das técnicas de manipulação e degrada o desempenho dos algoritmos de mineração. Visando diminuir tais efeitos, a redução de dimensionalidade representa um conjunto de dados num espaço de dimensão menor mantendo as mesmas características do conjunto de dados original (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

A respeito da tarefa de classificação, trata-se de uma das tarefas mais utilizadas no âmbito da saúde e segurança ocupacional e também da medicina ocupacional, pois objetiva identificar a classe à qual determinado indivíduo integra (SARKAR; VERMA; MAITI, 2018; ZHAO *et al.*, 2019). Assim sendo, o modelo realiza a análise do conjunto de registros

fornecidos, conjunto esse onde cada registro já comporta a indicação de qual classe pertence, e, aprende com esses registros para que posteriormente tenha condições de classificar automaticamente um novo registro (OBERMEYER; EMANUEL, 2016).

A tarefa de regressão determina o valor de uma variável- alvo ou variável dependente a partir das observações das demais variáveis (YOO; RAMIREZ; LUIZZI, 2014). A regressão, portanto, atribui peso às variáveis independentes para então determinar o valor da variável dependente.

Na tarefa de predição os registros apresentam dados temporais e objetiva descobrir o valor futuro de um determinado atributo (RAJKOMAR; DEAN; KOHANE, 2019; WU *et al.*, 2008). Dados históricos são utilizados, portanto, para construir um modelo capaz de explicar um comportamento observado (WUEST *et al.*, 2016).

2.7 TÉCNICAS DE CLASSIFICAÇÃO

2.7.1 Definições Gerais

As técnicas de mineração de dados são as diferentes classes de algoritmos e respectivos conceitos matemáticos que os descrevem, bem como a forma como ocorrem os processos de mineração e recuperação dos dados (MEHTA *et al.*, 2019). A principal característica das técnicas de classificação é a coleta de amostras e tais amostras são observadas e posteriormente atribuídas classes a essas amostras (OBERMEYER; EMANUEL, 2016). Existem duas categorias principais ao classificar algoritmos de *machine learning*: clássicas e *ensemble.*

Técnicas clássicas são as que realizam predições por si só enquanto técnicas *ensemble* realizam a combinação de diversas técnicas comparando-as com algum estimador individual. Técnicas *ensemble*, portanto, apresentam o intuito de abrandar suas limitações para produzir classificadores mais poderosos sendo subdivididos em *bagging* e *boosting* (WU *et al.*, 2008). Em *bagging*, as técnicas consideram diversos estimadores individuais e realizam a combinação dos mesmos ao fazer uso da média e como resultados tem-se estimadores superiores do que quando comparados com técnicas individuais um exemplo de técnicas de *bagging* é *Random Forest* (BREIMAN, 1996). Técnicas de *boosting* alcançam resultados poderosos realizando a combinação de muitos estimadores fracos o que reduz o viés e como exemplos tem-se *Adaboost*, *Catboost* e *Extreme Gradient Boosting* (HANCOCK; KHOSHGOFTAAR, 2020).

2.7.1 Regressão Logística

A técnica de Regressão Logística segundo Marucci-Wellman, Corns e Lehto (2017) refere-se a um modelo não paramétrico, cuja variável dependente configura-se como binária. Dessa forma, a regressão logística pode ser descrita conforme a Equação 2.1.

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}} \qquad (2.1)$$

Em que $g(x) = B_0 + B_1X_1 + \cdots + B_nX_n$. Os coeficientes (de B_0 até B_n) são estimados pelo método da verossimilhança onde a partir das variáveis independentes (de X_1 até X_n) atribui-se pesos as mesmas para determinação dos parâmetros para então categorizar a variável dependente. A representação gráfica dessa técnica é definida pela curva logística e, apresenta os melhores resultados quando as variáveis são contínuas e a suposição de linearidade é verdadeira (MEHTA *et al.*, 2019).

2.7.2 *Neural Network*

Neural Network (que traduzindo para a língua portuguesa se refere à rede neural) é constituída por um conjunto de algoritmos projetados para reconhecer padrões, interpretar os dados recebidos por um tipo de percepção de máquina e então os rotular ou agrupá-los ao reconhecer seus padrões, simulando assim, o comportamento do cérebro humano. Uma rede neural simples compõe-se de uma camada de entrada, uma camada de saída, e, dentre estas, uma camada oculta. As camadas, por sua vez, são compostas por nós e conectadas por meio de conexões sinápticas que constituem os pesos de ligação entre 2 nós. De forma simplista os dados são inseridos na rede neural por meio de uma camada de entrada que se comunica com as camadas ocultas, e, o processamento acontece nessas camadas por meio de um sistema de conexões ponderadas. Os nós existentes nas camadas ocultas, realizam a combinação dos dados de entrada com um conjunto de coeficientes e atribui distintos pesos as entradas que são então, somados e na sequência, a soma é processada pela função de ativação de um nó, responsável por determinar a extensão a qual um resultado deve progredir na rede para influenciar o resultado final. Para finalizar, as camadas ocultas se ligam a camada de saída (PARODI, 2012; WUEST *et al.*, 2016; MEHTA *et al.*, 2019).

2.7.3 *Random Forest*

Random Forest é um algoritmo de classificação e também de regressão composto por um conjunto de árvores de decisão que para realizar a classificação realiza o cálculo da média das

árvores de decisão, minimizando o componente de variação do modelo, tornando a aproximação mais próxima de um modelo ideal. Dessa forma, concentra-se na amostragem de observações e das variáveis de dados de treinamento para desenvolver árvores de decisão independentes e obter votação majoritária para então realizar a classificação (BREIMAN, 2001; XU *et al.*, 2012).

2.7.4 *Adaboost*

A ideia principal do *Adaboost* é focar em instâncias que foram classificadas incorretamente ao realizar o treinamento. Dessa forma, o nível de foco é determinado por um peso atribuído a cada instância no conjunto de treinamento, e, na primeira iteração, o mesmo peso é atribuído a todas as instâncias. Nas iterações subsequentes, por sua vez, os pesos das classificações incorretas são aumentados enquanto o das corretas diminuídos. Sendo assim, as deficiências são identificadas por pontos nos dados de alto peso e as perdas de função são classificadas como exponenciais (FREUND; SCHAPIRE, 1995; SAGI; ROKACH, 2018; PALIT; REDDY, 2012).

2.7.5 *Extreme Gradient Boosting*

Extreme Gradient Boosting é um modelo de *boosting* diferenciado das demais técnicas de *boosting* pois para apresentar o melhor desempenho faz uso da formalização do modelo para realizar a regularização e controle do sobreajuste (VERMA; PAL; KUMAR, 2019b). Além de realizar a regularização, XGB admite recursos esparsos para entradas automáticas lidando assim com distintos padrões de dispersão eficientemente e apresenta um método de validação cruzada embutido em cada iteração (SHERIDAN *et al.*, 2016).

2.7.6 *Catboost*

Catboost, por sua vez, como próprio nome sugere realiza a combinação de ‘categoria’ e ‘*boosting*’ uma vez que lida com categorias por conta própria e baseia-se no algoritmo de aumento de gradiente (GHORI *et al.*, 2020). CAT foi criado em meados de 2018 com o propósito de lidar com facilidade com grandes conjuntos de dados e também com dados heterogêneos (bancos de dados heterogêneos apresentam diferentes tipos de dados que comumente constituem bancos de dados relacionais) (PROKHORENKOVA *et al.*, 2018). Ao contrário de grande parte dos algoritmos de mineração que requerem várias etapas de pré-processamento, o CAT requer apenas índices de recursos categóricos para então realizar a

transformação das categorias em dados numéricos (HANCOCK; KHOSHGOFTAAR, 2020). Outra característica relevante do CAT é que o mesmo não requer grandes conjuntos de dados para treinamento extensivo mesmo que utilize uma diversidade de parâmetros (GHORI *et al.*, 2020).

Ainda em relação ao *Catboost*, é importante ressaltar que se trata de um algoritmo *open source* que traz consigo a inovação do ordenamento do *boosting* (HANCOCK; KHOSHGOFTAAR, 2020). Para ordenar o *boosting* os desenvolvedores de *Catboost* se inspiraram em algoritmos de aprendizagem *online*. Tais algoritmos, obtêm exemplos de treinamento em ordem sequencial. Para simular tal efeito no ambiente *offline* foi desenvolvida uma variável de permutação randômica a partir dos exemplos de treinamento. Para cada nova observação durante a fase de treinamento, *Catboost* utiliza todo o histórico disponível aplicando-o à nova observação (PROKHORENKOVA *et al.*, 2018).

Como se trata de uma ferramenta relativamente nova, Hancock e Khoshgoftaar (2020) realizaram uma revisão de literatura onde procuraram entender em quais segmentos a técnica CAT estava sendo utilizada e quando aplicável o que era medido quando a mesma era comparada com outras técnicas. Em relação aos segmentos, *Catboost* se apresentou como grande aliado a estudos interdisciplinares tendo aplicações na astronomia, finanças, medicina, biologia, fraude de utilidades, meteorologia, bioquímica, *marketing* e *cyber* segurança. Quando comparado com outras técnicas de mineração de dados, CAT mostrou-se superior em vários contextos e inferior em outros. Tratando-se exclusivamente da área médica, o estudo de Hancock e Khoshgoftaar (2020) encontrou apenas três estudos utilizando *Catboost* em comparação com outras técnicas e em apenas um dos três estudos em questão CAT foi o vencedor.

2.8 *MACHINE LEARNING* EM SSO

A subseção 2.8 apresenta trabalhos relacionados à Saúde e Segurança Ocupacional abrangendo aspectos relativos a diagnósticos clínicos, medidas de controle e proteção, perfil de acidentes estando disposta em 3 subseções. Na subseção 2.8.1.encontra-se estudos direcionados ao tipo de aprendizado, ferramentas, tarefas e algoritmos enquanto as subseções 2.8.2 e 2.8.3 abordam trabalhos cuja metodologia ou escopo de estudo se assemelharam ao proposto na presente dissertação.

O escopo da presente pesquisa abrange aspectos relacionados à SSO e o conjunto de dados utilizado origina-se de prontuários médicos de dermatologia do trabalho. Desta forma, a

conjuntura dos estudos que apresentam relação com o tema de pesquisa é encontrada em periódicos consideravelmente distintos. Por esta razão, na subseção 2.8.2 observa-se estudos cuja aplicação em SSO analisou doenças do trabalho enquanto a subseção 2.8.3 abrangeu pesquisas de aplicações de *machine learning* para aprimorar a assertividade de diagnósticos médicos em dermatologia.

2.8.1 Aplicações de *Data Mining* no Âmbito de Doenças e Acidentes Ocupacionais

No âmbito da SSO o trabalho de Cheng, Yao e Chih (2013) utilizou a tarefa de descrição para analisar as causas dos acidentes de trabalho na indústria petroquímica. Como achados da pesquisa foi constatado que os fatores determinantes para acidentes na indústria em questão são oriundos de atividades relacionadas à manutenção tais como necessidade de substituições de tubos, estancamento de vazamentos, etc. Fatores contribuintes aos acidentes citados por Cheng, Yao e Chih (2013) estão a implementação de medidas de segurança direcionadas à educação e treinamento das equipes de trabalho, projetos eficazes de proteção contra incêndios e planos de manutenção assertivos.

Ainda em relação à técnica de descrição Akboğa e Baradan (2017) direcionaram seus esforços para realizar a investigação dos acidentes de trabalho no setor da construção civil. Foi verificado que os fatores de maior influência para essa classe de eventos foram queda de elevação de origem não estrutural (23,6%), colapso de objetos iniciados com falha estrutural (21,2%) e colisão durante a movimentação de objetos (13,4%).

Em relação à tarefa de agrupamento de posse de um banco de dados oriundo de uma indústria madeireira Comberti *et al.* (2015) agrupou os dados referentes aos acidentes de trabalho e posteriormente os classificou segundo sua similaridade. Desse estudo, foram encontrados 13 grupos a saber: movimentos com manuseio de marcenaria e quedas em materiais diversos, trabalho com ferramentas manuais e danos por impacto com peças projetadas, perda de controle e impacto da carga de transporte com algo em movimento, movimentos de abertura/fechamento de portas e impacto de esmagamento, movimentos incorretos e danos por contato com peças de corte, movimentos incorretos e danos por contato com máquinas de corte, manuseio de objetos e danos por perda de controle e movimentos incorretos, transporte manual e tensões para ações incorretas, ação dinâmica ou estática com movimento torcional incorreto, perda de controle e danos por contato com elementos de corte, carga e transporte e danos por esmagamento ou impacto, perda de controle ou movimentos

incorretos e grandes danos de corte, perda de controle ou movimentos incorretos e grandes danos de corte.

Ainda em relação à tarefa de agrupamento, Valêncio *et al.* (2011) ocuparam-se em fazer uso da mesma na área da saúde ocupacional para auxiliar a equipe de atendimento à emergências numa cidade do interior do estado de São Paulo no Brasil. O propósito deste estudo consistiu na alocação da equipe de socorro, na priorização dos atendimentos e distribuição dos recursos humanos e materiais. Para realizar tal atividade foram relacionados dados geográficos e informações fornecidas durante a chamada telefônica de solicitação de socorro.

A respeito da tarefa de associação foram verificados o estudo de Shin *et al.* (2018) e o de Nenonem (2013). Shin *et al.* (2018) utilizaram regras de associação para descrever a relação entre as causas que provocaram acidentes de trabalho na construção civil na Coréia do Sul. Nenonen (2013), por sua vez, direcionou seus esforços para associar os eventos que acarretaram em acidentes ocupacionais na Finlândia que resultaram em longos períodos de afastamento.

Saâdaoui *et al.* (2015) fizeram uso da tarefa de redução da dimensão para selecionar as variáveis quantitativas e qualitativas que foram utilizadas para descrever acidentes de trabalho no âmbito da medicina ocupacional. As variáveis quantitativas que melhor descreveram os acidentes de trabalho para Saâdaoui *et al.* (2015) foram pressão sanguínea diastólica, índice de massa corporal e colesterol enquanto as qualitativas foram lesões por esforços repetitivos, trabalho noturno e trabalho em turnos. Ainda em relação à redução da dimensão, Chokor *et al.* (2016) analisou os eventos reportados a OSHA *(Occupational Safety and Health Administration*) nos Estados Unidos no estado do Arizona no setor da construção civil para determinar as variáveis de maior significância para os eventos registrados. Tais variáveis foram: causado por quedas (42,9%), atingidas por objetos (34,3%), eletrocuções (12,5%) e queda de trincheiras (10,3%).

Em relação à tarefa de classificação Kang e Ryu (2019) analisaram segundo a importância e a gravidade os fatores componentes dos acidentes ocupacionais registrados na indústria do aço do Irã por meio de árvores de classificação. Foram comparadas 2 técnicas de classificação com as mesmas variáveis e a que alcançou maior acuracidade atingiu 81,78%. Para a técnica que se mostrou superior, os fatores de maior destaque foram idade, causa do acidente e nível de escolaridade. A classificação foi empregada também por Rubaiyat *et al.* (2017) para detectar os trabalhadores que faziam o uso do capacete de segurança.

Tratando-se da tarefa de regressão, a mesma foi utilizada por Bohanec e Delibašić (2015) para elaborar um modelo de previsão de acidentes para instrutores e turistas durante a

operação de um *resort* utilizando dados do clima, sistema RFID (*Radio-Frequency IDentification*) e informações administrativas.

Em relação à tarefa de predição, Zhao *et al.* (2019) concentrou seus esforços para predizer as perdas auditivas associadas a pessoas expostas a ruídos industriais não gaussianos complexos. Para atingir tal feito, foi utilizado dados de trabalhadores de 17 indústrias e comparado 4 algoritmos de *machine learning*: *Support Vector Machine, Random Forest, Neural Network* e *Adaboost*. Para as técnicas de mineração em estudo, verificou-se uma acurácia que variou de 78,6% a 80,1% que são indícios de que todas as técnicas podem descrever satisfatoriamente a perda auditiva associada aos ruídos industriais não gaussianos, porém, o que mais teve destaque foi *Neural Network*.

O objetivo principal da pesquisa de Kakhki, Freeman e Mosher (2019) consistiu em realizar a comparação entre técnicas de mineração de dados para melhor representação dos eventos futuros em termos de acuracidade do algoritmo, direcionado aos acidentes do tipo "desliza-tropeça-cai" em agroindústrias. Para tal feito, foi utilizado a ferramenta Matlab e comparado os algoritmos Naïve Bayes, *Support Vector Machine* e *Adaboost*. O algoritmo de maior destaque foi *Support Vector Machine* contando com 98,44% de acuracidade.

2.8.2 Diagnósticos Clínicos com Aplicações em SSO

Minami *et al.* (2018) aplicaram a regressão logística para explorar a associação entre a gravidade do eczema de mãos (um tipo específico de dermatite de contato) e o risco de alergia alimentar em cozinheiros e manipuladores de alimentos (doença ocupacional) e donas de casa (doenças não ocupacionais). Como resultados, constatou-se que os profissionais que apresentaram eczema de mãos mais graves são mais propensos a desenvolver uma intolerância alimentar. Além disso, o eczema de mãos é mais comum nos indivíduos que exercem atividades profissionais do que nas pessoas que realizam somente atividades domésticas.

O estudo de Kraft *et al.* (2019) utilizou regressão logística para avaliar as doenças da pele e os padrões de sensibilização comuns entre músicos profissionais uma vez que os mesmos apresentam contato físico prolongado e intenso com seus instrumentos. Para atingir tal objetivo, coletaram dados de 1997 a 2017 da Rede de Informação dos Departamentos de Dermatologia da Alemanha sendo constatados 236 profissionais. Desse grupo foi verificado que 58,6% são do sexo masculino e 60,6% apresentam idade inferior a 40 anos. Em relação à frequência do diagnóstico, constatou-se as moléstias mais frequentemente encontradas foram respectivamente dermatite alérgica de contato, dermatite atópica e dermatite irritante de contato. Kraft *et al.*

(2019), dessa forma, concluíram que as dermatites não são doenças comuns nessa classe profissional e recomendaram que sejam realizados testes cutâneos além do tratamento com especialistas.

Hamnerius *et al.* (2018) avaliou com o uso da regressão logística as dermatites alérgicas de contato e também eczema de mãos em profissionais da saúde. Constatou-se que a alergia de contato a aditivos de borracha em luvas médicas é a causa mais comum de dermatite alérgica de contato nestes profissionais.

Srinivas, Rao e Govardhan (2010) realizaram a comparação de taxas de doenças cardiovasculares de uma região específica da Índia que não havia sido submetida a um controle de riscos ocupacionais com outras regiões as quais haviam sido aplicados medidas de controle. Para realizar tal comparação, fez-se uso das técnicas de árvores de decisão, redes neurais e Naïve Bayes. Como resultados, foi constatado que os trabalhadores das áreas onde não havia medidas de controle estabelecidas estavam mais propensos a desenvolver doenças de origem cardiovascular. Além disso, as três técnicas empregadas mostraram-se eficientes para predizer tais doenças de forma que em relação à acuracidade as árvores de decisão apresentaram 82,5%, redes neurais resultaram em 89,7% e Naïve Bayes 87%.

2.8.3 Comparação de Técnicas de Mineração de Dados em Diagnósticos Médicos de Dermatologia

O emprego de técnicas de mineração de dados para o tratamento e prevenção de doenças relacionadas à pele vêm sendo utilizado por vários pesquisadores bem como para a investigação e prevenção de acidentes no ambiente de trabalho. Dentre os estudos correlatos apresentados na presente subseção, não foram verificados trabalhos que abordassem conjuntamente as doenças da pele relacionadas com o trabalho e mineração de dados de maneira que as pesquisas aqui constantes foram selecionadas segundo a metodologia e variável de resposta quando aplicável. Quatro dos cinco estudos aqui presentes limitam-se à relação de doenças da pele com o emprego de técnicas *data mining* enquanto o trabalho restante diz respeito à aplicação de técnicas de mineração de dados para a predição de diversas doenças de relacionadas ao trabalho.

No Quadro 2.2 situam-se os estudos correlatos à metodologia proposta na presente dissertação. Dessa forma, classificou-os segundo a origem dos dados, o tamanho de amostra utilizado, quantidade de variáveis presente no banco, o tipo de tarefas de mineração, se a doença está ou não relacionada ao trabalho, a caracterização da variável de resposta quando aplicável e as técnicas de mineração empregadas.

Quadro 2.2 - Comparativo dos estudos relacionados

Autores	**Origem dos Dados**	**Amostras**	**Variáveis**	**Tarefa (s)**	**Doença do Trabalho**	**Variável de Resposta**	**Técnicas**
Verma, Pal e Kumar (2019, a)	Repositório de Dados	366	33	Classi-ficação	Não	Doenças da Pele: Psoríase, Dermatite Ceborréica, Dermatite Crônica, Pitiríase rósea, Pitiríase Rubra e Líquen Plano	PAC, LDA, RNC, BNB, NB e ETC
Verma, Pal e Kumar (2019, b)	Repositório de Dados	366	15	Classi-ficação	Não	Doenças da Pele: Psoríase, Dermatite Ceborréica, Dermatite Crônica, Pitiríase rósea, Pitiríase Rubra e Líquen Plano	CART, SVM, DT, RF, GBM
Yun *et al.* (2017)	Prontuários Médicos	66	33	Agrupa-mento	Não	Não Aplicável	K-MEANS
Di Noia *et al.* (2020)	Dados do INAL(National Institure for Insurance against Accidents at Work) da Itália	9.676	6	Agrupa-mento e Classi-ficação	Sim	Doenças Gerais: Perda auditiva, Doenças do Rachis, Doenças Osteomuscu-lares, tumor, Síndrome do Túnel do Carpo, Doenças da Pele	K-MEANS, KNN, SVM
Chang e Chen (2009)	Prontuários Médicos	366	34	Classi-ficação	Não	Doenças da Pele: Psoríase, Dermatite Ceborréica, Dermatite Crônica, Pitiríase rósea, Pitiríase Rubra e Líquen Plano	DT e NN

Fonte: Adaptado de Verma, Pal e Kumar (2019, a), Verma, Pal e Kumar (2019, b), Di Noia *et al.* (2020), Chang e Chen (2009) e Yun *et al.* (2017)

O trabalho de Verma, Pal e Kumar (2019, a) ocupou-se na aplicação de seis técnicas de mineração em três métodos *ensemble* (*AdaBoost*, *Bagging e Gradient Boosting*) para predizer as classes de doenças da pele e então propor um novo método para predizer as doenças da pele. Como resultados, os autores comparam a acuracidade dos métodos *ensemble* e também com a resposta individual. Verificou-se então que com a junção das técnicas no método *ensemble* obteve-se elevados níveis de acuracidade. O trabalho Verma, Pal e Kumar (2019, b) utilizou o mesmo banco de dados do de Verma, Pal e Kumar (2019, a) com menor quantidade de variáveis. Dessa forma, foram comparadas 6 técnicas de mineração os resultados mostraram-se similares ao primeiro estudo, porém com menor acuracidade.

O estudo realizado por Di Noia *et al.* (2020) objetivou demonstrar a melhor técnica capaz de prever o risco de doenças ocupacionais em trabalhadores da saúde. Para atingir seus objetivos, aplicou-se a técnica *K-means* antes das técnicas de classificação *K-Nearest Neighbors* e *Support Vector Machine* para seleção das variáveis mais importantes para realizar a predição. Como contribuições os autores concluíram que a aplicação de uma técnica de agrupamento

antes das técnicas de classificação colaborou para uma descoberta mais profunda de conhecimento, portanto, mais útil para a prevenção de riscos relacionados ao trabalho.

O trabalho apresentado por Chang e Chen (2009) conduziu cinco experimentos para avaliar seis tipos de doenças relacionadas a pele utilizando uma combinação de árvores de decisão com redes neurais e, assim comparar tais combinações com os resultados obtidos pelas técnicas individualmente. Os autores chegaram à conclusão de que o emprego de modelos mistos trouxera uma acuracidade inferior do que quando comparados com os algoritmos individuais, de maneira que a rede neural proporcionou o melhor resultado em relação à árvores de decisão.

Diferenciando-se dos estudos anteriormente apresentados que buscavam classificar os dados para predizer um diagnóstico, o trabalho de Yun *et al.* (2017) direcionou seus esforços para verificação da existência de grupos de pacientes com base nos sintomas apresentados, utilizando a tarefa de agrupamento. Como resultados, os grupos foram comparados pelo método de fatorização negativa e verificou-se que os mesmos se distinguiram de maneira considerável em relação aos sintomas e gravidade das doenças, viabilizando assim a padronização dos diagnósticos.

Na pesquisa de Melo, Villarinho e Leite (2019) foi realizado a identificação do perfil sociodemográfico e clínico dos pacientes de um serviço especializado em doenças do trabalho utilizando regressão logística com um viés predominantemente estatístico (sem a abordagem *machine learning*) com o mesmo banco de dados utilizado nesta dissertação. Como resultados, encontrou-se que dos 616 casos existentes no banco de dados 56 mostraram-se inconclusivos, 560 pacientes apresentaram teste de contato conclusivo, 289 desenvolveram dermatose ocupacional e os demais (271) contraíram dermatose de contato cuja origem não foi relacionada ao trabalho. Além disso, ocorreu o predomínio da dermatite do tipo alérgica em relação à irritativa nos casos ocupacionais e chance de contração de uma dermatose ocupacional foi maior entre homens e menor entre pacientes com idade maior ou igual a 50 anos e com nível de escolaridade mais elevado. Quanto à chance de apresentar dermatite de contato alérgica ocupacional, apenas a variável sexo mostrou-se como estatisticamente significativa.

Em relação às profissões, as atividades profissionais mais atendidas foram relacionadas à limpeza, pedreiro/servente, relacionadas a tintas, mecânico/metalúrgico e cozinheiro (MELO; VILLARINHO; LEITE, 2019). Tratando-se dos testes de contato da bateria padrão brasileira, os alérgenos mais incriminados foram: sulfato de níquel, bicromato de potássio, cloreto de cobalto, carba-mix e formaldeído. Os autores destacaram também que segundo estudos

anteriores dermatites de contato do tipo alérgica correspondem apenas a cerca de 20% das dermatoses de contato relativas ao trabalho, existindo, portanto, prevalência das dermatoses irritativas ao contrário dos achados em sua própria pesquisa.

Um aspecto adicional também abordado por Melo, Villarinho e Leite (2019) trata da proporcionalidade de casos oriundos do trabalho dos pacientes que adquiriram dermatoses em sua vida pessoal uma vez que tal proporção foi estabelecida próxima de 1 pra 1 (1 ocupacional pra 1 não ocupacional). Dessa forma, concluiu-se que os profissionais de saúde que não possuem formação específica em doenças do trabalho não conseguem distinguir facilmente uma doença do trabalho de uma doença adquirida na vida pessoal. Tal afirmação foi pautada no fato de que uma vez que o DRT é um serviço de dermatologia especializado em trabalho esperava-se encontrar uma quantidade consideravelmente maior de doenças ocupacionais quando comparado a doenças não ocupacionais.

PROCEDIMENTOS METODOLÓGICOS

3

Em relação a metodologia de pesquisa utilizada nesta dissertação, o estudo caracteriza-se por descritivo, exploratório e de natureza aplicada pois objetiva realizar a caracterização de um fenômeno específico. Tratando-se da abordagem, foi utilizada a abordagem quantitativa uma vez que é buscado traduzir dados de origem numérica em informações para classificação e análise. A respeito dos procedimentos de pesquisa, a mesma categoriza-se como experimental dado que serão selecionadas as variáveis capazes de influenciar um objeto.

Dessa forma, o objeto de estudo são as dermatites de contato relacionadas ao trabalho. Para condução da pesquisa procurou-se caracterizar os elementos que exercem influência na ocorrência de uma dermatite de contato partindo da comparação entre as técnicas de mineração de dados. O banco de dados utilizado apresenta 616 observações coletadas entre os anos de 2000 e 2014 de pacientes maiores de 18 anos que necessitaram de realizar a investigação com testes de pele pois apresentaram manifestações clínicas sugestivas a dermatite de contato. Cada paciente, portanto, corresponde a uma única observação e, cada um respondeu a um questionário contendo suas características sociodemográficos, queixas associadas, antecedentes familiares e pessoais, além dos tratamentos anteriormente realizados.

A pesquisa foi estruturada em 3 etapas principais: planejamento, execução e análise dos resultados. Em relação ao planejamento, o mesmo foi iniciado ao definir que seria utilizado informações relativas à saúde e segurança ocupacional e buscou-se encontrar um objeto que não havia até o momento sido estudado massivamente no âmbito da ciência. Realizado a leitura de diversos artigos, observou-se a oportunidade de estudar as dermatoses ocupacionais fazendo uso de um banco de dados de prontuários médicos e para condução das análises foi definido o uso de mineração de dados.

Determinados o objeto de estudo e o conjunto de dados, devido à natureza da informação foi necessário a submissão ao Comitê de Ética em Pesquisa, que por sua vez, conferiu um parecer favorável ao protocolo apresentado, parecer esse de CAAE número 25497719.0.0000.0104 que está disposto no Anexo C.

Após a aprovação da metodologia de pesquisa no Comitê de Ética, realizou-se a análise descritiva do banco de dados e posteriormente foram definidos os objetivos, ferramentas e tipos de aprendizado empregados. A respeito das ferramentas, optou-se por fazer uso do *software* R

versão 3.6.1 juntamente com o ambiente para desenvolvimento integrado RStudio *Desktop* versão 1.2.5019 se trata de uma ferramenta livre, possuidor de sua própria linguagem de programação, que dispõe uma vasta gama de funções analíticas, contando ainda com grande quantidade de pacotes disponíveis para *download.* A ferramenta utilizada foi instalada num computador com sistema operacional Microsoft Windows 10, operando com 8 GB de memória RAM, com processador Intel Core i5-6200, possuidor de 2 núcleos, 4 threads, 2,30 GHz.

A respeito do tipo de aprendizado foi empregado o aprendizado supervisionado pois as saídas são conhecidas e para ordenamento das variáveis muniu-se de tarefas de classificação. Para a etapa de Execução, o código foi desenvolvido e implementado na base de dados e como resultado desse processo foram gerados gráficos, tabelas e figuras que forneceram o suporte necessário para a as análises subsequentes.

3.1 PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO

Em relação à execução da pesquisa, a mesma foi demarcada com a categorização, análise e transformação das variáveis. No Quadro 3.1 estão presentes as variáveis constantes no banco de dados original, sua categorização e subclasses por variável quando aplicável e a Figura 3.1 dispõe o passo a passo da transformação das variáveis para o atingimento dos objetivos de pesquisa.

A respeito das particularidades relativas as variáveis, foi constatado que a base de dados original apresentou 84 fatores para análise, desse montante, estabeleceu-se que 14 compõe o perfil epidemiológico dos pacientes e 30 dizem respeito aos resultados da bateria padrão brasileira de testes de contato. Outras 30 variáveis, por sua vez, classificaram a relevância dos resultados dos testes de contato na determinação da dermatose ocupacional enquanto as 10 restantes apontam a parte do corpo lesionada.

As variáveis classificadas como epidemiológicas foram sexo, idade, etnia, escolaridade, profissão, atopia, síndrome da pele excitada, teste de luvas, dermatose ocupacional e diagnóstico final por paciente. A respeito do diagnóstico final, são 85 os tipos de doenças possíveis e cada paciente pode ser classificado com até 4 moléstias por exemplo: o diagnóstico 1 ser dermatite irritativa, o diagnóstico 2 ser líquen plano, o diagnóstico 3 ser eczema de estase e o diagnóstico 4 urticária crônica.

Em relação às variáveis da bateria padrão brasileira de testes de contato obteve-se que as subclasses constituintes são as mesmas constantes no teste de luvas estabelecendo-se,

portanto, em 7 categorias que vão de não realizado até positivo (+++). Se tratando do local das lesões, são 10 as áreas apontadas e cada área foi classificada como lesionada ou não lesionada.

Quadro 3.1 – Variáveis Constantes na Base de Dados

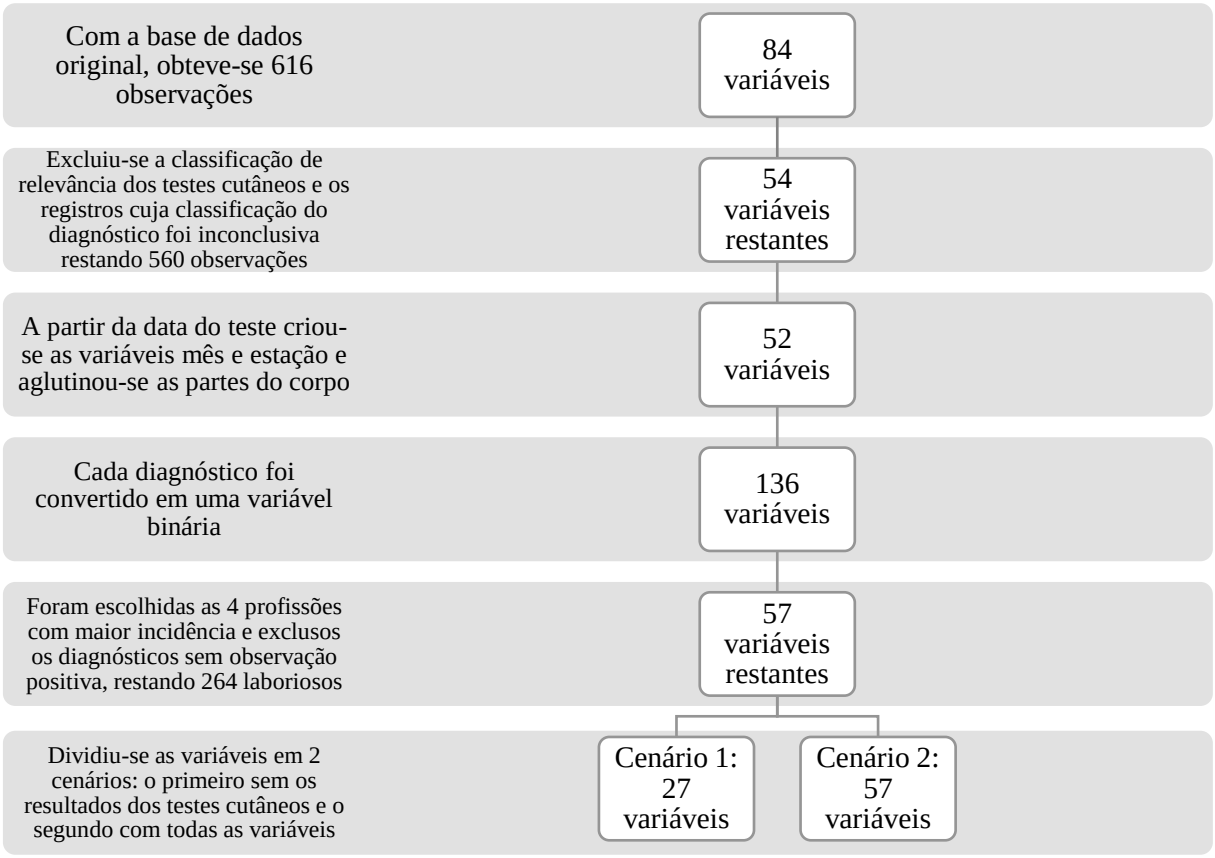
Variável	Total de Variáveis	Tipo	Subclasses
Sexo	1	Categórica	Masculino; Feminino
Profissão	1	Categórica	Costureiro/alfaiate; Balconista / vendedor; Relacionado a tintas; Cozinheiro; Lavrador; Marceneiro/carpinteiro; Manicure; Mecânico/ metalúrgico; Pedreiro/ servente; Químico; Relacionado a artes plásticas; Relacionado a limpeza; Cabelereiro; Trabalhador da saúde; Trabalhador de escritório; Outras; Lanterneiro; Atividades domésticas; Não especificado; Padeiro; Agente de endemias; Bombeiro Hidráulico; Motorista; Estudante; Cuidador de idoso; Professor/ auxiliar de educação; Porteiro; Cobrador de ônibus; Desempregado
Idade	1	Contínua	-
Etnia	1	Categórica	Não especificado; Parda; Negra; Amarela; Branca
Escolaridade	1	Categórica	Não sabe ler e escrever; Alfabetizado; Fundamental incompleto; Fundamental completo; Médio incompleto; Médio completo; Superior incompleto; Superior completo; Especialização/residência; Mestrado; Doutorado; Ignorado
Atopia	1	Categórica	Não especificado; Não; Sim; Inconclusivo
Local das Lesões	10	Binária	Face e pescoço; Dorso das Mãos; Palma das Mãos; Antebraço; Plantas; Pés e Coxas; Tórax Anterior; Abdome; Dorso; Dorso dos Pés
Data do teste	1	Data	DD/MM/AAAA
Bateria Padrão Brasileira de Testes de Contato	30	Categórica	Não Realizado; Duvidoso; Negativo; Irritativo; (+); (++); (+++)
Relevância dos Testes Cutâneos	30	Binária	1 classificação de relevância para cada composto dos testes cutâneos
Teste de Luvas	1	Categórica	Não Realizado; Duvidoso; Negativo; Irritativo; (+); (++); (+++)
Síndrome da Pele Excitada	1	Binária	Sim; Não
Dermatose Ocupacional	1	Categórica	Ocupacional; Não Ocupacional; Inconclusivo
Diagnóstico Final	4	Categórico	85 tipos de doenças possíveis

Fonte: Autora (2021)

Como é possível observar no Quadro 3.1, a base de dados original apresentou trabalhadores cuja moléstia foi classificada pela equipe médica como: ocupacional, não ocupacional e inconclusiva. Dessa forma, o primeiro passo da implementação foi estabelecido pela exclusão das observações que se mostraram inconclusivas seguido pela verificação de *missing values* e não foram verificadas informações faltantes no conjunto de dados.

Das 84 variáveis constantes na base original, foram exclusas as 30 variáveis que corresponderam à classificação dos resultados da bateria padrão por serem poucos os representantes classificados como relevantes, restando, portanto, 54 variáveis para estudo.

Figura 3.1 – Transformação das Variáveis



Fonte: Autora (2021)

Realizado a exclusão da classificação de relevância dos testes cutâneos, a partir da data de realização do teste criou-se as variáveis mês e estação e optou-se por excluir a data por ser considerada um fator de baixa relevância sob o objeto de estudo.

Após a criação das novas variáveis aglutinou-se as partes do corpo atingidas por meio de restrições condicionais (exemplo: se dorso das mãos ou palma das mãos é igual a 1 e antebraços igual a 1 então a nova variável mãos e antebraços é igual a 1 senão 0) e excluiu-se as variáveis originais correspondentes, resultando, dessa forma em 52 variáveis para análise. Como resultados da aglutinação estabeleceu-se apenas os fatores: mãos e antebraços, pés e coxas, face e pescoço e tórax e abdome.

Em relação aos 4 possíveis diagnósticos por trabalhador, foi criado um novo fator

nomeado de total de diagnósticos a partir da contagem de diagnósticos que cada pessoa recebeu. Além da variável de total de diagnósticos, cada uma das 85 moléstias presente nos 4 diagnósticos finais foi convertida com o uso de relações condicionais a uma variável binária. A título de ilustração, se uma pessoa recebeu o diagnóstico final 1 de dermatite alérgica de contato, o diagnóstico 2 de urticária e nenhum outro diagnóstico nos diagnósticos 3 e 4 foi apontado como 1 nas novas colunas dermatite alérgica de contato e urticária e 0 nas outras 83 colunas. Ao término da etapa de transformação a base de dados apresentou 136 variáveis.

Com o intuito de maior aprofundamento nas discussões, escolheu-se por manter para composição da referida pesquisa as profissões com maior incidência na base de dados a saber: Relacionado à Limpeza, Pedreiro/Servente, Atividades Domésticas e Trabalhador de Escritório. Definidos as profissões, a base de dados apresentou 264 observações e foram exclusos os diagnósticos que não apresentaram resultados positivos restando para aplicação das técnicas de *machine learning* 57 variáveis.

Escolhidos os fatores a serem considerados, visando atender o objetivo de utilizar ou não os testes cutâneos para classificar se uma doença é ou não oriunda do trabalho a base de dados foi disposta em dois cenários: o cenário 1 com 27 variáveis sem os resultados da Bateria Padrão Brasileira de Testes de Contato e o cenário 2 com os resultados dos testes cutâneos totalizando 57 fatores a serem explorados.

3.2 MINERAÇÃO

Após a transformação das variáveis foi elaborado a análise descritiva das observações restantes e posteriormente desenvolvido o código para comparação das técnicas de mineração de dados situado no Apêndice A.

Para particionamento do banco de dados, 80% dos registros foram empregadas para composição da base de treino totalizando 211 e os 20% restantes (53 trabalhadores) foram utilizados para teste por ser uma proporção usual em trabalhos científicos. Em relação ao controle do cenário de teste, foi empregado o método de validação cruzada com 10 repetições (10-*fold-cross-validation*), com o uso da acuracidade para escolha do melhor modelo, dessa forma, todas as técnicas passaram pelos mesmos critérios de controle. Além de similares critérios de controle, as técnicas foram também avaliadas pelas métricas constantes no Quadro 3.2 acrescido do índice *Kappa* e *Area Under Curve* (*AUC*).

Quadro 3.2 – Métricas utilizadas para problemas de duas classes

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

Fonte: Adaptado de Callahan e Shah (2017)

As métricas presentes do Quadro 3.2 originam-se da matriz de confusão que avalia a assertividade das técnicas de mineração em relação à classificação original do conjunto de dados conforme Sokolova e Lapalme (2009). Dessa forma, se a técnica de mineração classificou como positiva (classe de interesse) e no conjunto de dados original a classe também era positiva estabelece-se os verdadeiros positivos (*vp*) o que também ocorre para os negativos constituindo assim os verdadeiros negativos (*vn*) (KAKHKI; FREEMAN; MOSHER, 2019). Caso a classe de interesse seja classificada como negativa pela técnica de mineração ter-se-á um falso negativo (*fn*) e se a classe que não era de interesse for classificada como positiva ocorre um falso positivo (*fp*) (GUNS; LIOMA *et al.*, 2012).

Um outro indicador relevante para avaliar a performance preditiva é *AUC* que descreve o quão bem um algoritmo distingue os verdadeiros positivos selecionados aleatoriamente dos verdadeiros negativos, calculando a área existente abaixo da curva. No eixo das abcissas de *AUC* situa-se a especificidade e no das ordenadas a sensitividade, enquanto a curva propriamente dita, é referenciada como *ROC* (*Receiver Operating Characteristic*) (CALLAHAN; SHAH, 2017).

Há ainda, uma outra medida de avaliação da qualidade do ajuste do modelo nomeada de Índice *Kappa*, que avalia a reprodutibilidade entre dois conjuntos de dados. O índice *Kappa*

de acordo com Landis e Koch (1977) indica concordância quase perfeita quando seu valor se encontra na faixa de 0,80 a 1; de 0,60 a 0,79, possui-se uma concordância substantiva; e de 0,40 a 0,59, estabelece-se concordância moderada. Caso a concordância esteja entre 0,20 e 0,39, diz-se que a mesma é leve; de 0 a 0,19 concordância baixa; caso o índice *Kappa* seja menor que zero, estabelece-se que há ausência de concordância.

Tratando-se da escolha dos hiperparâmetros muniu-se da aplicação de *Grid Search* ou pesquisa em grade para determinação do conjunto de parâmetros capazes de proporcionar maior acuracidade à referida técnica de mineração. No Quadro 3.3 está disposto os nomes dos pacotes empregados no *RStudio* para avaliação das técnicas de *machine learning*.

Quadro 3.3 – Pacotes Utilizados

Pacote	Técnica	Finalidade	Fonte
mlbench	Todas	Converter objetos de *ML* para *dataframe*	Dimitriadou (2015)
caret	Todas	Partição do conjunto de dados, preparação e controle do cenário de teste e treino e construção da matriz de confusão	Kuhn (2019)
readxl	Todas	Importação de planilhas	Wickham (2019)
ROCR	Todas	Plotar a curva *ROC*	Sing (2020)
dplyr	Todas	Transformar dados	Package 'dplyr' (2020)
tidyr	Todas	Organizar e resumir dados	Package 'tidyr' (2020)
ggplot2	Todas	Plotar gráfico de barras	Wickham (2020)
pROC	Todas	Realizar predições	Robin (2020)
catboost	CAT	Aplicar a técnica *Catboost*	Gulin (2020)
glmnet	LRG	Aplicar a técnica Regressão Logística	Friedman (2020)
xgboost	XGB	Aplicar a técnica *Extreme Gradient Boosting*	Chen (2021)
fastAdaboost	ADA	Aplicar a técnica *Fast Implementation of Adaboost*	Chatterjee (2016)
randomForest	RF	Aplicar a técnica *Random Forest*	Wiener e Liaw (2018)
nnet	NN	Aplicar a técnica *Neural Network*	Ripley, Venables e Maintainer (2020)

Fonte: Autora (2021)

3.3 INTERPRETAÇÃO DOS RESULTADOS

Para realizar a comparação entre os cenários e entre técnicas fez-se uso dos testes de *t-student*, Tukey e também de Mann-Whitney (WINTER; DODOU, 2010; LEE; LEE, 2018). O teste de *t-student* foi aplicado para determinar se as médias do Cenário 1 e do Cenário 2 para a mesma

técnica são originadas de populações consideravelmente distintas a partir da parametrização dos dados bem como para distintas técnicas no mesmo cenário. Tratando-se do teste de Tukey o mesmo foi utilizado para avaliar a igualdade das médias das técnicas de mineração para o mesmo cenário. O teste de Mann-Whitney, por sua vez, foi empregado para avaliar se os valores resultantes do Cenário 2 distinguiram-se aos encontrados no Cenário 1 por meio do uso da igualdade das medianas utilizando a mesma sistemática do teste de *t-student* (cenários distintos para a mesma técnica e mesmo cenário para técnicas diferentes).

Além de comparar cenários e técnicas, discorreu-se a respeito da importância das variáveis para a técnica CAT em comparação com a que apresentou melhores resultados dentre as cinco restantes. A importância das variáveis foi calculada utilizando o cálculo da impureza de Gini que é realizado com base na redução da soma dos erros quadrados durante a etapa de treino o que requer um cálculo mínimo extra após o treinamento (NEMBRINI; KÖNIG; WRIGHT, 2018).

Ao findar das discussões, discorreu-se sobre as perspectivas de pesquisas futuras tanto em relação à Saúde e Segurança Ocupacional quanto possíveis aplicações em Medicina Ocupacional.

RESULTADOS

4

O objetivo deste Capítulo é apresentar os resultados da pesquisa e caracterizar o conjunto de dados. Sendo assim, dispõem-se na Seção 4.1 a análise descritiva da base de dados, na 4.2 encontra-se a síntese dos resultados obtidos com a implementação de técnicas de mineração e, na Seção 4.3 está disponível discussões no que se refere às pesquisas futuras.

4.1 BASE DE DADOS

As informações referentes ao banco de dados foram coletadas no Centro de Estudos da Saúde do Trabalhador e Ecologia Humana, situado na Escola Nacional de Saúde Pública Sérgio Arouca da Fiocruz entre 2000 e 2014. Os pacientes que compõem o banco de dados foram atendidos primeiramente em postos de saúde predominantemente da rede pública no estado do Rio de Janeiro e encaminhados ao Cesteh para determinação do agente causador da lesão.

4.1.1 Análise Descritiva da Base de Dados Original

Dentre as 616 observações constituintes do banco de dados, 288 dizem respeito a dermatoses ocupacionais (OD), 271 a dermatoses não ocupacionais (NOD) e em 56 indivíduos não foi possível estabelecer se a lesão foi ou não de originada no trabalho, as quais foram caracterizadas como inconclusivas (INC). A população estudada é composta por 249 pessoas do sexo masculino e 366 do sexo feminino. Em relação à atopia, 224 indivíduos relataram ocorrência de lesões em si próprio ou na família previamente ao atendimento, 309 pessoas não apresentaram sintomas antes do atendimento no Cesteh e em 27 casos não foi especificado se a atopia era preexistente.

Na Tabela 4.1 está disposta a quantidade de indivíduos para as variáveis profissão, idade, etnia e escolaridade em função da variável dermatose ocupacional.

Tabela 4.1 - Distribuição da população em função das variáveis dermatose ocupacional, idade, etnia, profissão e escolaridade

Idade	**OD**	**NOD**	**INC**	**Total**	**Profissão**	**OD**	**NOD**	**INC**	**Total**
18 a 27	25	36	5	66	Relacionado à Limpeza	89	48	16	153
28 a 37	67	54	13	134	Pedreiro / Servente	46	8	2	56
38 a 47	96	61	18	175	Outras	34	45	4	83
48 a 57	77	74	17	168	Relacionado a Tintas	16	3	1	20
58 a 67	20	32	2	54	Mecânico / Metalúrgico	14	6	2	22
67 ou mais	4	14	1	19	Cozinheiro	13	8	6	27
Total	289	271	56	616	Atividades Domésticas	11	26	3	40
Etnia					Balconista/ Vendedor	9	19	5	33
Branca	90	106	17	213	Trabalhador da Saúde	9	15	4	28
Não Especificado	36	52	11	99	Costureiro/ Alfaiate	7	12	1	20
Negra	63	48	6	117	Marceneiro / Carpinteiro	7	3	3	13
Parda	100	65	22	187	Manicure	6	3	1	10
Total	289	271	56	616	Agente de Endemias	5	3	3	11
Escolaridade					Químico	5	13	1	19
Não Sabe Ler e Escrever	8	1	1	10	Lanterneiro	3	0	0	3
Alfabetizado	5	5	1	11	Trabalhador de Escritório	3	33	2	38
Fundamental Incompleto	93	55	13	161	Bombeiro Hidráulico	2	2	0	4
Fundamental Completo	41	38	8	87	Cobrador de Ônibus	2	2	1	5
Médio Incompleto	33	9	4	46	Cuidador de Idosos	2	1	0	3
Médio Completo	58	66	18	142	Motorista	2	5	1	8
Superior Incompleto	8	22	0	30	Cabelereiro	1	3	0	4
Superior Completo	8	25	0	33	Padeiro	1	0	0	1
Mestrado	0	1	0	1	Porteiro	1	0	0	1
Doutorado	0	1	0	1	Desempregado	0	2	0	2
Ignorado	35	48	11	94	Estudante	0	9	0	9
Total	289	271	56	616	Não Especificado	0	1	0	1
					Professor /Aux. de Educação	0	1	0	1
					Total	289	271	56	616

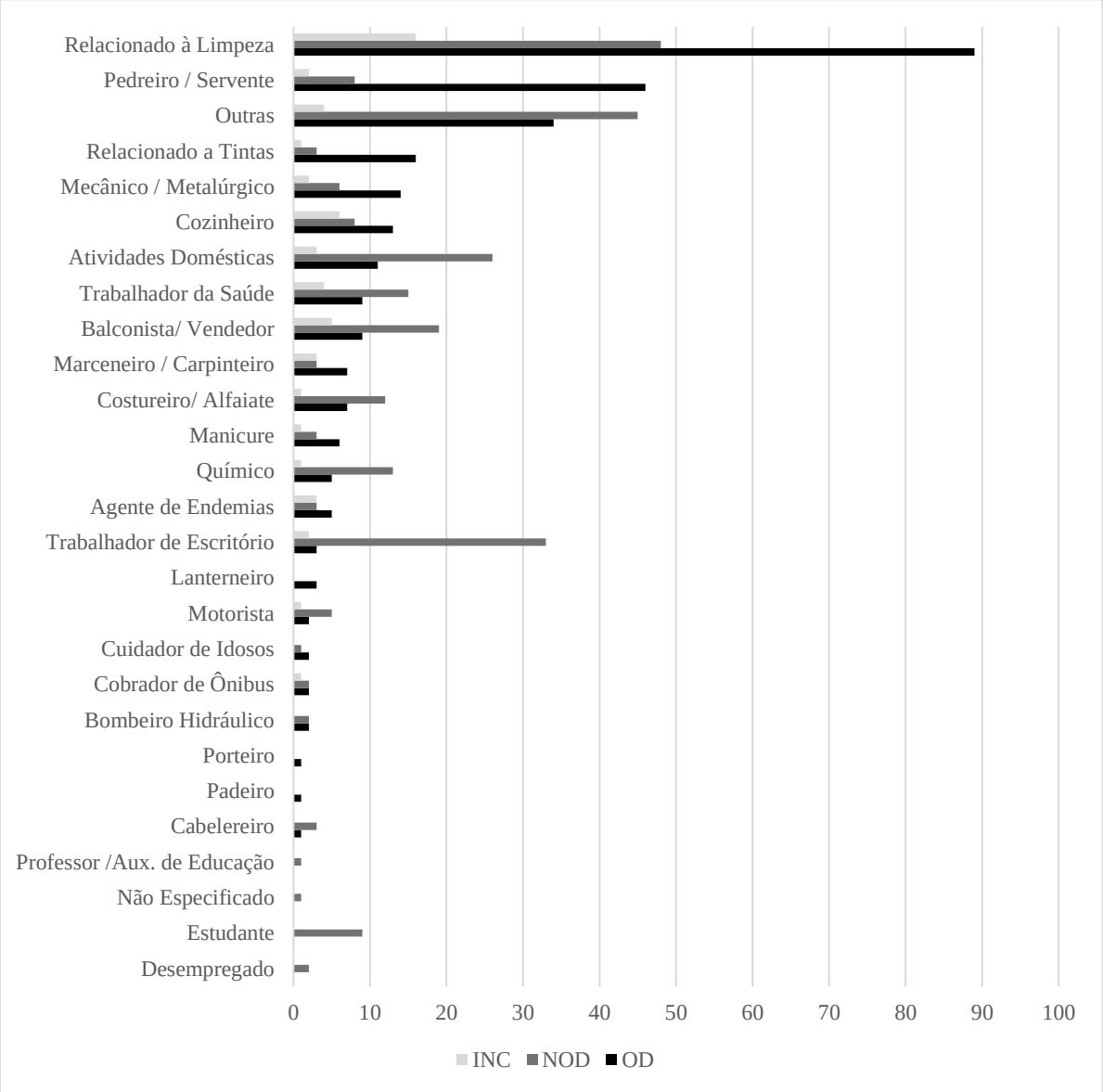
Fonte: Autora (2021)

Como é possível observar na Tabela 4.1, em relação à idade, a classe que apresentou maior ocorrência de dermatoses relacionadas ao trabalho foi a faixa etária compreendida entre os 38 e 47 anos. A respeito da etnia embora o grupo que apresentou no geral maior quantidade de pessoas tenha sido a etnia branca, não foi essa a classe que mais desenvolveu dermatoses ocupacionais, dessa forma, a categoria que mais apresentou dermatoses relacionadas ao trabalho foi a de pardos. Ao contrário do ocorrido com a etnia, em relação à escolaridade, o grupo que mais apresentou a existência na população também foi o que desenvolveu maior

quantidade de dermatoses ocupacionais que foi o das pessoas com ensino fundamental incompleto.

Tratando-se especificamente da categoria profissões, na Figura 4.1 apresenta-se um gráfico de barras que relaciona o tipo de dermatose com a profissão.

Figura 4.1 – Indivíduos por profissão e tipo de dermatose (OD, NOD e INC)



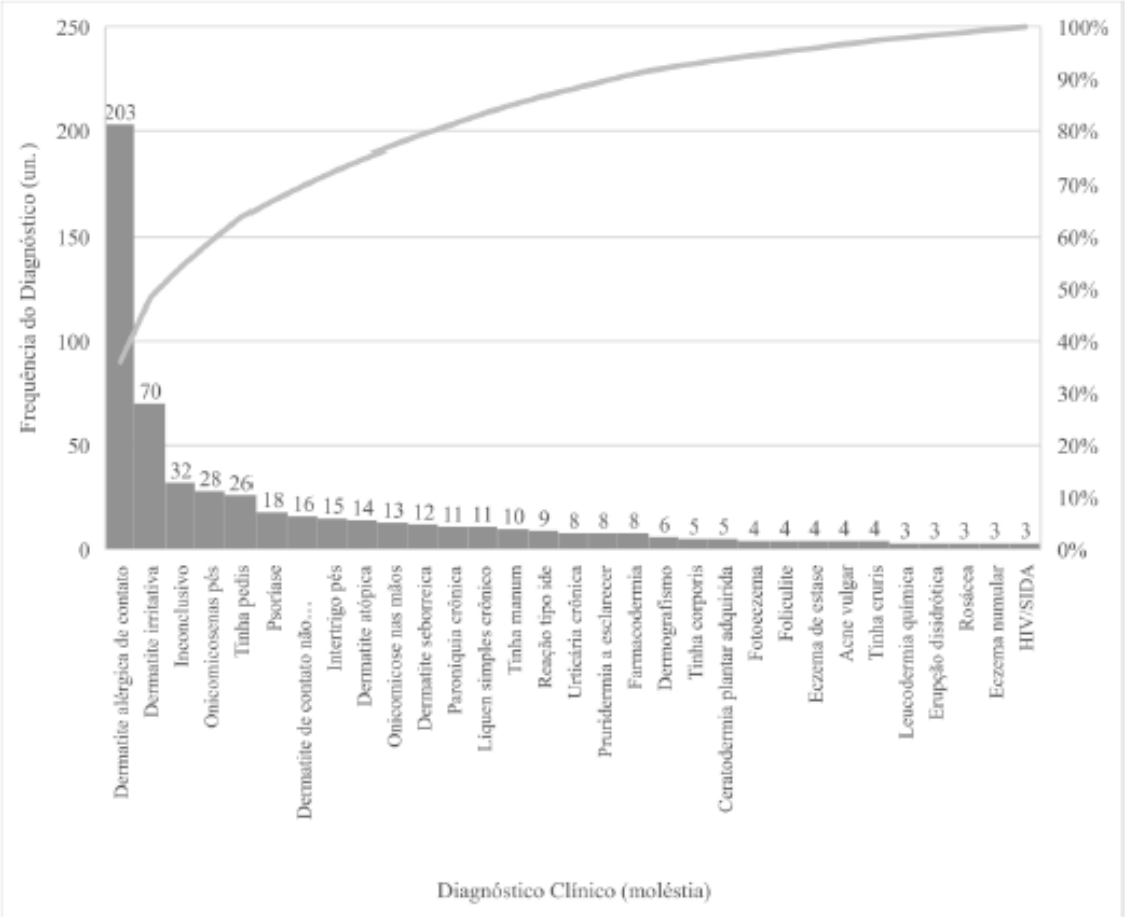
Fonte: Autora (2021)

Ao observar a Figura 4.1 e a Tabela 4.1 é possível estabelecer que a função laborativa que apresentou maior incidência foi o exercício profissional relacionado à limpeza seguido pelas funções de pedreiro/servente. Na maioria das categorias relativas à profissão ocorreu a incidência das 3 classes existentes para a dermatose ocupacional e somente as profissões de

lanterneiro, porteiro e padeiro apresentaram indivíduos exclusivamente com dermatoses ocupacionais.

Com o intuito de melhor compreender o diagnóstico clínico que compôs o montante de quatro variáveis, as informações foram agrupadas e ordenadas segundo sua distribuição de frequência relativa para verificação das moléstias com maior incidência que está alocado na Figura 4.2.

Figura 4.2 – Distribuição de Pareto dos Casos de Diagnóstico Clínico

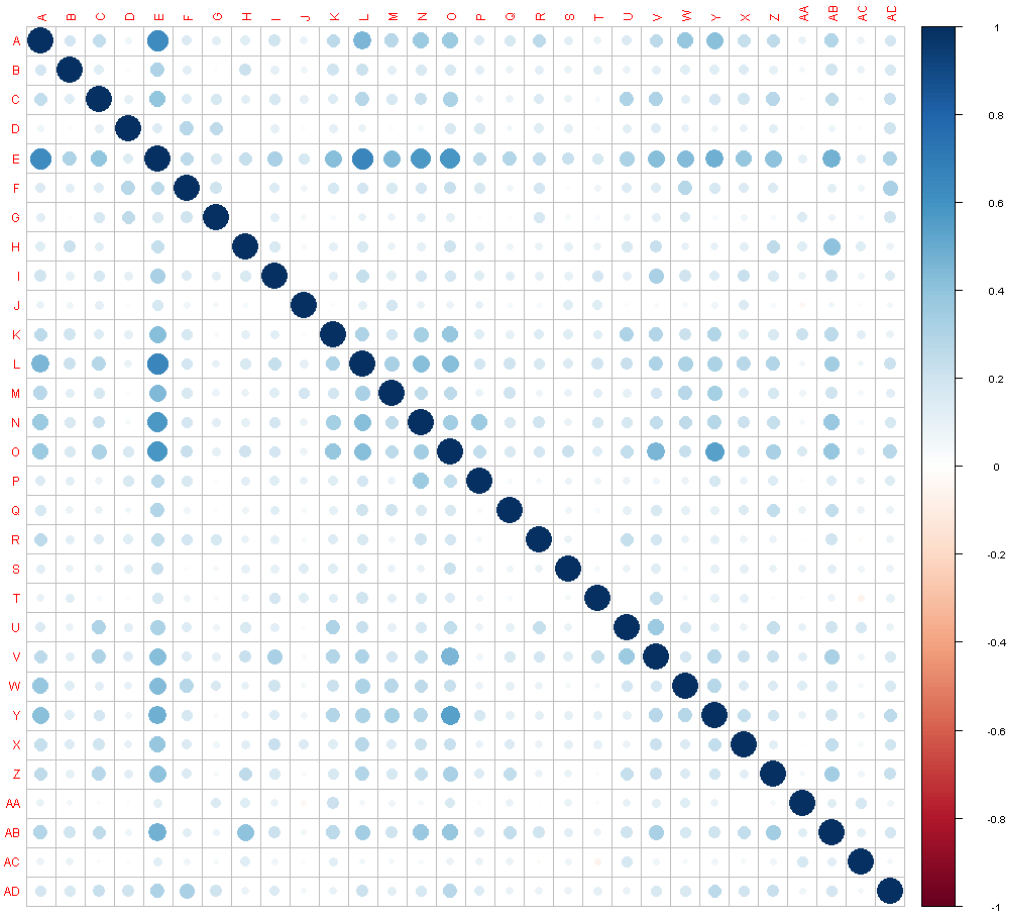


Fonte: Autora (2021)

Com a Figura 4.2 é possível estabelecer que o diagnóstico clínico de maior incidência foi o de dermatite alérgica de contato seguida pela dermatite irritativa. Dentre os 84 tipos de patologias existentes na base de dados somente 15 tipos de lesões compreendem cerca de 80% do quadro clínico apresentado para a população estudada.

Em relação aos testes de contato, realizou-se o teste de correlação entre os compostos que é apresentado na Figura 4.3 e, seus resultados no formato de tabela no Apêndice B.

Figura 4.3 – Gráfico de correlação para a bateria padrão brasileira de testes de contato



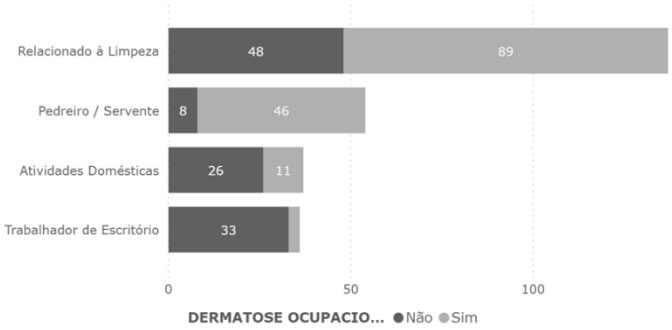
Fonte: Autora (2021)

A partir da Figura 4.3, pode-se estabelecer que a relação entre os compostos da bateria padrão brasileira de testes de contato varia de levemente fraca a moderadamente forte e, em sua grande maioria marcada como positiva. A relação mais forte existente foi entre os compostos irgasan (L) e butilfenol-p-terciário (E) com correlação atribuída de 0,972 seguidos por irgasan (L) e antraquinona (A) cuja correlação equivalente é de 0,241. O irgasan é um antisséptico amplamente utilizado em sabões desinfetantes e desodorantes enquanto a antraquinona é encontrada em corantes e laxativos. O butilfenol-p-terciário, por sua vez, é um componente da borracha que atua como cola em itens de couro. A correlação entre irgasan e butilfenol-p-terciário é relevante porque ao levar em consideração que o grupo de trabalhadores de maior incidência na base de dados é interessante pois de profissionais da limpeza que fazem uso tanto de produtos de higiene como de componentes da borracha nos equipamentos de proteção individual.

4.1.2 Análise Descritiva da Base de Dados Transformada

A base de dados utilizada para a aplicação das técnicas de *machine learning* passou pelas transformações descritas no tópico 3.1 e foi constituída pelas 4 profissões de maior incidência totalizando, portanto, 264 observações. A quantidade total de trabalhadores segundo sua atividade profissional foi de 137 relacionados à limpeza, 54 pedreiro/ servente, 37 atividades domésticas e 36 trabalhadores de escritório. Tais trabalhadores foram classificados segundo a classificação dermatose ocupacional de forma que "Sim" significa desenvolvimento de dermatose no trabalho e "Não" implica em origens distintas do trabalho para a dermatose.

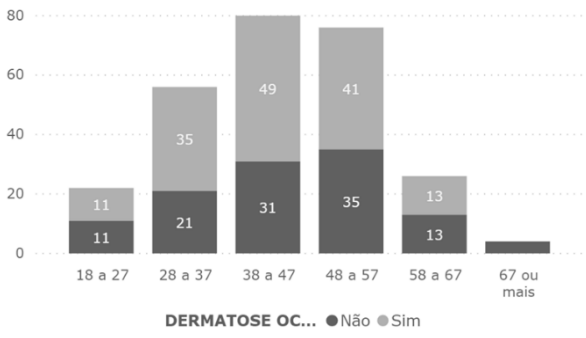
Figura 4.4 – Indivíduos por profissão e dermatose ocupacional



Fonte: Autora (2021)

Considerando a variável idade, foi verificado que 22 (8,33%) trabalhadores constituíram o grupo de 18 a 27 anos, 56 (21,21%) de 28 a 37 anos, 80 (30,30%) com faixa etária compreendida entre os 38 e 57 anos, são 76 (29,68%) as pessoas entre os 48 e 57 anos e 30 (11,36%) com 58 anos ou mais conforme Figura 4.5.

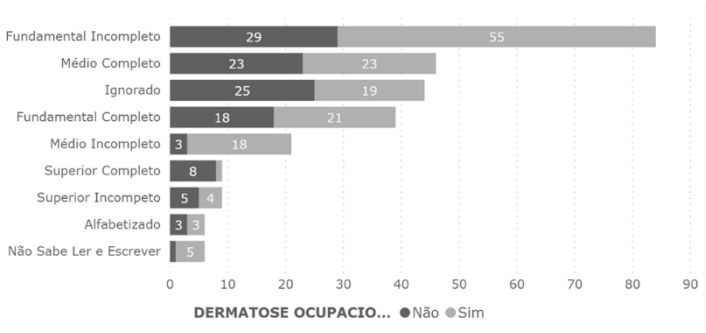
Figura 4.5 – Indivíduos por faixa etária e dermatose ocupacional



Fonte: Autora (2021)

A respeito do nível de escolaridade, foi verificado que 84 (31,81%) dos indivíduos presentes na base de dados possuíam o nível fundamental incompleto, 46 (17,42%) apresentaram nível médio completo e 39 (14,77%) de acordo com a Figura 4.6.

Figura 4.6 – Indivíduos por nível de escolaridade e dermatose ocupacional

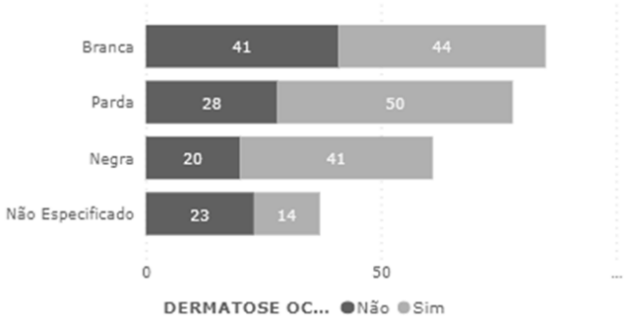


Fonte: Autora (2021)

Ao analisar a Figura 4.4 e a 4.6 é evidente que a desigualdade socioeconômica é uma característica da base de dados em estudo e também reflexo da realidade brasileira. A maior proporção de pessoas acometidas por dermatoses ocupacionais é ocupada por cidadãos que exercem funções que demandam maior esforço físico (pedreiro/servente e relacionado à limpeza) e que indivíduos com nível de escolaridade superior ao ensino médio constituem pequena quantidade de representantes.

Na figura 4.7 está situado a quantidade de trabalhadores segundo sua declaração de etnia. Dentre os 264 indivíduos, 85 (32,20%) se declararam como brancos, 78 (29,54%) pardos, 61 (23,11%) negros e 37 (14,01%) não especificaram a etnia a qual pertencem.

Figura 4.7 – Indivíduos por etnia e dermatose ocupacional



Fonte: Autora (2021)

Tratando da etnia, embora a maior quantidade de pessoas esteja presente na categoria branca a classe com maior incidência de indivíduos com doenças da pele relacionadas ao trabalho foi a de pardos. Em termos de proporção (ex.: 1 ocupacional pra 1 não ocupacional), a etnia composta por pessoas brancas apresentou frequência próxima de 1 pra 1 enquanto nas etnias negra e parda tal proporção foi estabelecida em torno da grandeza de 2 ocupacionais pra 1 não ocupacional. Em relação ao período do ano ao qual a dermatose foi desenvolvida, foi realizado a classificação por tipo de dermatose (do trabalho ou não) por mês que está presente na Tabela 4.2.

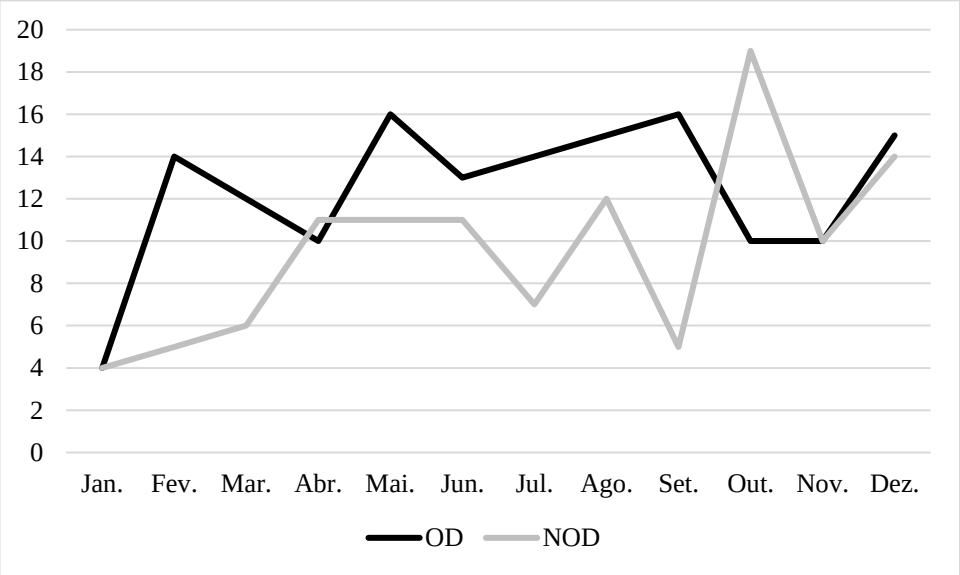
Tabela 4.2 – Dermatose Ocupacional e Não Ocupacional por Mês

Mês	OD	NOD	Total
Jan.	4	4	8
Fev.	14	5	19
Mar.	12	6	18
Abr.	10	11	21
Mai.	16	11	27
Jun.	13	11	24
Jul.	14	7	21
Ago.	15	12	27
Set.	16	5	21
Out.	10	19	29
Nov.	10	10	20
Dez.	15	14	29
Total	149	115	264

Fonte: Autora (2021)

Na Figura 4.8 localizam-se as informações constantes na Tabela 4.2 em forma gráfica.

Figura 4.8 – Dermatose Ocupacional e Não Ocupacional por Mês



Fonte: Autora (2021)

Como é possível observar na Figura 4.8, nos meses de janeiro e novembro a quantidade de pessoas acometidas com dermatose ocupacional foi a mesma da de não ocupacional. Os meses de maior incidência de dermatoses do trabalho foram fevereiro (14), maio (16) e setembro (16) enquanto o pico de dermatoses não ocupacionais foi estabelecido em outubro com 19 casos. O mês de dezembro, por sua vez, apresentou apenas um caso a mais ocupacional do que não ocupacional.

Para melhor compreender a ocorrência do diagnóstico clínico segundo uma perspectiva temporal, na Tabela 4.3 está presente os dois diagnósticos de maior incidência (dermatite irritativa e dermatite alérgica) por mês.

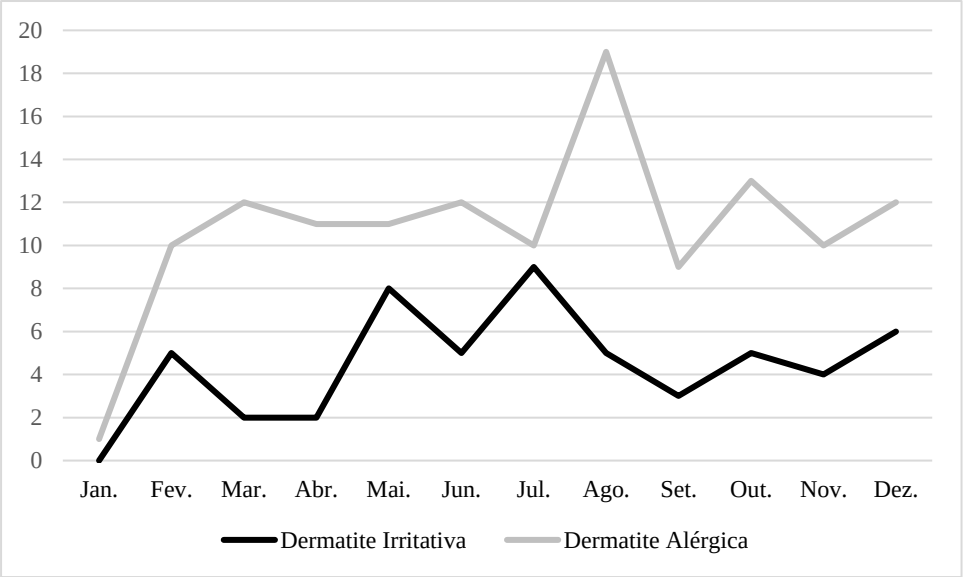
Tabela 4.3 – Dermatite Irritativa e Dermatite Alérgica por Mês

Mês	Dermatite Irritativa	Dermatite Alérgica	Total
Jan.	0	1	1
Fev.	5	10	15
Mar.	2	12	14
Abr.	2	11	13
Mai.	8	11	19
Jun.	5	12	17
Jul.	9	10	19
Ago.	5	19	24
Set.	3	9	12
Out.	5	13	18
Nov.	4	10	14
Dez.	6	12	18
Total	54	130	184

Fonte: Autora (2021)

Na Figura 4.9 está presente em forma gráfica as observações por tipo de dermatite ocupacional por mês.

Figura 4.9 – Dermatite Alérgica e Dermatite Irritativa por Mês



Fonte: Autora (2021)

Ao analisar a Figura 4.9 é possível observar que o comportamento das 2 moléstias é consideravelmente distinto de janeiro a agosto e assemelha-se de setembro a dezembro. A dermatite irritativa apresenta um aumento gradual de janeiro a julho enquanto a alérgica apresenta um crescimento exponencial de janeiro para fevereiro e praticamente se mantém

constante até julho. Em agosto ocorreu um aumento expressivo nos casos de dermatites alérgicas.

4.2 APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS

A presente Seção está disposta em 3 subseções: na subseção 4.2.1 dispõem-se a comparação das técnicas de mineração em estudo para o Cenário 1, na subseção 4.2.2 está localizado o cotejo das mesmas técnicas avaliadas na subseção anterior para o Cenário 2 enquanto na subseção 4.2.3 é avaliado a significância da Bateria Padrão Brasileira de Testes de Contato na determinação de dermatoses ocupacionais.

4.2.1 Comparação das Técnicas apenas com Variáveis Preditoras

Para composição do Cenário 1 (Comparação das Técnicas apenas com Variáveis Preditoras) foram consideradas 26 variáveis preditoras em função da variável dermatose ocupacional que foi convertida para binária. As variáveis preditoras são: mês, estação, dia da semana, escolaridade, atopia, sexo, idade, etnia, profissão, mãos e antebraço, pés e coxas, face e pescoço, tórax e abdome, teste de luvas, síndrome da pele excitada, total de diagnósticos, dermatite alérgica de contato, dermatite irritativa, líquen plano, psoríase, dermatite seborreica, urticária de contato, líquen simples plano, dermatite atópica, eczema numular e eczema de estase.

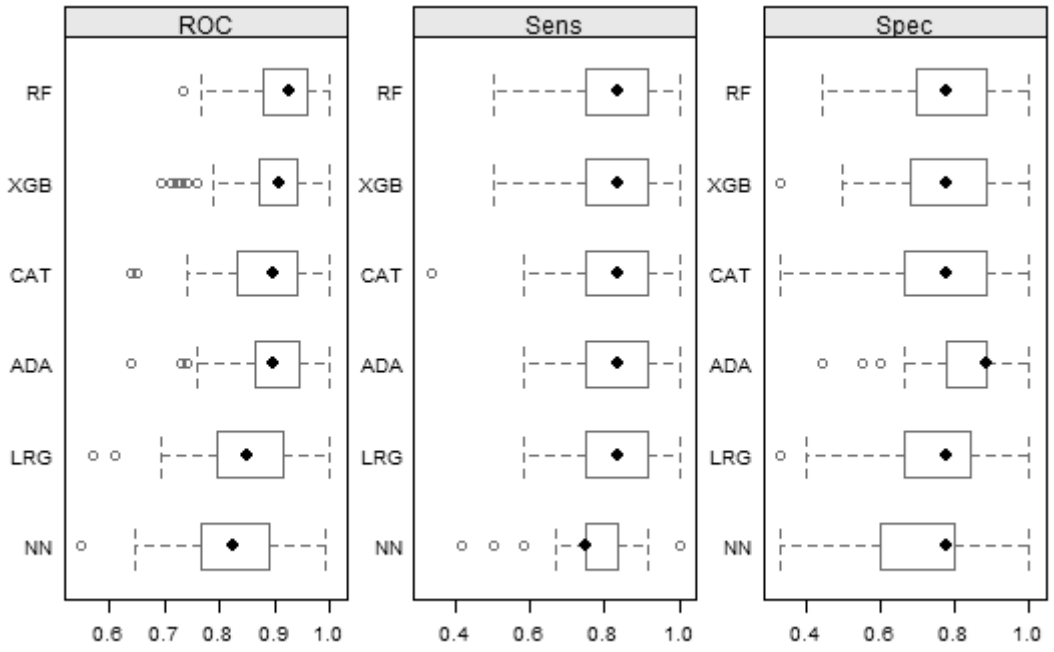
As métricas resultantes das predições estão presentes na Tabela 4.4 encontram-se em negrito os melhores indicadores para cada métrica. Na Tabela C.1 situada no Apêndice C está disposto as medidas de posição para a acuracidade, sensitividade e especificidade. Em relação à acuracidade, o algoritmo que apresentou melhores resultados foi RF com 84,62% em média e desvio padrão de 5,93%, entretanto, a acuracidade média entre todas as técnicas foi estabelecida em 81,73% e as piores acuracidades foram encontradas em CAT e LRG ambas com 78,85%, porém, quando analisa-se os desvios-padrões nota-se que o maior pertence à LRG. Tratando-se da sensitividade que avalia o acerto na classificação da dermatose ser do trabalho em ser positiva, as técnicas CAT e XGB lideraram e estando empatadas com 89,66% em média seguidos por RF, ADA e NN com 86,21%. Em termos da especificidade que avalia a classificação correta dos verdadeiros negativos as técnicas que lideraram o *ranking* foram RF e LRG com 82,61% e a pior foi CAT com 65,22% de acerto. Com o propósito de avaliar a distribuição empírica dos resultados das predições foi plotado gráficos *boxplot* para cada técnica e medidas acuracidade, sensitividade e especificidade situado na Figura 4.10.

Tabela 4.4 – Comparação das Técnicas apenas com Variáveis Preditoras

	RF	XGB	CAT	ADA	LRG	NN
Acuracidade	**0,8462 ± 0,05936**	0,8269 ± 0,0694	0,7885 ± 0,07176	0,8269 ± 0,07621	0,7885 ± 0,08677	0,8269 ± 0,08318
95% CI	**(0,7192; 0,9312)**	(0,6967; 0,9177)	(0,653; 0,8894)	(0,6967; 0,9177)	(0,653; 0,8894)	(0,6967;0,9177)
P-Valor	**9,76E-06**	3,94E-05	0,0004469	3,94E-05	0,0004469	3,94E-05
Hiperparâmetros	mtry = 2	nrounds = 50; max_depth = 3; eta = 0,3; gamma =0; colsample_bytree = 0,6; min_child_weight = 1; subsample = 1	depth = 6; learning_rate = 0,1353353; iterations=100; l2_leaf_reg = 1E-06; rsm = 0,9; border_count=255	nIter = 150; Method = Adaboost.M1	alpha = 1; lambda = 0,03674131	size = 3; decay = 0,1
Kappa	**0,6882**	0,6444	0,5613	0,6476	0,5769	0,6476
Sensitividade	0,8621	**0,8966**	**0,8966**	0,8621	0,7586	0,8621
Especificidade	**0,8261**	0,7391	0,6522	0,7826	**0,8261**	0,7826
Erro	**0,1538**	0,1731	0,2115	0,1731	0,2115	0,1731
Pos Pred Value	**0,8621**	0,8125	0,7647	0,8333	0,8462	0,8333
Neg Pred Value	0,8261	**0,85**	0,8333	0,8182	0,7308	0,8182
Precisão	**0,8621**	0,8125	0,7647	0,8333	0,8462	0,8333
Recall	0,8621	**0,8966**	**0,8966**	0,8621	0,7586	0,8621
F_1 *Score*	**0,8621**	0,8525	0,8254	0,8475	0,8	0,8475
Prevalência	0,5577	0,5577	0,5577	0,5577	0,5577	0,5577
Taxa de Detecção	0,4808	**0,5**	**0,5**	0,4808	0,4231	0,4808
Detecção da Prevalência	0,5577	0,6154	**0,6538**	0,5769	0,5	0,5769
Acuracidade Balanceada	**0,8441**	0,8178	0,7744	0,8223	0,7924	0,8223
AUC	**0,9213**	0,9175	0,9175	0,9115	0,8741	0,91

Fonte: Autora (2021)

Figura 4.10 – *Boxplot* dos resultados de acuracidade (ROC), sensitividade (Sens) e especificidade (Spec) para as técnicas no Cenário 1

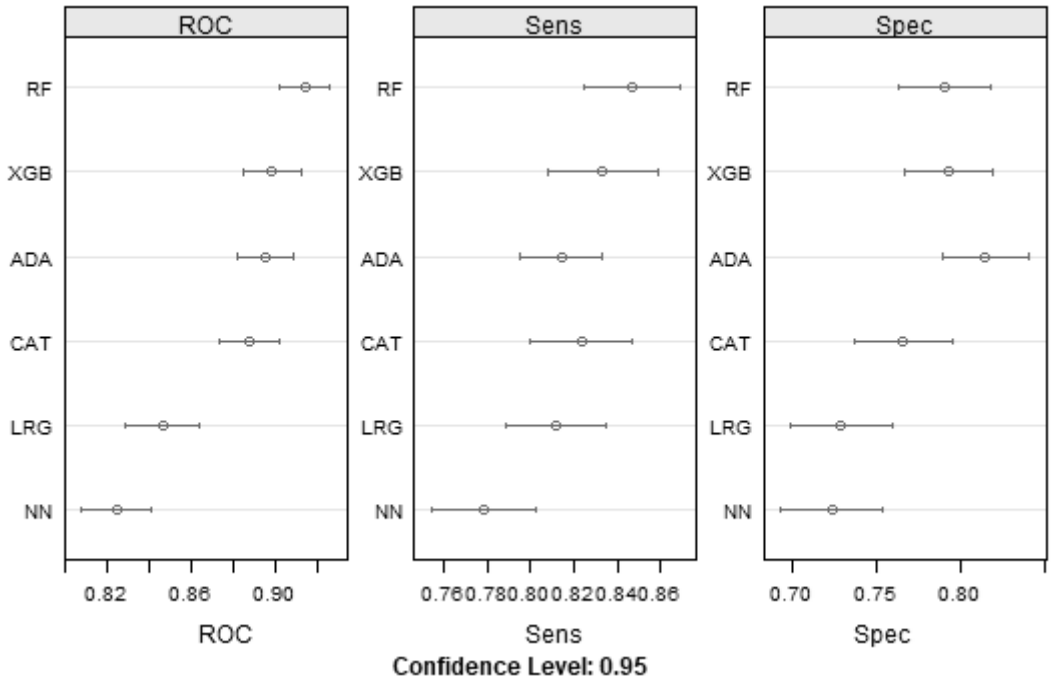


Fonte: Autora (2021)

A respeito da Figura 4.10, é possível estabelecer que a acuracidade apresentou *outliers* à esquerda para todas as técnicas e que a mediana permaneceu entre 80 e 93%. Em relação à sensitividade e a especificidade a amplitude das médias para todas as técnicas foi de 51,4% e 63% respectivamente, mostrando-se, portanto, superiores do que em relação à acuracidade cuja amplitude foi estabelecida em 36,1%.

Tratando-se da avaliação de p-valor situado na Tabela 4.4, em todas as técnicas o p-valor apresentado foi inferior à 0,05 que é um indicativo de que é possível sob determinado nível de confiabilidade rejeitar a hipótese nula quando a mesma estiver presente classificando uma nova observação corretamente quanto oriunda ou não do trabalho. No que diz respeito aos intervalos de confiança, os mesmos foram calculados para acuracidade, sensitividade e especificidade seguindo o método padrão de distribuição binomial com 95% para o nível de confiança conforme apresentado na Figura 4.11.

Figura 4.11 – Intervalos de confiança para acuracidade (*ROC*), sensitividade (Sens) e especificidade (Spec) no Cenário 1



Fonte: Autora (2021)

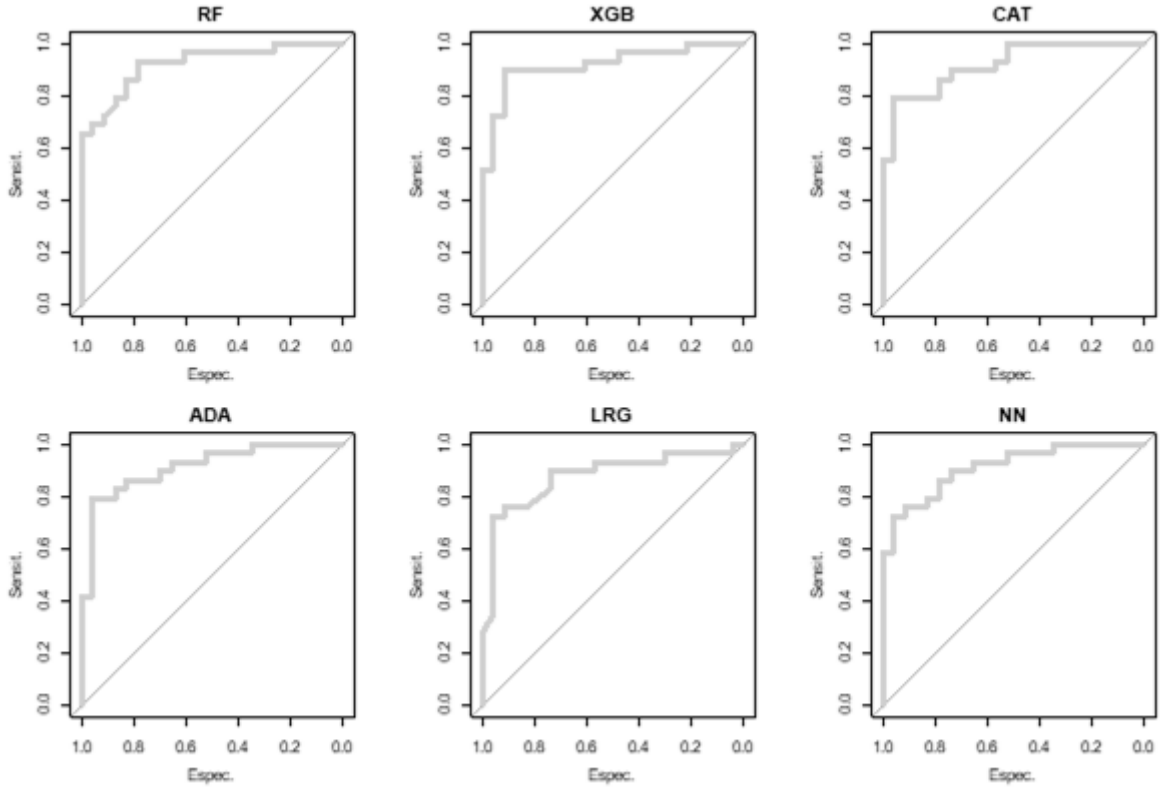
No que se refere aos hiperparâmetros os mesmos foram calculados conforme mencionado no Capítulo 3 por meio de *Grid Search* visando obter a maior acuracidade. O índice *Kappa*, por sua vez, mostrou-se com concordância substantiva para as técnicas RF, XGB, ADA e NN enquanto CAT e LRG obtiveram concordância moderada.

A precisão avalia a assertividade dos valores classificados como verdadeiros positivos visando fornecer uma métrica para a seguinte pergunta: qual o percentual de acerto dos classificados como verdadeiros positivos que realmente eram verdadeiros positivos? Dessa forma, foi verificado que 5 dos 6 algoritmos considerados tiveram precisão superior a 80% com destaque de RF cuja precisão correspondente foi 86,21%. O *recall* avalia a frequência com que o classificador classifica os dados da classe de interesse e, em relação a essa métrica destacaram-se as técnicas CAT e XGB com 89,66%.

O F_1 *Score* é uma combinação de precisão e *recall* propiciador da qualidade geral do modelo. Segundo tal métrica, o melhor avaliador da população em estudo é RF seguido por XGB. A precisão, por sua vez, compara os elementos considerados como verdadeiros positivos com os resultados de toda a população que compõe a matriz de confusão e, nesse caso as técnicas com maiores índices de detecção foram XGB e CAT. Em relação à prevalência que quantifica a superioridade da classe de interesse sobre o efeito indesejado o algoritmo CAT apresentou a melhor métrica com 65,38%.

A métrica *AUC* foi calculada para determinar o algoritmo que melhor distingue os verdadeiros positivos dos verdadeiros negativos e, nesse caso a técnica que realiza tal atividade com desempenho superior é RF conforme disponibilizado na Figura 4.12.

Figura 4.12 – Curva ROC Cenário 1



Fonte: Autora (2021)

Munidos da Figura 4.12 e da Tabela 4.4, verifica-se que de forma geral as técnicas conseguem distinguir de forma satisfatória em relação à *AUC* os verdadeiros positivos dos verdadeiros negativos de forma que somente a regressão logística ficou com *AUC* abaixo de 0,90. Empatados na segunda posição situou-se CAT e XGB com 0,9175 e o algoritmo de maior destaque foi RF com 0,9213.

O teste de Tukey foi utilizado para realizar a comparação das médias uma vez que o mesmo é baseado na diferença mínima significativa ao considerar os percentis do grupo. Dessa forma, na Tabela 4.5 situa-se o resultado da aplicação do teste de Tukey comparando, portanto, os pares de técnicas em termos de seus limites superior, inferior, centro e p-valor.

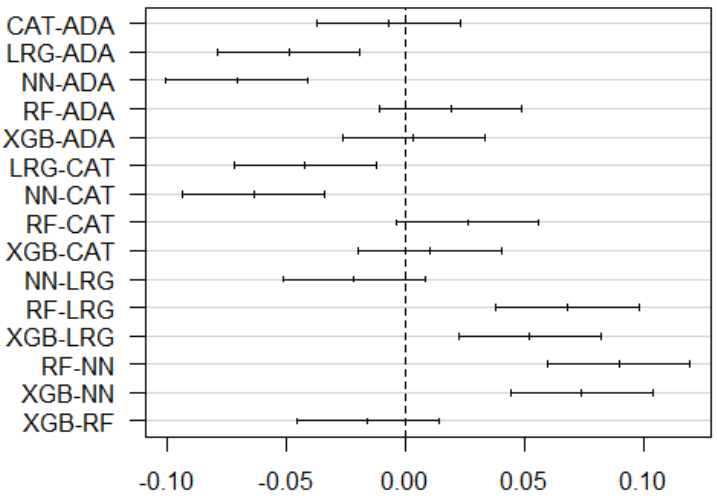
Tabela 4.5 – Teste de Tukey entre Técnicas no Cenário 1

	Centro	Limite Inferior	Limite Superior	P-valor
CAT-ADA	-0,006972	-0,036809	0,022864	0,985334
LRG-ADA	-0,048838	-0,078674	-0,019002	0,000052
NN-ADA	-0,070463	-0,100299	-0,040627	0,000000
RF-ADA	0,019116	-0,010721	0,048952	0,446039
XGB-ADA	0,003370	-0,026466	0,033207	0,999533
LRG-CAT	-0,041866	-0,071702	-0,012029	0,000957
NN-CAT	-0,063491	-0,093327	-0,033654	0,000000
RF-CAT	0,026088	-0,003748	0,055924	0,125727
XGB-CAT	0,010343	-0,019494	0,040179	0,920768
NN-LRG	-0,021625	-0,051461	0,008211	0,303350
RF-LRG	0,067954	0,038117	0,097790	0,000000
XGB-LRG	0,052208	0,022372	0,082045	0,000011
RF-NN	0,089579	0,059742	0,119415	0,000000
XGB-NN	0,073833	0,043997	0,103670	0,000000
XGB-RF	-0,015745	-0,045582	0,014091	0,658758

Fonte: Autora (2021)

Como é possível estabelecer na Tabela 4.5, 7 dos 15 pares de técnicas avaliados não se mostraram estatisticamente distintos. Os pares de técnicas que não apresentaram diferenças significativas foram: CAT e ADA, RF e ADA, XGB e ADA, RF e CAT, XGB e CAT, NN e LRG e XGB e RF. Os demais pares de técnicas se mostraram, portanto, estatisticamente distintos. Na Figura 4.13 está presente a representação gráfica do Teste de Tukey no Cenário 1.

Figura 4.13 – Comparação entre Médias Cenário 1



Fonte: Autora (2021)

Ao analisar a Figura 4.13 estabelece-se que as técnicas LRG e ADA, NN e ADA, LRG e CAT, NN e CAT, RF e LRG, XGB e LRG, RF e NN e XGB e NN na determinação de seus intervalos de confiança não incluem o valor zero que é um indicativo de diferença estatística significante. Tal afirmação também foi constatada na Tabela 4.5.

Para contrapor os resultados produzidos durante a fase de treinamento, realizou-se o teste de Mann-Whitney e o teste de *t-student* comparando técnica a técnica no Cenário 1. A comparação para o teste de Mann-Whitney situa-se na Tabela 4.6 enquanto o teste de *t-student* está na Tabela 4.7.

Tabela 4.6 – Teste de Mann-Whitney entre Técnicas no Cenário 1

Técnicas	**p-valor**
NN e RF	**1,218E-13**
NN e XGB	**3,093E-10**
NN e ADA	**1,839E-09**
LRG e RF	**1,189E-08**
NN e CAT	**5,839E-08**
LRG e XGB	**9,797E-06**
LRG e ADA	**3,741E-05**
LRG e CAT	**0,0107713**
RF e CAT	**0,0107713**
ADA e RF	**0,0363251**
LRG e NN	0,0501022
RF e XGB	0,1033975
CAT e XGB	0,3890849
ADA e XGB	0,6120964
ADA e CAT	0,6222873

Fonte: Autora (2021)

Ao observar a Tabela 4.6 é possível estabelecer que em apenas 5 das 15 comparações realizadas não apresentaram diferenças significativas nos resultados produzidos durante a fase de testes.

Tabela 4.7 – Teste de *t-student* entre Técnicas no Cenário 1

Técnicas	p-valor
NN e RF	**2,025E-08**
NN e XGB	**4,143E-06**
NN e ADA	**9,028E-06**
LRG e RF	**2,202E-05**
NN e CAT	**6,995E-05**
LRG e XGB	**0,00121**
LRG e ADA	**0,0022474**
LRG e CAT	0,0865814
RF e CAT	0,0865814
LRG e NN	0,1794616
ADA e RF	0,2023806
RF e XGB	0,2993311
CAT e XGB	0,5025657
ADA e CAT	0,6500105
ADA e XGB	0,8229618

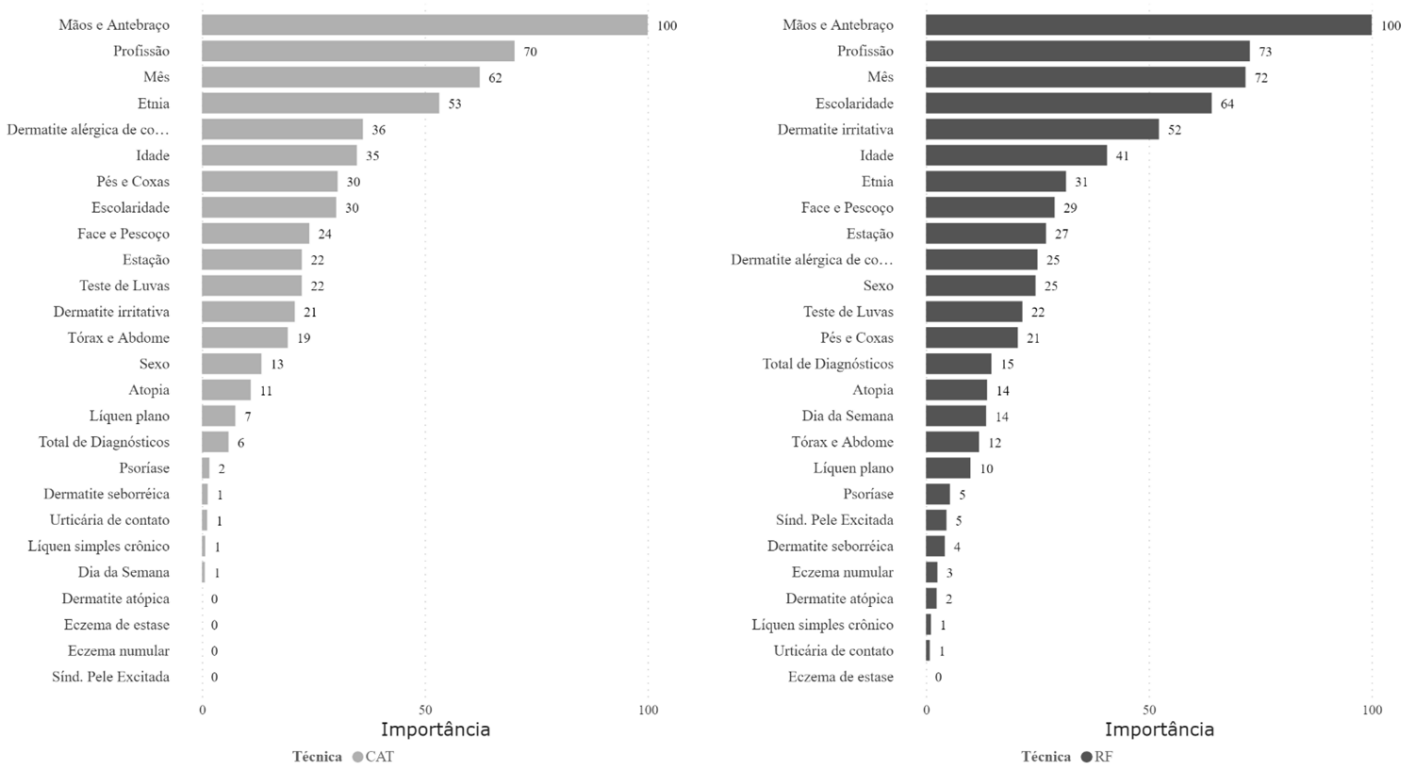
Fonte: Autora (2021)

A partir da análise das Tabelas 4.5, 4.6 e 4.7 é possível afirmar que os resultados fornecidos pelos pares de algoritmos NN e RF, NN e XGB, NN e ADA, NN e CAT, LRG e XGB, LRG e RF e LRG e ADA mostraram-se significantemente distintos nos testes de Tukey, de Mann-Whitney e no de *t-student*. Não se mostraram estatisticamente distintos nos 3 testes realizados os pares de técnicas ADA e CAT, ADA e XGB, CAT e XGB, LRG e NN e RF e XGB era de se esperar que os três primeiros pares os resultados alcançados fossem similares pois ambas constituem técnicas de *boosting*. Distinguiram-se no teste de Mann-Whitney e assemelharam-se no de *t-student* os pares de algoritmos LRG e CAT, RF e CAT e RF e ADA sendo que os 2 últimos também se mostraram distintos em Tukey. Tais resultados indicam que após parametrizar populações obtidas de métodos não parametrizados as observações de ambas as populações se tornam mais próximas. Tal consideração é importante ao tratar os termos práticos do uso de *machine learning* por conta do custo computacional que tende a ser maior em métodos não paramétricos do que em métodos parametrizados o que na presente dissertação não representa um problema a ser resolvido uma vez que o custo computacional não está sendo avaliado.

A partir da perspectiva das métricas, sobressaíram-se as técnicas RF, XGB e CAT. Como um dos objetivos específicos da presente dissertação é comparar a técnica que se mostrar superior dentre as 5 após a exclusão de *Catboost*, conclui-se que para os mesmos critérios de

controle do cenário de teste, ferramenta, base de dados e métricas a técnica de maior destaque é *Random Forest*. Sendo assim, está disposto no Apêndice D a importância das variáveis no Cenário 1 em forma de tabela na Tabela D.1 ordenados segundo a importância de *Catboost*. Na Figura 4.14 está exposto as variáveis com maior relevância na determinação de dermatose oriunda do trabalho para a técnica *Catboost* e *Random Forest* no Cenário 1.

Figura 4.14 – Importância das Variáveis entre CAT e RF no Cenário 1



Fonte: Autora (2021)

Ao fazer uso da Figura 4.14 é possível observar que os 3 fatores de maior influência são os mesmos para ambas as técnicas: Mãos e Antebraços, Profissão e Mês. RF atribuiu peso em 25 das 26 variáveis existentes enquanto CAT para determinar os fatores influenciadores não utilizou síndrome da pele excitada, eczema numular, eczema de estase e dermatite atópica. Em relação às diferenças de fatores que apresentaram peso em ambas as técnicas, RF atribuiu peso superior a 60% para escolaridade enquanto para CAT tal fator se estabeleceu em 30%, além disso, para CAT a etnia é um fator com determinação superior a 50% ao passo que para RF a mesma é inferior a 40%. Ainda no que diz respeito à diferença entre os dois algoritmos, para RF o diagnóstico Dermatite Irritativa é mais determinante que os demais diagnósticos enquanto para CAT o diagnóstico de maior influência é a Dermatite Alérgica de Contato.

4.2.2 Comparação das Técnicas com Variáveis Preditoras e Bateria de Testes

Para o segundo cenário (Comparação das Técnicas com Variáveis Preditoras e Bateria de Testes) fez-se uso das mesmas variáveis constantes no cenário 1 acrescido dos 30 compostos pertencentes à bateria padrão brasileira de testes de contato. Para o Cenário 2 as técnicas foram avaliadas por métricas de desempenho similares ao Cenário 1 e na Tabela 4.8 em negrito estão destacados os melhores resultados para cada métrica.

Ao analisar a acuracidade é possível estabelecer que a técnica com a métrica mais próxima de 1 é RF com 86,54% de média e desvio padrão de 6,285%. Ocupando a segunda posição ainda em relação à acuracidade situam-se empatadas em relação à média as técnicas CAT e ADA com 84,62% sendo que CAT apresentou desvio de 6,525% e ADA de 7,621% enquanto a pior técnica para o cenário em questão com 73,08% foi a NN. Em relação ao percentual de dermatoses ocupacionais classificados corretamente como dermatoses do trabalho (sensitividade) as técnicas de maior destaque que também ficaram empatadas foram RF e ADA com 93,1% seguidos por CAT e XGB ambas com 89,66%. Similarmente ao ocorrido com a sensitividade a especificidade ficou empatada para RF e CAT com 78,26% tendo como técnica de pior resultado NN com 65,22%.

No Apêndice C mais precisamente na Tabela C.2 está presente as medidas de posição para a acuracidade, sensitividade e especificidade.

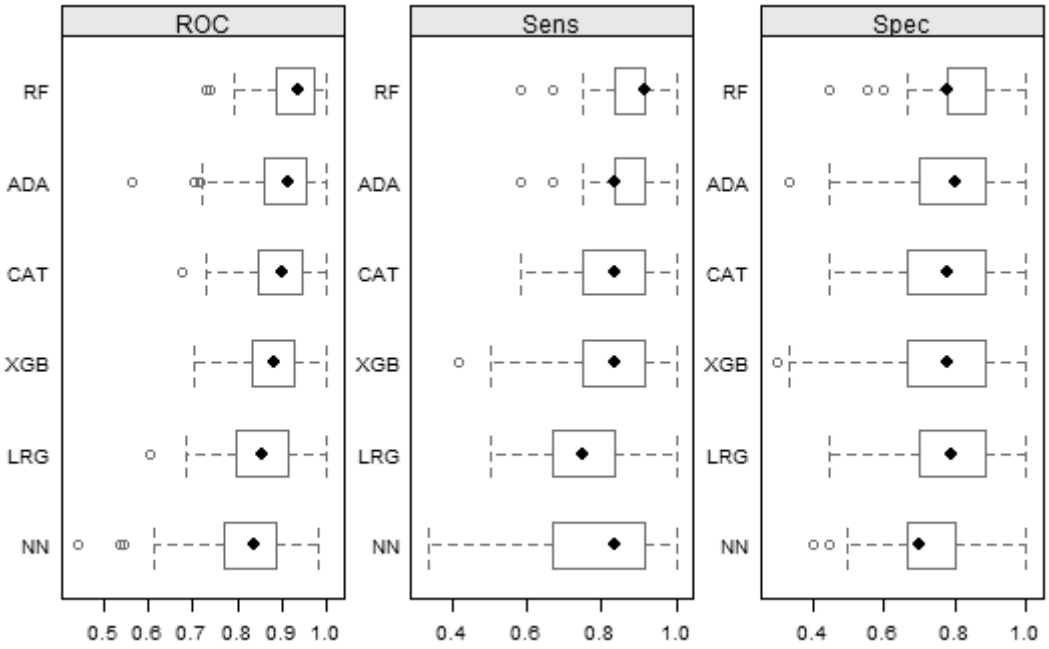
Tabela 4.8 – Comparação das Técnicas com Variáveis Preditoras e Bateria de Testes

	RF	CAT	ADA	XGB	LRG	NN
Acuracidade	**0,8654 ± 0,06285**	0,8462 ± 0,06525	0,8462 ± 0,07621	0,8077 ± 0,06328	0,7885 ± 0,07582	0,7308 ± 0,1001
95% CI	**(0,7421; 0,9441)**	(0,7192, 0,9312)	(0,7192; 0,9312)	(0,6747; 0,9037)	(0,653; 0,8894)	(0,5898; 0,8443)
P-Valor	**2,11E-06**	0,000009755	9,76E-06	1,41E-04	0,0004469	7,78E-03
Hiperparâmetros	mtry=29	depth=4; learning_rate= 0,04978707; iterations= 100; l2_leaf_reg=1e-06; rsm = 0,9; border_count=255	nIter = 150; Method = Adaboost.M1	nrounds = 50; max_depth = 1; eta = 0,3; gamma =0; colsample_bytree = 0,8; min_child_weight = 1; subsample = 1	alpha = 1; lambda = 0,03674131	size = 1; decay = 0,1
Kappa	**0,7234**	0,6853	0,6824	0,6031	0,5693	0,4493
Sensitividade	**0,931**	0,8966	**0,931**	0,8966	0,8276	0,7931
Especificidade	**0,7826**	**0,7826**	0,7391	0,6957	0,7391	0,6522
Erro	**0,1346**	0,1538	0,1538	0,1923	0,2115	0,2692
Pos Pred Value	**0,8438**	0,8387	0,8182	0,7879	0,8	0,7419
Neg Pred Value	**0,9**	0,8571	0,8947	0,8421	0,7727	0,7143
Precisão	**0,8438**	0,8387	0,8182	0,7879	0,8	0,7419
Recall	**0,931**	0,8966	**0,931**	0,8966	0,8276	0,7931
F_1 *Score*	**0,8852**	0,8667	0,871	0,8387	0,8136	0,7667
Prevalência	0,5577	0,5577	0,5577	0,5577	0,5577	0,5577
Taxa de Detecção	**0,5192**	0,5	**0,5192**	0,5	0,4615	0,4423
Detecção da Prevalência	0,6154	0,5962	**0,6346**	**0,6346**	0,5769	0,5962
Acuracidade Balanceada	**0,8568**	0,8396	0,8351	0,7961	0,7834	0,7226
AUC	**0,9273**	0,9265	0,9265	0,919	0,8568	0,8426

Fonte: Autora (2021)

Na Figura 4.15 está localizado o *boxplot* referente aos resultados encontrados no cenário 2 para a acuracidade, sensitividade e especificidade em relação à distribuição dos resultados apresentados nas predições.

Figura 4.15 – *Boxplot* dos resultados de acuracidade (*ROC*), sensitividade (Sens) e especificidade (Spec) para as técnicas no Cenário 2

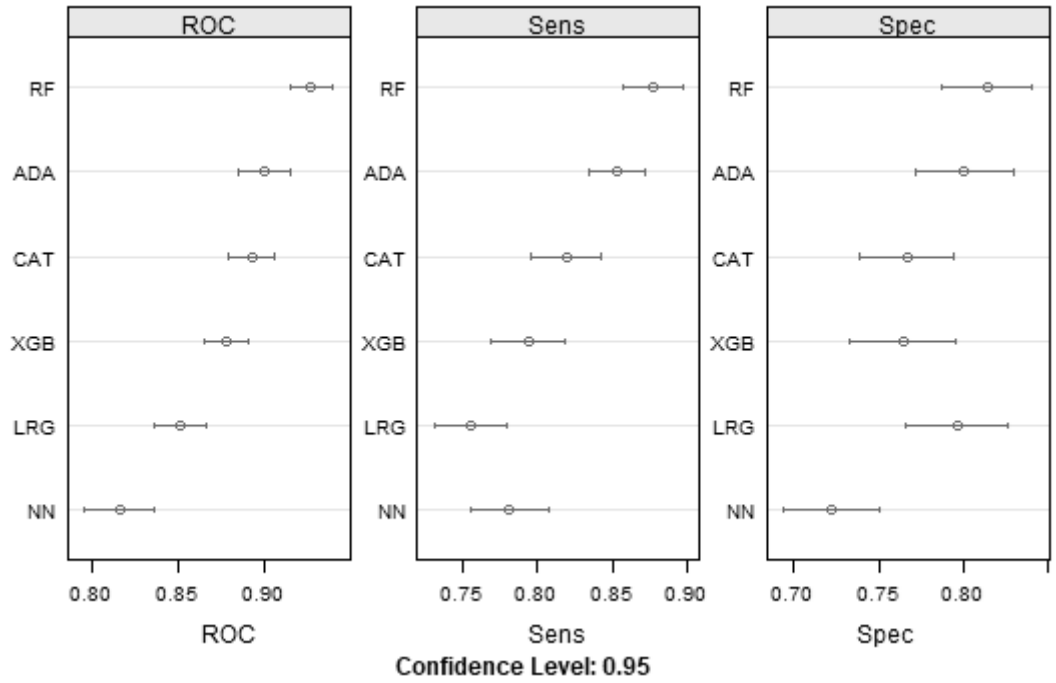


Fonte: Autora (2021)

Com a Figura 4.15 estabelece-se que no que diz respeito à acuracidade, o algoritmo com a mediana mais próxima de 1 foi RF enquanto o único que não apresentou *outliers* foi XGB. Dentre as 3 métricas avaliadas a técnica que se mostrou com maior esparsidade foi NN variando de 40 a 100% na distribuição. Em relação à sensitividade, 5 dos 6 algoritmos avaliados apresentaram mediana superior à 80% com grande variação comportamental. No que se refere à especificidade a variação foi mantida e somente LRG e CAT se isentaram da presença de pontos fora da curva.

No que se refere ao p-valor localizado na Tabela 4.8, as técnicas em estudo mostraram-se com significância estatística permanecendo inferior à 0,05 e os intervalos de confiança estimados a 95% de confiabilidade dispõem-se na Figura 4.16. Em relação à concordância entre conjuntos avaliada pelo índica *Kappa*, as técnicas RF, CAT, ADA e XGB apresentaram concordância substantiva enquanto LRG e NN resultaram em concordância moderada.

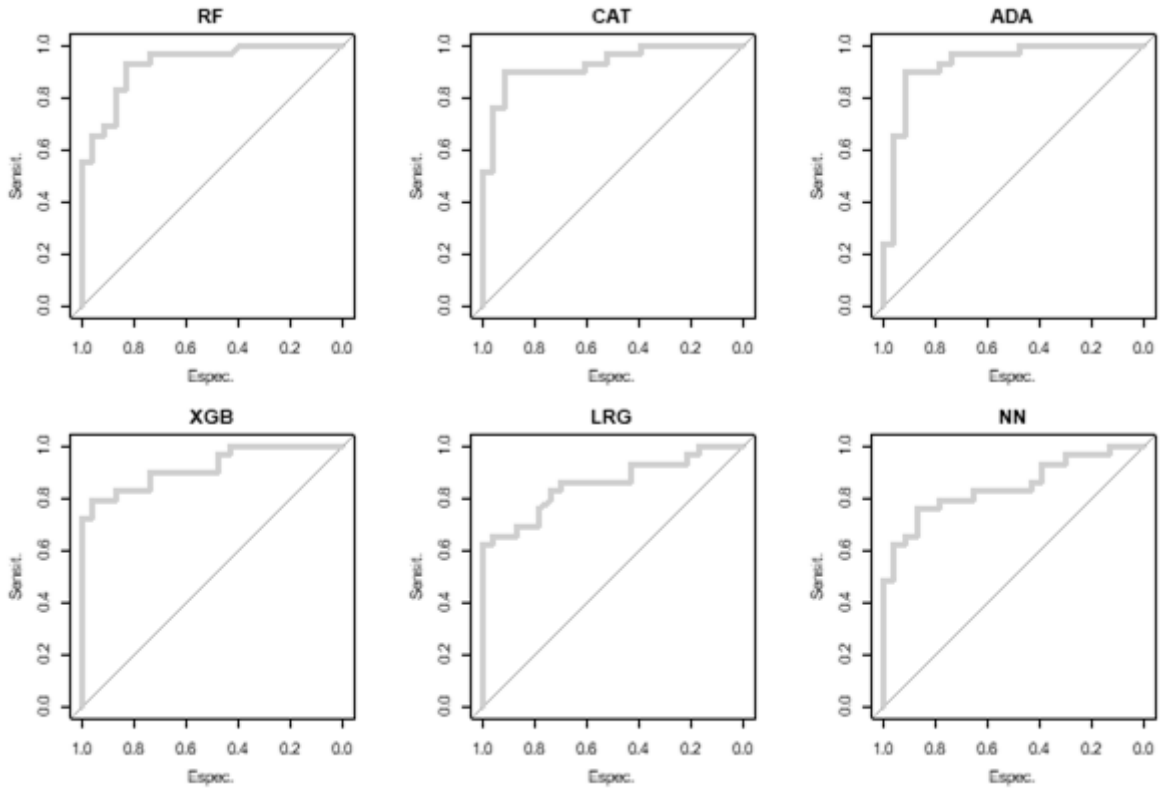
Figura 4.16 – Intervalos de confiança para acuracidade (*ROC*), sensitividade (Sens) e especificidade (Spec) no Cenário 2



Fonte: Autora (2021)

Em relação à AUC na Figura 4.17 está disposto a Curva *ROC* para cada técnica. RF destacou-se em primeiro lugar (0,9273), seguido por CAT e NN (0,9265).

Figura 4.17 – Curva *ROC* Cenário 2



Fonte: Autora (2021)

Analogamente ao ocorrido no primeiro cenário, foi realizado o Teste de Tukey para o Cenário de número 2 cujos resultados estão presentes na Tabela 4.9.

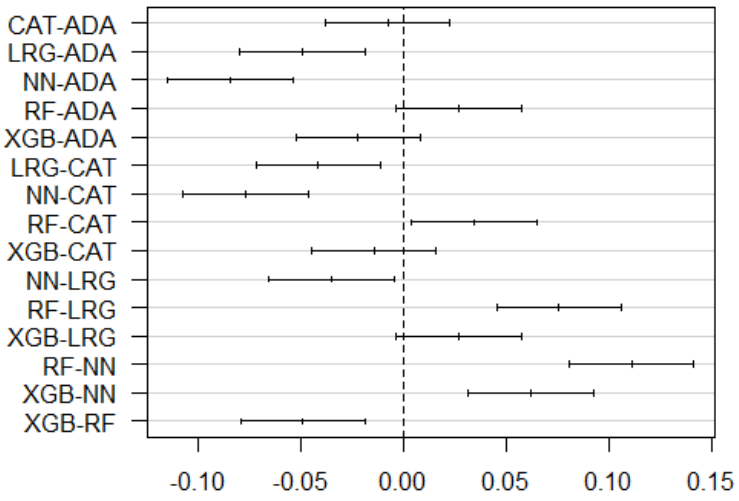
Tabela 4.9 – Teste de Tukey entre Técnicas no Cenário 2

	Centro	**Limite Inferior**	**Limite Superior**	**P-valor**
CAT-ADA	-0,007620	-0,038122	0,022881	0,980210
LRG-ADA	-0,049019	-0,079520	-0,018517	0,000077
NN-ADA	-0,084213	-0,114715	-0,053711	0,000000
RF-ADA	0,026856	-0,003645	0,057358	0,120648
XGB-ADA	-0,021986	-0,052488	0,008515	0,309473
LRG-CAT	-0,041398	-0,071900	-0,010897	0,001612
NN-CAT	-0,076593	-0,107094	-0,046091	0,000000
RF-CAT	0,034477	0,003975	0,064978	0,016315
XGB-CAT	-0,014366	-0,044867	0,016136	0,758755
NN-LRG	-0,035194	-0,065696	-0,004693	0,013106
RF-LRG	0,075875	0,045373	0,106377	0,000000
XGB-LRG	0,027032	-0,003469	0,057534	0,116015
RF-NN	0,111069	0,080568	0,141571	0,000000
XGB-NN	0,062227	0,031725	0,092728	0,000000
XGB-RF	-0,048843	-0,079344	-0,018341	0,000083

Fonte: Autora (2021)

No segundo cenário no Teste de Tukey 5 dos 15 pares de técnicas analisados apresentaram p-valor superior a 0,05 mostrando, portanto, que não há diferença estatística entre as médias que são: ADA e CAT, ADA e RF, ADA e XGB, CAT e XGB e LRG e XGB. Na Figura 4.18 está localizado que representa os resultados obtidos no Teste de Tukey no Cenário de número 2.

Figura 4.18 – Comparação entre Médias Cenário 2



Fonte: Autora (2021)

Similarmente ao ocorrido no primeiro cenário, foram realizados os testes de Mann-Whitney (Tabela 4.10) e de *t-student* (Tabela 4.11) nos resultados obtidos no cenário 2.

Tabela 4.10 – Teste de Mann-Whitney entre Técnicas no Cenário 2

Técnicas	p-valor
NN e RF	**8,706E-17**
LRG e RF	**1,763E-12**
NN e ADA	**3,263E-10**
NN e CAT	**1,782E-08**
RF e XGB	**8,077E-08**
LRG e ADA	**3,105E-06**
NN e XGB	**5,811E-06**
LRG e CAT	**6,682E-05**
RF e CAT	**6,682E-05**
ADA e XGB	**0,0056321**
ADA e RF	**0,0062228**
LRG e XGB	**0,011814**
LRG e NN	**0,0264727**
CAT e XGB	0,1095295
ADA e CAT	0,2090525

Fonte: Autora (2021)

Com o uso do teste de Mann-Whitney no contexto estudado situado na Tabela 4.10, pode-se observar que 13 dos 15 pares de algoritmos avaliados apresentaram resultados de predições significantemente diferentes e apenas CAT e XGB além de CAT e ADA não foi possível estabelecer presença de diferenciação estatística.

Tabela 4.11 – Teste de *t-student* entre Técnicas no Cenário 2

Técnicas	p-valor
NN e RF	**1,44201E-10**
NN e ADA	**1,11697E-06**
LRG e RF	**1,4751E-06**
NN e CAT	**5,08267E-06**
NN e XGB	**5,81059E-06**
RF e XGB	**0,001218676**
LRG e ADA	**0,001989532**
LRG e CAT	**0,02299116**
RF e CAT	**0,02299116**
LRG e NN	**0,03492447**
LRG e XGB	0,07442593
ADA e RF	0,08686979
ADA e XGB	0,1493813
CAT e XGB	0,3314969
ADA e CAT	0,6169438

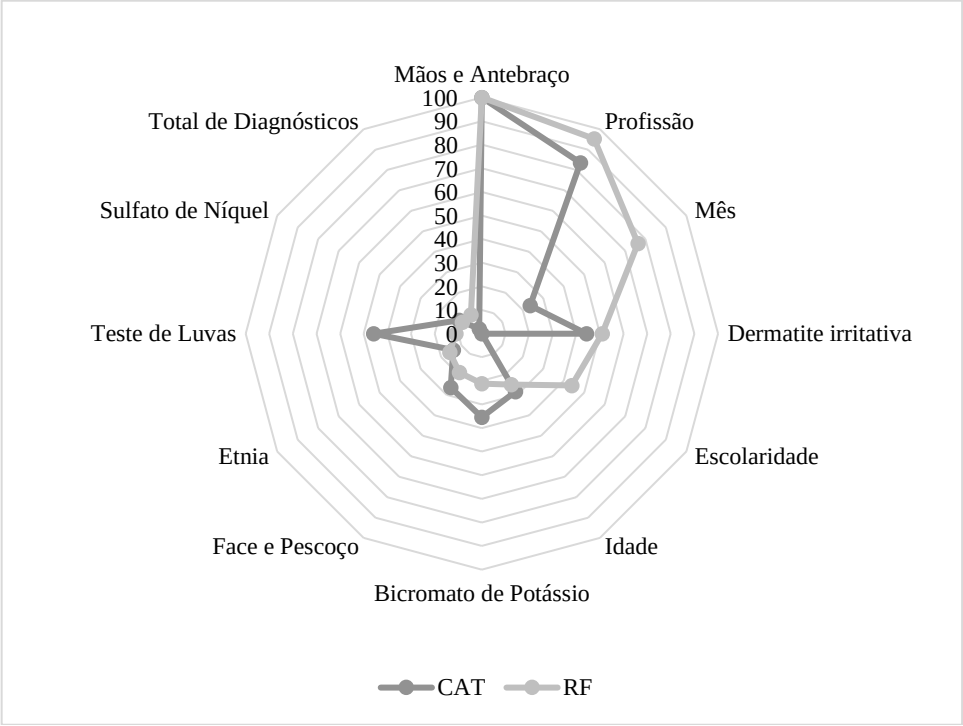
Fonte: Autora (2021)

No que se refere à comparação entre as técnicas nos testes de *t-student* e de Tukey para o cenário de número 2, foi verificado que LRG e XGB, ADA e RF, ADA e XGB, ADA e CAT e CAT e XGB apresentaram ausência de significância estatística enquanto os demais pares de técnicas se mostraram com distinção significante estatística.

Nos 3 testes as díades CAT e XGB bem como ADA e CAT não apresentaram significância estatística e mostraram-se significantemente distintos todas as técnicas pareadas à Rede Neural além de LRG e RF, LRG e ADA, LRG e CAT, RF e XGB e por fim RF e CAT. Os pares LRG e XGB, ADA e RF e ADA e XGB por sua vez, nos testes de Tukey e de *t-student* não denotaram distinção enquanto no teste de Mann-Whitney indicaram-se como distintos.

Ao que diz respeito à importância das variáveis, está exposto na Figura 4.19 (apenas as 12 de maior importância para CAT) e também na Tabela D.2 (todas as variáveis) situada no Apêndice D a importância das variáveis para as técnicas *Catboost* e *Random Forest*.

Figura 4.19 – Gráfico *Spider* da importância das variáveis



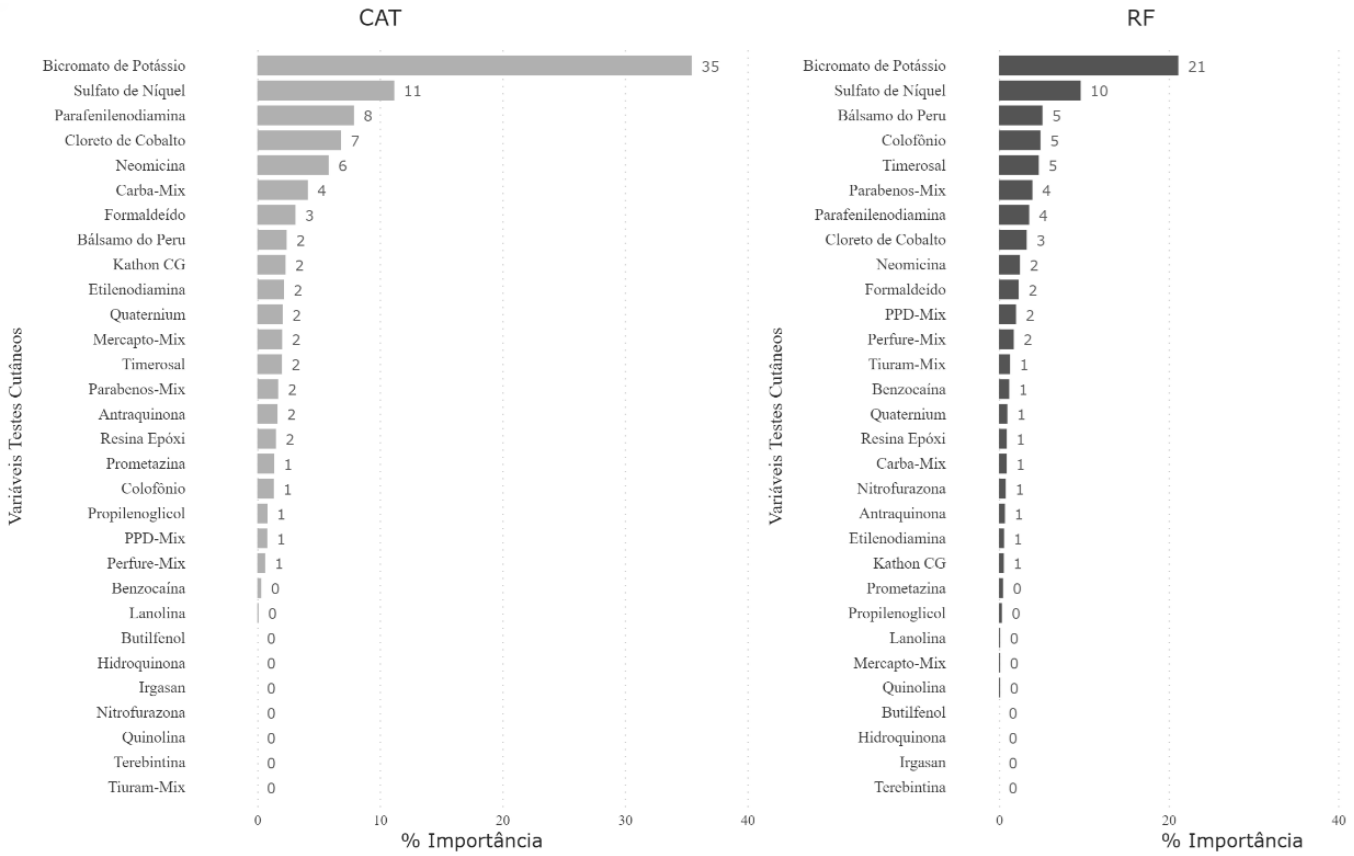
Fonte: Autora (2021)

Com a Figura 4.19 é possível estabelecer que para os fatores mão e antebraço, dermatite irritativa, etnia e idade ambas as técnicas apresentaram valores correspondentes à importância relativamente próximos enquanto para mês, escolaridade e teste de luvas os resultados obtidos foram consideravelmente distintos. Em relação aos compostos pertencentes aos testes cutâneos

apenas sulfato de níquel e bicromato de potássio ocuparam lugar dentre os 12 fatores de maior influência para *Catboost*.

Como o que distingue o cenário 1 do cenário 2 são os fatores pertencentes aos testes cutâneos, na Figura 4.20 está localizado a importância de todos os compostos segundo seu algoritmo correspondente (CAT e RF).

Figura 4.20 – Importância das Variáveis dos Testes Cutâneos no Cenário 2



Fonte: Autora (2021)

Ao realizar a observação da Figura 4.20, verifica-se que tanto para CAT quanto para RF os 2 compostos de maior influência na determinação de uma dermatose ocupacional são bicromato de potássio e sulfato de níquel. Os demais compostos em ambos os casos, mostraram-se com importância relativa inferior a 10% e butilfenol, hidroquinona, irgasan e terebintina não apresentaram influência alguma sobre a existência de dermatose oriunda do trabalho.

4.2.3 Comparação entre Cenários e Significância dos Testes Cutâneos

Na presente subseção são comparadas as métricas entre cenários para a mesma técnica bem como as variações no que se referem aos hiperparâmetros. Apresenta-se ainda as diferenças apresentadas na importância das variáveis para CAT e RF em ambos os cenários e é discutido a respeito da significância estatística dos testes cutâneos para a determinação de uma doença da pele relacionada ao trabalho.

Em relação às métricas, a média da acuracidade na comparação dos cenários para as 6 técnicas de ML avaliadas foi de -0,30%. Dessa forma, LRG permaneceu inalterada em relação à média e ainda, foram destaque na elevação da assertividade as técnicas CAT (5,80%), ADA (2,00%) e RF (1,90%) enquanto XGB e NN apresentaram redução na acurácia de suas predições de -1,90% e -9,60% respectivamente além de que LRG apresentou redução no que se refere aos acertos. A respeito dos desvios padrões correspondentes às acurácias, ocorreu variação para a mesma técnica entre cenários na seguinte ordem: LRG -1,10%, CAT -0,65%, XGB -0,61%, ADA 0,00%, RF 0,35% e NN 1,69%. No cenário 2, cinco dos seis algoritmos avaliados apresentaram sensitividade superior do que quando comparado com o primeiro cenário, e, em relação à especificidade todas as técnicas retornaram resultados inferiores quando pareadas ao cenário 1.

Ao avaliar separadamente a métrica acurácia constata-se que NN, XGB, ADA e RF no primeiro cenário conseguem distinguir uma dermatose ocupacional de uma não ocupacional de forma satisfatória. Quando se realiza a mesma análise para o segundo cenário, constata-se que CAT, ADA, XGB e RF mostraram conseguir diferenciar satisfatoriamente uma dermatose do trabalho de uma ocasionada na vida cotidiana.

No que se refere aos hiperparâmetros, verificou-se que apenas LRG e ADA expuseram em ambos cenários os mesmos hiperparâmetros responsáveis por conferir maior acuracidade ao passo que RF os alterou em sua totalidade e XGB, CAT e NN para atingir níveis superiores de acuracidade modificaram parcialmente seus preceitos.

Tratando-se da importância das variáveis, ao realizar a comparação entre os resultados do cenário 1 para o cenário 2 ao desconsiderar os testes cutâneos para RF apresentou 8 variáveis que permaneceram na mesma posição do *ranking* que são: mãos e antebraço (1º), profissão (2º), mês (3º), idade (6º), atopia (15º), tórax e abdome (17º), líquen plano (18º) e dermatite atópica (23º). As variáveis que ganharam 1 posição (ex. subindo de 5º lugar no cenário 1 para 4º lugar no cenário 2) foram dermatite irritativa, face e pescoço, pés e coxas, síndrome da pele excitada, urticária de contato e eczema de estase enquanto as que perderam 1 posição foram escolaridade,

etnia, dermatite alérgica de contato e dermatite seborreica. Os fatores que mais ganharam posições no segundo cenário para o primeiro em RF são no total de diagnósticos (4 posições), líquen simples crônico (4 posições), teste de luvas (3 posições) e dia da semana (3 posições) enquanto dos que mais sofreram de queda no posicionamento estação e sexo com decréscimo de 5 lugares cada, seguido por eczema numular (4 posições) e psoríase (2 posições).

Ao realizar o mesmo comparativo feito para RF para a técnica CAT obteve-se os seguintes achados: tórax e abdome, atopia e líquen plano perderam 1 posição, apenas dermatite seborreica ganhou uma posição e permaneceram no mesmo lugar no *ranking* mãos e antebraço (1º), profissão (2º) e urticária de contato (20º). Apresentado grandes variações de posicionamento, constatou-se que subiram no segundo cenário em comparação com o primeiro dia da semana (11 posições), teste de luvas (8 posições), dermatite irritativa (8 posições), eczema de estase (5 posições), psoríase (5 posições), estação (5 posições), eczema numular (3 posições), face e pescoço (2 posições), sexo (2 posições) enquanto diminuíram seu posicionamento os fatores escolaridade (16 posições), dermatite alérgica de contato (10 posições), etnia (6 posições), mês (5 posições), pés e coxas (2 posições), total de diagnósticos (2 posições), líquen simples crônico (2 posições), síndrome da pele excitada (2 posições) e dermatite atópica (2 posições).

A importância das variáveis para CAT, dessa forma, variou mais no segundo cenário quando comparado em relação ao primeiro que RF na mesma situação. Foi evidenciado, dessa forma, que CAT é mais sensível à adição de fatores do que RF no contexto em questão. Um aspecto importante a respeito da importância das variáveis e das características do conjunto de dados é que na base de dados ocorre com maior frequência a dermatite alérgica de contato e CAT e RF caracterizaram como mais relevante na determinação de dermatose ocupacional no segundo cenário a dermatite irritativa de contato.

Além de realizar a comparação entre variáveis, foi executado a comparação entre cenários para a mesma técnica com os testes de Mann-Whitney presente na Tabela 4.12 e de *t-student* constante na Tabela 4.13 em negrito em ambas as tabelas quando aplicável situa-se os valores cujo p-valor mostrou-se inferior à 0,05.

Tabela 4.12 – Teste de Mann-Whitney para o mesmo Algoritmo entre Cenários

Técnica	p-valor
XGB	**0,0163375**
RF	0,0618235
ADA	0,3370814
LRG	0,7772243
CAT	0,8528561
NN	0,9108391

Fonte: Autora (2021)

Ao realizar a observação dos resultados existentes na Tabela 4.12 referentes ao teste de Mann-Whitney, foi verificado que apenas a técnica XGB apresentou significância estatística na distinção dos resultados obtidos no cenário 1 em comparação com o cenário 2.

Tabela 4.13 – Teste de *t-student* para o mesmo Algoritmo entre Cenários

Técnica	p-valor
XGB	0,183623
RF	0,3776812
NN	0,6253666
ADA	0,7211421
LRG	0,7461535
CAT	0,7559203

Fonte: Autora (2021)

Em relação ao teste de *t-student* constante na Tabela 4.13, é evidenciado que quando realizado a comparação entre cenários para cada algoritmo em questão não foram verificados a presença de significância estatística entre os achados.

Após realizar a comparação da importância das variáveis foi constatado que *Catboost* se apresentou com mais sensibilidade depois de adicionado novas variáveis do que *Random Forest.* Em relação às métricas, a técnica com maior acréscimo de acuracidade no segundo cenário foi CAT e que a melhor técnica entre as cinco técnicas após a exclusão do *Catboost* foi *Random Forest* nos dois cenários.

Em relação aos testes estatísticos, não houve diferença entre realizar ou não os exames complementares (testes cutâneos) pra determinar se a dermatite é ocupacional ou não em 5 dos 6 algoritmos em estudo para Mann-Whitney (apenas XGB mostrou-se com diferenciação estatística) e nenhuma técnica mostrou-se significantemente distinta em *t-student*. Dessa forma, segundo os resultados encontrados não há indícios de que seja necessário realizar a bateria

padrão brasileira de testes de contato para determinar se a dermatose existente apresenta origem em atividades relativas ao trabalho o que implica em substancial economia de recursos aos hospitais e postos de saúde que fazem uso de tais testes para determinação do nexo causal entre a doença e o trabalho. Apesar desse resultado, não significa que não seja necessário realizar os testes cutâneos para a determinação de diagnósticos de doenças ou ainda tratamentos específicos de moléstias. Nesta pesquisa, portanto, foi detectado apenas que não há indícios que os testes cutâneos possam colaborar para estabelecer o nexo entre a doença apresentada e o trabalho.

4.3 PERSPECTIVAS FUTURAS

Esta Seção está destinada à discussão quanto às perspectivas futuras a partir dos resultados encontrados na Seção anterior e com base na revisão de literatura. A subseção 4.3.1 está destinada para tratar das aplicações no âmbito da Saúde e Segurança Ocupacional enquanto a subseção 4.3.2 se orienta para a Medicina Ocupacional e ambas estão direcionadas para discutir conforme encontrado na revisão bibliográfica. A subseção 4.3.3 orienta-se para pesquisas futuras a partir dos resultados encontrados na pesquisa desta dissertação.

4.3.1 Pesquisas Futuras em SSO – Revisão de Literatura

Em termos da Saúde e Segurança Ocupacional, foi verificado por meio da revisão de literatura que as ações principais para a prevenção de dermatoses ocupacionais são: legislação, vigilância, medidas de controle de exposição, educação e treinamento e abordagens multifacetadas resultadas da combinação de múltiplos métodos (KEEFE *et al.*, 2020). Nos parágrafos a seguir discorre-se a respeito desses 5 tópicos visando elucidar a aplicabilidade e também direcionar estudos posteriores.

No que se refere à legislação, no Brasil todos os trabalhadores e empregadores devem seguir as normas regulamentadoras de forma que conforme a norma regulamentadora de número 1, os empregadores devem cumprir e fazer cumprir as disposições legais a respeito de SSO (BRASIL, 2019). Além de cumprir as leis, os empregadores são responsáveis por informar à sua equipe de trabalho os riscos existentes no local de trabalho, as medidas que são adotadas para controle desses riscos, elaborar procedimentos para eliminação/mitigação das doenças ocupacionais e implementar medidas de proteção em relação aos perigos encontrados.

Um tópico a ser estudado de maneira mais aprofundada é a legislação pertinente e também buscado entender como as organizações desenvolvem, controlam e melhoram suas políticas de prevenção aos acidentes e doenças ocupacionais. Um exemplo de desenvolvimento, controle e melhoria é que cada empresa pode criar suas regras de ouro segundo os perigos existentes nas instalações e tê-las expostas para a equipe de trabalho bem como as implicações do não cumprimento das mesmas na sua própria saúde. Além do desenvolvimento, controle e melhoria outro tópico que pode ser explorado mais a fundo é a respeito da periodicidade da revisão das políticas bem como entender as práticas realizadas pelas empresas do mesmo ramo dentro e fora do Brasil.

No que diz respeito à vigilância, recomenda-se que seja estudado o perfil da liderança na prevenção de dermatoses do trabalho uma vez que a figura do supervisor hierárquico é fundamental para garantir o cumprimento das regras internas e por verificar a conservação e uso dos equipamentos de proteção. Além da vigilância por parte do supervisor, é interessante que seja melhor explorado o cuidado ativo onde cada colaborador independente do seu nível hierárquico tem autonomia de cobrar o e deixar ser cobrado no que concerne à observação dos regimentos e perigos existentes.

Tratando-se das medidas de controle de exposição, recomenda-se que sejam desenvolvidos estudos para compreender o relacionamento entre as medidas de controle de exposição com a legislação e também com a vigilância. Tais medidas de controle e exposição se relacionam com a legislação uma vez que todas as empresas têm a obrigação de seguir as normas e regulamentações dos locais aos quais estão instaladas bem como compreender como os procedimentos internos suportam o atendimento da regulamentação governamental. Em relação à vigilância das medidas de controle de exposição no que se refere às dermatoses ocupacionais, uma vertente de pesquisa poderia ser orientada para determinação das práticas de controle, como são medidas e à periodicidade de monitoramento das mesmas.

Em termos de educação e treinamento, propõem-se que sejam melhor compreendidos as metodologias de treinamento que apresentaram êxito na prevenção das doenças da pele relativas ao trabalho recomendadas pela literatura e as mesmas sejam adaptadas para a realidade brasileira.

4.3.2 Pesquisas Futuras em Medicina Ocupacional – Revisão de Literatura

No que faz referência à Medicina Ocupacional, para pesquisas futuras é recomendado que o questionário nórdico situado no Anexo A seja adaptado para a realidade brasileira. Além de

adaptá-lo a realidade vivenciada no Brasil, sugere-se que o mesmo apresente uma redução na quantidade total de questões e ainda seja acrescentado o tempo em que exerce a profissão até o aparecimento da primeira lesão para que seja possível a realização de análise de sobrevivência.

O propósito do uso dessa melhoria do questionário orienta-se para proporcionar a utilização durante a prática clínica de forma que os médicos ao realizarem o atendimento de um indivíduo com suspeita de dermatose ocupacional respondessem o questionário e recebessem uma classificação a partir de um algoritmo a respeito da dermatose (se é do trabalho ou não).

4.3.3 Pesquisas Futuras - Resultados da Pesquisa

A respeito dos resultados encontrados na presente pesquisa, foram observadas duas principais oportunidades para pesquisas futuras: sazonalidade das classificações de ocorrências e a respeito da sensibilidade das métricas bem como dos hiperparâmetros conforme a disponibilidade de variáveis além da quantidade de observações.

Para pesquisas futuras em relação à sazonalidade das classificações de ocorrências recomenda-se que seja melhor compreendido a relação entre o tipo da doença (exemplo: dermatite irritativa) *versus* o período do tempo. Além de realizar a análise do diagnóstico clínico ao longo do tempo, recomenda-se que seja também melhor compreendido o padrão de desenvolvimento de doenças ocupacionais sob a perspectiva temporal. Em engenharia de produção, sobremaneira quando se trata de planejamento de produção tal tipo de estudo é consideravelmente comum estando inclusive no escopo das disciplinas básicas de formação do curso. Pesquisas futuras dessa natureza mostram-se como relevantes porque uma das variáveis classificadas pelos algoritmos de mineração de dados como fortemente influenciadoras na ocorrência da dermatose ocupacional é o mês que consiste num atributo temporal.

No que faz referência ao aprendizado de máquina no âmbito das dermatoses ocupacionais, foi evidenciado certa sensibilidade das métricas e dos hiperparâmetros. Recomenda-se, portanto, que seja realizado estudos para avaliar se tal sensibilidade observada foi estabelecida pela metodologia de pesquisa, por fatores intrínsecos às próprias técnicas de aprendizado de máquina ou ainda, por características do conjunto de dados. Essa avaliação crítica da sensibilidade é relevante para viabilizar pesquisas terciárias de inferência populacional que possibilitem replicação dos resultados a uma quantidade maior de pessoas com níveis de confiabilidade estatística conhecidos.

CONCLUSÕES

5

5.1 CONSIDERAÇÕES FINAIS

As dermatoses ocupacionais são caracterizadas por se apresentarem como resposta imunológica em decorrência do contato com substâncias irritantes constantes no local de trabalho ou durante a execução das atividades comumente apresentadas nas formas irritativa e alérgica. Por se tratar de uma classe de doenças relativamente comuns que em menos de dois anos a previdência social brasileira concedeu mais de dois milhões de reais em benefícios aos segurados é interessante compreender os fatores contribuintes para o desenvolvimento desse tipo de lesão.

Na última década técnicas de aprendizado de máquina têm sido empregadas para auxiliar os governos no estabelecimento de políticas públicas. Com o intuito de apresentar uma metodologia alternativa para a compreensão das dermatoses ocupacionais fez-se uso do processo *KDD* aplicado no conjunto de dados disponibilizado pela Fiocruz referente à pacientes atendidos no Serviço de Dermatoses Relacionadas ao Trabalho. Tais trabalhadores foram submetidos à bateria padrão brasileira de testes de contato, responderam a um questionário para composição do perfil epidemiológico, a equipe médica realizou o registro da doença e classificou a dermatose como ocupacional ou não ocupacional.

De posse do conjunto de dados, determinou-se como objetivos: identificar padrões de comportamento entre as variáveis, determinar os fatores de influência e avaliar a relevância dos testes cutâneos da Bateria Padrão Brasileira de Testes de Contato na incidência de dermatoses ocupacionais dos pacientes atendidos na Fiocruz no DRT. Como objetivos específicos estabeleceu-se: a comparação das técnicas Regressão Logística, *Adaboost*, *Neural Network*, *Random Forest*, *Extreme Gradient Boosting* e *Catboost;* fazer uso das métricas Acuracidade, Sensitividade, Especificidade, Erro, *Recall*, Precisão, F_1 *Score*, Prevalência, índice *Kappa* e *Area Under the Curve* para comparação dos algoritmos; e realizar o confrontamento da importância das variáveis de *Catboost* com a técnica de mineração que apresentar os melhores resultados dentre as demais técnicas.

As seis técnicas de mineração passaram pelos mesmos critérios de limpeza, transformação e avaliação em dois cenários: o primeiro apresentando 27 variáveis no total e o segundo cenário fazendo uso das mesmas 27 do primeiro cenário acrescido dos testes cutâneos. Sendo assim, foi verificado no cenário 1 que a técnica que mais se destacou foi RF, com 84,62% de acuracidade seguido por XGB que apresentou 82,69% enquanto CAT apresentou acuracidade de 78,85%. No cenário 2 a técnica com o maior percentual de acurácia continuou sendo RF, porém, o segundo lugar ficou empatado entre CAT e ADA. Em relação à acuracidade, foi possível estabelecer que no primeiro cenário RF, ADA, XGB e NN conseguem distinguir satisfatoriamente uma dermatose ocupacional de uma não ocupacional apresentando acuracidade média superior a 80% enquanto no segundo cenário RF, CAT, ADA, XGB o fazem.

A técnica que apresentou as melhores métricas foi *Random Forest* e a mesma foi comparada com *Catboost* em relação à importância das variáveis. Quando pareado os resultados dos dois cenários verifica-se que *Random Forest* apresenta maior estabilidade do que *Catboost* para a adição de novas variáveis uma vez que a posição no *ranking* de importância variou menos em RF. Para RF as 5 variáveis mais influentes em ordem decrescente foram mãos e antebraço, profissão, mês, dermatite irritativa e escolaridade enquanto para CAT as 5 com mais importância também em ordem decrescente são mãos e antebraço, profissão, teste de luvas, dermatite irritativa e estação.

Além dos resultados apresentados no *KDD*, foi realizado a comparação dos resultados entre técnicas para o mesmo cenário e entre cenários para a mesma técnica com o intuito de avaliar a similaridade entre os resultados. Na comparação entre técnicas no cenário 1 constatou-se que os pares NN e RF, NN e XGB, NN e ADA, NN e CAT, LRG e XGB, LRG e RF e LRG e ADA mostraram-se significantemente distintos nos testes de Tukey, de Mann-Whitney e no de *t-student*. Ainda no primeiro cenário foi verificado distinção no teste de Mann-Whitney e semelhança em *t-student* os pares de algoritmos LRG e CAT, RF e CAT e RF e ADA sendo que os 2 últimos também se mostraram distintos em Tukey.

Ao parear as técnicas no cenário 2 foi observado que nos testes de Tukey, Mann-Whitney e *t-student* as díades CAT e XGB bem como ADA e CAT não apresentaram significância estatística e mostraram-se significantemente distintos todas as técnicas pareadas à Rede Neural além de LRG e RF, LRG e ADA, LRG e CAT, RF e XGB e por fim RF e CAT. Os pares LRG e XGB, ADA e RF e ADA e XGB por sua vez, nos testes de Tukey e de *t-student* não denotaram distinção enquanto no teste de Mann-Whitney indicaram-se como distintos.

Quando realizado a comparação de distintos cenários para a mesma técnica foi verificado que apenas XGB apresentou resultados significantemente distintos exclusivamente no teste de Mann-Whitney enquanto para as demais técnicas e testes não foram constatadas variações significativas. De acordo com os resultados encontrados, não há indícios de que seja necessário realizar a bateria padrão brasileira de testes de contato para determinar se a dermatose apresentada é originada de atividades relativas ao trabalho.

5.2 LIMITAÇÕES DO TRABALHO

A partir das contribuições e da revisão de literatura constantes nesta dissertação foi possível estabelecer que a fronteira entre a medicina ocupacional, a saúde e segurança do trabalho e as práticas de gestão são limítrofes. Apesar dessa característica poder ser considerada um desafio, quão melhor for o entendimento sobre o que a escolha de uma determinada área impacta na outra ter-se-á maiores chances de que as decisões tomadas sejam as mais assertivas possíveis.

Uma das dificuldades encontradas no desenvolvimento dessa pesquisa foi intrínseco à formação da discente uma vez que durante a jornada acadêmica de engenharia pouco se aprende sobre saúde. Dessa forma, as discussões podem não se apresentar de maneira facilmente entendível e ainda a terminologia técnica utilizada não ser a mais coerente para o contexto.

Em relação às limitações da pesquisa, tem-se que os achados são restritos à população avaliada podendo não representar a realidade brasileira uma vez que todos os participantes são oriundos da mesma região do país.

5.3 PESQUISAS FUTURAS

A presente pesquisa apresenta o potencial de servir como direcionador para uma diversidade de pesquisas futuras, entretanto, dois pontos carecem ser melhor compreendidos segundo os resultados apresentados: sazonalidade das classificações das ocorrências e a respeito da sensibilidade das métricas bem como dos hiperparâmetros.

Para pesquisas futuras em relação às oportunidades observadas durante a revisão de literatura, detectou-se oportunidades de pesquisa em Medicina Ocupacional e também em Saúde e Segurança Ocupacional. No que se refere à Medicina Ocupacional, recomenda-se que seja desenvolvido e validado uma plataforma a qual um médico diante de um paciente possa responder as questões constantes na plataforma e a disponibilize a categorização do mesmo. Em relação às práticas em Saúde e Segurança Ocupacional aconselha-se que seja estudado de

forma mais aprofundada a legislação, vigilância, medidas de controle de exposição, educação e treinamento e abordagens multifacetadas.

5.4 RESULTADOS DA PESQUISA

Durante a participação da discente no programa de pós-graduação foram produzidas algumas pesquisas que contribuíram direta ou indiretamente para o desenvolvimento da presente dissertação. Tais trabalhos estão dispostos a seguir:

1. ROSA, Ana Caroline Francisco Da *et al*.. REGRESSÃO LOGÍSTICA APLICADA A DERMATOSES OCUPACIONAIS: UMA ANÁLISE EXPLORATÓRIA.. *In*: **Anais** do Simpósio de Engenharia, Gestão e Inovação. Anais...São Paulo(SP) USP, 2020. Disponível em: <https//www.even3.com.br/anais/sengi2020/270867-REGRESSAO-LOGISTICA-APLICADA-A-DERMATOSES-OCUPACIONAIS--UMA-ANALISE-EXPLORATORIA>. Acesso em: 11/02/2021 20:19.
2. REIS, Beatriz Lavezo dos *et al*.. ANÁLISE DO ABASTECIMENTO EM UMA REDE SUPERMERCADISTA UTILIZANDO O PROBLEMA DE TRANSPORTE E MÉTODO DAS P-MEDIADAS.. *In*: **Anais** do Simpósio de Engenharia, Gestão e Inovação. Anais...São Paulo(SP) USP, 2020. Disponível em: <https//www.even3.com.br/anais/sengi2020/271108-ANALISE-DO-ABASTECIMENTO-EM-UMA-REDE-SUPERMERCADISTA-UTILIZANDO-O-PROBLEMA-DE-TRANSPORTE-E-METODO-DAS-P-MEDIADAS>. Acesso em: 11/02/2021 20:22.
3. ROSA, Ana Caroline Francisco Da *et al*.. Risk Management in Occupational Safety: A Systematic Mappping. **Work.** *Status:* Aguardando Publicação.
4. ROSA, Ana Caroline Francisco Da *et al*.. Identificação dos fatores de influência em doenças ocupacionais utilizando técnicas de machine learning em aprendizado supervisionado com dados de saúde e segurança do trabalho. **Revista Brasileira de Segurança Ocupacional**. *Status:*Em Avaliação.
5. ROSA, Ana Caroline Francisco Da *et al*.. How can data mining contribute to OSH? Systematic mapping and roadmap. **Archives of Computational Methods in Engineering (ARCO).** *Status:*Em Avaliação.

REFERÊNCIAS

ALANAZI, Hamdan O.; ABDULLAH, Abdul Hanan; QURESHI, Kashif Naseer. A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. **Journal of Medical Systems** v. 41, n. 4, p. 1–10 , 1 abr. 2017.

ANDRUP, Lars. *Det Nationale Forskningscenter for Arbejdsmiljø (NFA)*. Disponível em: <https://nfa.dk/da/Uddannelse/Uddannelse/Vejledning>. Acesso em: 1 out. 2021.

BADRI, Adel; BOUDREAU-TRUDEL, Bryan; SOUISSI, Ahmed Saâdeddine. Occupational health and safety in the industry 4.0 era: A cause for major concern? .**Safety Science**, v.109, n.01, p.403 - 411, 2018.

BAINS, Sonia N.; NASH, Pembroke; FONACIER, Luz. Irritant Contact Dermatitis. **Clinical Reviews in Allergy and Immunology** v. 56, n. 1, p. 99–109 , 2019.9781118441213.

BEHROOZY, Ali; KEEGEL, Tessa G. Wet-work exposure: A main risk factor for occupational hand dermatitis.**Safety and Health at Work,** v.5, n.4, p. 175-180, 2014.

BOHANEC, Marko; DELIBAŠIĆ, Boris. Data-mining and expert models for predicting injury risk in ski resorts. **Lecture Notes in Business Information Processing** v. 216, p. 46–60 , 2015.

BRASIL. *CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988*. Disponível em: <https://www.planalto.gov.br/ccivil_03/Constituicao/Constituicao.htm>. Acesso em: 13 mar. 2021.

BRASIL. **Dermatoses ocupacionais**. 1. ed. Brasília: Editora do Ministério da Saúde, 2006. 92 p. Disponível em: <http://www.saude.gov.br/editora>. Acesso em: 12 abr. 2020.

BRASIL. **Lei n. 8.213, de 24 de jul. de 1991.** *Lei de Planos e Benefícios da Previdência Social.* Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8213cons.htm>. Acesso em: 12 abr. 2020.

BREIMAN, Leo. Bagging predictors. **Machine Learning,** v. 24, n. 2, p. 123–140 , 1996.

BREIMAN, Leo. Random forests. **Machine Learning** v. 45, n. 1, p. 5–32 , out. 2001.

BREWER, Gayle; HOLT, Barry; MALIK, Shahzeb. Workplace bullying in risk and safety professionals. **Journal of Safety Research** v. 64, p. 129–133 , 1 fev. 2018.

BUCZAK, Anna L.; GUVEN, Erhan. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. **IEEE Communications Surveys and Tutorials** v. 18, n. 2, p. 1153–1176 , abr. 2016.

CALLAHAN, Alison; SHAH, Nigam H. Machine Learning in Healthcare. **Key Advances in Clinical Informatics:** Transforming Health Care through Health Information Technology, v. 1, p. 279-291, 2017.

CHANG, Chun Lang; CHEN, Chih Hao. Applying decision tree and neural network to increase quality of dermatologic diagnosis. **Expert Systems with Applications** v. 36, n. 2 PART 2, p. 4035–4041 , 2009.

CHATTERJEE. Sourav. **Package "fastAdaboost"**. 2016. Disponível em: <https://github.com/souravc83/fastAdaboost/issues>. Acesso em: 9 jan. 2021.

CHEN, Hong *et al.* Comparative study on the strands of research on the governance model of international occupational safety and health issues. **Safety Science** v. 122, n. August 2019, p. 104513 , 2020. Disponível em: <https://doi.org/10.1016/j.ssci.2019.104513>.

CHEN, Tianqi. *XGBoost*. Disponível em: <https://cran.r-project.org/web/packages/xgboost/xgboost.pdf>. Acesso em: 9 jan. 2021.

COMBERTI, Lorenzo; BALDISSONE, Gabriele; DEMICHELA, Micaela. Workplace accidents analysis with a coupled clustering methods: S.O.M. and K-means algorithms. **Chemical Engineering Transactions** v. 43, p. 1261–1266 , 20 maio 2015.

COMBERTI, Lorenzo; DEMICHELA, Micaela; BALDISSONE, Gabriele. A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making. **Safety Science** v. 106, p. 191–202 , jul. 2018.

COMITE DE DADOS ABERTOS DO INSS. *Benefícios Concedidos*. Disponível em: <http://dadosabertos.dataprev.gov.br/>.Acesso em: 20 abr. 2020.

DALENOGARE, Lucas Santos *et al.* The expected contribution of Industry 4.0 technologies for industrial performance. **International Journal of Production Economics** v. 204, p. 383–394 , 1 out. 2018.

DAVOUDI KAKHKI, Fatemeh; FREEMAN, Steven A.; MOSHER, Gretchen A. Evaluating machine learning performance in predicting injury severity in agribusiness industries. **Safety Science** v. 117, p. 257–262 , ago. 2019.

DEO, Rahul C. Machine learning in medicine. **Circulation** v. 132, n. 20, p. 1920–1930 , 17 nov. 2015.

DERMATOLOGIA, DEPARTAMENTO ESPECIALIZADO DE ALERGIA EM. Anais Brasileiros de Dermatologia - Estudo multicêntrico para elaboração de uma bateria-padrão brasileira de teste de contato. **An Bras Dermatol** v. 75 , 2000. Disponível em:

<http://www.anaisdedermatologia.org.br/detalhe-artigo/10194/Estudo-multicentrico-para-elaboracao-de-uma-bateria-padrao-brasileira-de-teste-de-contato>. Acesso em: 20 abr. 2020.

DI NOIA, Antonio *et al.* Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. **Soft Computing** v. 24, n. 6, p. 4393–4406 , 2020.

DIMITRIADOU. Friedrich Leisch Evgenia. **Package "mlbench" Machine Learning Benchmark Problems**. [S.l: s.n.], 2015.

DOLTSINIS, Stefanos; FERREIRA, Pedro; LOHSE, Niels. Reinforcement learning for production ramp-up: A Q-batch learning approach. 1, 2012. **Anais Proceedings - 2012 11th International Conference on Machine Learning and Applications**, **ICMLA 2012**, 2012. p. 610–615.

DOS SANTOS, Bruno Samways *et al.* Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. **Computers and Industrial Engineering** v. 138, p. 106120 , 2019.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine** v. 17, n. 3, p. 37–53 , 1996.

FERNÁNDEZ-MUÑIZ, Beatriz; MONTES-PEÓN, José Manuel; VÁZQUEZ-ORDÁS, Camilo José. Occupational risk management under the OHSAS 18001 standard: Analysis of perceptions and attitudes of certified firms. **Journal of Cleaner Production** v. 24, p. 36–47 , 1 mar. 2012.

FORSTING, Michael. *Machine learning will change medicine* .**Journal of Nuclear Medicine,** v. 58, n.3, p. 357-358, 2017.

FREUND, Yoav; SCHAPIRE, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences.** p.23–37, 1995.

FREUND, Yoav; SCHAPIRE, Robert E. Experiments with a New Boosting Algorithm. **Proceedings of the 13th International Conference on Machine Learning** p. 148–156 , 1996.

FRIEDMAN, Jerome. *Package 'glmnet'*. Disponível em: <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>. Acesso em: 9 jan. 2021.

GAWKRODGER, D. J. Patch testing in occupational dermatology. **Occupational and Environmental Medicine** v. 58, n. 12 , 2001.

GHORI, Khawaja Moyeezullah *et al.* Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection. **IEEE Access** v. 8, p. 16033–16048 , 2020.

GONZÁLEZ-LÓPEZ, G. *et al.* Difficulties Coding Dermatological Disorders Using the ICD-10: The DIADERM Study. **Actas Dermo-Sifiliográficas (English Edition)** v. 109, n. 10, p. 893–899 , 2018.

GULIN, Andrey. *R package installation - CatBoost. Documentation*. Disponível em:

<https://catboost.ai/docs/concepts/r-installation.html>. Acesso em: 9 jan. 2021.

GUNS, Raf; LIOMA, Christina; LARSEN, Birger. The tipping point: F-score as a function of the number of retrieved items. **Information Processing and Management** v. 48, n. 6, p. 1171–1180 , 2012.

HAMNERIUS, Nils *et al.* Hand eczema and occupational contact allergies in healthcare workers with a focus on rubber additives. **Contact Dermatitis** v. 79, n. 3, p. 149–156 , 2018.

HANCOCK, John T.; KHOSHGOFTAAR, Taghi M. CatBoost for big data: an interdisciplinary review. **Journal of Big Data** v. 7, n. 1 , 1 dez. 2020.

HOLNESS, D. Linn. Occupational skin allergies: Testing and treatment (the case of occupational allergic contact dermatitis) topical collection on occupational allergies. **Current Allergy and Asthma Reports** v. 14, n. 2 , 2014.

JOHANSEN, Jeanne Duus *et al.* Prevention of hand eczema among Danish hairdressing apprentices : an intervention study. **Occup Environ Med** v. 69, p. 310–316 , 2012.

JORDAN, M. I.; MITCHELL, T. M. *Machine learning: Trends, perspectives, and prospects*. **Science**, v. 349, n.6245, p. 255-260, 2015.

KANG, Kyungsu; RYU, Hanguk. Predicting types of occupational accidents at construction sites in Korea using random forest model. **Safety Science** v. 120, p. 226–236 , 2019.

KARPATNE, Anuj *et al.* Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. **IEEE Transactions on Knowledge and Data Engineering** v. 29, n. 10, p. 2318–2331 , 2017.

KEEFE, Anya R. *et al.* A scoping review to identify strategies that work to prevent four important occupational diseases. **American Journal of Industrial Medicine** v. 63, n. 6, p. 490–516 , 2020.

KEEGEL, Tessa *et al.* The epidemiology of occupational contact dermatitis (1990-2007): A systematic review. **International Journal of Dermatology** v. 48, n. 6, p. 571–578 , 2009.

KRAFT, Magdalena *et al.* Contact dermatitis and sensitization in professional musicians. **Contact Dermatitis** v. 80, n. 5, p. 273–278 , 2019.

KUHN, Max. *The caret Package*. Disponível em: <https://topepo.github.io/caret/index.html>. Acesso em: 15 jul. 2020.

LACHAPELLE, J-M.; MAIBACH, H. I. **Patch testing and prick testing: a pratical guide official publication of the ICDRG**. 3. ed. Berlin, Heidelberg: Springer, 2012. 147–157 p. .

LAMPEL, Heather P.; POWELL, Helen B. Occupational and Hand Dermatitis: a Practical Approach. **Clinical Reviews in Allergy and Immunology** v. 56, n. 1, p. 60–71 , 2019.

LANDIS, J Richard; KOCH, Gary G. The Measurement of Observer Agreement for Categorical Data. **Biometrics** v. 33, n. 1, p. 159 , 1977.

LEE, Sangseok; LEE, Dong Kyu. What is the proper way to apply the multiple comparison test? **Korean Journal of Anesthesiology** v. 71, n. 5, p. 353–360 , 2018.

LESO, Veruscka; FONTANA, Luca; IAVICOLI, Ivo. The occupational health and safety dimension of Industry 4.0. **La Medicina del lavoro** v. 110, n. 5, p. 327–338 , 2018.

LEVTEROVA, Boryana. KNOWLEDGE – International Journal Vol. 40.4 APPLIED EPIDEMIOLOGY AND PUBLIC HEALTH. **KNOWLEDGE –International Journal** v. 40, n. 4, p. 687–690 , 2014.

LISE, Michelle Larissa Zini *et al.* Occupational dermatoses reported in Brazil from 2007 to 2014. **Anais Brasileiros de Dermatologia** v. 93, n. 1, p. 27–32 , 2018.

LIU, Zimei *et al.* A paradigm of safety management in Industry 4.0. **Systems Research and Behavioral Science** v. 37, n. 4, p. 632–645 , 2020.

LUO, Qi. Advancing knowledge discovery and data mining. **Proceedings - 1st International Workshop on Knowledge Discovery and Data Mining, WKDD** p. 3–5 , 2008.

MARUCCI-WELLMAN, Helen R.; CORNS, Helen L.; LEHTO, Mark R. Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. **Accident Analysis and Prevention** v. 98, p. 359–371, 2017.

MEHTA, Pankaj *et al. A high-bias, low-variance introduction to Machine Learning for physicists.* **Physics Reports**, v. 810, p. 1-124, 2019.

MELO, Maria das Graças Mota; VILLARINHO, Ana Luiza Castro Fernandes; LEITE, Iuri da Costa. Sociodemographic and clinical profile of patients with occupational contact dermatitis seen at a work-related dermatology service, 2000 - 2014. **Anais brasileiros de dermatologia** v. 94, n. 2, p. 147–156 , 2019.

MILAM, Emily C.; COHEN, David E. Contact Dermatitis: Emerging Trends. **Dermatologic Clinics** v. 37, n. 1, p. 21–28 , 2019.

MINAMI, Takafumi *et al.* Hand eczema as a risk factor for food allergy among occupational kitchen workers. **Allergology International** v. 67, n. 2, p. 217–224 , 2018. Disponível em: <http://dx.doi.org/10.1016/j.alit.2017.08.005>.

MOURA, Raphael *et al.* Learning from accidents: Interactions between human factors, technology and organisations as a central element to validate risk studies. **Safety Science** v. 99, p. 196–214 , 1 nov. 2017.

NEMBRINI, Stefano; KÖNIG, Inke R.; WRIGHT, Marvin N. The revival of the Gini importance? **Bioinformatics** v. 34, n. 21, p. 3711–3718 , 2018.

NENONEN, Noora. Analysing factors related to slipping, stumbling, and falling accidents at work: Application of data mining methods to Finnish occupational accidents and diseases statistics database. **Applied Ergonomics** v. 44, n. 2, p. 215–224 , 2013.

OBERMEYER, Ziad; EMANUEL, Ezekiel J. *Predicting the future-big data, machine learning, and clinical medicine* .**New England Journal of Medicine**, v. 376, n. 13, p. 1216-1219, 2016.

ORGANIZATION, World Health. *Frequently asked questions - Classification of Diseases (ICD)*. Disponível em: <https://www.who.int/standards/classifications/frequently-asked-questions>. Acesso em: 14 mar. 2021a.

ORGANIZATION, World Health. *Saúde do Trabalhador*. Disponível em: <https://www.paho.org/bra/index.php?option=com_content&view=article&id=378:saude-do-trabalhador&Itemid=685>. Acesso em: 13 mar. 2021b.

PACHECO, Karin A. *Occupational dermatitis: How to identify the exposures, make the diagnosis, and treat the disease* .**Annals of Allergy, Asthma and Immunology**, v. 120, n. 6, p. 583-591, 2018.

Package "dplyr" Type Package Title A Grammar of Data Manipulation Depends R (¿= 3.2.0). [S.l: s.n.], 2020.

Package "tidyr" Title Tidy Messy Data Version 1.1.2. [S.l: s.n.], 2020.
PALIT, Indranil; REDDY, Chandan K. Scalable and parallel boosting with mapReduce. **IEEE Transactions on Knowledge and Data Engineering** v. 24, n. 10, p. 1904–1916 , out. 2012. Disponível em: <http://ieeexplore.ieee.org/document/6035709/>. Acesso em: 16 abr. 2020.

PARODI, Pietro. Computational intelligence with applications to general insurance: a review: I – The role of statistical learning. **Annals of Actuarial Science** v. 6, n. 2, p. 307–343 , 2012.

PAZMINÑO-MAJI, Rubén A.; GARCÍA-PEÑALVO, Francisco J; CONDE-GONZÁLEZ, Miguel A. Statistical implicative analysis approximation to KDD and Data Mining: A systematic and mapping review in Knowledge Discovery Database framework. **Ninth International Conference on Advances in Databases, Knowledge, and Data Applications** n. c, p. 70–77 , 2017.

PROKHORENKOVA, Liudmila; GUSEV, Gleb; VOROBEV, Aleksandr; DOROGUSH, Anna Veronika; GULIN, Andrey. Catboost: Unbiased boosting with categorical features. **Advances in Neural Information Processing Systems**.v. 31, p.6638–6648, 2018.

PROVAN, David J.; RAE, Andrew J.; DEKKER, Sidney W.A. An ethnography of the safety professional's dilemma: Safety work or the safety of work? **Safety Science** v. 117, p. 276–289, 2019.

RAJKOMAR, Alvin; DEAN, Jeffrey; KOHANE, Isaac. *Machine learning in medicine* .**New England Journal of Medicine**. v. 380, n. 14, p. 1347-1358, 2019. Disponível em: <http://www.nejm.org/doi/10.1056/NEJMra1814259>. Acesso em: 11 mar. 2020.

REIS, Felipe Rovere Diniz; DE OLIVEIRA, José Inácio; FESTINO, Fernando Simões. Perfil clínico-ocupacional e resultados dos testes de contato dos pacientes atendidos em um ambulatório terciário de medicina do trabalho. **Revista Brasileira de Medicina do Trabalho** v. 10, n. 1, p. 1–5 , 2012.

RIPLEY. Brian; VENABLES. William; MAINTAINER.]. **Package "nnet" NeedsCompilation yes**. [S.l: s.n.], 2020.

ROBIN, Xavier. *Display and Analyze ROC Curves*. Disponível em: <https://cran.r-project.org/web/packages/pROC/pROC.pdf>. Acesso em: 15 jul. 2020.

RUBAIYAT, Abu H. M. *et al.* Automatic Detection of Helmet Uses for Construction Safety. 17 jan. 2017, Institute of Electrical and Electronics Engineers (IEEE), 17 jan. 2017. p.135–142.

SAÂDAOUI, Foued *et al.* A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining. **IEEE Transactions on Nanobioscience** v. 14, n. 7, p. 707–715, 2015.

SAGI, Omer; ROKACH, Lior. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery** v. 8, n. 4 , 27 jul. 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249>. Acesso em: 15 abr. 2020.

SHERIDAN, Robert P *et al.* Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. **Journal of Chemical Information and Modeling,** v. 56, n. 12, p. 2353–2360 , 2016. Disponível em: <https://github.com/dmlc/>. Acesso em: 19 abr. 2020.

SHIN, Dong Pil *et al.* Association Rules Mined from Construction Accident Data. **KSCE Journal of Civil Engineering** v. 22, n. 4, p. 1027–1039, 2018.

SIDEY-GIBBONS, Jenni A.M.; SIDEY-GIBBONS, Chris J. Machine learning in medicine: a practical introduction. **BMC Medical Research Methodology** v. 19, n. 1 , 2019. Disponível em: <https://doi.org/10.1186/s12874-019-0681-4>. Acesso em: 11 mar. 2020.

SING, Tobias. *Package 'ROCR'*. Disponível em: <https://cran.r-project.org/web/packages/ROCR/ROCR.pdf>. Acesso em: 9 jan. 2021.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information Processing and Management** v. 45, n. 4, p. 427–437 , 2009. Disponível em: <http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.>. Acesso em: 26 mar. 2020.

SRINIVAS, K.; RAGHAVENDRA RAO, G.; GOVARDHAN, A. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. **ICCSE 2010 - 5th International Conference on Computer Science and Education, Final Program and Book of Abstracts** p. 1344–1349 , 2010.

SZELKA, J.; WRONA, Z. Knowledge Discovery in Data in Construction Projects. **Archives of Civil Engineering** v. 62, n. 2, p. 217–228 , 2016.

TAMERS, Sara L. *et al.* Total worker health® 2014–2018: The novel approach to worker safety, health, and well-being evolves. **International Journal of Environmental Research and Public Health** v. 16, n. 3 , 2019.

UHRENHOLDT MADSEN, Christian; HASLE, Peter; LIMBORG, Hans Jørgen. Professionals

without a profession: Occupational safety and health professionals in Denmark. **Safety Science** v. 113, p. 356–361 , 1 mar. 2019.

VALÊNCIO, Carlos Roberto *et al.* Spatial clustering applied to health area. **Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings,** 2011. p. 427–432. Disponível em: <https://www.researchgate.net/publication/221396446>. Acesso em: 24 mar. 2020.

VAN GILS, Robin F. *et al.* Effectiveness of prevention programmes for hand dermatitis: A systematic review of the literature. **Contact Dermatitis** v. 64, n. 2, p. 63–72 , 2011.

VERMA, Anurag Kumar; PAL, Saurabh; KUMAR, Surjeet. Classification of skin disease using ensemble data mining techniques. **Asian Pacific Journal of Cancer Prevention** v. 20, n. 6, p. 1887–1894 , 2019a.

VERMA, Anurag Kumar; PAL, Saurabh; KUMAR, Surjeet. Comparison of skin disease prediction by feature selection using ensemble data mining techniques. **Informatics in Medicine Unlocked** v. 16 , 2019b.

WEISTENHÖFER, W. *et al.* How to quantify skin impairment in primary and secondary prevention? HEROS: A proposal of a hand eczema score for occupational screenings. **British Journal of Dermatology** v. 164, n. 4, p. 807–813 , 2011.

WICKHAM, Hadley. *Create Elegant Data Visualisations Using the Grammar of Graphics*. Disponível em: <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>. Acesso em: 15 jul. 2020.

WICKHAM, Hadley. *Read Excel Files*. Disponível em: <https://cran.r-project.org/web/packages/readxl/readxl.pdf>. Acesso em: 15 jul. 2020.

WIENER, Matthew; LIAW, Andy. *Package 'randomForest'*. Disponível em: <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>. Acesso em: 9 jan. 2021.

WINTER, Joost C.F.; DODOU, Dimitra. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon. **Practical Assessment, Research & Evaluation** v. 15, p. 1–16 , 2010.

WU, Xindong *et al.* Top 10 algorithms in data mining. **Knowledge and Information Systems** v. 14, n. 1, p. 1–37 , 4 jan. 2008.

WUEST, Thorsten *et al.* Production & Manufacturing research: an oPen access. **Journal** v. 4, n. 1, p. 23–45 , 2016. Disponível em: <http://dx.doi.org/10.1080/21693277.2016.1192517>. Acesso em: 22 mar. 2020.

XU, Baoxun *et al.* Classifying very high-dimensional data with random forests built from small subspaces. **International Journal of Data Warehousing and Mining** v. 8, n. 2, p. 44–63 , 2012.

YANAR, Basak; LAY, Morgan; SMITH, Peter M. The Interplay Between Supervisor Safety Support and Occupational Health and Safety Vulnerability on Work Injury. **Safety and Health at Work** v. 10, n. 2, p. 172–179 , jun. 2019.

YOO, Changwon; RAMIREZ, Luis; LIUZZI, Juan. *Big data analysis using modern statistical and machine learning methods in medicine* .**International Neurourology Journal**, v. 18, n.2, p. 50-57, 2014.

YUN, Younghee *et al.* Exploring syndrome differentiation using non-negative matrix factorization and cluster analysis in patients with atopic dermatitis. **Computers in Biology and Medicine** v. 87, n. May, p. 70–76 , 2017.

ZHAO, Yanxia *et al.* Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. **Ear and Hearing** v. 40, n. 3, p. 690–699 , 1 maio 2019. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/30142102>. Acesso em: 24 mar. 2020.

APÊNDICE A – *Script* em linguagem R

```
#Subindo Pacotes
library(mlbench)
library(caret)
library(readxl)
library(ROCR)
library(dplyr)
library(tidyr)
library(ggplot2)
library(pROC)
library(knitr)

#Importar banco de dados
banco <- read_excel("C:/Users/anaca/OneDrive/Área de Trabalho/Banco.xlsx")

#Convertendo variáveis para fator
banco$SEXO<-as.factor(banco$SEXO)
banco$PROFISSAO<-as.factor(banco$PROFISSAO)
banco$Etnia<-as.factor(banco$Etnia)
banco$ESCOLARIDA<-as.factor(banco$ESCOLARIDA)
banco$ATOPIA<-as.factor(banco$ATOPIA)
banco$FACEEPESCO<-as.factor(banco$FACEEPESCO)

...

banco$Ocupacional<-as.factor(banco$Ocupacional)

#Particionando o banco de dados em teste e treino
trainIndex<-createDataPartition(banco$Ocupacional, p=0.8, list = FALSE)
trainData <- banco[trainIndex,-c(1,1)]
testData  <- banco[-trainIndex,-c(1,1)]

trainData$Ocupacional<-as.factor(trainData$Ocupacional)
trainData$Ocupacional <-relevel(trainData$Ocupacional, ref = "OD")
testData$Ocupacional<-as.factor(testData$Ocupacional)
testData$Ocupacional <-relevel(testData$Ocupacional, ref= "OD")

# Tirando a variável dependente
trainX <-trainData[,-19]
testX <- testData[,-19]
sapply(trainX,summary)

#Preparando o cenário de teste
ctrl <- trainControl(method = "repeatedcv",
                     number = 10, summaryFunction=twoClassSummary,
                     repeats = 10, classProbs=TRUE, allowParallel = TRUE)

# Treinando o modelo
fit.RF <- train(x=trainX,y=trainData$Ocupacional, method = "rf",
                metric = "Accuracy",trControl = ctrl)
```

```
# Apresentar Hiperparâmetros
fit.RF$bestTune

#Apresentar resultados
resRF <- fit.RF$results
resRF

# Mostrar Importância das Variáveis
varimp.rf<-varImp(fit.RF)
varimp.rf

# Fazendo Predições com a Base de Teste
RF.pred <- predict(fit.RF,testX)

#Matriz de Confusão
cf.RF<-confusionMatrix(RF.pred,testData$Ocupacional, mode = 'everything')

#Desenhar a Curva ROC
RF.probs <- predict(fit.RF,testX,type="prob")
RF.ROC <- roc(predictor=RF.probs$OD, response=testData$Ocupacional,
          levels=rev(levels(testData$Ocupacional)))

#Cálculo de AUC
RF.ROC$auc

#Plotar Curva ROC
RF.plot<-plot(RF.ROC,xlab = "Especificidade", ylab = "Sensitividade", main = "RF", col =
"gold", lwd =5)

#Coleta dos Resultados
results<-resamples(list(RF= fit.RF, NN= fit.nn, XGB = fit.xgb, ADA = fit.ada, LRG = fit.lrg,
CAT = fit.CAT))

# Diferença entre os modelos de Predição
diffs<-diff(results)

# Resumo de p-valores para comparações pareadas
tb.diffs<-summary(diffs)

#Plotar todos os graficos ROC juntos
png("C:/Users/anaca/OneDrive/Área de Trabalho/Resultados Dissertação/ ROC.png",
width=600, height=600, units="px")
par(mfrow = c(3,3))
plot(RF.ROC,xlab = "Espec.", ylab = "Sensit.", main = "RF", col = "gold", lwd =3)
plot(xgb.ROC,xlab = "Espec.", ylab = "Sensit.", main = "XGB", col = "gold", lwd =3)
plot(CAT.ROC, xlab ="Espec.", ylab = "Sensit.", main = "CAT", col = "gold", lwd =3)
plot(ada.ROC,xlab = "Espec.", ylab = "Sensit.", main = "ADA", col = "gold", lwd =3)
plot(lrg.ROC,xlab = "Espec.", ylab = "Sensit.", main = "LRG", col = "gold", lwd =3)
plot(nn.ROC,xlab = "Espec.", ylab = "Sensit.", main = "NN", col = "gold", lwd =3)
dev.off()
```

#Comparar os modelos usando boxplot

```
png("C:/Users/anaca/OneDrive/Área de Trabalho/Resultados Dissertação/BOXPLOT.png",
width=547, height=357, units="px")
scales <- list(x=list(relation="free"), y=list(relation="free"))
bwplot(results, scales=scales)
dev.off()
```

 # Gráfico do Intervalo de Confiança

```
png("C:/Users/anaca/OneDrive/Área de Trabalho/Resultados Dissertação/ACURAC.png",
width=547, height=357, units="px")
scales <- list(x=list(relation="free"), y=list(relation="free"))
dotplot(results, scales=scales)
dev.off()
```

APÊNDICE B – Tabela de Correlação das Variáveis da Bateria Padrão Brasileira de Testes de Contato

Tabela B.1 – Correlação entre as Variáveis da Bateria Padrão

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1,000	0,056	0,076	0,034	0,303	0,042	0,080	0,049	0,123	0,014	0,080	0,241	0,098	0,146	0,175	-0,008	0,173	0,174	0,038	0,014	0,185	0,090	0,247	0,282	0,075	0,105	0,019	0,106	0,087	0,075
0,056	1,000	0,011	0,051	0,141	0,052	0,008	0,139	0,144	0,076	0,085	0,111	0,025	0,057	0,076	0,065	0,057	0,159	0,097	-0,020	0,028	0,065	0,122	0,036	0,010	0,033	0,055	0,032	0,036	0,035
0,076	0,011	1,000	0,045	0,180	0,001	0,094	0,092	0,055	0,058	0,112	0,143	0,042	0,079	0,266	-0,026	0,099	0,028	0,101	0,082	0,170	0,127	0,167	0,053	0,113	0,266	0,068	0,160	-0,021	0,132
0,034	0,051	0,045	1,000	0,102	0,217	0,138	0,033	0,048	0,061	0,163	0,079	0,038	0,025	0,084	0,158	0,070	0,087	0,076	-0,011	0,185	0,088	0,122	0,053	0,058	0,062	-0,006	0,011	0,040	0,059
0,303	0,141	0,180	0,102	1,000	0,117	0,069	0,123	0,267	0,064	0,187	0,972	0,224	0,315	0,374	0,008	0,368	0,085	0,110	0,065	0,134	0,204	0,213	0,240	0,176	0,235	0,098	0,235	0,074	0,176
0,042	0,052	0,001	0,217	0,117	1,000	0,051	0,042	0,027	0,037	0,059	0,092	0,033	0,099	0,050	0,055	0,038	0,060	0,036	0,043	0,050	0,006	0,191	0,152	0,031	0,084	0,001	0,113	-0,027	0,145
0,080	0,008	0,094	0,138	0,069	0,051	1,000	-0,042	-0,007	-0,006	0,118	0,052	0,012	0,082	0,015	-0,033	0,016	0,085	0,045	0,070	0,147	0,062	0,161	0,062	0,084	0,054	0,060	0,024	0,007	0,098
0,049	0,139	0,092	0,033	0,123	0,042	-0,042	1,000	0,032	0,043	0,112	0,097	0,021	0,134	0,284	0,182	0,050	0,105	0,166	0,132	0,022	0,094	0,135	0,072	0,007	0,185	0,104	0,229	-0,003	0,054
0,123	0,144	0,055	0,048	0,267	0,027	-0,007	0,032	1,000	0,050	0,069	0,213	0,084	0,127	0,154	-0,025	0,154	0,078	0,162	0,069	0,137	0,201	0,082	0,094	0,169	0,076	0,069	0,079	0,011	0,166
0,014	0,076	0,058	0,061	0,064	0,037	-0,006	0,043	0,050	1,000	0,012	0,049	0,066	0,012	0,026	0,065	0,014	0,024	0,106	0,137	0,039	0,010	0,011	-0,003	0,081	0,035	0,027	-0,002	0,015	-0,025
0,080	0,085	0,112	0,163	0,187	0,059	0,118	0,112	0,069	0,012	1,000	0,148	0,046	0,229	0,105	0,051	0,103	0,095	0,117	0,087	0,209	0,091	0,104	0,047	0,087	0,054	0,072	0,167	-0,010	0,174
0,241	0,111	0,143	0,079	0,972	0,092	0,052	0,097	0,213	0,049	0,148	1,000	0,178	0,251	0,298	0,005	0,294	0,065	0,086	0,038	0,105	0,162	0,169	0,176	0,140	0,187	0,070	0,187	0,050	0,140
0,098	0,025	0,042	0,038	0,224	0,033	0,012	0,021	0,084	0,066	0,046	0,178	1,000	0,101	0,126	-0,029	0,285	0,065	0,005	0,058	0,122	0,055	0,048	0,071	0,037	0,067	0,051	0,069	-0,023	0,041
0,146	0,057	0,079	0,025	0,315	0,099	0,082	0,134	0,127	0,012	0,229	0,251	0,101	1,000	0,182	0,107	0,180	0,016	0,157	0,014	0,052	0,093	0,097	0,112	0,077	0,109	0,023	0,288	0,108	0,077
0,175	0,076	0,266	0,084	0,374	0,050	0,015	0,284	0,154	0,026	0,105	0,298	0,126	0,182	1,000	0,054	0,215	0,132	0,161	0,083	0,053	0,284	0,241	0,101	0,098	0,344	0,119	0,344	0,023	0,099
-0,008	0,065	-0,026	0,158	0,008	0,055	-0,033	0,182	-0,025	0,065	0,051	0,005	-0,029	0,107	0,054	1,000	0,038	-0,012	0,038	0,020	0,058	0,027	-0,025	-0,024	0,013	-0,025	0,018	0,072	-0,065	0,025
0,173	0,057	0,099	0,070	0,368	0,038	0,016	0,050	0,154	0,014	0,103	0,294	0,285	0,180	0,215	0,038	1,000	0,145	0,125	0,013	0,232	0,099	0,119	0,124	0,098	0,135	0,078	0,088	0,075	0,075
0,174	0,159	0,028	0,087	0,085	0,060	0,085	0,105	0,078	0,024	0,095	0,065	0,065	0,016	0,132	-0,012	0,145	1,000	0,110	0,057	0,253	0,207	0,141	0,078	0,031	0,177	0,126	0,061	-0,010	0,086
0,038	0,097	0,101	0,076	0,110	0,036	0,045	0,166	0,162	0,106	0,117	0,086	0,005	0,157	0,161	0,038	0,125	0,110	1,000	0,094	0,033	0,108	0,089	0,060	0,012	0,133	-0,005	0,136	-0,032	0,048
0,014	-0,020	0,082	-0,011	0,065	0,043	0,070	0,132	0,069	0,137	0,087	0,038	0,058	0,014	0,083	0,020	0,013	0,057	0,094	1,000	-0,016	0,136	0,014	0,171	0,078	0,033	0,043	0,146	0,005	0,122
0,185	0,028	0,170	0,185	0,134	0,050	0,147	0,022	0,137	0,039	0,209	0,105	0,122	0,052	0,053	0,058	0,232	0,253	0,033	-0,016	1,000	0,226	0,120	0,139	0,052	0,137	0,173	0,024	0,015	0,112
0,090	0,065	0,127	0,088	0,204	0,006	0,062	0,094	0,201	0,010	0,091	0,162	0,055	0,093	0,284	0,027	0,099	0,207	0,108	0,136	0,226	1,000	0,159	0,187	0,130	0,182	0,074	0,295	0,028	0,132
0,247	0,122	0,167	0,122	0,213	0,191	0,161	0,135	0,082	0,011	0,104	0,169	0,048	0,097	0,241	-0,025	0,119	0,141	0,089	0,014	0,120	0,159	1,000	0,137	0,043	0,144	0,038	0,146	0,031	0,028
0,282	0,036	0,053	0,053	0,240	0,152	0,062	0,072	0,094	-0,003	0,047	0,176	0,071	0,112	0,101	-0,024	0,124	0,078	0,060	0,171	0,139	0,187	0,137	1,000	0,159	0,078	0,041	0,166	0,118	0,149
0,075	0,010	0,113	0,058	0,176	0,031	0,084	0,007	0,169	0,081	0,087	0,140	0,037	0,077	0,098	0,013	0,098	0,031	0,012	0,078	0,052	0,130	0,043	0,159	1,000	0,116	-0,024	0,155	-0,029	0,190
0,105	0,033	0,266	0,062	0,235	0,084	0,054	0,185	0,076	0,035	0,054	0,187	0,067	0,109	0,344	-0,025	0,135	0,177	0,133	0,033	0,137	0,182	0,144	0,078	0,116	1,000	0,063	0,201	-0,013	0,036
0,019	0,055	0,068	-0,006	0,098	0,001	0,060	0,104	0,069	0,027	0,072	0,070	0,051	0,023	0,119	0,018	0,078	0,126	-0,005	0,043	0,173	0,074	0,038	0,041	-0,024	0,063	1,000	0,100	0,034	0,069
0,106	0,032	0,160	0,011	0,235	0,113	0,024	0,229	0,079	-0,002	0,167	0,187	0,069	0,288	0,344	0,072	0,088	0,061	0,136	0,146	0,024	0,295	0,146	0,166	0,155	0,201	0,100	1,000	0,024	0,156
0,087	0,036	-0,021	0,040	0,074	-0,027	0,007	-0,003	0,011	0,015	-0,010	0,050	-0,023	0,108	0,023	-0,065	0,075	-0,010	-0,032	0,005	0,015	0,028	0,031	0,118	-0,029	-0,013	0,034	0,024	1,000	0,070
0,075	0,035	0,132	0,059	0,176	0,145	0,098	0,054	0,166	-0,025	0,174	0,140	0,041	0,077	0,099	0,025	0,075	0,086	0,048	0,122	0,112	0,132	0,028	0,149	0,190	0,036	0,069	0,156	0,070	1,000

Fonte: Autora (2021)

APÊNDICE C – Medidas de Posição das Técnicas de Mineração nos 2 Cenários

Tabela C.1 – Medidas de Posição das Técnicas de Mineração no Cenário 1

Acuracidade	**Min.**	**1º Quad.**	**Mediana**	**Média**	**3º Quad.**	**Máx.**	**NA's**
RF	0,733	0,880	0,926	0,914	0,959	1,000	0,000
NN	0,550	0,771	0,824	0,824	0,889	0,991	0,000
XGB	0,694	0,870	0,907	0,898	0,942	1,000	0,000
ADA	0,639	0,868	0,898	0,895	0,944	1,000	0,000
LRG	0,569	0,799	0,851	0,846	0,917	1,000	0,000
CAT	0,638	0,833	0,898	0,887	0,938	1,000	0,000
Sensitividade	**Min.**	**1º Quad.**	**Mediana**	**Média**	**3º Quad.**	**Máx.**	**NA's**
RF	0,500	0,750	0,833	0,847	0,917	1,000	0,000
NN	0,417	0,750	0,750	0,778	0,833	1,000	0,000
XGB	0,500	0,750	0,833	0,833	0,917	1,000	0,000
ADA	0,583	0,750	0,833	0,814	0,917	1,000	0,000
LRG	0,583	0,750	0,833	0,812	0,917	1,000	0,000
CAT	0,333	0,750	0,833	0,823	0,917	1,000	0,000
Especificidade	**Min.**	**1º Quad.**	**Mediana**	**Média**	**3º Quad.**	**Máx.**	**NA's**
RF	0,444	0,700	0,778	0,791	0,889	1,000	0,000
NN	0,333	0,600	0,778	0,724	0,800	1,000	0,000
XGB	0,333	0,692	0,778	0,793	0,889	1,000	0,000
ADA	0,444	0,778	0,889	0,814	0,889	1,000	0,000
LRG	0,333	0,667	0,778	0,729	0,822	1,000	0,000
CAT	0,333	0,667	0,778	0,766	0,889	1,000	0,000

Fonte: Autora (2021)

Tabela C.2 – Medidas de Posição das Técnicas de Mineração no Cenário 2

Acuracidade	**Min.**	**1º Quad.**	**Mediana**	**Média**	**3º Quad.**	**Máx.**	**NA's**
RF	0,731	0,889	0,938	0,927	0,973	1,000	0,000
NN	0,444	0,769	0,838	0,816	0,889	0,981	0,000
XGB	0,704	0,833	0,883	0,878	0,927	1,000	0,000
ADA	0,565	0,861	0,913	0,900	0,955	1,000	0,000
LRG	0,602	0,800	0,857	0,851	0,912	1,000	0,000
CAT	0,676	0,850	0,898	0,893	0,944	1,000	0,000
Sensitividade	**Min.**	**1º Quad.**	**Mediana**	**Média**	**3º Quad.**	**Máx.**	**NA's**
RF	0,583	0,833	0,917	0,877	0,917	1,000	0,000
NN	0,333	0,667	0,833	0,782	0,917	1,000	0,000
XGB	0,417	0,750	0,833	0,794	0,917	1,000	0,000
ADA	0,583	0,833	0,833	0,853	0,917	1,000	0,000
LRG	0,500	0,667	0,750	0,756	0,833	1,000	0,000
CAT	0,583	0,750	0,833	0,819	0,916	1,000	0,000
Especificidade	**Min.**	**1º Quad.**	**Mediana**	**Média**	**3º Quad.**	**Máx.**	**NA's**
RF	0,44	0,78	0,78	0,81	0,89	1,00	0,00
NN	0,40	0,67	0,70	0,72	0,80	1,00	0,00
XGB	0,30	0,67	0,78	0,76	0,89	1,00	0,00
ADA	0,33	0,70	0,80	0,80	0,89	1,00	0,00
LRG	0,44	0,70	0,79	0,80	0,89	1,00	0,00
CAT	0,44	0,67	0,78	0,77	0,89	1,00	0,00

Fonte: Autora (2021)

APÊNDICE D – Tabelas de Importância das Variáveis para os 2 Cenários

Tabela D.1 – Importância das Variáveis para o Cenário 1

Variável	CAT	RF
Mãos e Antebraço	100,00	100,00
Profissão	70,12	72,66
Mês	62,32	71,71
Etnia	53,23	31,42
Dermatite alérgica de contato	36,12	25,03
Idade	34,77	40,63
Pés e Coxas	30,45	20,60
Escolaridade	30,12	64,11
Face e Pescoço	24,08	28,86
Estação	22,39	26,93
Teste de luvas	22,38	21,63
Dermatite irritativa	20,82	52,27
Tórax e abdome	19,26	11,88
Sexo	13,33	24,58
Atopia	10,94	13,70
Líquen plano	7,49	9,99
Total de diagnósticos	5,96	14,70
Psoríase	1,66	5,37
Dermatite seborreica	1,28	4,23
Urticária de contato	1,15	0,84
Líquen simples crônico	0,70	1,08
Dia da semana	0,64	13,51
Síndrome pele excitada	0,00	4,59
Eczema numular	0,00	2,56
Dermatite atópica	0,00	2,40
Eczema de estase	0,00	0,00

Fonte: Autora (2021)

Tabela D.2 – Importância das Variáveis para o Cenário 2

Variável	CAT	RF	Variável	CAT	RF
Mãos e Antebraço	100,00	100,00	Urticária de contato	2,22	0,74
Profissão	83,46	95,20	Etilenodiamina	2,18	0,60
Teste de Luvas	45,80	11,04	Quaternium	2,07	1,00
Dermatite irritativa	44,26	50,86	Mercapto-Mix	2,04	0,09
Bicromato de Potássio	35,46	21,16	Timerosal	2,00	4,70
Estação	28,91	7,10	Parabenos-Mix	1,71	3,95
Idade	28,36	24,92	Antraquinona	1,65	0,71
Face e Pescoço	26,29	18,97	Resina Epóxi	1,52	0,92
Mês	23,65	76,28	Prometazina	1,37	0,49
Pés e Coxas	22,00	8,28	Colofônio	1,34	4,89
Etnia	14,01	15,74	Propilenoglicol	0,82	0,36
Sulfato de Níquel	11,18	9,63	PPD-Mix	0,81	2,02
Dia da Semana	10,84	7,14	Eczema de estase	0,70	0,31
Parafenilenodiamina	7,89	3,57	Eczema numular	0,69	0,04
Sexo	6,86	5,71	Perfure-Mix	0,64	1,75
Cloreto de Cobalto	6,83	3,29	Líquen simples crônico	0,60	2,44
Psoríase	6,60	1,38	Benzocaína	0,32	1,21
Neomicina	5,83	2,50	Lanolina	0,01	0,14
Tórax e Abdome	5,78	5,51	Escolaridade	0,00	43,92
Carba-Mix	4,12	0,90	Sínd. Pele Excitada	0,00	2,82
Formaldeído	3,12	2,35	Tiuram-Mix	0,00	1,32
Dermatite alérgica de contato	2,94	8,42	Dermatite atópica	0,00	1,24
Atopia	2,88	5,88	Nitrofurazona	0,00	0,82
Líquen plano	2,69	3,95	Quinolina	0,00	0,02
Dermatite seborréica	2,69	1,34	Hidroquinona	0,00	0,00
Bálsamo do Peru	2,39	5,13	Butilfenol	0,00	0,00
Total de Diagnósticos	2,30	9,22	Irgasan	0,00	0,00
Kathon CG	2,30	0,57	Terebintina	0,00	0,00

Fonte: Autora (2021)

ANEXO A – Benefícios Concedidos por Estado

Tabela AA.1 – Benefícios Concedidos por UF

Estado	Quantidade	%Freq.	Acumulado
Minas Gerais	405	18,27%	18,27%
São Paulo	367	16,55%	34,82%
Rio Grande do Sul	193	8,71%	43,53%
Santa Catarina	147	6,63%	50,16%
Rio de Janeiro	140	6,31%	56,47%
Bahia	103	4,65%	61,12%
Paraná	94	4,24%	65,36%
Maranhão	93	4,19%	69,55%
Distrito Federal	88	3,97%	73,52%
Pernambuco	71	3,20%	76,73%
Ceará	66	2,98%	79,70%
Piauí	57	2,57%	82,27%
Goiás	49	2,21%	84,48%
Rio Grande do Norte	47	2,12%	86,60%
Pará	44	1,98%	88,59%
Paraíba	39	1,76%	90,35%
Espírito Santo	38	1,71%	92,06%
Mato Grosso do Sul	37	1,67%	93,73%
Mato Grosso	30	1,35%	95,08%
Alagoas	23	1,04%	96,12%
Rondônia	22	0,99%	97,11%
Sergipe	19	0,86%	97,97%
Amazonas	15	0,68%	98,65%
Tocantins	10	0,45%	99,10%
Amapá	7	0,32%	99,41%
Roraima	7	0,32%	99,73%
Acre	6	0,27%	100,00%
Total	2.217		100,00%

Fonte: Comitê de Dados Abertos do INSS (2020)

ANEXO B – Questionário Nórdico de Doenças Ocupacionais Relativas à Pele

Nome _____

G1 Local de Trabalho _____

G1 Departamento _____

G2 Sexo Masculino( ) Feminino( )

G3 Ano de Nascimento _____

G4 No momento você está (marque somente 1 alternativa) Empregado () Autônomo ( ) Doméstico ( ) Desempregado ( ) Estudante ( ) Aprendiz ( ) Licença Maternidade/ Paternidade ( ) Aposentado ou pensionista ( ) Outro _____

G5 Qual é sua ocupação no momento? _____

G5 Desde que ano você exerce essa ocupação? _____

G6 Qual é a sua principal atividade no trabalho? _____

G6 Desde que ano você exerce essa atividade? _____

G7 Quantas horas por semana em média você trabalha?

G8 Você exerce outro trabalho remunerado? Não ( ) Sim ( )
Qual?_____

A1 Você tem alguma erupção na pele que vem e vai há pelo menos 6 meses? Não ( ) Sim ( ) Não sei ( )

A2 Você já teve febre do feno ou outros sintomas de alergia nasal depor exemplo pólens ou animais? Não ( ) Sim ( ) Não sei ( )

A3 Seus olhos sempre apresentam sintomas alérgicos por exemplo de pólens ou animais?
Não ( ) Sim ( ) Não sei ( )

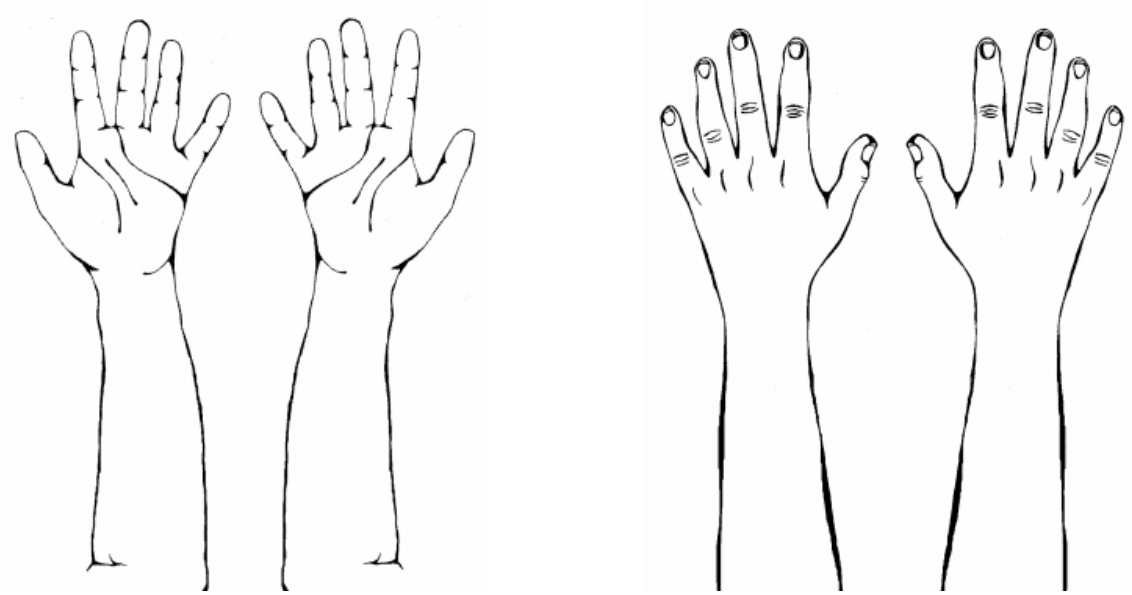
A4 Você já teve asma alguma vez? Não ( ) Sim ( )

A4 Se já teve asma, já foi diagnosticada por um médico? Quando? Não ( ) Sim ( )
Ano _____

D1 Você já teve eczema de mãos? Não ( ) Sim ( )

D2 Você já teve eczema nos pulsos ou antebraços (excluindo a frente dos cotovelos)?
Não ( ) Sim ( )

D3 Marque as áreas onde você comumente apresenta eczema de mãos



D4 Com que frequência você apresenta eczema de mãos?
Apenas uma vez e por menos de duas semanas ( ) Apenas uma vez e por duas semanas ou mais ( ) Mais de uma vez ( ) Em todo o tempo ( )

D4 Com que frequência você apresenta eczema no pulso/antebraço?
Apenas uma vez e por menos de duas semanas ( ) Apenas uma vez e por duas semanas ou mais ( ) Mais de uma vez ( ) Em todo o tempo ( )

D5 Qual foi a última vez que você teve eczema de mãos?
Estou neste momento ( ) Não agora, mas há menos de 3 meses ( )
Entre 3 e 12 meses ( ) Mais de 12 meses ( ) Em que ano? _____

D5 Qual foi a última vez que você teve eczema no pulso/antebraço?
Estou neste momento ( ) Não agora, mas há menos de 3 meses ( )
Entre 3 e 12 meses ( ) Mais de 12 meses ( ) Em que ano? _____

D6 Quando você apresentou seu primeiro eczema nas mãos?
Até 6 anos de idade ( ) Entre 6 e 14 anos ( ) Entre 15 e 18 anos ( ) Acima de 18 anos ( ) Em que ano? _____

D6 Quando você apresentou seu primeiro eczema no pulso/antebraço?
Até 6 anos de idade ( ) Entre 6 e 14 anos ( ) Entre 15 e 18 anos ( )
Acima de 18 anos ( ) Em que ano? _____

D7 Você conhece a causa do seu eczema de mãos? Não ( ) Sim ( )
Causa_____

D7 Você conhece a causa do seu eczema no pulso/ antebraço? Não ( ) Sim ( )
Causa_____

D8 Qual era o seu trabalho quando o eczema começou? _____

D9 Quais eram as principais atividades no trabalho que você exercia quando o eczema começou? _____

D10 Você se consultou com um médico quando o eczema de mãos começou? Não ( ) Sim ( ) Ano _____

D10 Você se consultou com um médico quando o eczema no pulso/antebraço começou?
Não ( ) Sim ( ) Ano _____

D11 Em qual estação do ano você apresenta mais problemas de eczema nas mãos?
Sem distinção entre as estações ( ) Inverno ( ) Primavera ( ) Verão ( ) Outono ( )

D11 Em qual estação do ano você apresenta mais problemas de eczema no pulso/antebraço?
Sem distinção entre as estações ( ) Inverno ( ) Primavera ( ) Verão ( ) Outono ( )

D12 Numa escala de 0 a 10 onde 10 significa extremamente ruim e 0 sem eczema como você classifica seu eczema hoje?

D12 Numa escala de 0 a 10 onde 10 significa extremamente ruim e 0 sem eczema como você classifica seu eczema no pior momento?

F1 Você teve contato com certos tipos de materiais, químicos ou qualquer coisa no seu trabalho que faça o eczema piorar nas mãos? Não ( ) Sim ( ) Causa_____

F1 Você teve contato com certos tipos de materiais, químicos ou qualquer coisa no seu trabalho que faça o eczema piorar no pulso e antebraço? Não ( ) Sim ( )
Causa_____

F2 Você teve contato com certos tipos de materiais, químicos ou qualquer coisa fora do trabalho que faça o eczema piorar nas mãos? Não ( ) Sim ( ) Causa_____

F2 Você teve contato com certos tipos de materiais, químicos ou qualquer coisa fora do trabalho que faça o eczema piorar no pulso e antebraço? Não ( ) Sim ( )
Causa_____

F3 O que você considera como sendo mais importantes dentro do seu ambiente de trabalho que fazem seu eczema nas mãos piorar? Marque até 5 opções
sabonete, sabonete líquido, xampu e outros produtos de higiene pessoal ( )
detergentes e outros produtos de limpeza e lavanderia domésticos ( )
manipulação de alimentos ( ) trabalhar com as mãos molhadas ( )
lavar as mãos com frequência ( ) luvas de proteção ( )
manutenção de máquinas (manuseio de óleos) ( )
trabalho na construção, pintura, colocação de papel de parede, reforma e decoração ( )
jardinagem, manuseio de plantas, soja, vegetais, frutas. Etc.

infecções (resfriado, gripe ou febre) ( ) humor, estresse ( )

F3 O que você considera como sendo mais importantes fora do seu ambiente de trabalho que fazem seu eczema no pulso/antebraço piorar? Marque até 5 opções
sabonete, sabonete líquido, xampu e outros produtos de higiene pessoal ( )
detergentes e outros produtos de limpeza e lavanderia domésticos ( )
manipulação de alimentos ( ) trabalhar com as mãos molhadas ( )
lavar as mãos com frequência ( ) luvas de proteção ( )
manutenção de máquinas (manuseio de óleos) ( )
trabalho na construção, pintura, colocação de papel de parede, reforma e decoração ( )
jardinagem, manuseio de plantas, soja, vegetais, frutas. Etc.
infecções (resfriado, gripe ou febre) ( ) humor, estresse ( )

F4 Seu eczema nas mãos melhora quando você fica afastado do seu local de trabalho (por semanas ou longos períodos)?
Não ( ) Sim às vezes ( ) Sim sempre ( ) Não sei ( )

F4 Seu eczema no pulso/antebraço melhora quando você fica afastado do seu local de trabalho (por semanas ou longos períodos)?
Não ( ) Sim às vezes ( ) Sim sempre ( ) Não sei ( )

C1 O eczema em suas mãos, pulsos ou antebraços afetou de alguma forma suas atividades diárias em sua ocupação? Qual das seguintes afirmações são verdadeiras:
( ) Por conta do meu eczema eu tenho que usar luvas de proteção
( ) Por conta do meu eczema minhas tarefas de trabalho foram alteradas
( ) Por conta do meu eczema mudei de emprego
( ) Por conta do meu eczema tenho tido dificuldades em conseguir um emprego
( ) Por conta do meu eczema meus colegas de trabalho ou empregador (es) têm uma atitude negativa em relação a mim
( )Por conta do meu eczema minha escolha de trabalho ou ocupação foi afetada
( )Por conta do meu eczema minha renda diminuiu por conta do meu eczema Eu estive doente ou fora do trabalho
( )Por conta do meu eczema me aposentei
( )Por conta do meu eczema perdi um trabalho
( )Outras consequências _____

C2 Como o eczema afetou sua vida nos últimos 12 meses? (Responda numa escala de 0 a 4 onde 0 = sem relevância; 1 = não afetou; 2 = pouco afetou; 3 = afetou moderadamente; 4 = afetou muito) vida ocupacional ( ) atividades domésticas ( ) esportes e atividades similares ( ) outros hobbies ou atividades ( ) sono ( ) viagens ( ) atividades sociais ( ) relacionamentos próximos ( )vida sexual ( ) humor ( )

C3 A existência do eczema atrapalhou sua vida financeira? Não tive efeitos financeiros ( ) Tive efeitos mas não o suficiente para atrapalhar minha vida substancialmente ( )Tive percas razoáveis ( ) Perca financeira substancial ( )

U1 Você já teve pápulas com coceira aparecendo e desaparecendo rapidamente (em poucas horas) nas mãos, pulsos ou antebraços (urticária ou erupção cutânea com urtiga)?
Não ( ) Sim ( )

U2 Essas pápulas com coceira (urticária) nas mãos, pulsos ou antebraços foram causadas pelo contato da pele com frutas, vegetais, luvas de borracha, animais, etc.? Não ( ) Sim ( ) Depois do contato com a pele, qual? _____

U3 Com que frequência você teve essas pápulas com coceira (urticária) nas mãos, pulsos ou antebraços? Marque apenas 1 Uma vez ( ) De 2 a 5 vezes ( ) Mais de 5 vezes ( )

U4 Quando foi a última vez que você teve essas pápulas com coceira (urticária) nas mãos, pulsos ou antebraços? Marque apenas 1 Durante os 7 últimos dias ( ) Há mais de 7 dias e menos de 3 meses ( ) De 3 a 12 meses ( ) Um ano atrás ( ) Ano _____

U5 Quando você teve urticária nas mãos, pulsos ou antebraços a primeira vez? Até 6 anos de idade ( ) Entre 6 e 14 anos ( ) Entre 15 e 18 anos ( ) Acima de 18 anos ( ) Em que ano? _____

U6 Qual era o seu trabalho quando a urticária começou? _____

U7 Quais eram as principais atividades no trabalho que você exercia quando a urticária começou? _____

U8 Você se consultou com um médico quando a urticária começou? Não ( ) Sim ( ) Ano _____

U9 Numa escala de 0 a 10 onde 10 significa extremamente ruim e 0 sem urticária como você classifica sua urticária hoje? _____

U9 Numa escala de 0 a 10 onde 10 significa extremamente ruim e 0 sem urticária como você classifica sua urticária no pior momento? _____

S1 Você teve algum dos seguintes sintomas nas mãos nos últimos 12 meses?
( ) nenhum sintoma durante os últimos 12 meses
( ) vermelhidão
( ) pele seca com escamação / descamação
( ) fissuras ou rachaduras, chorando ou crostas
( ) pequenas bolhas de água (vesículas) pápulas
( ) Pápulas / vergões com coceira de aparecimento rápido (urticária)
( ) coceira ardor, formigamento ou picada
( ) ternura

S1 Você teve algum dos seguintes sintomas no pulso / antebraços nos últimos 12 meses?
( ) nenhum sintoma durante os últimos 12 meses
( ) vermelhidão
( ) pele seca com escamação / descamação
( ) fissuras ou rachaduras, chorando ou crostas
( ) pequenas bolhas de água (vesículas) pápulas
( ) Pápulas / vergões com coceira de aparecimento rápido (urticária)
( ) coceira ardor, formigamento ou picada
( ) ternura

S2 Você fica com erupção na pele por causa de botões de metal, fechos de metal, bijuterias de metal (por exemplo, brincos) ou outros objetos de metal próximos à pele?

Não ( ) Sim ( )

S3 Você tem pele seca? Não ( ) Sim ( )

S4 Sua pele coça quando sua? Não ( ) Sim ( )

T1 Algum médico já te diagnosticou com alergia? Não ( ) Sim ( ) Não sei ( )

T2 Especifique suas alergias _____

T3 A alergia foi diagnosticada com
Testes de contato ( ) Testes cutâneos de picada ( ) Exames de sangue ( )
Outros _____ Não sei ( )
E1 Você usa ou já usou luvas no seu trabalho?
Não, nunca ( ) Sim, todos os dias ( ) Sim, mas não no momento ( )

E2 Que tipo de luvas você usa no trabalho?
( ) borracha natural de látex
( ) borracha sintética (por exemplo, nitrila, neoprene etc.)
( ) plástico (por exemplo, vinil, PVC, polietileno)
( ) luvas de algodão por baixo de luvas de borracha ou plástico
( ) couro
( ) pano
( ) outro _____
( ) não sei

E2 Que tipo de luvas você usou no trabalho?
( ) borracha natural de látex
( ) borracha sintética (por exemplo, nitrila, neoprene etc.)
( ) plástico (por exemplo, vinil, PVC, polietileno)
( ) luvas de algodão por baixo de luvas de borracha ou plástico
( ) couro
( ) pano
( ) outro o quê? _____
( ) não sei

E3 Você teve sintomas de pele como resultado do uso de luvas de proteção?
Não ( ) Sim ( )

E4 Você trocou de luvas ou parou de usar por conta dos sintomas na pele?
Não ( ) Sim ( )

E5 Quanto tempo em horas você está fazendo ou lidando com seu trabalho nas atividades a seguir no momento?
( ) trabalho molhado
( ) preparar comida / manusear comida
( ) plantas
( ) animais
( ) agentes de limpeza

( ) solventes
( ) óleos, fluidos de corte, etc.
( ) tintas, lacas, revestimentos, etc.
( ) colas, adesivos, etc.
( ) selantes, massa, gesso, agentes de piso, cimento etc.
( ) pó (pó de madeira, pó de moagem, pó de papel etc.)

E6 Quantas horas por dia você gasta preparando a comida fora do seu trabalho?
0 ( ) menos de 30 min. ( ) de 30 min. A 2 horas ( ) mais de 2 horas ( )

E6 Quantas horas por dia você gasta limpando ou lavando fora do seu trabalho?
0 ( ) menos de 30 min. ( ) de 30 min. A 2 horas ( ) mais de 2 horas ( )

E6 Quantas horas por dia você gasta cuidando de crianças menores que 4 anos fora do seu trabalho?
0 ( ) menos de 30 min. ( ) de 30 min. A 2 horas ( ) mais de 2 horas ( )

E7 Com que frequência você realizou jardinagem nos últimos 12 meses?
Diariamente ( ) Uma vez por semana ( ) Uma vez por mês ( ) Menos de uma vez por mês ( ) Esporadicamente ( )

E7 Com que frequência você realizou reparo no carro nos últimos 12 meses?
Diariamente ( ) Uma vez por semana ( ) Uma vez por mês ( ) Menos de uma vez por mês ( ) Esporadicamente ( )

E7 Com que frequência você realizou reparo construção nos últimos 12 meses?
Diariamente ( ) Uma vez por semana ( ) Uma vez por mês ( ) Menos de uma vez por mês ( ) Esporadicamente ( )

E7 Com que frequência você realizou outras atividades com exposição a químicos nos últimos 12 meses?
Diariamente ( ) Uma vez por semana ( ) Uma vez por mês ( ) Menos de uma vez por mês ( ) Esporadicamente ( )

E8 Quantas vezes você lava as suas mãos durante um dia normal de trabalho?
0 - 5 vezes por dia ( )0 - 5 vezes por dia ( )6 - 10 vezes por dia ( ) 11 - 20 vezes por dia ( ) mais de 20 vezes por dia ( )

H1 O que você diria no geral da sua saúde considerando sua idade?
Excelente ( ) Muito boa ( ) Boa ( ) Razoável ( ) Pobre ( )

H2 Quantas pessoas vivem na sua casa incluindo você?

H2 Quantas crianças menores de 4 anos vivem na sua casa?

ANEXO C – Parecer Consubstanciado do Comitê de Ética em Pesquisa

UNIVERSIDADE ESTADUAL DE MARINGÁ

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: AVALIAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS EM UMA BASE DE DADOS DE DERMATITE OCUPACIONAL OBTIDA A PARTIR DE UM SERVIÇO ESPECIALIZADO

Pesquisador: EDWIN VLADIMIR CARDOZA GALDAMEZ

Área Temática:

Versão: 2

CAAE: 25497719.0.0000.0104

Instituição Proponente: CTC - Centro de Tecnologia

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 3.794.265

Apresentação do Projeto:

Trata-se de resposta à pendência de projeto de pesquisa proposto por pesquisador vinculado à Universidade Estadual de Maringá.

Objetivo da Pesquisa:

O objetivo deste projeto é comparar técnicas de mineração de dados para extração de conhecimento a partir dos dados disponíveis com potencial para o eventual estabelecimento de ações preventivas à saúde do trabalhador.

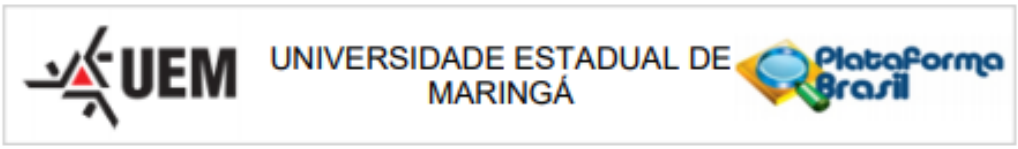
Avaliação dos Riscos e Benefícios:

Avalia-se que os possíveis riscos a que estarão submetidos os sujeitos da pesquisa serão suportados pelos benefícios apontados.

Comentários e Considerações sobre a Pesquisa:

A metodologia escolhida para a condução da pesquisa é pautada no processo KDD (Knowledge Discovery in Databases) que em português significa descoberta de conhecimento em base de dados e, em tal metodologia a mineração de dados é uma etapa constituinte do processo de descoberta. Metodologia Proposta: 1. Seleção: a seleção consiste na escolha do banco de dados a ser utilizado, que, conforme apresentado anteriormente, será o disponibilizado pela Fiocruz referente ao Serviço Especializado em doenças do Trabalho; 2. Pré-processamento: Implica na

Endereço: Av. Colombo, 5790, UEM-PPG, sala 4
Bairro: Jardim Universitário **CEP:** 87.020-900
UF: PR **Município:** MARINGA
Telefone: (44)3011-4597 **Fax:** (44)3011-4444 **E-mail:** copep@uem.br



exclusão dos dados que não se referem a doenças do trabalho, ou possam ser inconclusivos (se relacionados ou não com o trabalho); 3. Transformação: a etapa de transformação diz respeito à ordenação das informações, conforme requerido pela técnica a ser utilizada na etapa de mineração; 4. Mineração: na fase de mineração serão escolhidas 4 técnicas de mineração de dados, e, cada técnica será aplicada nos dados transformados; 5. Interpretação: durante a etapa de interpretação, as técnicas de mineração serão comparadas em termos de critérios de desempenho. A análise dos resultados será medida em termos de critérios de desempenho, como acurácia preditiva e custo computacional. Todas as etapas serão documentadas e, a síntese de cada uma delas será disposta à sociedade na forma de publicações.

Considerações sobre os Termos de apresentação obrigatória:

Apresenta Informações Básicas do Projeto; Termo de anuência da FIOCRUZ, assinado pela Drª Maria das Graças Mota Melo, dermatologista responsável (CRM 5235468-7); Dicionário de dados; Resposta a pendência documental; Termo de Confidencialidade; Solicitação de dispensa do TCLE; Folha de rosto assinada por Gislaine Leal, chefe do Dep. Engenharia de Produção; Projeto de pesquisa; Cronograma de execução. Justifica a dispensa do TCLE, já que na pesquisa serão utilizados somente dados obtidos a partir do estudo de material (dados) já coletados e de investigação de prontuários do serviço especializado com as informações referentes aos pacientes.

Conclusões ou Pendências e Lista de Inadequações:

O Comitê Permanente de Ética em Pesquisa Envolvendo Seres Humanos da Universidade Estadual de Maringá é de parecer favorável à aprovação do protocolo de pesquisa apresentado.

Considerações Finais a critério do CEP:

Face ao exposto e considerando a normativa ética vigente, este Comitê se manifesta pela aprovação do protocolo de pesquisa em tela. Alerta-se a respeito da necessidade de apresentação de relatório final no prazo de 30 dias após o término do projeto.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1459451.pdf	12/12/2019 17:08:27		Aceito
Outros	ANEXO1_TERMO_DE_ANUENCIA.pdf	12/12/2019 17:06:50	ANA CAROLINE FRANCISCO DA ROSA	Aceito
Outros	Respostas_a_pendencias_Plataforma_	12/12/2019	ANA CAROLINE	Aceito

Endereço: Av. Colombo, 5790, UEM-PPG, sala 4
Bairro: Jardim Universitário **CEP:** 87.020-900
UF: PR **Município:** MARINGA
Telefone: (44)3011-4597 **Fax:** (44)3011-4444 **E-mail:** copep@uem.br

UNIVERSIDADE ESTADUAL DE MARINGÁ

Continuação do Parecer: 3.794.265

Outros	Brasil.docx	17:03:57	FRANCISCO DA ROSA	Aceito
Outros	Dicionario_de_dados.xlsx	07/11/2019 18:55:45	ANA CAROLINE FRANCISCO DA ROSA	Aceito
Outros	RESPOSTA.docx	07/11/2019 18:44:43	ANA CAROLINE FRANCISCO DA ROSA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TERMO_DE_CONFIDENCIALIDADE_doc.docx	07/11/2019 18:41:32	ANA CAROLINE FRANCISCO DA ROSA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_doc.doc	07/11/2019 18:40:34	ANA CAROLINE FRANCISCO DA ROSA	Aceito
Folha de Rosto	Folha_de_Rosto.pdf	29/10/2019 12:06:47	ANA CAROLINE FRANCISCO DA ROSA	Aceito
Projeto Detalhado / Brochura Investigador	Projeto_de_Pesquisa.pdf	25/10/2019 12:33:52	ANA CAROLINE FRANCISCO DA ROSA	Aceito
Cronograma	CRONOGRAMA.pdf	25/10/2019 12:33:27	ANA CAROLINE FRANCISCO DA ROSA	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

MARINGA, 06 de Janeiro de 2020

Assinado por:
Ricardo Cesar Gardiolo
(Coordenador(a))

Endereço: Av. Colombo, 5790, UEM-PPG, sala 4
Bairro: Jardim Universitário **CEP:** 87.020-900
UF: PR **Município:** MARINGA
Telefone: (44)3011-4597 **Fax:** (44)3011-4444 **E-mail:** copep@uem.br