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

BEATRIZ LAVEZO DOS REIS

Análise exploratória e aplicação de técnicas de classificação em dados de acidentes de trabalho da indústria de transformação

Maringá
2021

BEATRIZ LAVEZO DOS REIS

Análise exploratória e aplicação de técnicas de classificação em dados de acidentes de trabalho da indústria de transformação

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

Orientadora: Profa. Dra. Gislaine Camila Lapasini Leal
Coorientador: Prof. Dr. Rodrigo Clemente Thom de Souza

Maringá
2021

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

R375a

Reis, Beatriz Lavezo dos

Análise exploratória e aplicação de técnicas de classificação em dados de acidentes de trabalho da indústria de transformação / Beatriz Lavezo dos Reis. -- Maringá, PR, 2021.
114 f.: il. color., tabs.

Orientadora: Profa. Dra. Gislaine Camila Lapasini Leal.
Coorientador: Prof. Dr. Rodrigo Clemente Thom de Souza.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Engenharia de Produçao, Programa de Pós-Graduação em Engenharia de Produção, 2021.

1. Mineração de dados. 2. Segurança e saúde no trabalho. 3. Inteligência artificial. 4. Comunicação de acidentes de trabalho. I. Leal, Gislaine Camila Lapasini , orient. II. Souza, Rodrigo Clemente Thom de, coorient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Engenharia de Produçao. Programa de Pós-Graduação em Engenharia de Produção. IV. Título.

CDD 23.ed. 620.0044

Síntique Raquel Eleutério - CRB 9/1641

FOLHA DE APROVAÇÃO

BEATRIZ LAVEZO DOS REIS

Análise exploratória e aplicação de técnicas de classificação em dados de acidentes de trabalho da indústria de transformação

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

Profa. Dra. Orientadora Gislaine Camila Lapasini Leal
Universidade Estadual de Maringá – DEP/UEM

Prof. Dr. Coorientador Rodrigo Clemente Thom de Souza
Universidade Estadual de Maringá – DEP/UEM
Universidade Federal do Paraná – UFPR/Jandaia do Sul

Prof. Dr. Rafael Henrique Palma Lima
Universidade Estadual de Maringá – DEP/UEM

Documento assinado digitalmente
gov.br
Denise Fukumi Tsunoda
Data: 20/09/2021 16:33:16-0300
Verifique em https://verificador.iti.br

Profa. Dra. Denise Fukumi Tsunoda
Universidade Federal do Paraná – PPGGI/UFPR

Aprovada em: 28 de junho de 2021.
Local da defesa: meet.google.com/pew-ppkq-jor, conforme, PORTARIA CAPES Nº 36, DE 19 DE MARÇO DE 2020 e Ofício Circular nº 10/2020-DAV/CAPES.

DEDICATÓRIA(S)

Aos meus pais, Marcia e Reinaldo, que sempre acreditaram em mim, apoiaram e continuam apoiando cada passo da minha caminhada.

AGRADECIMENTO(S)

Agradeço primeiramente a Deus que me sustentou durante essa caminhada e direcionou meus passos, sem a Sua força teria sido impossível.

Aos meus pais que acreditaram e apoiaram essa trajetória, desde quando era apenas um sonho.

Ao Pedro que esteve ao meu lado, dando força e apoio em todos os momentos, sendo fundamental para me manter firme.

À professora e orientadora Camila que acreditou em mim desde o início, dando suporte em todas as etapas desse processo e oportunidades de desenvolvimento em muitos âmbitos.

Ao professor e coorientador Rodrigo, por me direcionar nos caminhos da mineração de dados e ser fundamental para meu desenvolvimento.

Aos demais professores do Programa de Pós-Graduação em Engenharia de Produção, por auxiliarem na minha formação. Em especial à professora Daniele Cristina Tita Granzotto (em memória), por ter cultivado a semente da análise de dados e por ser exemplo de docência para mim.

Aos colegas de turma, de grupo de pesquisa e em especial aos bolsistas, que acompanharam cada passo e tornaram o dia a dia mais leve.

Aos demais amigos que me acompanharam nessa caminhada e me apoiaram, mesmo que de longe.

À CAPES pelo apoio financeiro e por possibilitar a realização da pesquisa.

EPÍGRAFE

Assim, fixamos os olhos,
não naquilo que se vê,
mas no que não se vê,
pois o que se vê é transitório,
mas o que não se vê é eterno.

(2 Coríntios 4, 18)

Análise exploratória e aplicação de técnicas de classificação em dados de acidentes de trabalho da indústria de transformação

RESUMO

Preocupar-se com a saúde e segurança do trabalhador é uma ação necessária, que impacta organizações públicas, privadas e principalmente a vida do trabalhador. Ao redor do mundo são registrados anualmente milhões de acidentes, doenças e óbitos em função do trabalho ou ambiente ao qual o trabalhador está inserido. Estas ocorrências abalam a vida do trabalhador e seus familiares, assim como afetam a produtividade das organizações, seus recursos materiais e financeiros e sua imagem perante o mercado. Diante deste contexto, torna-se fundamental avaliar e se aprofundar no cenário de saúde e segurança do trabalho a fim de buscar alternativas que reduzam ou cessem esses impactos. Este estudo busca prever a ocorrência de óbitos do trabalhador, utilizando dados de Comunicação de Acidente de Trabalho na indústria de transformação brasileira. Para este fim, a pesquisa é desenvolvida em três fases, primeiro um estudo exploratório, seguido pelo desenvolvimento e concluindo com o refinamento na pesquisa. Além disso, o trabalho está estruturado com a apresentação de dois artigos: um descrevendo a análise exploratória dos dados selecionados e outro com a aplicação de mineração de dados e inteligência artificial explicável. O primeiro artigo descreve o processo de seleção, limpeza e pré-processamento dos dados, com a indicação de categorias mais recorrentes nos registros, tais como sexo masculino, idade entre 26 e 35 anos e indústria de transformação como setor industrial com maior número de ocorrências. Também é desenvolvida uma análise exploratória específica para a indústria de transformação, além da categorização e preparação do conjunto de dados para a etapa seguinte. Posteriormente, no segundo artigo desta pesquisa, realiza-se a mineração de dados, utilizando doze técnicas e cinco métricas para comparação de seus resultados. Dessa etapa indica-se as duas técnicas com melhores resultados: *Naïve Bayes* e *Random Forest*. Estas, por sua vez, destinam-se à aplicação de um algoritmo de inteligência artificial explicável, o *SHapley Additive exPlanations* (SHAP), que apontou os atributos de parte do corpo atingida, natureza da lesão e agente causador, como os mais significativos para a decisão dos modelos. Esta pesquisa, devido ao seu cunho multidisciplinar, apresenta contribuições ao meio científico, mas não apenas a ele, para as organizações e sociedade fornece conhecimento sobre a saúde e segurança. Os resultados da pesquisa envolvem a síntese do cenário de acidentes e doenças do trabalho no Brasil, assim como apresenta uma classificação e indicação dos fatores com maior relação ao óbito na

indústria de transformação.

Palavras-chave: Mineração de dados. Saúde e segurança do trabalho. Comunicação de Acidentes de Trabalho. Inteligência Artificial Explicada.

Exploratory analysis and application of classification techniques in occupational accident data from the manufacturing industry

ABSTRACT

Worrying about the health and safety of workers is a necessary action, which impacts public and private organizations and especially society. Around the world, millions of accidents, diseases and deaths are registered annually due to the work or environment in which the worker is inserted. These occurrences affect the workers lives and their families, as well as affect the organizations productivity, their material and financial resources and their image on the market. In this context, it is essential to assess and deepen the occupational health and safety scenario in order to seek alternatives that reduce or stop these impacts. Based on this motivation, the present study seeks to predict the worker deaths, using data from the Work Accident Report in the Brazilian manufacturing industry. To this end, the research is developed in three phases, first an exploratory study, followed by development and concluding with research refinement. In addition, the work is structured with the presentation of two articles: one describing the exploratory analysis of selected data and the other with the data mining application and explainable artificial intelligence. The first article describes the data selection process, cleaning and pre-processing, indicating the most recurrent categories in the records, such as male gender, age between 26 and 35 years and manufacturing industry as the industrial sector with the highest number of occurrences. A specific exploratory analysis for the manufacturing industry is also developed, in addition to the categorization and dataset preparation for the next step. Later, in the second article of this research, data mining is performed, using twelve techniques and five metrics to compare their results. From this stage, the two techniques with the best results are indicated: Naïve Bayes and Random Forest. These, in turn, are intended to apply an explainable artificial intelligence algorithm, SHapley Additive exPlanations (SHAP) which highlights the attributes of the affected body part, injury nature and causative agent, as the most significant for the model decision. This research, due to its multidisciplinary nature, presents contributions to the scientific community, but not only to it, also for organizations and society, it provides knowledge about health and safety. The research results involve a synthesis of the scenario of occupational accidents and diseases in Brazil, as well as presenting a classification and indication of the factors most related to death in the manufacturing industry.

***Keywords*:** Data mining. Occupational safety and health. Artificial Intelligence Explained.

LISTA DE QUADROS E TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	*Associação Brasileira de Normas Técnicas*
AM	*Aprendizado de Máquina*
AUC	*Area Under Curve*
BA	*Bagging*
CAT	*Comunicação de Acidente de Trabalho*
CBO	*Classificação Brasileira de Ocupações*
CID	*Código Internacional de Doenças*
CNAE	*Classificação Nacional de Atividades Econômicas*
CONCLA	*Comissão Nacional de Classificação*
CSV	*Comma-Separated Values*
DATAPREV	*Empresa de Tecnologia e Informações da Previdência*
DSPC	*Dipartimento Servizi Alla persona e Alla Comunità*
DT	*Decision Trees*
EPC	*Equipamento de Proteção Coletiva*
EPI	*Equipamento de Proteção Individual*
ET	*Extra Trees*
FN	*False Negative*
FP	False Positive
IA	*Inteligência Artificial*
IBGE	*Instituto Brasileiro de Geografia e Estatística*
ILO	*International Labour Organization*
INAIL	*Istituto nazionale per l'assicurazione contro gli infortuni sul lavoro*
INSS	*Instituto Nacional do Seguro Social*
KDD	*Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbors*
LIME	*Local Interpretable Model-agnostic Explanations*
LR	*Logistic Regression*
MD	*Mineração de Dados*
MTE	*Ministério do Trabalho e Emprego*
NB	*Naïve Bayes*
NN	*Neural Networks*
NR	*Norma Regulamentadora*
OIT	*Organização Internacional do Trabalho*
OSHA	*Occupational Safety and Health Administration*

PRISMA	*Projeto de Regionalização de Informações e Sistemas*
RAIS	*Relação Anual de Informações Sociais*
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
SHAP	*SHapley Additive exPlanations*
SIM	*Sistema de Informação sobre Mortalidade*
SINAN	*Sistema de Informação de Agravos de Notificação*
SST	*Saúde e Segurança do Trabalho*
ST	*Stacking*
SUS	*Sistema Único de Saúde*
SVM	*Support Vector Machine*
TN	*True Negative*
TP	*True Positive*
UF	*Unidade Federativa*
UTI	*Unidade de Terapia Intensiva*
VO	*Voting*
XAI	*eXplainable Artificial Intelligence*
XGB	*XGBoost*

SUMÁRIO

1

INTRODUÇÃO

A seção inicial desta pesquisa apresenta o contexto de sua realização, conectando com as justificativas e contribuições a partir de sua execução e dos resultados. Na introdução também é apresentada a questão de pesquisa e os objetivos que serão alcançados, bem como uma síntese dos artigos e capítulos da dissertação.

1.1 RELEVÂNCIA DO TEMA

Acidentes, doenças e óbitos ligados ao trabalho são motivos de atenção em organizações privadas e governamentais. Cada vez mais preocupar-se com a Saúde e Segurança do Trabalho (SST) têm sido fator decisivo na gestão, concorrência do mercado e reputação das empresas. Isto porque, segundo a *International Labour Organization* (ILO) todos os trabalhadores devem atuar em locais seguros, sem exposição a riscos ou perigos. No entanto, a imprudência e falta de preocupação com a SST são causadoras de mais de 370 milhões acidentes não fatais e quase 2,8 milhões óbitos ao redor do mundo anualmente (ILO, 2020).

No Brasil, também são registrados dados alarmantes, para os mais de 46 milhões de empregos formais no ano de 2018, foram registrados aproximadamente 577 mil acidentes de trabalho (MINISTÉRIO DA ECONOMIA, 2020; MINISTÉRIO DA FAZENDA, 2018). Este número revela que a cada 80 trabalhadores um será vítima de acidente de trabalho durante o ano, considerando os acidentes com e sem registros Comunicação de Acidente de Trabalho (CAT).

A comunicação do acidente por meio do registro de CAT é uma obrigação do trabalhador, que deve comunicar o acidente ou doença, um dia após a sua ocorrência e em casos de óbito a comunicação deve ser imediata (BRASIL, 1999). Estes registros são utilizados para controle e acompanhamento dos incidentes ligados à SST no Brasil, mas também podem servir de apoio para estudos. Nesse contexto, podem ser desenvolvidos estudos que analisam o histórico de acontecimentos ou mesmo pesquisas que fazem a previsão de novos casos, baseados nos dados existentes.

Utilizar a mineração de dados para esta finalidade é uma opção válida, pois a aplicação de algoritmos a grandes conjuntos de dados, objetiva a geração de conhecimento a partir de padrões encontrados na mineração (LIU *et al*., 2019; ZHAO *et al*., 2020). As aplicações de mineração de dados são amplas e podem ser voltadas para diversas áreas como a medicina (ARJI *et al*., 2019), a educação (ALDOWAH; AL-SAMARRAIE; FAUZY, 2019), o gerenciamento de processos produtivos complexos (SCHUH *et al*., 2019), para melhoria de sistemas energéticos em edifícios (ZHAO *et al*., 2020), para a saúde pública (DOS SANTOS *et al*., 2019) e em saúde e segurança do trabalho (CHOI *et al*., 2020; SANNI-ANIBIRE *et al*., 2020).

Aplicações em mineração de dados estão geralmente envolvidas com atividades de baixo custo, mas com grande valor em seus resultados, pois podem ser utilizadas como subsídio para tomada de decisão (DEL POZO-ANTÚNEZ *et al*., 2018; TORRECILLA; ROMO, 2018). Também podem facilitar a interpretação de dados complexos, como informações geoquímicas espaciais (HOOD; CRACKNELL; GAZLEY, 2018) ou dados de transferência de UTI (Unidade de Terapia Intensiva) (CHOU *et al*., 2020). Além disso, são utilizados na compreensão e aprendizado de conjuntos de dados, pois a mineração possibiliza a organização das informações (HOOD; CRACKNELL; GAZLEY, 2018).

No entanto, apenas a aplicação de técnicas de mineração de dados pode não ser suficiente para a análise dos dados, principalmente quando esses resultados impactam na vida do ser humano, como no diagnóstico de doenças (BULLOCK *et al*., 2020). Nesses casos, é necessário entender detalhadamente como o modelo de mineração de dados tomou sua decisão e para isso a Inteligência Artificial (IA) pode ser utilizada, em especial a IA explicável, conhecida mundialmente como *eXplainable Artificial Intelligence* (XAI). A aplicação de XAI tem o objetivo de garantir, não apenas o bom desempenho do modelo, mas também o entendimento e confiabilidade pelo tomador de decisão que irá se basear no resultado da mineração de dados (ARRIETA *et al*., 2020).

Neste contexto a presente pesquisa foi desenvolvida, buscando entender o cenário de SST no Brasil e classificar a ocorrência de óbitos por meio de técnicas de mineração de dados. Com destaque para as etapas de mineração e interpretação dos dados, propostas por Fayyad *et al.* (1996) no processo de KDD (*Knowledge Discovery in Databases*). Primeiramente com uma análise exploratória, a pesquisa utiliza a totalidade de dados com CAT disponibilizados publicamente, até o momento. Já nas etapas de classificação, é utilizado apenas um recorte dos dados referente ao setor industrial com maior número de ocorrências: a indústria de transformação.

Mas não apenas apresentar uma classificação, esta pesquisa busca entender quais as melhores técnicas para esta aplicação e quais os atributos que influenciam na sua decisão, utilizando para isso aplicações de XAI. Com a realização da pesquisa espera-se apresentar contribuições para a área de Engenharia de Produção, para o meio científico, para organizações privadas, governos e sociedade.

Para a área de Engenharia de Produção e para o meio científico as contribuições estão relacionadas à realização de um estudo que reúne SST, mineração de dados e inteligência artificial, tornando-se uma pesquisa multidisciplinar, pois envolve temas de engenharia, estatística, computação e saúde. Além disso, o estudo desenvolvido cobre uma lacuna na pesquisa, identificada a partir de um mapeamento sistemático desenvolvido e citado nas seções seguintes. Nesta ocasião foram observados poucos estudos com a utilização de dados de abrangência nacional e técnicas de mineração de dados de forma conjunta.

Em relação às contribuições para organizações privadas, este estudo pode auxiliar nas estratégias e tomada de decisão por parte da equipe de SST e gestores. Tanto a compilação e apresentação dos dados, por meio da análise exploratória, como a apresentação dos fatores que influenciam na ocorrência do óbito. Além disso, esses pontos de destaque indicados na pesquisa, tais como: idade, sexo, setor industrial e demais atributos, indicam qual o perfil do trabalhador que requem mais atenção no combate aos acidentes ou ainda, quais funções e postos de trabalho precisam ser verificados com maior frequência.

Semelhante às contribuições para organizações privadas, para os governos a pesquisa também pode fornecer subsídios para criação de novas políticas e regulamentações específicas para os pontos de destaque, visando reduzir tanto o número de acidentes de maneira geral, como principalmente acidentes fatais. Como reflexo das contribuições das organizações públicas e privadas, estão as contribuições para a sociedade. Promoção de um ambiente de trabalho mais seguro, preocupação com a saúde e segurança do trabalhador, fiscalizações e acompanhamento do dia a dia no trabalho, entre outras medidas que podem ser tomadas. Essas ações objetivam

reduzir os acidentes, doenças e óbitos da população, garantindo maior confiança no desempenho de sua rotina de trabalho e tranquilidade para o trabalhador e sua família.

Apresentar uma pesquisa relevante e de qualidade, auxiliar organizações privadas e governamentais no cuidado com a saúde e segurança do trabalhador e principalmente, fornecer subsídios para garantir um melhor ambiente de trabalho para a sociedade de maneira geral são as motivações pessoais para realização dessa pesquisa.

Diante disso, o trabalho está estruturado com este capítulo de introdução, com a contextualização do tema, objetivos e justificativas, seguido pelo capítulo 2 com a metodologia utilizada na pesquisa. O capítulo 3 apresenta o primeiro artigo que compõe o modelo *multipaper*, descrevendo uma análise exploratória dos dados da pesquisa. O segundo artigo está descrito no capítulo 4, com a aplicação da mineração de dados e também o pós-processamento. O último capítulo representa as considerações finais da pesquisa, com contribuições e indicações futuras.

1.2 OBJETIVOS

Este estudo visa responder a seguinte questão de pesquisa: **Como a mineração de dados pode contribuir para a saúde e segurança do trabalho no Brasil?** Assim, relacionado a essa questão, está o objetivo geral da pesquisa de prever a ocorrência de óbitos a partir de registros de doenças e acidentes de trabalho no Brasil. Para atingir este objetivo geral, alguns objetivos específicos também foram delimitados:

- Conduzir um mapeamento sistemático sobre saúde e segurança do trabalho e mineração de dados;
- Definir uma base de dados de acidentes de trabalho no Brasil para utilização no estudo;
- Realizar tratamentos e adequações, bem como análises exploratórias sobre os dados selecionados;
- Delimitar um setor industrial específico para refinar a pesquisa;
- Aplicar técnicas de mineração de dados sobre os dados do setor industrial selecionado;
- Interpretar os resultados da mineração de dados, utilizando para isso inteligência artificial explicável.

1.3 SÍNTESE DOS ARTIGOS

Esta dissertação foi estruturada utilizando o modelo *multipaper*, onde dois artigos, ligados às seções de introdução, metodologia e considerações finais, compõem a apresentação da pesquisa. No Quadro 1 é apresentada uma síntese dos artigos, com seus títulos e respectivos objetivos, metodologias e contribuições.

Quadro 1 – Síntese dos artigos

	Artigo 1	**Artigo 2**
Título	Uma análise exploratória das doenças e acidentes de trabalho no Brasil a partir dos dados da Comunicação de Acidente de Trabalho (CAT).	Análise preditiva de óbitos por acidentes de trabalho na indústria de transformação baseada em inteligência artificial explicável.
Objetivo	Analisar os registros de doenças, acidentes e óbitos relacionados ao trabalho no Brasil.	Prever a ocorrência de óbitos em função de acidentes e doenças ocupacionais com registro no Brasil na indústria de transformação.
Metodologia	Seleção do conjunto de dados; Limpeza e pré-processamento; Análise exploratória para avaliação dos atributos; Análise exploratória detalhada no setor de indústria de transformação.	Seleção e pré-processamento dos dados da indústria de transformação; Realização de três experimentos com aplicação de técnicas de mineração de dados; Comparação dos resultados das métricas; Aplicação de IA explicável.
Contribuições	Apresentação do cenário de doenças, acidentes e óbitos relacionados ao trabalho no Brasil, especificamente no setor de indústria de transformação, assim como a análise de todos os atributos do conjunto de dados da CAT.	Previsão de óbitos na indústria de transformação brasileira, assim como a apresentação de quais são os atributos e categorias que mais impactam nesta ocorrência.

Fonte: A autora (2021)

Como é possível observar, o primeiro artigo descreve uma análise exploratória dos dados selecionados, com foco na seleção e pré-processamento dos dados. Este artigo inclui ainda um destaque para o setor industrial da indústria de transformação, que é responsável pelo maior percentual de registros de acidentes e doenças do trabalhador. O segundo artigo, por sua vez, utiliza o recorte da indústria de transformação para aplicações de mineração de dados, buscando prever a ocorrência de acidentes por meio de doze algoritmos. Ao final, dois

algoritmos são selecionados e passam por uma etapa de pós-processamento, utilizando inteligência artificial explicável.

2

METODOLOGIA

Este capítulo apresenta a metodologia utilizada na pesquisa, primeiramente detalhando a sua caracterização e estrutura. Dividida em três partes, a estrutura da pesquisa é descrita por (i) um estudo exploratório, (ii) desenvolvimento e (iii) refinamento da pesquisa. A primeira etapa serviu de subsídio para realização das etapas seguintes, que apresentam foco na seleção e pré-processamento dos dados, na mineração de dados e o pós-processamento, envolvendo a explicação dos resultados encontrados por meio de algoritmos.

2.1 CARACTERIZAÇÃO DA PESQUISA

Segundo Gil (2017), uma pesquisa pode ser descrita a partir de quatro aspectos, sendo: natureza, abordagem do problema, objetivos e procedimentos. Em relação a natureza, esta pesquisa se caracteriza como aplicada, pois busca gerar conhecimento e soluções voltadas a SST, utilizando dados e modelos já conhecidos. Considerando a abordagem dos problemas da pesquisa, pode se identificar como quantitativa, visto que busca classificar as informações de forma quantificável, utilizando estatística e mineração de dados, para posterior análise.

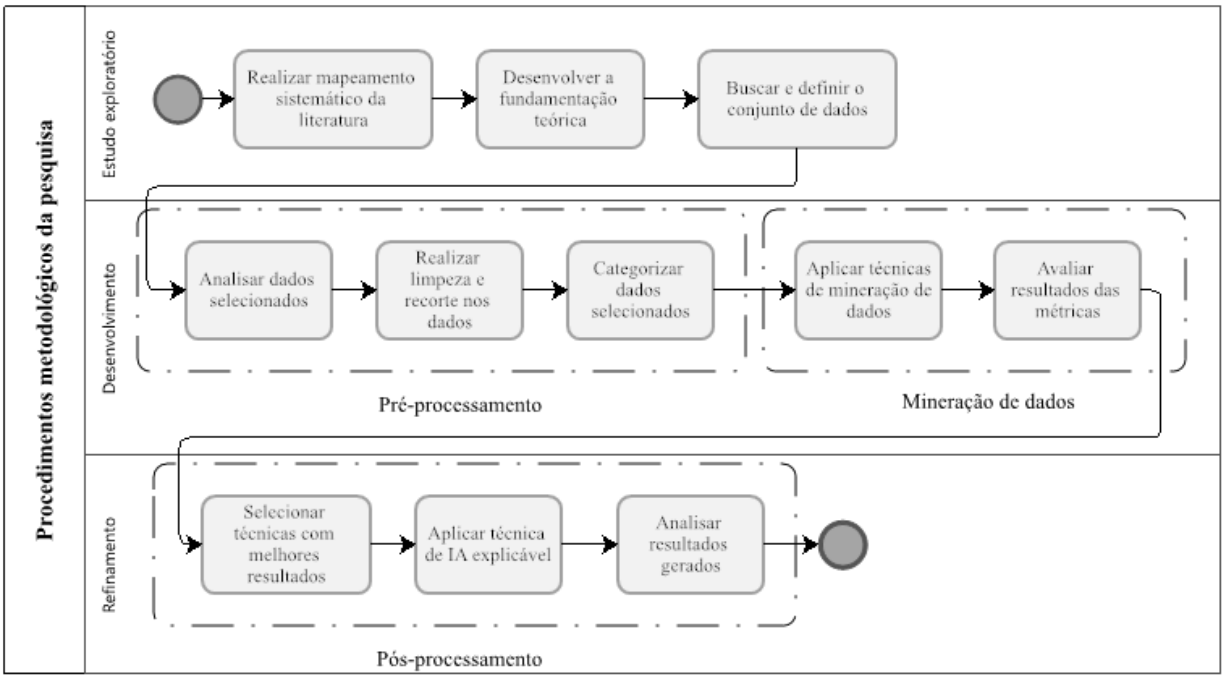
Quanto aos objetivos da pesquisa, se define como exploratória e descritiva. É exploratória devido ao processo de levantamento bibliográfico da literatura, que identificou as lacunas na pesquisa e guiou o escopo do trabalho. É descritiva, pois busca estabelecer conexões entre as variáveis do conjunto de dados, prevendo a ocorrência de óbitos baseados em dados de acidentes de trabalho na indústria de transformação do Brasil.

Em relação aos procedimentos utilizados para a condução da pesquisa, inicialmente foi realizada a pesquisa bibliográfica, descrevendo a fundamentação teórica que auxiliou a definição da proposta. As etapas seguintes, no desenvolvimento do trabalho, se definem como uma pesquisa *ex-post facto*, visto que após a ocorrência de um fato (acidente de trabalho), busca-se avaliar as relações entre as variáveis para então prever o óbito. É de responsabilidade do pesquisador buscar condições naturais desse conjunto de dados e realizar verificações como se fossem situações de controle (GIL, 2017).

2.2 ESTRUTURA DA PESQUISA

Os procedimentos metodológicos adotados são divididos em três fases: estudo exploratório, desenvolvimento e refinamento, apresentados pela Figura 1. Também são descritas e sequenciadas as etapas que compõem cada uma das fases, destacando as etapas de pré-processamento, mineração de dados e pós-processamento.

Figura 1 – Procedimentos metodológicos da pesquisa



Fonte: A autora (2021)

Na fase de **estudo exploratório** foi realizado inicialmente o mapeamento sistemático da literatura. Este mapeamento foi conduzido pelo grupo de pesquisa da área e serviu de embasamento não apenas para essa pesquisa, mas também para os demais discentes envolvidos. A partir das lacunas identificados no mapeamento sistemático, foi possível delimitar o tema de estudo na área de aplicações de mineração de dados à conjuntos de dados nacionais, seguindo

então para fundamentação teórica, buscando o entendimento dos temas da pesquisa e do contexto a ser estudado.

Também foi necessário buscar por conjuntos de dados disponíveis, assim como selecionar e analisar o conjunto de dados escolhido. A partir desse ponto, no **desenvolvimento** da pesquisa, se inicia a etapa conhecida como pré-processamento dos dados, onde optou-se pelo conjunto de dados disponibilizado pela Previdência Social do Brasil, com os dados de CAT, no horizonte temporal de julho/setembro de 2018 a julho/setembro de 2020. Em seguida, após o agrupamento dos conjuntos de dados, uma análise exploratória foi realizada.

Ainda no desenvolvimento, a etapa seguinte foi de execução da limpeza dos dados, que consiste na remoção de dados e atributos repetidos e *outliers*, ou seja, dados que podem causar um desvio na pesquisa. Em sequência, os dados foram categorizados e foi selecionado o recorte para dar continuidade na pesquisa: o setor de indústria de transformação. Este setor foi escolhido por ser o maior causador de acidentes e doenças no Brasil, segundo os dados escolhidos. Por fim, este estudo exploratório originou o primeiro artigo que compõe essa dissertação.

A aplicação da mineração de dados foi etapa seguinte, onde foram utilizadas doze técnicas para comparação de seus resultados. As técnicas selecionadas foram: *Bagging* (BREIMAN, 1996), *Extra Trees* (GEURTS, 2006), *Random Forest* (BREIMAN, 2001), *Stacking* (WOLPERT, 1992), *Voting* (BAUER; KOHAVI, 1999), *XGBoost* (CHEN, GUESTRIN, 2016), *Decision Trees* (QUINLAN, 1986), *K-Nearest Neighbors* (SARKAR *et al.*, 2000), *Logistic Regression* (LIAO *et al*., 2005), *Naïve Bayes* (ZHANG, 2004), *Neural Networks* (ZHANG, 2000) e *Support Vector Machine* (BOSER *et al*., 1992). A base de dados foi submetida a três experimentos em cada uma das técnicas para comparação de seus resultados, que foram validados a partir de cinco métricas (acurácia, precisão, *recall*, F1 score e curva ROC/AUC (*Receiver Operating Characteristic/Area Under Curve*).

Após aplicar cada modelo para os três experimentos, os dois que obtiveram melhores resultados passaram pela etapa de pós-processamento, já na fase de **refinamento** da pesquisa. Para o pós-processamento dos dados foi selecionada uma técnica de inteligência artificial explicável, o *SHapley Additive exPlanations* (SHAP). Proposto por Lundberg e Lee (2017), na presente dissertação, o SHAP tem o objetivo de entender a previsão de óbitos que foi realizada pelas técnicas na mineração de dados. A realização da mineração e explicação dos resultados foram utilizados para construção do segundo artigo desta dissertação. As etapas de seleção e pré-processamento dos dados, assim como a mineração e pós-processamento, são destacadas nas seções posteriores.

2.3 SELEÇÃO E PRÉ-PROCESSAMENTO

A partir das lacunas encontradas na pesquisa com aplicação de dados de abrangência nacional, optou-se por utilizar um conjunto de dados que representasse ocorrências registradas no Brasil. Considerando as possibilidades disponibilizadas publicamente, como os conjuntos de dados da Empresa de Tecnologia e Informações da Previdência (DATAPREV), foram escolhidos os dados com abertura de Comunicação de Acidentes de Trabalho (CAT). Estes acontecimentos são registrados por meio do CATWEB, um sistema de Comunicação de Acidentes do Trabalho do Instituto Nacional do Seguro Social (INSS).

Após o registro, as ocorrências são agrupadas trimestralmente e disponibilizadas virtualmente no formato *Comma-Separated Values* (*.CSV*). Até o momento desta pesquisa, foram apresentados dez conjuntos trimestrais, referentes ao período de julho de 2018 a dezembro de 2020. No entanto, para esta pesquisa foi considerado o período de dois anos completos (julho/setembro de 2018 a julho/setembro de 2020) e os dados desse período foram agrupados em único documento, utilizando uma planilha. Essa união resultou em um conjunto de dados com 25 atributos e 990.870 instâncias. Um resumo dos atributos, sua definição e informações complementares é apresentado no Quadro 2.

Quadro 2 – Atributos do conjunto de dados e informações complementares

Nome	**Descrição**	**Exemplo de preenchimento**
Agente causador do acidente	Coisa, substância ou condição do ambiente de trabalho que causou o acidente	Engrenagem - Disposi
Data do acidente	Data em que aconteceu o acidente	201801
		03/01/2018
CBO	Ocupação vinculada ao trabalhador segundo a Classificação Brasileira de Ocupações (CBO)	784205
		784205-Alimentador d
CID-10	Classificação da doença ou lesão que o trabalhador foi submetido, segundo o Código Internacional de Doenças (CID-10)	S681
		S68.1 Amput Traum de
CNAE	Classificação Nacional de Atividades Econômicas do empregador	4763
		Comercio Varejista d
Emitente CAT	Quem foi o responsável pela emissão da CAT	Empregador
Espécie do benefício	Qual benefício o trabalhador ou seu dependente recebeu após a ocorrência	Afastamento Até 15 D
Filiação	Classificação do trabalhador perante à Previdência Social	Empregado
Óbito	Existência de óbito associado à ocorrência	Não
Município do empregador	Município onde o empregador é registrado	350570-Barueri
Natureza da lesão	Classificação da lesão segundo sua natureza	Amputacao ou Enuclea
Origem da CAT	Qual foi a origem do cadastramento da CAT	Internet
Parte do corpo atingida	Qual parte do corpo foi atingida pelo acidente ou doença	Dedo
Sexo	Sexo do trabalhador	Masculino
Tipo do acidente	Classificação do acidente	Típico
UF do acidente	Estado onde ocorreu o acidente	Maranhão
UF do empregador	Estado onde o empregador é registrado	São Paulo
Data do afastamento	Data em que o trabalhador foi afastado de sua função	2018/06
Data de despacho do benefício	Data em que o trabalhador ou seu dependente recebeu o benefício	201803 ou 2018/03
Data de nascimento	Data de nascimento do trabalhador	10/12/1982
Data de emissão da CAT	Data em que a CAT foi emitida	10/07/2018

Fonte: A autora (2021)

O primeiro atributo é o agente causador do acidente, definido pela Norma técnica Brasileira (NBR) 14280 como “Coisa, substância ou ambiente que, sendo inerente à condição ambiente de insegurança, tenha provocado o acidente” (ABNT, 2001). No conjunto de dados estudado são apresentadas descrições textuais com no máximo 20 caracteres (considerando espaços), e descreve 273 possibilidades de causas associadas ao acidente ou doença registrada. De acordo com a instrução normativa para preenchimento do CATWEB existem 241 classes de agente causador para acidentes e 58 para doenças.

Em seguida, o segundo atributo apresentado tem relação à data do acidente, expresso por seis caracteres numéricos que descrevem ano e mês do incidente. Outro atributo do conjunto de dados também apresenta informações sobre a data do acidente, por sua vez mais completas, apresentando o dia, mês e ano de ocorrência, separados por barras.

Considerando o terceiro e quarto atributos, ambos descrevem a Classificação Brasileira de Ocupações (CBO) no momento do acidente. A CBO define quais são as ocupações através do seu nome e descrição, que são agrupadas por famílias em função da sua semelhança, e podem ser expressas por um código numérico de até seis caracteres ou pela descrição da classe (MINISTÉRIO DO TRABALHO E EMPREGO, 2010). No conjunto de dados são apresentados em ambos os formatos, primeiro apenas o código e em seguida com o código mais uma descrição inicial, pois o máximo são de 20 caracteres.

Similarmente, o conjunto de dados também apresenta informações duplicadas em relação ao Código Internacional de Doenças (CID-10), que descreve as classificações das doenças ou lesões as quais o trabalhador foi submetido. O CID-10 refere-se à 22 classes, definidas como capítulos do código, que nos dados em questão são expressos por um código de números e letras de até quatro caracteres (quinta coluna), além desse código somado a uma breve descrição de até 20 caracteres no total.

O próximo atributo apresentado no conjunto de dados é a Classificação Nacional de Atividades Econômicas (CNAE), onde a Comissão Nacional de Classificação (CONCLA) descreve quais os setores industriais, divididos por seções. A primeira seção da CNAE descreve 21 grupos, nomeados da letra A ao U de acordo com sua categoria. Para os dados utilizados, duas colunas apresentam informações da CNAE: uma com códigos numéricos de no máximo quatro caracteres e outra com breves descrições textuais de até 20 caracteres.

Além disso, algumas informações ligadas à CAT e ao benefício fornecido também são apresentadas. O responsável por realizar a emissão da CAT é descrito na coluna nove e apresenta cinco possibilidades de emitente: autoridade pública, empregador, médico, segurado/dependente e sindicato. A origem do cadastramento também é um atributo do

conjunto de dados, podendo ser classificado como originário da internet ou do Projeto de Regionalização de Informações e Sistemas (PRISMA). Por fim, a data de emissão da CAT e data do afastamento são apresentadas, uma indicada pelo dia, mês e ano da ocorrência e a outra apenas por ano e mês.

A espécie do benefício apresenta qual benefício foi direcionado ao trabalhador ou seus dependentes após a ocorrência do incidente. Analisando o conjunto de dados, foram encontradas sete possibilidades diferentes de resposta, todas com descrição de no máximo 20 caracteres, dificultando o entendimento de algumas delas. As classes de espécie de benefício observadas são: “Afastamento até 15 D”, “Auxílio Acidente”, “Auxílio Doenca por A”, “Auxílio Doenca Previ”, “Pa”, “Pensão por Morte Aci” e “Pensão por Morte Pre”. Não foram encontradas definições exatas para o que expressa cada uma das categorias anteriores. Ainda em relação ao benefício é apresentada a data de despacho do benefício, expressa por ano e mês respectivamente, com seis caracteres numéricos sem espaço.

Tomando a variável que apresenta a filiação do segurado, esta é descrita pelo tipo de filiação à Previdência Social que compreende o trabalhador e pode ser definida como: empregado, segurado especial ou trabalhador avulso. Em seguida é apresentada a variável do óbito relacionado ao acidente, indicada pelos valores sim ou não. Também em relação ao acidentado, o conjunto de dados apresenta o atributo de sexo, descrito por masculino, feminino, indeterminado e não informado, e sua data de nascimento, com dia, mês e ano, separado por barras.

Em relação à localização, são descritos três atributos para este fim, dois ligados ao empregador e um relacionado ao local de ocorrência do incidente. Inicialmente é apresentado um código da cidade, com seis caracteres numéricos complementados ao nome da cidade, desde que somados acumulem 20 caracteres. Esse atributo descreve a cidade em que o empregador está inscrito, assim como sua respectiva unidade federativa, a partir da descrição do nome do estado. No entanto, também é apresentado um atributo que descreve qual unidade federativa realmente aconteceu o acidente ou doença registrada com a CAT.

Quanto ao atributo de natureza da lesão, que descreve quais as características principais da lesão, o conjunto de dados apresenta 28 categorias para associação ao incidente. São apresentados por até 20 caracteres textuais descrevendo as especificidades de cada categoria, como por exemplo: fratura, luxação e lesão imediata. Assim como a natureza da lesão, a parte do corpo atingida pelo acidente ou doença também é um atributo descrito por até 20 caracteres textuais, que apresentam 42 possibilidades de partes corporais que foram atingidas no momento do incidente.

Por fim, em relação ao incidente, é apresentada a data de sua ocorrência, descrita por dia, mês e ano separado por barras, assim como o tipo do acidente. O tipo descreve a classificação da ocorrência, que pode ser atribuída a um acidente típico, de trajeto ou a uma doença.

O Quadro 2 apresenta os atributos encontrados no conjunto de dados da CAT, assim como uma descrição do que representa esse atributo e um exemplo de preenchimento dos dados, como é originalmente encontrado no conjunto selecionado. Também é possível observar que alguns atributos apresentam informações duplicadas, como data do acidente, CBO, CID-10 e CNAE, mesmo que descritas de forma diferentes, possuem mesmo significado.

Após a primeira análise dos dados, foram removidos *outliers*, valores faltantes, atributos repetidos e aqueles que não apresentavam confiabilidade nos dados. Com o pré-processamento também foi possível identificar quais categorias de dados eram mais significativas com relação às ocorrências. Uma dessas indicações foi quanto ao setor industrial, que apresentou destaque para indústria de transformação, responsável por aproximadamente 30% dos registros de CAT. A fim de refinar a pesquisa e buscar *insights* pontuais para o setor industrial, foi selecionado apenas este setor industrial para as etapas de mineração e pós-processamento de dados.

2.4 MINERAÇÃO DE DADOS E PÓS-PROCESSAMENTO

Após o pré-processamento e preparação do conjunto de dados, a etapa seguinte foi a aplicação das técnicas de mineração para previsão da ocorrência de óbitos. Neste momento foram selecionadas doze técnicas, sendo seis delas caracterizadas como *ensemble* e as demais não *ensemble*. Os algoritmos *ensemble* são considerados mais robustos e para esta aplicação foram escolhidos: *Bagging*, *Voting*, *Stacking*, *Extra Trees*, *Random Forest* e *XGBoost*. Algoritmos não *ensemble*, por sua vez, são métodos mais simples, sendo selecionados: *Support Vector Machine*, *Logistic Regression*, *K-Nearest Neighbors*, *Naïve Bayes*, *Decision Trees* e *Neural Networks*.

Como ferramentas para realização, tanto da etapa de mineração de dados como pós-processamento, foi utilizada a linguagem Python, por meio da distribuição Anaconda no ambiente *Jupyter Notebook.* As técnicas foram executadas em sistema Intel Core i3-5005U, com memória RAM (*Random Access Memory*) de 4,00 GB (*Gigabyte*) e uma unidade de armazenamento (SSD - *Solid State Drive*) de 120 GB. Para a aplicação das técnicas de mineração, métricas e para explicabilidade dos resultados, foram utilizadas funções do *scikit-learn* no modelo desenvolvido.

Cada uma das técnicas citadas foi submetida à três experimentos para avaliar seus resultados, segundo cinco métricas: acurácia, precisão, *recall*, F1 score e ROC/AUC. O primeiro experimento utilizou todos os parâmetros no padrão do modelo e um particionamento de dados para 70% treino e 30% para teste. Já o segundo experimento utilizou a validação cruzada para divisão do seu conjunto de dados, onde nove subconjuntos (*folds*) foram utilizados para treino e apenas um para teste, com os parâmetros também em *default*. O último experimento, além de utilizar a validação cruzada para o conjunto de dados, também aplicou o método de *Grid Search* a fim de encontrar a melhor combinação dos parâmetros.

Os resultados de todas as métricas em cada experimento foram avaliados e as técnicas que apresentaram melhor resultado no último experimento (mais robusto) foram selecionadas. Uma técnica *ensemble* e uma não *ensemble* foram examinadas, buscando entender quais fatores influenciaram a previsão do modelo. Para isso, foi aplicado o algoritmo *SHAP*, que indicou quais atributos e categorias estavam mais relacionados com a previsão de óbito.

2.5 CONSIDERAÇÕES FINAIS

Neste tópico foi apresentada a metodologia utilizada na pesquisa, com uma categorização segundo sua natureza, abordagem, objetivos e procedimentos. Além disso, também foi apresentada a estrutura detalhada da pesquisa, dividida em: estudo exploratório, desenvolvimento e refinamento da pesquisa. Assim como as fases do estudo, também foram apresentadas as etapas principais, sendo a seleção e pré-processamento dos dados, com o detalhamento do conjunto de dados escolhido, e a mineração e pós-processamento dos dados, com as técnicas utilizadas na previsão do óbito e sua explicação.

No próximo tópico é apresentado o detalhamento da etapa de seleção e pré-processamento dos dados, por meio do primeiro artigo que compõe esta dissertação. Neste artigo é apresentada uma análise exploratória do conjunto de dados da CAT, assim como uma descrição do recorte de dados na indústria de transformação. Em seguida, na seção 4, consta o segundo artigo da pesquisa que descreve as etapas de mineração de dados e pós-processamento, com a aplicação de técnicas para previsão e explicação do óbito na indústria de transformação brasileira. Por fim, são destacadas as considerações finais da pesquisa, assim como os elementos pós-textuais.

3

ARTIGO 1

Este capítulo apresenta o primeiro artigo, que descreve uma análise exploratória dos dados selecionados para a pesquisa, em sua totalidade, assim como um recorte dos dados buscando explorar o setor industrial de maior incidência para acidentes e doenças do trabalho: o setor de indústria de transformação.

UMA ANÁLISE EXPLORATÓRIA DAS DOENÇAS E ACIDENTES DE TRABALHO NO BRASIL A PARTIR DOS DADOS DA COMUNICAÇÃO DE ACIDENTE DE TRABALHO (CAT)

Resumo

O tema de saúde e segurança do trabalho tem apresentado crescimento, tanto no âmbito acadêmico como empresarial. Este fator se justifica não somente devido ao aumento no número dos casos de doenças, acidentes e óbitos ligados ao trabalhador, mas também pela preocupação com os impactos gerados por estes acontecimentos. Além de impactar o bem-estar do trabalhador, os acidentes e doenças também afetam a produtividade das organizações, sua reputação perante o mercado e sua condição financeira. Isso se deve ao fato de que, após as ocorrências, podem ocorrer instabilidades na produção, afastamentos e pagamento de benefícios ao trabalhador ou seu dependente, ações que impactam diretamente nos indicadores econômicos de organizações púbicas e privadas. Em vista das implicações resultantes de doenças, acidentes e óbitos ligados ao trabalho, é necessário acompanhar e controlar essas

ocorrências. No Brasil, estes acontecimentos devem ser registrados com a abertura de Comunicação de Acidente de Trabalho (CAT), que posteriormente são agrupadas por trimestres e disponibilizadas publicamente pela Previdência Social do Brasil. No trabalho em questão, estes dados são submetidos a um estudo exploratório, buscando analisar casos de doenças, acidentes e óbitos no Brasil, assim como os atributos do conjunto de dados. Para isso é utilizada toda a disponibilidade de dados até o momento, no horizonte temporal de junho de 2018 a setembro de 2020. Além disso, outra análise foi desenvolvida focando no setor industrial com maior número de registros: a indústria de transformação, comparando suas características aos dados totais. Como principais contribuições da pesquisa, busca-se analisar as categorias de maior incidência para cada atributo, assim como identificar valores faltantes e *outliers,* a fim de construir um novo conjunto de dados que possibilite futuras aplicações.

Palavras-chave:
Saúde e segurança do trabalho; Comunicação de acidente de trabalho; Indústria de transformação.

1. Introdução

A Organização Internacional do Trabalho (OIT) estima que anualmente são registradas 2,78 milhões de mortes no mundo todo em decorrência do trabalho e que desse número 2,4 milhões estão relacionados especificamente a doenças ocupacionais. Como resultado desses incidentes, aproximadamente 3,94% do PIB mundial é gasto com assistência médica, benefícios aos acidentados e familiares, treinamentos para mitigação de novos acidentes, dias de afastamento e redução da produtividade das organizações, resultando em um impacto significativo em diversos âmbitos da sociedade (ILO, 2019a).

No Brasil, os números em relação aos acidentes de trabalho também são alarmantes. Segundo os registros do Anuário Estatístico de Acidentes de Trabalho, os anos de 2016, 2017 e 2018 representaram, juntos, mais 1,7 milhões de acidentes no país. Em relação aos óbitos ligados ao trabalho, foram registrados no ano de 2016 um total de 2.288 mortes no país, em 2017 foram 2.132 e 2018 com 2.098 (MINISTÉRIO DA FAZENDA, 2018). Em vista do impacto negativo na economia e sociedade, causado por acidentes, doenças e óbitos em decorrência do trabalho (ILO, 2006), é necessário entender e analisar quais os conceitos, dados e consequências desses incidentes, a fim de que as organizações se mantenham atuantes e representativas em seu mercado (SANNI-ANIBIRE *et al*., 2020).

O termo Saúde e Segurança do Trabalho (SST) está diretamente relacionado com a análise, prevenção e mitigação de acidentes e doenças causadas ao trabalhador, em decorrência da sua profissão ou função que ele executa (CHEN *et al*., 2020; MUTLU; ALTUNTAS, 2019). Além de se caracterizar por uma necessidade das organizações, também existem imposições legais dos governos para implantação de medidas relacionadas a saúde e segurança do trabalhador, por meio de leis e regulamentos (AZIZ; OSMAN, 2019). Essas obrigações legais incluem especificações mais comuns como salários, licenças e férias, mas também envolvem prevenção de acidentes e doenças, políticas contra assédio e especificações para cada classe trabalhadora (ILO, 2019).

Como a saúde e segurança do trabalhador tornaram-se pilares no desenvolvimento econômico e social (CHEN *et al*., 2020), é necessário entender também as definições que permeiam o termos relacionados a SST. Os riscos estão associados a resultados prejudiciais ao indivíduo a partir de alguma ação tomada em seu ambiente de trabalho (MUTLU; ALTUNTAS, 2019). Esses resultados podem estar relacionados a doenças, acidentes ou mortes, que quando causados devido ao exercício específico que o trabalhador executa são caracterizados como doença do trabalho, acidente de trabalho ou de trajeto e óbito, e devem ser registrados através de uma Comunicação de Acidentes de Trabalho (CAT) (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2001).

Além dos dados com abertura de CAT outros conjuntos de dados de domínio público também são divulgados, como Sistema de Informação de Agravos de Notificação (SINAN) e Sistema de Informação sobre Mortalidade (SIM) (SANNI ALI *et al*., 2019) e dados utilizados em estudos específicos (DAS CHAGAS MOURA et al., 2016; TOMIAZZI *et al*., 2018). Em vista do grande número de dados associados a SST gerados diariamente, é necessário buscar por métodos e análises que facilitem a compreensão dessas informações. Nesse contexto esta pesquisa se desenvolve, com o objetivo de analisar os registros de doenças, acidentes e óbitos relacionados ao trabalho no Brasil.

Para atingir esse objetivo foi selecionado um conjunto de dados abrangente para todo território nacional e disponibilizado publicamente: o conjunto de dados disponibilizado pela Previdência Social do Brasil, com abertura de CAT. Após a seleção, os dados foram analisados por meio de um estudo exploratório, buscando entender e avaliar as informações coletadas. Por fim, para o refinamento da análise exploratória, o setor industrial sobressalente também passou por um estudo exploratório, buscando comparar o conjunto de dados total, com os seus elementos.

2. Referencial teórico

2.1 Saúde e Segurança do Trabalho (SST)

As informações envolvendo saúde e segurança do trabalhador tornaram-se mais representativas nas últimas décadas, em função do aumento de registros de acidentes e mortes ao redor do mundo (SÁNCHEZ-HERRERA; DONATE, 2019). A saúde e segurança do trabalho atua nesse âmbito, buscando identificar e gerenciar os riscos associados aos trabalhadores, sendo essencial para o desenvolvimento social, a permanência das organizações e seu destaque perante o mercado (CHEN *et al*., 2020; MUTLU; ALTUNTAS, 2019).

Com a industrialização, mecanização e avanços tecnológicos, os processos tornaram-se mais ágeis e eficientes, no entanto, elevaram os riscos de danos ao trabalhador. O risco pode ser definido como a possibilidade de ocorrência de um incidente que causará um resultado negativo, sendo ele perda, ferimento, ou qualquer evento desfavorável (MUTLU; ALTUNTAS, 2019). Analisar e avaliar os riscos no local de trabalho é útil para prevenir lesões e doenças ao trabalhador, perdas financeiras e prejuízos à reputação das organizações (ILO, 2006; BEVILACQUA; CIARAPICA; GIACCHETTA, 2008; MUTLU; ALTUNTAS, 2019). Para realizar esse processo, métodos já foram desenvolvidos, alguns com foco específico para determinado setor industrial, outros utilizando a junção de várias técnicas, voltados à prevenção de acidentes e lesões (SANNI-ANIBIRE *et al*., 2020).

Envolvidos com o conceito de SST estão algumas consequências de atos inseguros, dentre elas: doenças, acidentes e óbitos. As doenças ocupacionais, do ponto de vista normativo, podem ser classificadas como profissionais ou do trabalho (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2001). Doenças profissionais estão relacionadas a profissão ou função que o trabalhador exerce. Em contrapartida, as doenças do trabalho são associadas às condições especiais que o trabalhador é submetido, como seu ambiente, por exemplo (AZZOLIN *et al*., 2012).

Os acidentes associados ao trabalho também possuem subdivisões, que segundo ABNT (2001) com a NBR 14280:2001 podem ser acidentes de trabalho, sem lesão, de trajeto, impessoal e pessoal. Os destaques são para a diferenciação entre acidentes de trabalho e de trajeto. O primeiro é descrito como situação indesejável ou imprevista que possa resultar em alguma lesão ao trabalhador, imediata ou não (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2001). Nenonen (2013) completa que se o colaborador estiver em outra localidade, mas em função do seu ofício, também deve ser considerado acidente de trabalho. O acidente de

trajeto é definido como a ocorrência durante o deslocamento para o trabalho ou de retorno para casa (NENONEN, 2013).

O resultado mais grave de atos inseguros no trabalho é o óbito do colaborador, que se associa à sua profissão, independe do tempo decorrido desde o episódio da lesão (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2001). As ocorrências relacionadas a SST resultam em um grande número de informações. Considerando apenas os casos de óbitos, são estimados 2,78 milhões de mortes por ano ao redor do mundo (ILO, 2019b). Esses dados são registrados não apenas pela Organização Mundial da Saúde, mas também por programas de iniciativa pública e privada dos países.

Em pesquisas de Taiwan há um destaque para o conjunto de dados do *Council of Labor Affairs* (Executive Yuan), com registros de acidentes e óbitos (CHENG *et al*., 2012; CHENG; LIN; LEU, 2010; CHENG; YAO; WU, 2013; LIAO; PERNG, 2008). Nos Estados Unidos alguns estudos utilizam os dados da *Occupational Safety and Health Administration* (OSHA) (CHOKOR *et al*., 2016; GOH; UBEYNARAYANA, 2017; MISTIKOGLU *et al*., 2015; SHIN *et al*., 2018; TIXIER *et al*., 2017) e *Bureau of Labor Statistics* (NANDA *et al*., 2016).

Na Itália são utilizados registros de acidentes do *Istituto nazionale per l'assicurazione contro gli infortuni sul lavoro* (INAIL) (COMBERTI; BALDISSONE; DEMICHELA, 2015; COMBERTI; DEMICHELA; BALDISSONE, 2018; PALAMARA; PIGLIONE; PICCININI, 2011) e ferimentos ocupacionais do *Dipartimento Servizi Alla persona e Alla Comunità* (DSPC), que são acidentes com morte ou que gere incapacidade parcial ou total do trabalhador (CIARAPICA; GIACCHETTA, 2009).

No Brasil há pesquisas que utilizam conjuntos de dados regionais, como os dados do Centro de Saúde Ocupacional de Presidente Prudente que foram aplicados para análise da saúde de trabalhadores rurais (TOMIAZZI *et al*., 2018, 2019). Também há conjuntos de dados mais amplos, como o Sistema de Informação de Agravos de Notificação (SINAN) relacionado às doenças e o Sistema de Informações de Mortalidade (SIM) com a descrição de óbitos do país (DOS SANTOS *et al*., 2019; SANNI ALI *et al*., 2019). Outro conjunto de dados nacional é disponibilizado pela Previdência Social do Brasil, com informações específicas para SST, relativas à abertura de Comunicação de Acidente de Trabalho (CAT).

Para análise e investigação de informações relacionadas à SST é necessário utilizar métodos que viabilizem esse estudo. Algumas pesquisas utilizam questionários e processamento simples para as respostas, buscando entender o cenário estudado (BARLAS; IZCI, 2018; RAMOS; AFONSO; RODRIGUES, 2020). Outros apresentam a aplicação de um modelo teórico para auxiliar na tomada de decisões na área de saúde e segurança (GYEKYE;

SALMINEN; OJAJARVI, 2012). O maior destaque está nos estudos que utilizam ferramentas estatísticas e aprendizagem de máquina para entender o comportamento dos acidentes (ANYFANTIS; BOUSTRAS, 2020), prever novas ocorrências (KANG; RYU, 2019) e visualizar dados em larga escala (COMBERTI; DEMICHELA; BALDISSONE, 2018).

Explorar e acompanhar as informações relacionadas a SST pode beneficiar organizações públicas e privadas, mas principalmente o capital humano. Dentre os benefícios, são destacados o apoio a tomada de decisão por gerentes e líderes nas empresas (DEL POZO-ANTÚNEZ *et al*., 2018; YANAR; LAY; SMITH, 2019), o embasamento para criação de novas políticas de SST (COMBERTI; DEMICHELA; BALDISSONE, 2018), atenuação da queda de produtividade devido a afastamentos e preservação da integridade física e psicológica do trabalhador (RAMOS; AFONSO; RODRIGUES, 2020).

2.2 Saúde e Segurança do Trabalho no Brasil

O Brasil tem apontado crescimento no número de empregos formais com vínculos ativos nos últimos anos, chegando à marca de 47.554.211 empregos no ano de 2019, segundo a Relação Anual de Informações Sociais (RAIS) (2020). Esse número corresponde a um percentual de crescimento de 1,98% em relação ao ano anterior. No entanto, se comparado aos empregos formais de 20 anos atrás, o percentual salta para mais de 47% de aumento em ralação ao ano base de 2019.

Acompanhando o crescimento nos empregos com vínculos formais, também se observa um crescimento em casos de doenças e acidentes relacionados ao trabalho. Estima-se que a cada minuto um trabalhador é vítima de um acidente de trabalho no Brasil, resultado de 576.951 registros no ano de 2018, apresentando 3,35% de aumento em relação ao anterior (MINISTÉRIO DA FAZENDA, 2018). Contudo, esse número se relaciona apenas aos trabalhadores com vínculos formais, pois, se expandido aos trabalhadores informais e autônomos os registros podem se aproximar da marca de 4 milhões (JUSTIÇA DO TRABALHO, 2020).

Dessa forma, não diferente do restante do mundo, no Brasil também se faz necessário atentar-se a saúde e segurança do trabalhador. Esse cuidado pode ser oriundo de políticas públicas, assim como de normas internas de organizações privadas. Segundo o Artigo 336 do Decreto nº 3.048 de 06 de Maio de 1999, o empregador é obrigado a comunicar o acidente de trabalho ocorrido, dentro do prazo de um dia útil, ou para óbito, instantaneamente, sob pena de multa. Assim, a abertura da CAT torna-se uma obrigação legal às organizações.

A CAT é um documento emitido com intuito de formalizar a ocorrência de um acidente ou doença relacionada ao trabalho, que deve ser informada em primeira instância pelo empregador, mas caso não ocorra pode também ser registrada pelo trabalhador ou seu dependente, médico do trabalho, sindicato ou autoridade pública (BRASIL, 1999).

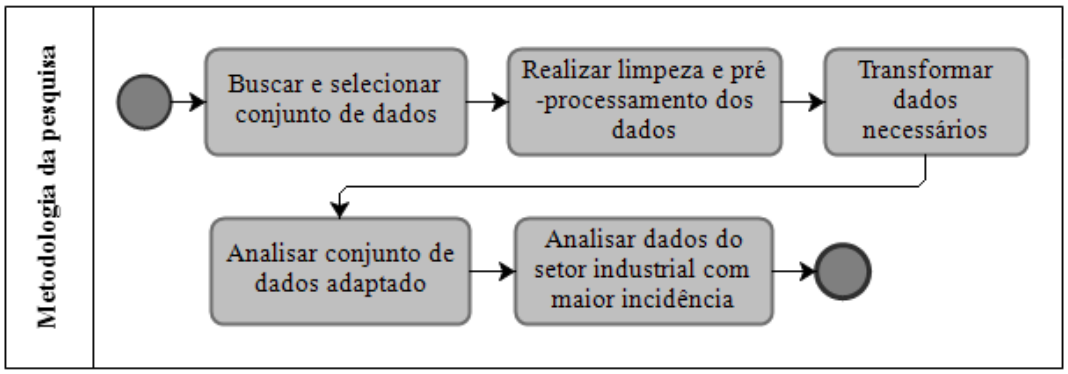
As informações estatísticas relacionadas a acidentes de trabalho no Brasil são comumente um resultado dos dados divulgados a partir da abertura da CAT (HENNINGTON; MONTEIRO, 2006). Essas informações são utilizadas em pesquisas específicas da área de SST, centradas em algumas localidades (HENNINGTON; MONTEIRO, 2006), ou ainda analisando causas específicas das ocorrências (CHIODI *et al.*, 2010) e podem servir de subsídios, tanto para criação de novas políticas, mas principalmente para redução ou mitigação de novos acidentes de trabalho.

Pesquisas com foco em estudos de caso e dados nacionais, voltadas ao cuidado com a SST, estão sendo desenvolvidas. São exemplos: aplicação de mineração em dados públicos de acidentes de trabalho, buscando identificar seus comportamentos e associações (SIEMINKOSKI, 2017); análise dos fatores que impactam a saúde de professores universitários atuantes na grande área da saúde (LEITE; NOGUEIRA, 2017); o uso de aprendizado de máquina para avaliar a saúde respiratória de trabalhadores do setor agrícola (TOMIAZZI *et al.*, 2018); e o desenvolvimento de recursos educativos digitais com foco na saúde do trabalhador (ANTONIOLLI *et al*., 2021).

3. Metodologia

Neste estudo a metodologia se dividiu em cinco etapas distintas, representadas por: planejamento, seleção do conjunto de dados, pré-processamento e transformação dos dados, análise exploratória do conjunto de maneira geral e uma análise específica para o setor de indústria de transformação. As etapas da pesquisa e sua ordenação são representadas pela Figura 1.

Figura 1 – Etapas da metodologia de pesquisa



Fonte: A autora (2021)

A etapa inicial, descrita pelo planejamento, refere-se a uma contextualização da pesquisa no tema de SST, buscando por estudos e referências na área para embasar teoricamente o estudo. Em seguida foi realizada uma busca em conjuntos de dados nacionais, optando pelos dados de abertura de CAT, devido à sua disponibilização pública, abrangência nacional e quantidade de atributos para observação e análise.

Com a etapa de pré-processamento e transformação dos dados, foi possível conhecer as instâncias e atributos do conjunto de dados selecionado. Além disso, também foram realizados exclusões, adaptações e agrupamentos nos dados, a fim de facilitar as etapas seguintes. A partir disso foi realizada a análise exploratória dos dados, com o auxílio dos *softwares* Microsoft Excel® e Tableau®, buscando avaliar os atributos restantes. Por fim, como resultado na análise anterior, o setor industrial com maior incidência de registros de CAT foi a indústria de transformação, motivando uma análise específica em seus elementos.

4. Resultados e discussões

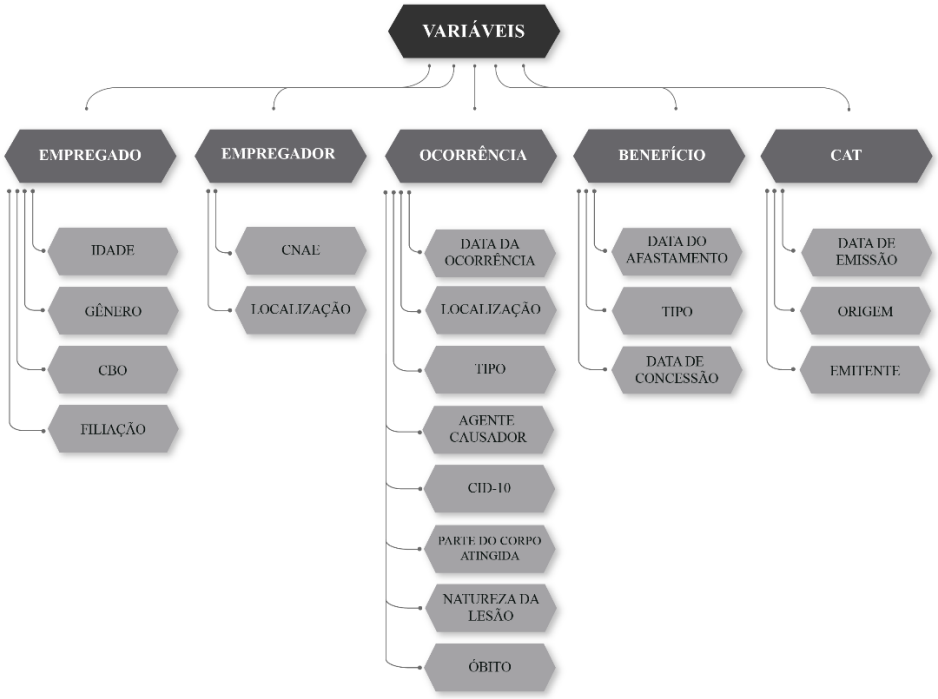
4.1 Seleção do conjunto de dados

Em meio as possibilidades de dados nacionais para uso na pesquisa, foi selecionado um conjunto disponibilizado publicamente, através dos dados abertos do governo federal. Dentre os 14 conjuntos de dados disponibilizados pela DATAPREV[1] encontram-se as Comunicações de Acidente de Trabalho (CAT), utilizadas no trabalho em questão. Esses registros são equivalentes ao horizonte temporal de julho de 2018 a setembro de 2020 e apresentam os casos em que acidentes, doenças ou óbitos associados ao trabalho foram registrados por meio do sistema de Comunicação de Acidentes do Trabalho do INSS (CATWEB) ou ainda, nas situações em que ocorre concessão de benefício em virtude de incapacidade acidentária.

Os dados são disponibilizados por meio de documentos no formato *Comma-Separated Values* (*.CSV*) e agrupados a cada trimestre, portanto, até o momento da pesquisa, foram disponibilizados nove conjuntos de dados. Para este estudo todos os conjuntos de dados foram considerados e agrupados em um único documento, utilizando como ferramenta o Microsoft Excel®. Como resultado da união, o novo conjunto de dados apresentou 25 atributos e 990.870 instâncias em sua totalidade.

Buscando facilitar a interpretação dos dados e reduzir atributos em duplicidade, os 25 atributos foram desdobrados em cinco categorias distintas, sendo elas associadas: ao empregado, ao empregador, à ocorrência, ao benefício e à CAT. A divisão de cada classe é apresentada pela Figura 2 e a definição de cada variável é descrita na sequência.

Figura 2 – Variáveis do conjunto de dados



Fonte: A autora (2021)

Em relação aos atributos associados ao trabalhador, é apresentada sua data de nascimento, expressa pelo dia, mês e ano, assim como seu **gênero**, especificado como masculino, feminino, indeterminado ou não identificado. A partir da data de nascimento e da data do acidente, que também é apresentada, é possível calcular a **idade** do trabalhador. A Classificação Brasileira de Ocupações (**CBO**) descreve a ocupação ao qual o colaborador estava associado no momento do acidente, dividida em dez famílias no primeiro nível. Ainda para variáveis do empregado, a **filiação** é descrita pela ligação do mesmo com a Previdência Social, que pode ser definida como empregado, segurado especial ou trabalhador avulso.

Considerando as variáveis do empregador, a Classificação Nacional de Atividades Econômicas (**CNAE**) descreve o setor industrial associado ao empregador. O CNAE é dividido em seções e subseções, sendo o primeiro nível com 21 categorias, nomeadas pelas letras A à U e sua respectiva descrição. Já a **localização do empregador** é composta pela cidade e unidade federativa que a empresa está localizada.

Similarmente, a **localização do acidente** representa a unidade federativa onde o incidente realmente aconteceu. Além disso, nos atributos relacionados a ocorrência, encontra-se a **data da ocorrência** que é descrita por seu dia, mês e ano. O **tipo** de acidente define se a ocorrência é associada a um acidente típico, de trajeto ou a uma doença. Já o atributo de **agente causador** do acidente, com 241 categorias de agente causador para acidentes e 58 categorias para doenças, apresenta 273 opções nos dados selecionados.

Outro atributo ligado à ocorrência é Código Internacional de Doenças (**CID-10**), que apresenta 22 capítulos e busca classificar as doenças ou lesões as quais o trabalhador foi submetido. A **parte do corpo atingida** descreve 42 membros ou partes do corpo possíveis para associar ao acidente, assim como a **natureza da lesão** nomeia 28 características que explicam o tipo de lesão associada ao incidente ocorrido. Por fim, no que tange os atributos da ocorrência, também é descrito o **óbito**, que indica se houve morte ou não quando associada ao evento registrado pela CAT.

São três os atributos ligados ao benefício, sendo descritos pela **data do afastamento**, a **data de concessão do benefício** e seu tipo. Ambas as datas são definidas por mês e ano, respectivamente, quando o trabalhador foi afastado e quando o benefício foi concedido ao trabalhador ou seu dependente. Quanto ao **tipo de benefício**, são encontradas sete possibilidades no conjunto de dados: “Afastamento até 15 D”, “Auxílio Acidente”, “Auxílio Doenca por A”, “Auxílio Doenca Previ”, “Pa”, “Pensão por Morte Aci” e “Pensão por Morte Pre”. Devido a limitação de caracteres dos dados apresentados, algumas das categorias apresentadas anteriormente não são totalmente interpretáveis.

Por fim, a última classe de atributos são aqueles associados à CAT, descrevendo qual sua **data de emissão**, por meio de qual origem foi registrada e quem foi o responsável pela emissão. A data da emissão apresenta o dia, mês e ano em que foi registrada e a **origem** expõe qual o meio em que aconteceu o registro, que pode acontecer através da Internet ou do Projeto de Regionalização de Informações e Sistemas (PRISMA). O último atributo é o **emitente** da CAT, que pode ser de responsabilidade do empregador, do médico, do sindicato, de uma autoridade pública, ou mesmo do segurado e seu dependente.

4.2 Pré-processamento e transformação dos dados

Em busca de aprofundamento nos atributos dos dados e para eliminar *outliers*, uma etapa fundamental é o pré-processamento e limpeza dos dados. Para realizar essa etapa cada atributo foi analisado separadamente, utilizando como apoio a ferramenta Microsoft Excel®, a fim de identificar as categorias de cada atributo, bem como os valores faltantes. Para tanto, foi utilizado o conjunto de dados completo, com 990.870 e 20 atributos, removendo os quatro que apresentavam informações duplicadas no conjunto de dados inicial (CBO, CID-10, CNAE e data do acidente) e agrupando as informações de localização do empregador em apenas um.

A primeira variável adaptada foi a **idade** do trabalhador, que não era descrita no conjunto de dados inicial, mas era passível de cálculo utilizando a data do acidente e a data de nascimento. Este cálculo foi realizado fazendo a subtração das duas datas no Microsoft Excel® e utilizando a função de arredondamento, buscando obter um valor inteiro da idade do acidentado. Como resultado da operação foram observados 217 valores com data de nascimento ou do acidente inexistentes ou iguais a zero, impossibilitando o cálculo e possibilitando sua remoção deste atributo.

Além disso, foram observados valores para idade abaixo do permitido pela legislação trabalhista, como crianças de 8 anos, e valores de idade acima do esperado, como são os casos de trabalhadores com 97 anos. Para identificação dos *outliers* foi considerado o Art. 403 da Lei nº 10.097 de 19 de dezembro de 2000, que prevê trabalho legal apenas para indivíduos com mais de 16 anos. Sendo assim, foram encontrados e removidos 145 registros vinculados a idades inferiores a 16 anos.

Como idade máxima o limitante considerado foi de 75 anos, levando em conta que deve ser realizada aposentadoria compulsória para servidores públicos com essa idade (BRASIL, 2015). É fato que existem registros no conjunto dados que se caracterizam tanto como trabalhadores de organizações públicas como privadas, mas em vista da falta de legislação limitante para trabalhadores na segunda condição, foi considerada a lei vigente para servidores públicos. Desta situação de análise foram localizadas 329 ocorrências que correspondiam a idades superiores a 75 anos, resultando no total de 691 ocorrências (0,07%) no conjunto de dados com valores discrepantes ou inexistentes para idade.

Em relação ao **gênero** indicado nos registros foram destacados masculino, feminino, indeterminado ou não informado. Como gênero indeterminado foram identificadas 21 ocorrências e como não informado 954, resultando em 975 casos que podem ser removidos segundo a variável gênero do trabalhador (0,10%). Considerando a filiação do trabalhador, as

categorias apresentadas foram: empregado, segurado especial e trabalhador avulso, com apenas 16 instâncias preenchidas como "*ñ class*" descartadas do conjunto. Ainda nas variáveis relacionadas ao empregado, a **CBO** também é analisada, mas apenas em seu primeiro nível. Para este atributo foram encontradas 404 instâncias com valores não informados ou iguais a zero, representando 0,04% dos dados selecionados.

Considerando as variáveis do empregador, 9.609 registros apresentaram valores de **CNAE** que não estavam presentes na estrutura detalhada de código e denominações desenvolvida pela Comissão Nacional de Classificação (CONCLA). Esses valores foram identificados através da coluna de descrição da CNAE no conjunto de dados, que apresentava valores descritos como "*ñ class*", que ao serem confrontados com seus respectivos códigos, eram inexistentes na estrutura da CNAE. As ocorrências removidas, segundo o atributo de CNAE do empregador, representam 0,97% do conjunto de dados.

Quanto à **localização do empregador**, essa variável é representada pela cidade e estado que a empresa está inscrita. Analisando as cidades observadas no conjunto de dados, são destacadas 4.582 cidades diferentes com registros da CAT, do total 5.570 cidades do país indicadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Ainda em relação as cidades, apenas 85 ocorrências não apresentavam registros, podendo ser retiradas do conjunto de dados. No que se refere as unidades federativas todas as instâncias possuem preenchimento com o respectivo nome do estado, assim como todos os estados do país são observados.

Similarmente, em relação à **localização da ocorrência**, também é apresentado o nome do estado ao qual foi registrado o incidente. No entanto, para essa variável 309.992 instâncias estão registradas com valor "*ñ class*", resultando em mais de 31% dos dados sem apontamentos para este atributo. No contexto de variáveis relacionadas à ocorrência, em relação ao atributo de **data do acidente**, apenas sete instâncias apresentavam valores faltantes para a data do acontecimento. Já em relação ao **tipo** da ocorrência, que pode ser registrada como acidente típico, de trajeto ou doença, 16 instâncias apresentaram como categorização "ignorado", sendo removidos deste atributo.

O **agente causador** do acidente, conforme já mencionado, apresenta 273 categorias no conjunto de dados selecionado. No entanto, mesmo com diversas categorias para adequação, 13.748 instâncias (1,39%) são categorizadas como "*ñ class*" e estas se destacam no recorte dos últimos trimestres de dados disponibilizados. Pôde-se observar que no intervalo de julho a setembro de 2020 aproximadamente 50% dos registros "*ñ class*" são contemplados e de abril a junho de 2020 cerca de 26%.

Com relação ao atributo **CID-10**, descrito por um código, que representa cada capítulo de doenças ou lesão que o trabalhador sofreu, foram localizados 4.542 valores faltantes (0,46%) e 12.674 classificados como "*ñ class*" (1,28%). Considerando a **parte do corpo atingida** pelo acidente ou doença, 497 instâncias estão categorizadas como "*ñ class*" e outras 273 caracterizadas como "Localizacao da Lesao", em ambos os casos foram considerados *outliers* do atributo. Além disso, para a variável de **natureza da lesão** foram localizados 1.681 valores "*ñ class*", que não estavam relacionados a uma das 28 categorias. O último atributo ligado a ocorrência é o óbito, onde 4.406 acidentes e doenças foram associados a morte do trabalhador, correspondendo a 0,44% dos registros da CAT.

A classe de atributos relacionados ao benefício é a que apresenta maior número de dados faltantes. Para a **data do afastamento** 222.670 instâncias (22,47%) apresentaram a data 0000/00, sem informações para análise. Já em relação a **data de despacho do benefício** foram 979.177 registros com valor igual a zero, representando 98,82% dos dados impossibilitados de avaliação. Quanto ao **tipo de benefício** oferecido ao trabalhador ou a seu dependente, não existem dados faltantes, porém 98,81% das instâncias classificam o benefício como "Pa" e não foram encontradas definições para as categorias de espécie do benefício, impossibilitando a interpretação.

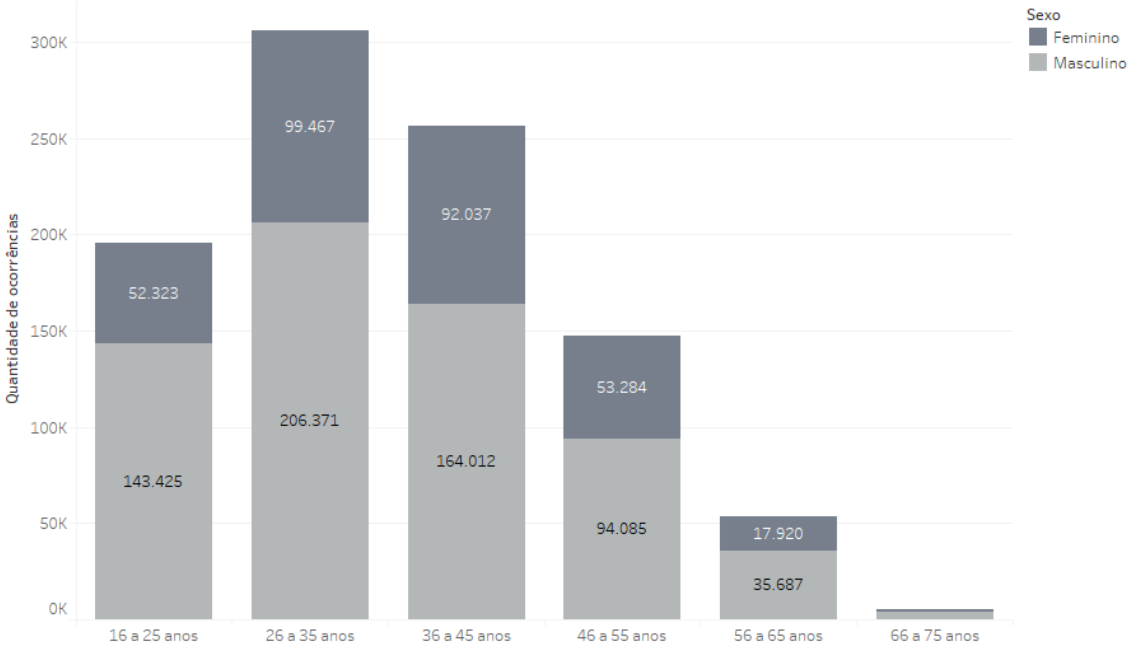
Por fim, a última classe de atributos considera a relação com a CAT, onde a **data de registro** é descrita pelo dia, mês e ano que foi formalizada e não apresenta valores faltantes. Já a **origem** do cadastramento da CAT pode ser classificada como Internet ou PRISMA, mas a última categoria só é observada em três instâncias da totalidade do conjunto de dados e não apresenta valores faltantes. O último atributo é o **emitente**, responsável pela emissão da CAT, representado por cinco possíveis categorias e também sem nenhuma instância com valores faltantes.

4.3 Análise exploratória dos dados

Após a coleta, pré-processamento e limpeza dos dados, a etapa seguinte é focada no entendimento e avaliação do conjunto de dados selecionado, assim como seus atributos. Para tanto foi desenvolvida uma análise exploratória dos dados, com foco em cada um dos atributos e suas categorias, considerando a remoção dos *outliers* e de atributos com informações duplicadas. Nesta etapa da pesquisa foram utilizadas como ferramentas de apoio os *softwares* Tableau® e Microsoft Excel®. Além disso, a fim de relacionar o conjunto de dados desta pesquisa com dados de controle, foram utilizadas informações da Relação Anual de Informações Sociais (RAIS), equivalente ao ano base de 2019 e 2018.

O primeiro grupo a ser avaliado são os dados relacionados ao empregado, representado por quatros atributos do conjunto de dados: idade, gênero, filiação e CBO. A Figura 3 apresenta a relação dos atributos **idade** e **gênero** do trabalhador com a quantidade de acidentes e doenças registradas. A fim de refinar a análise dos dados, o atributo idade foi agrupado em seis faixas etárias.

Figura 3 – Ocorrências relacionadas com gênero e idade do empregado



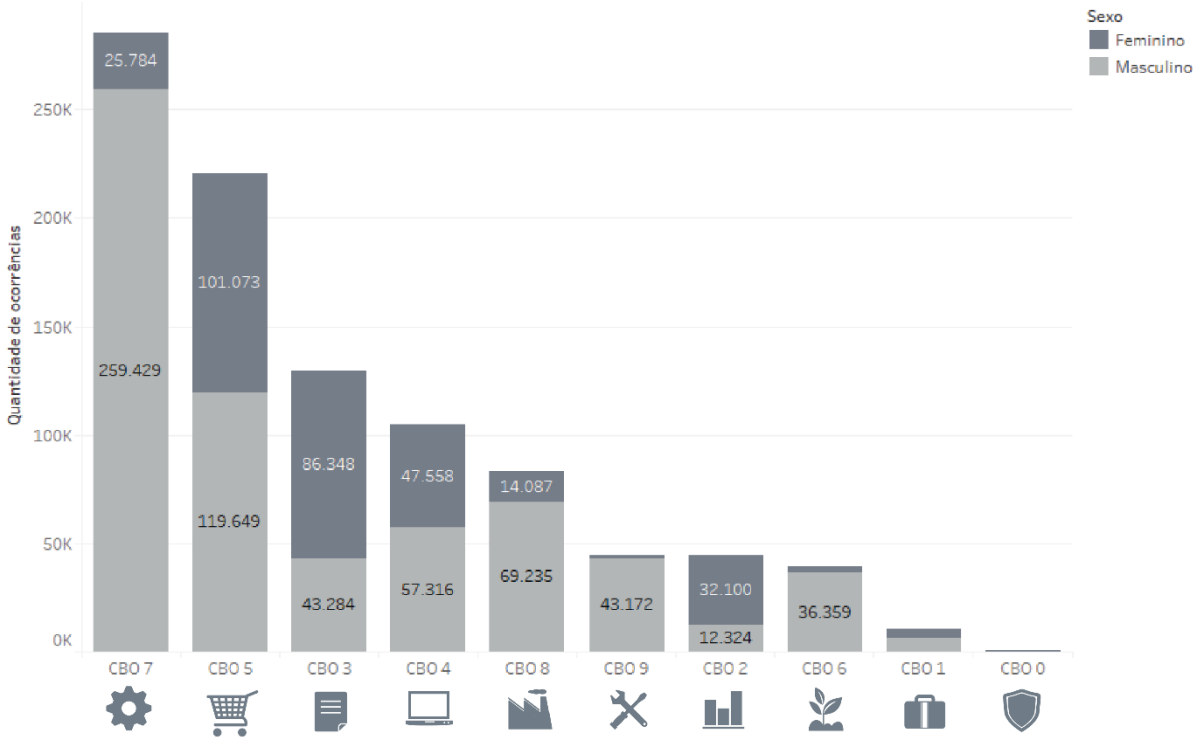
Fonte: A autora (2021)

A faixa etária com maior número de registros de CAT é a segunda, evidenciando trabalhadores de 26 a 35 anos, que representam 31% dos dados. Este grupo é seguido pela terceira faixa etária com 26% dos registros para trabalhadores com idades entre 36 e 45 anos. Juntas, as duas faixas etárias representam mais da metade do conjunto de dados selecionado. Esta tendência de registros da CAT acompanha os dados da RAIS nos últimos dez anos, onde a faixa etária com maior número de vínculos empregatícios é entre 30 e 39 anos.

Outro fator evidenciado pela Figura 3 é sua relação com o **gênero** do trabalhador, que é frequentemente associado ao sexo masculino, em mais de 67% dos casos de doenças e acidentes do trabalho. No entanto, quando comparado aos dados de controle da RAIS o percentual de trabalhadores do sexo masculino é de 56%, implicando que homens se acidentam mais no trabalho em comparação ao sexo oposto.

Para o atributo de **filiação**, a categoria de empregado registra 99,77% dos dados, restando apenas 262 casos de segurados especiais e 1.981 trabalhadores avulsos. A última variável do grupo relacionado ao empregado é a **CBO**, representada pela Figura 4, relacionando a quantidade de ocorrências com o gênero do trabalhador e a CBO característica da sua função.

Figura 4 - Ocorrências relacionadas com gênero e CBO do empregado



Fonte: A autora (2021)

A categoria com maior número de registros é a CBO 7, que equivale aos trabalhadores da produção de bens e serviços industriais e abrange aproximadamente 30% dos dados, em sua maioria composta por trabalhadores do sexo masculino. A segunda categoria em destaque é a CBO 5, descrita por trabalhadores dos serviços, vendedores do comércio em lojas e mercados, registrando aproximadamente 23% dos casos.

A partir da Figura 4 ainda é possível analisar quais CBOs são significativos para o sexo feminino, que se sobressai ao masculino apenas em dois casos: CBO 2 e 3. O CBO 2 é descrito por profissionais das ciências e das artes, enquanto o CBO 3 caracteriza-se por técnicos de nível médio. A RAIS ano base 2019 não apresenta informações detalhadas sobre o primeiro nível de CBO, portanto, foram feitos comparativos em relação ao ano base 2018.

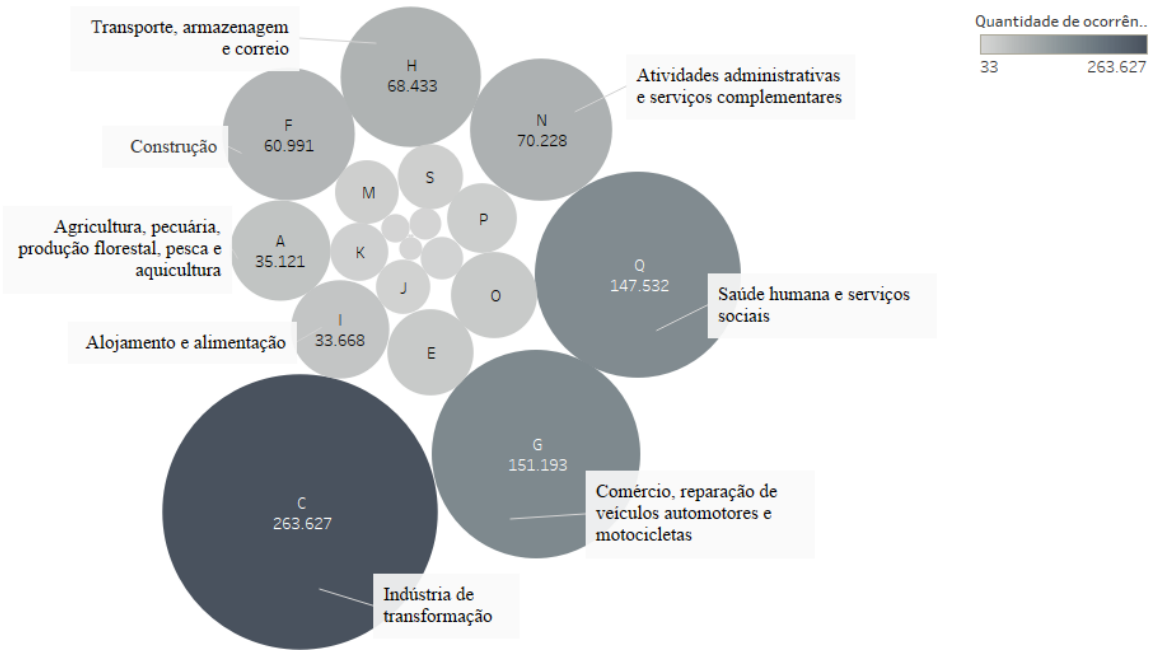
O CBO que possui mais trabalhadores associados segundo a RAIS é o grupo 5, que corresponde ao segundo lugar em registros do conjunto de dados da CAT. Em relação ao gênero, para as ocupações que empregam mais mulheres do que homens, as categorias de CBO 2 e 3 ocupam a primeira e terceira posição, respectivamente. De acordo com os dados da RAIS 2018 a representatividade de mulheres nestes setores é de 62% para trabalhadoras do CBO 2 e 57% para o CBO 3.

Em relação aos atributos associados ao empregador, a análise ocorreu em função das cidades e estados ligados ao registro da empresa, e o CNAE relacionado. Ao avaliar as **cidades**

com maior número de ocorrências o ranking foi liderado apenas por capitais, descritas por São Paulo, com 98.125 ocorrências, Rio de Janeiro (36.550), Belo Horizonte (21.464), Curitiba (19.211) e Porto Alegre (18.007). Considerando os **estados** com maior número de registros, São Paulo também está à frente, representando mais de 36% do conjunto de dados. Também possuem significância os estados de Minas Gerais (10,42%), Rio Grande do Sul (8,32%), Paraná (8,05%), Rio de Janeiro (6,47%) e Santa Catarina (6,16%). Todos os demais estados apresentam percentual inferior a 3%.

O atributo **CNAE** do empregador foi avaliado segundo as 21 seções previstas pelo CONCLA e são detalhadas na Figura 5, que descreve, por meio de um gráfico de bolhas, a quantidade de ocorrências relacionadas a cada setor industrial. Os setores com o maior número de registros estão detalhados na figura.

Figura 5 - Ocorrências relacionadas ao CNAE do empregador



Fonte: A autora (2021)

A indústria de transformação, caracterizada pela seção C do CNAE, é o setor industrial responsável pelo maior número de acidentes e doenças ligadas ao trabalhador, ocupando aproximadamente 27% da totalidade de registros. É seguido pela seção G, de comércio, reparação de veículos automotores e motocicletas com 15,26% e seção Q, de saúde humana e social responsável por 14,89% das ocorrências. As seções C, G e Q representam, juntas, mais de 56% dos registros de CAT realizados.

Considerando as variáveis relacionadas à ocorrência, o **óbito** foi associado a 4.406 registros de acidentes e doenças do conjunto de dados. Ainda em relação as variáveis do grupo da ocorrência, a **localização** representa o estado onde o acidente ocorreu, e neste caso apresenta

mais de 31% de registros sem classificação, ou seja, valores faltantes nesta categoria. Para este mesmo atributo, também foi observado que não há registros ligados a nenhum estado da região sul, sudeste, centro-oeste e ao estado da Bahia. Essas duas informações podem estar associadas, com os 31% de dados faltantes distribuídos nos estados sem registros, mas essa hipótese não é confirmada. Ainda em relação ao estado que ocorreu o acidente, o Maranhão é o mais evidente, acumulando 36% dos registros, seguido por Rondônia (10%) e Roraima (8%).

Quanto ao **tipo** da ocorrência, mais de 77% dos incidentes são registrados como acidentes típicos, seguidos por acidentes de trajeto com 19%, restando pouco mais de 3% dos casos registrados como doenças. Associando as ocorrências ao seu **agente causador**, havia inicialmente 273 categorias para associação, mas a fim de facilitar a análise essas categorias foram agrupadas em 32 novas classes, de acordo com a proximidade de utilização. O agrupamento das categorias é descrito no Apêndice A deste artigo. A Tabela 1 apresenta as dez classes com maior número de registros de CAT, assim como sua distribuição de acordo com o tipo de acidente.

Tabela 1 – Agentes causadores relacionados aos tipos de acidentes

Agente causador	**Total**	**Típico**	**Trajeto**	**Doença**
Veículos, meios de transporte e equipamentos de transporte	189.622	60.250	129.271	101
Superfícies e equipamentos utilizados para sustentar pessoas	155.998	115.037	40.878	83
Máquinas	92.520	90.795	1.540	185
Ferramentas manuais sem força motriz	83.973	83.084	842	47
Metais e minerais	53.433	52.426	977	30
Agente infeccioso, produto biológico e medicamentos	50.160	40.363	90	9.707
Aprisionamento, atrito, abrasão, impacto ou queda	39.964	33.847	4.485	1.632
Embalagem ou recipiente (vazio ou cheio)	31.342	31.052	215	75
Mobiliário e acessórios	31.274	30.678	508	88
Ser vivo	26.909	22.821	2.114	1.174

Fonte: A autora (2021)

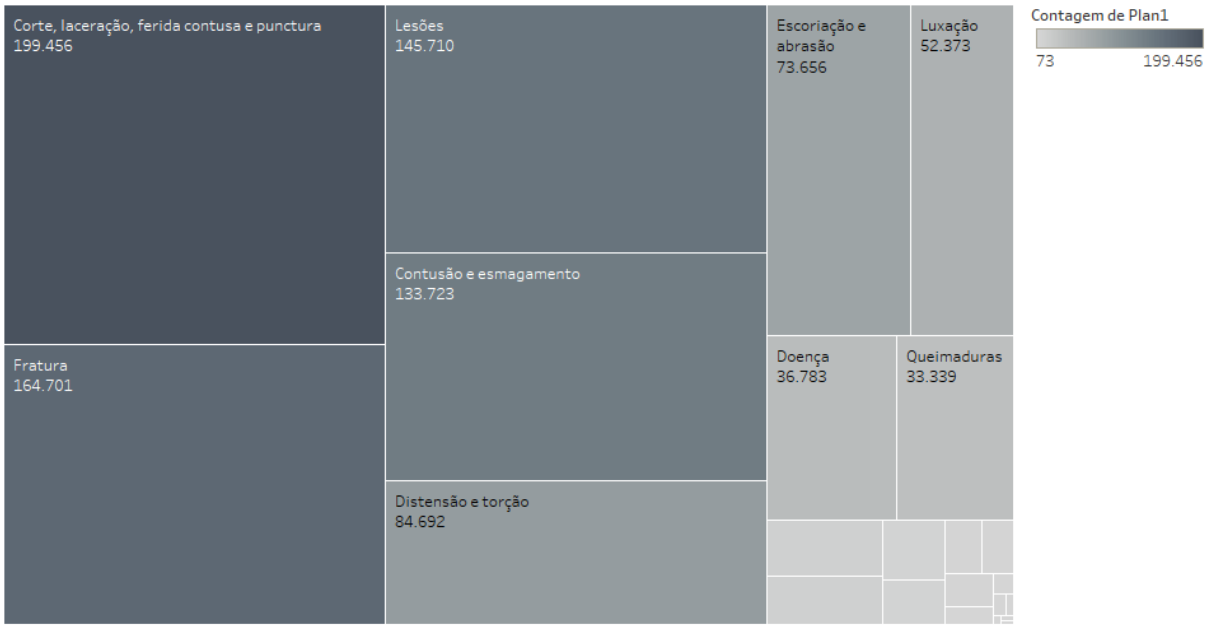
Os maiores agentes causadores de incidentes com registro de CAT estão associados a categoria de veículos, meios de transporte e equipamentos de transporte. Esta é a única classe

que apresenta maior percentual relacionado a acidentes de trajeto (68%), pois em todas as demais classes de agente causador se sobressai com os acidentes típicos.

Com relação ao **CID-10**, atributo que descreve a classificação de doenças e lesões, a predominância dos registros é relacionada ao capítulo XIX. Este capítulo caracteriza-se por lesões, envenenamentos e algumas outras consequências de causas externas e evidencia 74% dos dados selecionados. Em destaque também estão capítulos XIII - Doenças do sistema osteomuscular e do tecido conjuntivo (6,3%), XXI - Fatores que influenciam o estado de saúde e o contato com os serviços de saúde (6,2%) e XX - Causas externas de morbidade e de mortalidade (6%).

Considerando a **natureza da lesão**, os 28 grupos apresentados no conjunto de dados foram agrupados em 23 categorias. A Figura 6 apresenta todas as categorias e quantidade de ocorrências associadas e apresenta a descrição daquelas com maior incidência.

Figura 6 – Ocorrências associadas a natureza da lesão



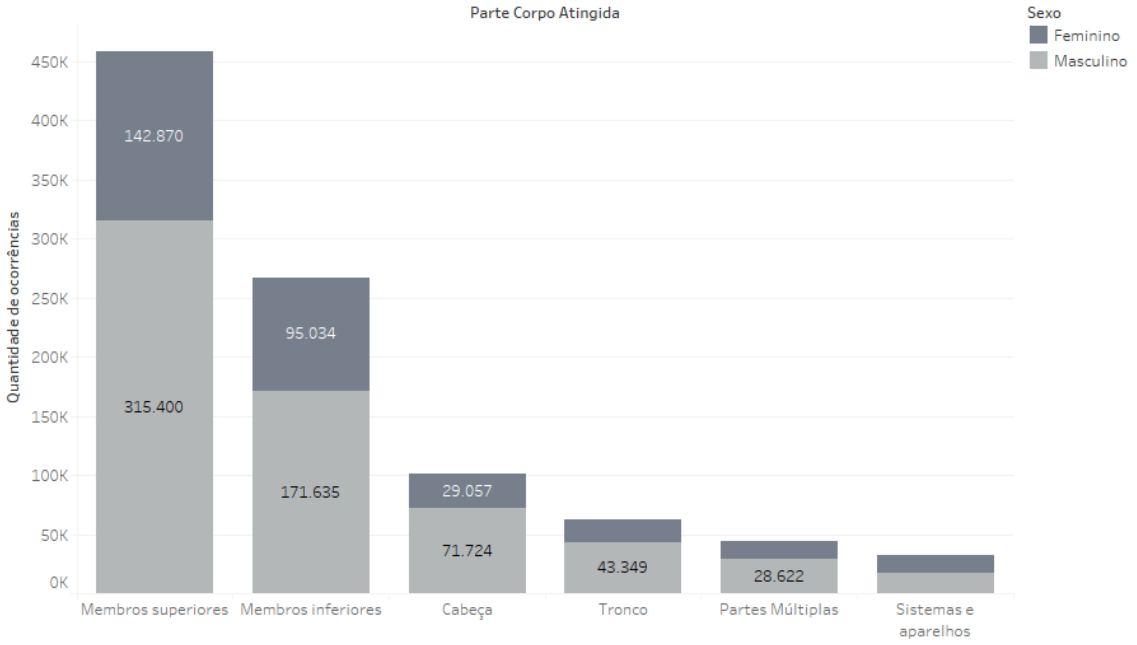
Fonte: A autora (2021)

A classe de lesão com maior número de registros é representada por corte, laceração, ferida contusa e punctura, englobando mais de 20% dos dados selecionados. Essa categoria é seguida por fraturas, com aproximadamente 17% dos casos, lesões com 14,7% e contusão e esmagamento representando 13,5%. Juntas, as cinco classes de lesões representam 65% do conjunto de dados selecionado.

O último atributo ligado à ocorrência é a **parte do corpo atingida** pelo acidente ou doença. No conjunto de dados selecionado eram observadas 41 categorias, que foram agrupadas em seis novas possibilidades de classificação, sendo elas: cabeça, membros superiores,

membros inferiores, partes múltiplas, sistemas e aparelhos, e tronco. A Figura 7 cada uma das partes do corpo, assim como sua incidência segundo o gênero do trabalhador.

Figura 7 – Ocorrências relacionadas com a parte do corpo atingida e gênero do empregado



Fonte: A autora (2021)

Membros superiores são o grupo com maior número de registros de acidentes e doenças, representando mais de 46% do conjunto de dados selecionado. Neste grupo são englobados: braço, antebraço, cotovelo, punho, mão, dedo, ombros e membros superiores. O segundo grupo de maior incidência são os membros inferiores (27%), que correspondem à: artelho, articulação do torno, coxa, joelho, perna, pé e membros inferiores. Os demais agrupamentos referentes a variável de parte do corpo atingida são apresentados no Apêndice B.

Quanto as variáveis relacionadas ao benefício, duas delas são descritas por datas, onde ocorreu o **afastamento** e a **concessão do benefício**. No entanto, estas apresentam grande percentual de dados faltantes, que dificultam a análise. De maneira análoga, o atributo de **espécie de benefício** impede sua avaliação, pois aproximadamente 99% dos dados estão relacionados a categoria “Pa”, que não apresenta descrição. Além disso, as outras seis categorias de espécie de benefício também apresentam descrição incompleta, dificultando a diferenciação entre as categorias.

Por fim, o último grupo de atributos está ligado a CAT e são detalhados através da data de registro, origem e emitente. Conforme apresentado previamente, a **origem** apresenta apenas três casos com registros pelo PRISMA, dentre a totalidade do conjunto de dados. Já o **emitente** é representado em maior parcela pela classe do empregador (98,43%), seguido por sindicato,

segurado ou seu dependente, médico e autoridade pública, com respectivamente 7.806 casos, 4.590, 1.866 e 1.296.

Ao final da análise exploratória foi possível identificar variáveis duplicadas e não confiáveis, que poderiam ser removidas do conjunto de dados a fim de realizar futuras aplicações. As variáveis com informações duplicadas e que foram apresentadas apenas uma vez no conjunto de dados, são: data do acidente, CBO e CID-10, CNAE. Já as variáveis consideradas não confiáveis e que poderiam ser retiradas de análises posteriores foram relacionadas à ocorrência, ao benefício e à CAT.

Segundo a localização da ocorrência, sua justificativa para exclusão está relacionada aos dados faltantes, que representam mais de 31% e aos estados inexplorados de algumas regiões do país. Em relação à data de afastamento e de concessão do benefício, assim como seu tipo, a justificativa se embasa na quantidade de dados faltantes e na dificuldade de interpretação deles. Já em relação aos atributos do grupo da CAT, a origem não apresenta significância para análise, pois apenas três instâncias descrevem um valor diferente.

De acordo com a seção anterior, onde também foram destacados os *outliers* que poderiam ser retirados, elaborou-se um novo conjunto de dados, a partir das adaptações realizadas. Após a transformação, o conjunto de dados que poderia ser utilizado em futuras aplicações conta com 948.405 instâncias e 12 atributos.

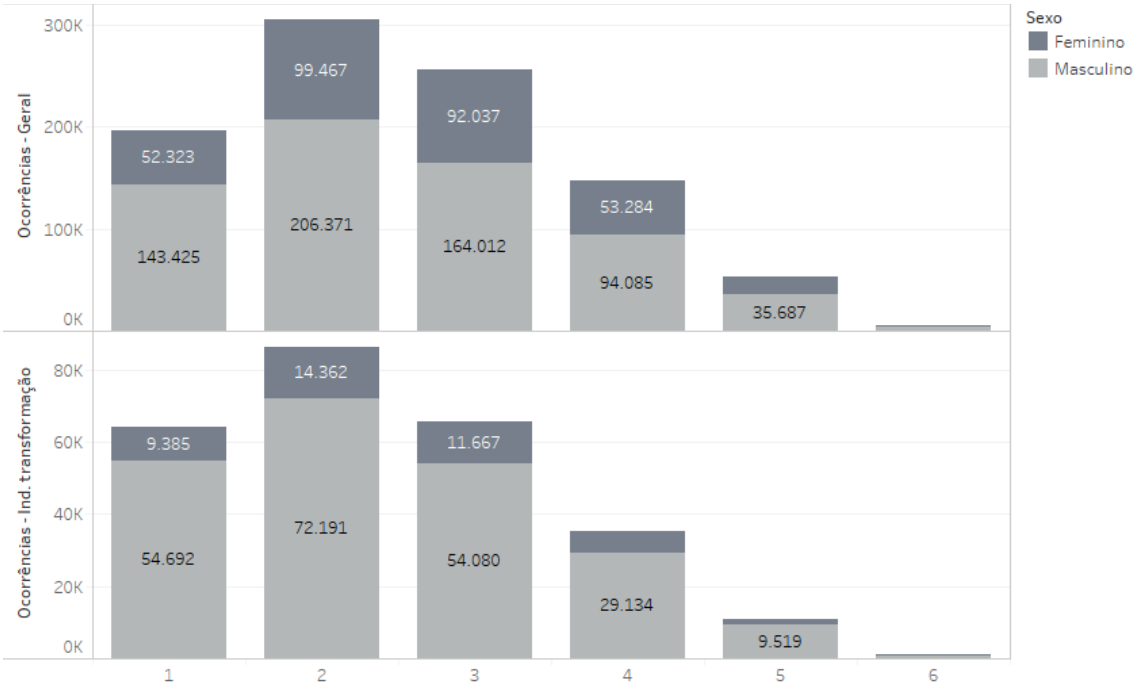
4.4 Análise exploratória da indústria de transformação

A partir da análise exploratória do conjunto de dados total, foi possível identificar as categorias de cada atributo que eram mais observadas. Em relação ao atributo de **CNAE** do trabalhador, o setor industrial que apresentou maior número de registros foi a indústria de transformação, que neste tópico passará também por uma análise exploratória. Esta avaliação busca identificar o comportamento desse setor em específico para os atributos, em relação ao conjunto de dados de maneira geral.

Foram observadas 267.616 instâncias associadas à indústria de transformação, que após a remoção dos *outliers* resultou em 263.627. Dessa totalidade, 750 registros estavam relacionados a casos com **óbito** do trabalhador (0,28%), assemelhando a uma letalidade de 0,44% do conjunto de dados de maneira geral. Se comparado a outros setores industriais, é o que representa maior número de óbitos, seguido por transporte, armazenagem e correio (668) e comércio, veículos automotores e motocicletas (614).

No entanto, o número absoluto de óbitos não é único indicador que pode ser analisado. Mesmo representando o maior número de registros, a indústria de transformação não é o setor com maior percentual de letalidade em seus acidentes. Este, conforme já apresentado, apresenta 0,28% de ocorrências associadas ao óbito do trabalhador, enquanto o setor de transporte, armazenagem e correio descreve 0,98% dos registros com essa característica, liderando essa categoria. Considerando o **gênero** e **faixa etária** predominantes também se mantiveram em relação a análise geral, conforme representação por meio da Figura 8.

Figura 8 – Ocorrências gerais e da indústria de transformação relacionadas com gênero e idade do empregado



Fonte: A autora (2021)

Mesmo com a predominância do gênero masculino e segunda faixa etária (26 a 35 anos), houve variações nos percentuais de cada atributo. No conjunto de dados completo o percentual de trabalhadores do **sexo** masculino era de 67% para essa faixa etária, passando para 84% na análise da indústria de transformação. Em comparação a outra setores industriais, esta não descreve o maior percentual de acidentes em homens, pois quem ocupa essa posição é o setor de construção, com mais de 96% de trabalhadores do sexo masculino. Em contrapartida, o setor com maior percentual feminino é a saúde humana e serviços sociais, onde ocupa 80% dos registros.

Ainda em relação a Figura 8, a faixa etária de 26 a 35 anos equivalia a 31% dos registros, passando à 33% para a indústria de transformação. Analisando outros setores, nem sempre se mantém a segunda faixa etária como predominante. Em oito dos 21 setores a faixa etária com

maior número de registros é a terceira, descrita por trabalhadores de 36 a 45 anos. Essa categoria representa a segunda em número de registros, tanto para o conjunto geral, como para o recorte da indústria de transformação.

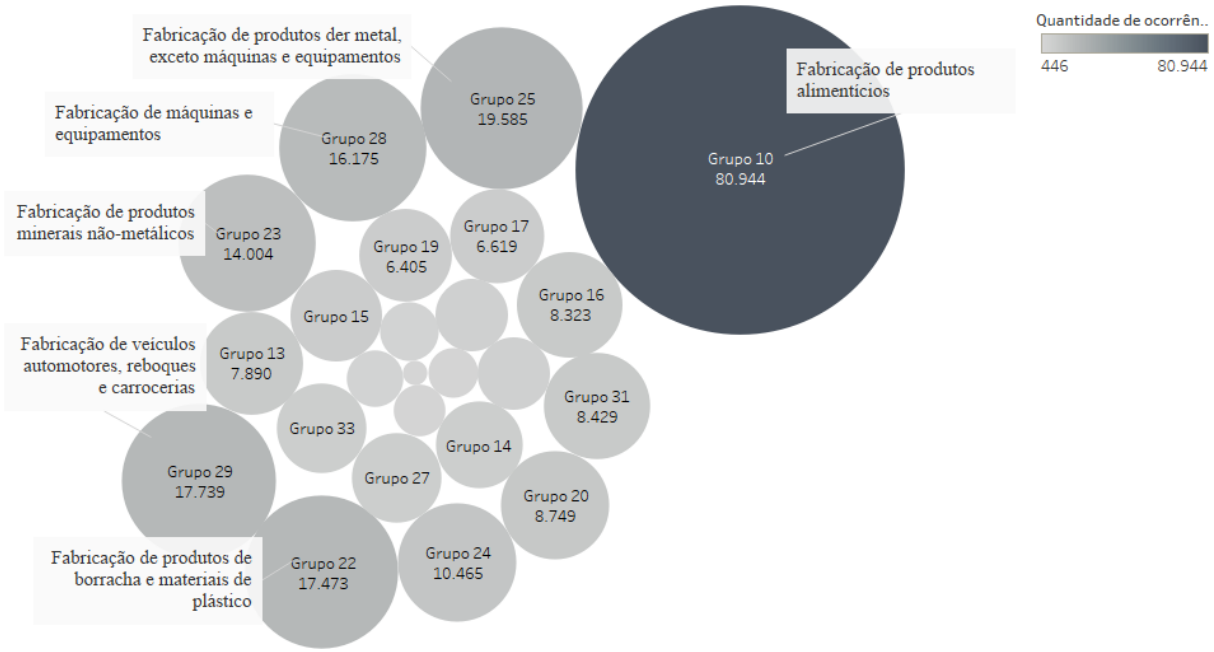
Em relação a **filiação**, 98,5% são empregados e o restante são trabalhadores avulsos, sem nenhum registro classificado como segurado especial. Essa predominância se mantém também em relação aos demais setores industriais. No que diz respeito ao **CBO**, a classe predominante continua sendo o CBO 7, que representa trabalhadores da produção de bens e serviços industriais, antes com 30% dos registros e aqui descrevendo 53%. Em seguida, o segundo CBO em destaque é o 8 (22%) que reflete também trabalhadores da produção de bens e serviços industriais, mas com diferentes subníveis. Esta predominância de CBOs é esperada, devido ao CNAE que está sendo analisado, de forma que, avaliando outros setores industriais, esse atributo muda de comportamento.

O CBO 7 ainda é predominante em outros sete setores industriais, com maior destaque para a construção e o transporte, armazenagem e correio. Já o CBO 5 (trabalhadores dos serviços, vendedores do comércio em lojas e mercados) é o responsável pelo maior percentual de acidentes em dez CNAEs, cuja maior representatividade está no setor de comércio, reparação de veículos automotores e motocicletas e atividades administrativas e serviços complementares. O CBO 3 (técnicos de nível médio) também apresenta relevância, e seu destaque está no setor de saúde humana e serviços sociais. Em todos os casos de associação do CNAE ao CBO predominante, é possível observar a relação entre o setor industrial analisado e as ocupações esperadas para seus trabalhadores.

Quanto a **localização do empregador**, São Paulo continua sendo o estado com maior número de registros (35%), seguido por Rio Grande do Sul (11%), Santa Catarina (10%), Paraná (10%) e Minas Gerais (9%). Similarmente, os demais setores industriais têm seus registros associados ao estado de São Paulo, com exceção das indústrias extrativistas cujo estado predominante é Minas Gerais (1.357 ocorrências) e organismos internacionais e outras instituições extraterritoriais, onde todos seus registros estão ligados a organizações do Distrito Federal (35 ocorrências). Em ambos os casos, é possível validar essa predominância, de acordo com as características locais dos empregos nestes estados.

Passando a análise ao **CNAE** do empregador em segundo nível, pois o primeiro nível é a indústria de transformação, selecionada como um recorte dos dados nesta etapa, são descritos 24 grupos. Estes grupos e sua incidência no conjunto de dados selecionado são detalhados na Figura 9.

Figura 9 - Ocorrências relacionadas aos subgrupos de CNAE do empregador



Fonte: A autora (2021)

Neste segundo nível, o grupo de CNAE que apresenta o maior número de registros é a fabricação de produtos alimentícios, com mais de 30% dos casos, representada pelo grupo 10. Também se destacam a fabricação de produtos de metal, exceto máquinas e equipamentos (grupo 25) com 7,29%, fabricação de veículos automotores, reboques e carrocerias (grupo 29) com 6,64% e fabricação de produtos de borracha e material plástico (grupo 22) com 6,52%.

Considerando o **tipo** do acidente registrado 84% se caracterizam como acidente típico, percentual superior se comparado ao conjunto de dados total que apresentou 77%. Em contrapartida o percentual de acidentes de trajeto caiu, de 19% para 14%, e as doenças de 3% para 2% de incidência. Além da indústria de transformação, para todos os demais CNAEs o tipo predominante são os acidentes típicos, seguidos por acidentes de trajeto, com exceção de atividades financeiras, de seguros e serviços relacionados, onde as doenças estão em segundo lugar de ocorrências e acidentes de trajeto em terceiro.

Quanto ao **CID-10**, o mesmo capítulo predominante no conjunto de dados total se mantém para a indústria de transformação, representado por lesões, envenenamentos e algumas outras consequências de causas externas (capítulo XIX), antes com 74% dos dados selecionados e agora com aproximadamente 85%. Para o atributo de **natureza da lesão**, as categorias com maior número de registros continuam sendo a de corte, laceração, ferida contusa e punctura (27%), fratura (17%), contusão e esmagamento (16%) e lesões (11%). Com esses percentuais é possível observar uma permanência das quatro categorias mais incidentes, mas com alterações na sua ordenação e percentuais, em comparação com o total.

No entanto, mesmo representando uma predominância em número absoluto, a categoria de corte, laceração, ferida contusa e punctura é destaque em apenas quatro setores, além da indústria de transformação. A fratura como categoria de natureza da lesão, é a mais observada nessa comparação, pois está em evidência no maior número de setores industriais (dez), principalmente representada pelo CNAE de comércio, reparação de veículos automotores e motocicletas.

Em relação as **partes do corpo atingidas**, agora observa-se um destaque ainda maior para os membros superiores, que no geral representavam 46% dos registros e agora saltaram para 57%. Os membros inferiores aparecem em seguida, mas estes apresentam uma queda, antes com 27% e agora com 22%. Essa ordenação de categorias passa por uma inversão para os setores de educação, informação e comunicação e eletricidade e gás, onde os membros inferiores possuem maios número de registros do que membros superiores.

O último atributo a ser analisado é o **agente causador** do acidente, que no conjunto de dados geral predominavam os registros associados à veículos, meios de transporte e equipamentos de transporte. Aqui, para a indústria de transformação, essa categoria está na segunda posição, com apenas 11%, pois o agente causador mais recorrente são as máquinas, representando 20% dos dados. Também se destacam as categorias de ferramentas manuais sem força motriz (11%), superfícies e equipamentos utilizados para sustentar pessoas (9,4%) e metais e minerais (9,2%).

O comportamento de predominância de veículos, meios de transporte e equipamentos de transporte como principal agente causador, é observado em parte dos setores industriais (dez). As superfícies e equipamentos utilizados para sustentar pessoas são destaque em outros seis setores industriais, sendo assim, os dois últimos agentes causadores citados são aqueles com maior impacto nos CNAEs. No entanto, algumas exceções são observadas, como o caso da indústria de transformação já citado, o setor de saúde humana e serviços sociais com predominância de agente infeccioso, produto biológico e medicamentos, e o setor de agricultura, pecuária, produção florestal, pesca e agricultura, cuja predominância é a classe de agente causador de produtos alimentícios e/ou de origem animal.

5. Considerações finais

Este trabalho analisou os casos de doenças, acidentes e óbitos relacionados ao trabalho no cenário brasileiro. Mediante esse objetivo e em meio as possibilidades de conjuntos de dados nacionais, optou-se pelos dados com abertura de CAT, devido a sua abrangência para todos os

tipos de incidentes e por sua extensão em todo território nacional. Após a seleção do conjunto de dados, com a realização da análise exploratória, foi possível entender as instâncias e atributos representados nos dados selecionados.

Os resultados dessa pesquisa, derivados da análise exploratória dos dados, contribuíram para o estudo das perspectivas e extração de informações relacionadas às doenças e acidentes de trabalho no Brasil. Dessa forma, apresentando relevância para o ambiente acadêmico, empresarial e governamental, por meio da avaliação do cenário nacional a partir de dados atuais e exibição das categorias com maior destaque em cada atributo. Ações estas, que auxiliam na criação de políticas e diretrizes com enfoque em SST.

As primeiras conclusões com relação a essa análise foram as variáveis não confiáveis do conjunto, que apresentavam grande percentual de informações incompletas, como o estado em que ocorreu o acidente ou as variáveis relacionadas ao benefício em sua totalidade. Também foram considerados os *outliers* presentes no conjunto de dados, como os casos de idade discrepantes com a faixa etária legal para trabalho ou as ocorrências com registro "*ñ class*" para algumas categorias.

Dentre as variáveis com registros válidos, foi possível constatar que o gênero mais representativo é o mais masculino nas ocorrências e que idades entre 20 e 40 anos correspondem a faixa etária com maior número de casos. A ocupação profissional que mais registra acidentes também foi avaliada e está relacionada ao grupo de trabalhadores da produção de bens e serviços industriais, assim como o setor industrial com maior número de registros é a indústria de transformação. Esses empregadores estão localizados em sua grande maioria no estado de São Paulo e considerando as cidades com maior destaque para registros de acidentes, as mais representativas são capitais.

Considerando as variáveis relativas à ocorrência, aproximadamente 77% registravam acidentes típicos. Em relação ao agente causador, a categorias com maior número de registros foi a de veículos, meios de transporte e equipamentos de transporte, enquanto para classificação das doenças, o capítulo XIX descrito por lesões, envenenamentos e algumas outras consequências de causas externas, foi o destaque. Quanto a parte do corpo atingida com maior número de ocorrências, se destacou membros superiores, seguidos por membros inferiores. Para a natureza da lesão, a classe de corte, laceração, ferida contusa e punctura foi frequentemente registrada. Também foi possível observar através dos dados o responsável pela abertura da CAT concluindo que na maioria das vezes os comunicados são registrados pelo próprio empregador.

Com o estudo exploratório realizado foi possível observar os dados de maneira geral e ainda tecer um estudo exploratório específico para o CNAE de indústria de transformação. Este,

por sua vez, apresentou similaridade em grande parte dos atributos de destaque, variando a ordenação e percentual de incidência em relação ao conjunto de dados total. Também foi possível observar que no segundo nível do CNAE, aquele com maior número de registros é responsável pela fabricação de produtos alimentícios.

Por fim, esta pesquisa permitiu realizar uma análise mais atenta ao conjunto de dados da CAT, identificação de dados faltantes e *outliers*, assim como o agrupamento e exclusão de alguns atributos. Essa ação permite a construção de um novo conjunto de dados, com instâncias e atributos adaptados, que possibilitem a aplicação de métodos para estudos mais aprofundados. Como limitações do estudo são pontuadas a dificuldade de encontrar definição para algumas categorias de dados, obscurecendo a análise, assim como dados faltantes e conflitantes, que provocam a criação de hipóteses, mas estas, sem confirmação na pesquisa.

São recomendações de pesquisas futuras, estudos que apliquem métodos estatísticos e de mineração de dados, buscando encontrar padrões, prever novos acidentes ou até mesmo a letalidade das ocorrências. Também são indicações, o desenvolvimento de ações que busquem melhorar a coleta de dados, evitando erros de preenchimento e facilitando a interpretação dos dados. A presente pesquisa, por meio da apresentação de dados e análises de SST, poderá motivar discussões acerca da criação de novas políticas e normas voltadas a SST, que auxiliem na tomada de decisão das organizações.

Referências

ANTONIOLLI, S. A. C.; ASSENATO, A. P. R.; ARAÚJO, B. R.; LAGRANHA, V. E. C.; SOUZA, L. M.; PAZ, A. A. Construção e validação de recursos educativos digitais para a saúde e segurança do trabalhador. **Revista Gaúcha de Enfermagem**, v. 42, 2021.

ANYFANTIS, I. D.; BOUSTRAS, G. The effects of part-time employment and employment in rotating periods on occupational accidents. The case of Greece. **Safety Science**, v. 121, p. 1–4, 2020.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14280: cadastro de acidente do trabalho - Procedimento e classificação**. Rio de Janeiro, p. 1-94, 2001.

AZIZ, S. F. A.; OSMAN, F. Does compulsory training improve occupational safety and health implementation? The case of Malaysian. **Safety Science**, v. 111, p. 205–212, 2019.

AZZOLIN, K.; SOUZA, E. N.; RUSCHEL, K. B.; MUSSI, C. M.; LUCENA, A. F; RABELO, E. R. Consenso de diagnósticos, resultados e intervenções de enfermagem para pacientes com insuficiência cardíaca em domicílio. **Revista Gaúcha de Enfermagem**, v. 33, no. 4, p. 56–63, 2012.

BARLAS, B.; IZCI, F. B. Individual and workplace factors related to fatal occupational

accidents among shipyard workers in Turkey. **Safety Science**, v. 101, p. 173–179, 2018.

BEVILACQUA, M.; CIARAPICA, F. E.; GIACCHETTA, G. Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach. **Accident Analysis & Prevention**, v. 40, no. 4, p. 1468–1479, 2008.

BRASIL. Decreto nº 3.048 de 06 de Maio de 1999. **Regulamento da Previdência Social**. Brasília, DF, 6 mai. 1999. 178º da Independência e 111º da República. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/d3048.htm>. Acesso em: 10 mar. 2021.

CHEN, H.; HOU, C.; ZHANG, L.; LI, S. Comparative study on the strands of research on the governance model of international occupational safety and health issues. **Safety Science**, v. 122, p. 104513, 2020.

CHENG, C. W.; LIN, C. C.; LEU, S. SEN. Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. **Safety Science**, v. 48, no. 4, p. 436–444, 2010.

CHENG, C. W.; YAO, H. Q.; WU, T. C. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. **Journal of Loss Prevention in the Process Industries**, v. 26, no. 6, p. 1269–1278, 2013.

CHENG, C.-W.; LEU, S.-S.; CHENG, Y.-M.; WU, T.-C.; LIN, C.-C. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. **Accident Analysis & Prevention**, v. 48, p. 214–222, 2012.

CHIODI, M. B.; MARZIALE, M. H. P.; MONDADORI, R. M.; ROBAZZI, M. L. C. C. Acidentes registrados no centro de referência em saúde do trabalhador de Ribeirão Preto, São Paulo. **Revista Gaúcha de Enfermagem**, v. 31, no. 2, p. 211-217, 2010.

CHOKOR, A.; NAGANATHAN, H.; CHONG, W. K.; ASMAR, M. E. Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning. **Procedia Engineering**, v. 145, p. 1588–1593, 2016.

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

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

DEL POZO-ANTÚNEZ, J. J.; ARIZA-MONTES, A.; FERNANDÉZ-NAVARRO, F.; MOLINA-SANCHÉZ, H. Effect of a job demand-control-social support model on accounting professionals' health perception. **International Journal of Environmental Research and Public Health**, v. 15, no. 11, 2018.

DOS SANTOS, B. S.; STEINER, M. T. A.; FENERICH, A. T.; LIMA, R. H. P. 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.

GOH, Y. M.; UBEYNARAYANA, C. U. Construction accident narrative classification: An evaluation of text mining techniques. **Accident Analysis & Prevention**, v. 108, p. 122-130, 2017.

GYEKYE, S. A.; SALMINEN, S.; OJAJARVI, A. A theoretical model to ascertain determinates of occupational accidents among Ghanaian industrial workers. **International Journal of Industrial Ergonomics**, v. 42, no. 2, p. 233–240, 2012.

HENNINGTON, É. A.; MONTEIRO, M. O perfil epidemiológico dos acidentes de trabalho no Vale dos Sinos e o sistema de vigilância em saúde do trabalhador. **História, Ciências, Saúde-Manguinhos**, v. 13, no. 4, p. 865-876, 2006.

ILO - INTERNATIONAL LABOUR ORGANIZATION. **C187 - Promotional Framework for Occupational Safety and Health Convention**, 2006. Disponível em: <https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO::P12100_ILO_CODE:C187>. Acesso em: 10 mar. 2021.

ILO - INTERNATIONAL LABOUR ORGANIZATION. **ILO 100: Law for Social Justice**. Geneva, 2019a. Disponível em: <https://www.ilo.org/global/about-the-ilo/how-the-ilo-works/departments-and-offices/jur/law-for-social-justice/WCMS_730958/lang--en/index.htm>. Acesso em: 10 mar. 2021.

ILO - INTERNATIONAL LABOUR ORGANIZATION. **Rules of the game: An introduction to the standards-related work of the International Labour Organization**, 2019b. Disponível em: <https://www.ilo.org/global/standards/information-resources-and-publications/publications/WCMS_672549/lang--en/index.htm>. Acesso em: 10 mar. 2021.

JUSTIÇA DO TRABALHO. Número de acidentes de trabalho no Brasil e no RS segue alto. **TRT da 4ª Região (RS)**, Ago. 2020. Disponível em: <https://www.trt4.jus.br/portais/trt4/modulos/noticias/305976>. Acesso em: 10 mar. 2021.

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

LEITE, A. F.; NOGUEIRA, J. A. D. Fatores condicionantes de saúde relacionados ao trabalho de professores universitários da área da saúde: uma revisão integrativa. **Revista Brasileira de Saúde Ocupacional**, v. 42, 2017.

LIAO, C.-W.; PERNG, Y.-H. Data mining for occupational injuries in the Taiwan construction industry. **Safety Science**, v. 46, no. 7, p. 1091–1102, 2008

MINISTÉRIO DA ECONOMIA. **Relação Anual de Informações Sociais – RAIS 2019**, 2020. Disponível em: <https://static.poder360.com.br/2020/10/Sumario-Executivo_RAIS-2019.pdf>. Acesso em: 10 mar. 2021.

MINISTÉRIO DA FAZENDA. **Anuário Estatístico de Acidentes de Trabalho - AEAT 2018**, 2018. Disponível em: <https://www.gov.br/previdencia/pt-br/assuntos/previdencia-social/saude-e-seguranca-do-trabalhador/dados-de-acidentes-do-trabalho/arquivos/aeat-2018.pdf>. Acesso em: 10 mar. 2021.

MINISTÉRIO DO TRABALHO E EMPREGO. **Classificação Brasileira de Ocupações: CBO – 2010**. 3ª ed. Brasília: MTE, SPPE, 2010, v.1, p. 828, 2010. Disponível em: <https://wp.ufpel.edu.br/observatoriosocial/files/2014/09/CBO-Livro-1.pdf>. Acesso em: 10 mar. 2021.

MISTIKOGLU, G.; GEREK, I. H.; ERDIS, E.; USMEN, P. E. M.; CAKAN, H.; KAZAN, E. E. Decision tree analysis of construction fall accidents involving roofers. **Expert Systems with Applications**, v. 42, no. 4, p. 2256–2263, 2015.

MUTLU, N. G.; ALTUNTAS, S. Risk analysis for occupational safety and health in the textile industry: Integration of FMEA, FTA, and BIFPET methods. **International Journal of Industrial Ergonomics**, v. 72, p. 222–240, 2019.

NANDA, G.; GRATTAN, K. M.; CHU, M. T.; DAVIS, L. K.; LEHTO, M. R. Bayesian decision support for coding occupational injury data. **Journal of Safety Research**, v. 57, p. 71–82, 2016.

NENONEN, N. 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, no. 2, p. 215–224, 2013.

PALAMARA, F.; PIGLIONE, F.; PICCININI, N. Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. **Safety Science**, v. 49, no. 8–9, p. 1215–1230, 2011.

RAMOS, D.; AFONSO, P.; RODRIGUES, M. A. Integrated management systems as a key facilitator of occupational health and safety risk management: A case study in a medium sized waste management firm. **Journal of Cleaner Production**, v. 262, p. 121346, 2020.

SÁNCHEZ-HERRERA, I. S.; DONATE, M. J. Occupational safety and health (OSH) and business strategy: The role of the OSH professional in Spain. **Safety Science**, v. 120, p. 206–225, 2019.

SANNI ALI, M.; ICHIARA, M. Y.; LOPES, L. C.; BARBOSA, G. C. G.; PITA, R.; CARREIRO, R. P.; SANTOS, D. B.; RAMOS, D.; BISPO, N.; RAYNAL, F.; CANUTO, V.; ALMEIDA, B. A.; FIACCONE, R. L.; BARRETO, M. E.; SMEETH, L.; BARRETO, M. T. Administrative data linkage in Brazil: Potentials for health technology assessment. **Frontiers in Pharmacology**, v. 10, 2019.

SANNI-ANIBIRE, M. O.; MAHMOUD, A. S.; HASSANAIN, M. A.; SALAMI, B. A. A risk assessment approach for enhancing construction safety performance. **Safety Science**, v. 121, p. 15–29, 2020.

SHIN, D.-P.; PARK, Y.-J.; SEO, J.; LEE, D.-E. Association Rules Mined from Construction Accident Data. **KSCE Journal of Civil Engineering**, v. 22, no. 4, p. 1027-1039, 2018.

SIEMINKOSKI, T. **Data mining: clustering aplicado a banco de dados de acidentes de trabalho**. 2017. 38 f. Trabalho de Conclusão de Curso (Especialização) - Universidade Tecnológica Federal do Paraná, Pato Branco, 2017. Disponível em: < http://repositorio.roca.utfpr.edu.br/jspui/handle/1/9841>. Acesso em: 20 mar. 2021.

TIXIER, A. J.-P.; HALLOWELL, M. R.; RAJAGOPALAN, B.; BOWMAN, D. Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining. **Automation in Construction**, v. 74, p. 39, 2017.

TOMIAZZI, J. S.; JUDAI, M. A.; NAI, G. A.; PEREIRA, D. R.; ANTUNES, P. A.; FAVARETO, A. P. A. Evaluation of genotoxic effects in Brazilian agricultural workers exposed to pesticides and cigarette smoke using machine-learning algorithms. **Environmental Science and Pollution Research,** v. 25, no. 2, p. 1259–1269, 2018.

TOMIAZZI, J. S.; PEREIRA, D. P.; JUDAI, M. A.; ANTUNES, P. A.; FAVARETO, A. P. A. Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. **Environmental Science and Pollution Research**, v. 26, no. 7, p. 6481–6491, 2019.

YANAR, B.; LAY, M.; SMITH, P. M. The Interplay Between Supervisor Safety Support and Occupational Health and Safety Vulnerability on Work Injury. **Safety and Health at Work**, v. 10, no. 2, p. 172–179, 2019.

Apêndices

Apêndice A – Divisão e categorização do atributo agente causador.

Código	Divisão agente causador	Base da CAT
0	Veículos, meios de transporte e equipamentos de transporte	Aeronave
		Veiculo Aquatico
		Veiculo de Terraplen
		Veiculo de Tracao An
		Veiculo Deslisante
		Veiculo Funicular (T
		Veiculo Rodoviario M
		Veiculo Sobre Trilho
		Veiculo, Nic
		Trator
		Transportador com Fo
		Transportador por Gr
		Transportador, Nic
		Carro de mao
		Empilhadeira
		Rebocador Mecanico,
		Triciclo
		Motocicleta, Motonet
		Bicicleta
1	Produtos alimentícios e/ou de origem animal	Carne e derivados -
		Cereal e derivados

		Fruta e Derivados
		Produto Alimenticio
		Produtos Alimenticio
		Osso - Produto Anima
		Couro cru ou curtido
		Legume, Verdura e De
		Leite e Derivados -
		Produto Animal, Nic
		Animal vivo
		Pele, crina, Pelo, L
		Pena - Produto Anima
2	Ferramentas manuais sem força motriz	Alavanca, Pe-de-Cabr
		Alicate, Torques, Te
		Chave de parafuso - f
		Chave de porca ou de
		Corda, Cabo, Corrent
		Enxada, Enxadão, Sac
		Faca, Facao - Ferramenta
		Ferramenta Manual se
		Formao, Cinzel-Ferr
		Garfo, Ancinho, Forc
		Lima, Grossa - Ferrame
		Machadinha, Enxo-Fe
		Machado - Ferramenta
		Martelo, Malho, Marr
		Pa, Cavadeira - Ferra
		Picareta - Ferramenta
		Plaina - Ferramenta M
		Pua, Trado, Verruma,
		Puncao, Serrote - Ferr
		Serra, Serrote - Ferr
		Tesoura, Tesourao - F
3	Ferramentas portáteis com força motriz ou aquecimento	Cortadeira, Guilhoti
		Esmeril - Ferramenta
		Ferramenta Acionada
		Ferramenta de Soldag
		Ferramenta Portatil
		Ferro de Passar - Fer
		Jato de Areia- Ferra
		Maquina de Aparafusa
		Martelete, Socador -
		Masarico - Ferrament
		Perfuratriz - Ferrame
		Politriz, Enceradeir
		Rebitadeira - Ferrame

		Serra - Ferramenta Po
		Talhadeira - Ferramen
4	Mobiliário e acessórios	Arquivo, fichario, e
		Balcao, Bancada - MO
		Cadeira Banco - Mobi
		Luminaria, Globo, La
		Mesa Elastica Desmon
		Mesa, Carteira, Exce
		Mobiliario e Acessor
		Tapete, Forracao de
5	Superfícies e equipamentos utilizados para sustentar pessoas	Asfalto, Alcatrao, P
		Calcada ou caminho p
		Escada Movel ou Fixa
		Chao - superficie ut
		Escada Permanente Cu
		Passarela ou Platafo
		Piso de Andaime e PI
		Piso de Edificio - S
		Piso de Mina - Super
		Piso de Veiculo - Su
		Rampa - Superficie U
		Rua e Estrada - Supe
		Superficie de Susten
6	Edifício ou estrutura	Andaime, Plataforma
		Arquibancada, estadi
		Cais, doca - edifici
		Deposito fixo (tanqu
		Dique, barragem - Ed
		Edificio - Edificio
		Edificio ou estrutur
		Escavacao (Para Edif
		Ponte, Viaduto - Edi
		Torre, Poste - Edifi
		Superficie e Estrutu
		Canal, fosso
		Escavacao, Fosso, Tu
		Poco, Entrada, Galer
		Telhado
		Tunel
7	Embalagem ou recipiente (vazio ou cheio)	Barril, Barrica, Bar
		Caixa, Engradado, ca
		Embalagem e recipien
		Frasco, Garrafa - Em
		Tanque, Cilindro (Tr
8		Correia - dispositiv

	Dispositivo de transmissão de energia mecânica	Corrente, Corda, Cab
		Dispositivo de trans
		Embreagem de friccao
		Engrenagem - disposi
		Tambor, Polia, Rolsa
9	Condições ambientais ou do ambiente	Aerodispersoides
		Area ou ambiente de
		Exposicao a Pressao
		Particulas - não Ide
		Pressao Ambiente Alt
		Pressao Ambiente Bai
		Temperatura Muito Alt
		Temperatura Muito Bai
		Temperatura Ambiente
		Ruido
		Ruido, Exposicao A
		Vibracao, Exposicao
		Absorcao de Substanc
		Ingestao de Substanc
		Inalacao, Ingestao o
		Poluicao, Nic, Expos
		Poluicao da Agua, Ac
		Poiluicao do Ar
		Neblina
		Inalacao de Substanc
10	Compostos e/ou subtâncias químicas	Acido
		Alcali
		Alcool
		Cianeto ou composto
		Composto aromatico
		Composto de arsenio
		Composto de fosforo
		Composto metalico (d
		Composto organico Há
		Dissulfeto de carbon
		Gas Cabonico (Dioxi
		Hidrocarboneto Gasos
		Monoxido de Carbono
		Nafta e Solvente de
		Oxidos de Nitrogenio
		Substancia Quimica,
11	Equipamentos de guindar	Elevador - Equip. de
		Elevador de cacamba
		Equip. de Guindar, N
		Guincho Eletrico - E

		Guincho Pneumatico -
		Guindaste - Equip. d
		Macaco (Mecanico, Hi
		Pa Mecanica, Draga -
		Pau de Carga - Equip
		Ponte Rolante - Equi
		Talha - Equip. de Gu
12	Equipamentos elétricos	Condutor - equip. El
		Equip. de Aqueciment
		Equip. Eletrico, Nic
		Equip. Eletrolitico
		Equip. Magnetico - E
		Energia
		Eletrica, Exposicao
		Gerador - Equip. Ele
		Painel de Controle,
		Reostato, Dispositiv
		Transformandor, Conv
13	Equipamentos ou subtâncias emissoras de radiação (ionizante e não ionizante)	Equip. de Iluminacao
		Equip. de Raio X -
		Equip. Emissor de Ra
		Equip. ou Substancia
		Arco eletrico
		Fonte de Radioisotop
		Radiacao nao Ionizan
		Reator (Inclui Combu
14	Petróleo, combustíveis e derivados	Carvao
		Coque
		Gas Encanado de Carv
		Gasolina (exceto Qua
		Gasoleo, Oleo Diesel
		Oleo Combustivel
		Parafina, Oleo Lubri
		Produto de Petroleo
		Petroleo Bruto, Brut
		Querosene
15	Equipamentos para trabalho em ambiente de pressão anormal	Caixao pneumatico
		Equip. de Mergulho
		Equip. para Trabalho
		Escafandro - Equip.
16	Máquinas	Britador, Moinho - M
		Ferramenta, Maquina,
		Furadeira, Broqueade
		Laminadora, Calandra
		Maquina Agricola

		Maquina de Costurar
		Maquina de Embalar o
		Maquina de Escritori
		Maquina de Fundir, d
		Maquina de Imprimir
		Maquina de Mineracao
		Maquina de Terraplen
		Maquina Textil
		Maquina, Nic
		Misturador, Batedeir
		Peneira Mecanica, Ma
		Plaina, Tupia - Maqu
		Politriz, Lixadora,
		Prensa - Maquina
		Serra - Maquina
		Tesoura, Guilhotina,
17	Cerâmica, utensílios e materiais derivados	Ceramica
		Ceramica, Nic
		Louca de Mesa e Outr
		Louca Sanitaria (Pia
		Tubo, Manilha - Cera
		Tijolo e Telha - Cer
		Revestimento Ceramic
18	Fornos e caldeiras	Caldeira
		Caldeira, Vaso sob p
		Forno, Estufa, Retor
19	Vestuário e têxteis	Texteis - Inclui Fib
		Vestuario, Nic
20	Bombas, motores e turbinas	Bomba
		Motor (Combustao Int
		Motor Eletrico - Que
		Motor, Bomba, Turbin
		Turbina
21	Agente infeccioso, produto biológico e medicamentos	Produto Biologico (S
		Medicamento em Geral
		Agente infeccioso ou
22	Metais e minerais	Metal - Inclui Liga
		Produto Mineral Meta
		Produto Mineral não
23	Equipamentos sob pressão	Vaso Sob Pressao (Pa
		Tubo Sob Pressao (Ma
24	Ser vivo	Ataque de Ser Vivo,
		Ataque de Ser Vivo p
		Ataque de Ser Vivo c
		Ser Vivo, Nic

25	Aprisionamento, atrito, abrasão, impacto ou queda	Aprision. Em, Sob ou
		Aprision. Em, Sobre
		Atrito ou Abrasao, N
		Atrito ou Abrasao po
		Impacto de Pes. Cont
		Impacto Sofrido por
		Queda de Pes. em Mes
		Queda de Pes. com Di
26	Madeira	Madeira (Toro, Madei
27	Água e líquidos	Agua - Usar quando o
		Imersao
		Liquido, Nic
28	Vidro	Vidraria, Fibra de V
29	Fogo e materiais inflamáveis	Fogo-Chama, Material
30	Esfoço excessivo e movimentos involuntários	Esforco Excessivo ao
		Esforco Excessivo, N
		Reacao do Corpo a Mo
31	Outros	Agente do Acidente,
		Agente do Acidente I
		Gas e Vapor
		Papel e Pasta para P
		Plastico - Inclui Po
		Produto de Limpeza,
		Sucata, Entulho, Res
		Vegetal - Planta, Ar

Apêndice B – Divisão e categorização do atributo parte do corpo atingida.

Divisão parte do corpo atingida	Descrição na base de dados CAT
Cabeça	Boca (Inclusive Labi
	Cabeca, Nic
	Cabeca, Partes Multi
	Cranio (Inclusive En
	Face, Partes Multipl
	Mandibula (Inclusive
	Nariz (Inclusive Fos
	Olho (Inclusive Nerv
	Ouvido (Externo, Med
	Pescoco
Membros Inferiores	Artelho
	Articulacao do Torno
	Coxa
	Joelho
	Membros Inferiores,
	Pe (Exceto Artelhos)
	Perna (Do Tornozelo,
	Perna (Entre O Torno
Membros Superiores	Antebraco (Entre O P
	Braco (Acima do Coto
	Braco (Entre O Punho
	Cotovelo
	Dedo
	Mao (Exceto Punho ou
	Membros Superiores,
	Ombro
	Punho
Partes Múltiplas	Partes Multiplas - A
Sistemas e aparelhos	Aparelho Circulatori
	Aparelho Digestivo
	Aparelho Genito-Urin
	Aparelho Respiratori
	Sistema Musculo-Esqu
	Sistema Nervoso
	Sistemas e Aparelhos
Tronco	Abdome (Inclusive Or
	Dorso (Inclusive Mus
	Quadris (Inclusive P
	Torax (Inclusive Org
	Tronco, Nic
	Tronco, Parte Multip

4

ARTIGO 2

Este capítulo apresenta o segundo artigo, em que são apresentadas aplicações de técnicas de mineração de dados, utilizando o conjunto de dados com abertura de CAT. Além disso, são comparadas as métricas de cada técnica e aquelas com melhores resultados são submetidas à um algoritmo de inteligência artificial explicável.

ANÁLISE PREDITIVA DE ÓBITOS POR ACIDENTES DE TRABALHO NA INDÚSTRIA DE TRANSFORMAÇÃO BASEADA EM INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Resumo

Um maior compromisso com a saúde e segurança do trabalhador permite não somente mais garantias a sua integridade física e mental, como também favorece o desempenho produtivo nas organizações. Este compromisso também traz como consequências a diminuição de gastos de empresas e governos com a concessão de benefícios previdenciários, quedas de produtividade e aplicações de ações corretivas. No Brasil, grande parte dos índices de acidentes e de doenças ocupacionais estão relacionados ao setor da indústria de transformação, que é o maior responsável por registros de Comunicação de Acidentes de Trabalho (CAT). Diante deste cenário, monitorar, explorar e, principalmente, interpretar os dados de ocorrências ocupacionais registrados, podem ser fatores críticos de sucesso para o processo de tomada de decisão. Uma maneira elegante e eficiente de realizar essa investigação é por meio de algoritmos de mineração

de dados e pela inteligência artificial, que buscam facilitar a compreensão de conhecimentos valiosos “ocultos” nos dados. Com esta finalidade, o presente artigo compara a performance de doze diferentes algoritmos de mineração de dados, por meio de cinco métricas de desempenho, quanto à capacidade preditiva da ocorrência de óbitos por acidentes de trabalho, tendo como base os registros disponíveis no CAT referentes ao setor de indústria de transformação. Os dois algoritmos que apresentaram melhor capacidade preditiva (*Random Forest* e *Naïve Bayes*) foram então submetidos a uma técnica de pós-processamento baseada em *eXplainable Artificial Intelligence* (*XAI*), para interpretação e discussão acerca dos principais fatores que implicaram na determinação dos óbitos por acidente de trabalho. Esta aplicação evidenciou quais foram os atributos e categorias mais influentes na predição dos óbitos, destacando parte do corpo atingida, natureza da lesão e agente causador do acidente.

Palavras-chave:
Saúde e Segurança do Trabalho; Comunicação de Acidentes de Trabalho; Aprendizado de Máquinas; Inteligência Artificial Explicável.

1. Introdução

A preocupação com a saúde e segurança do trabalhador tem sido pauta em discussões com abrangência mundial, devido ao número de doenças, acidentes e óbitos que acometem os trabalhadores todos os anos. A *International Labour Organization* (ILO, 2020) estima que 374 milhões de pessoas sofrem acidentes não fatais em decorrência do seu trabalho por ano, ao redor do mundo. Essas ocorrências resultam em prejuízos ao trabalhador que vão muito além de custos ou dias inativos em seu posto de trabalho, mas impactam a saúde, bem-estar ou mesmo a vida do trabalhador.

No Brasil, grande parte dos registros de acidentes e doenças ocupacionais são associados à indústria de transformação, setor industrial responsável por atividades de transformação de materiais, com o objetivo de gerar um novo produto. Essas indústrias promovem transformações que podem ser químicas, físicas ou biológicas (IBGE, 2021) e acumulam quase 30% das ocorrências com registro no país. A Comunicação de Acidentes de Trabalho (CAT) é uma obrigação legal dos empregadores, que devem registrar a ocorrência de acidentes e doenças ocupacionais de todos os trabalhadores formais associados à sua empresa (BRASIL, 1999).

O registro das ocorrências a partir da CAT resultam em um conjunto de dados disponibilizado pelo Governo Federal, que assim como outros dados de saúde, são de acesso público. O estudo, transformação e aplicação desses dados pode ser utilizado para mitigar o

número de ocorrências, promovendo benefícios ao trabalhador e às organizações públicas e privadas (SANNI-ANIBIRE *et al*., 2020). Em vista do alto volume de dados associados à Saúde e Segurança do Trabalho (SST) gerados diariamente, é necessário buscar métodos que facilitem a compreensão e análise dessas informações. Nesse contexto, a inteligência artificial e mineração de dados podem ser aplicadas, em busca desses mesmos objetivos.

A mineração de dados tem recebido destaque nos últimos anos, utilizada com o intuito de coletar, processar, administrar, analisar e visualizar grandes quantidades de informações, a mineração de dados pode ser útil em diversas áreas e aplicações (LIU *et al.*, 2019; HAJAKBARI; MINAEI-BIDGOLI, 2014). São exemplos de áreas que podem utilizar a mineração de dados: educação (ALDOWAH; AL-SAMARRAIE; FAUZY, 2019), construção civil (ZHAO *et al*., 2020), medicina (ARJI *et al*., 2019) e saúde e segurança do trabalho (CHOI *et al*., 2020).

Aprofundando nos estudos de mineração de dados e SST, algumas pesquisas apresentam foco em casos de óbitos relacionados ao trabalho. Essas pesquisas possuem grande importância, pois estudam os resultados mais graves dos atos inseguros no ambiente de trabalho: os óbitos. Sendo assim, algumas pesquisas já foram desenvolvidas ao redor do mundo, avaliando quais fatores podem influenciar esses acontecimentos (CHOI *et al*., 2020; CIARAPICA; GIACCHETTA, 2009; JOCELYN; OUALI; CHINNIAH, 2018; LIAO; PERNG, 2008; MISTIKOGLU *et al*., 2015; SHIN *et al*., 2018; SHIRALI; NOROOZI; MALEHI, 2018). Mesmo apresentando estudos na área de mineração de dados e SST, inclusive com dados brasileiros, há uma escassez de pesquisa avaliando o óbito do trabalhador no Brasil.

Aliada a mineração de dados, buscando entender e explicar o funcionamento de alguns métodos, principalmente os caracterizados como caixa-preta, tem sido utilizada a Inteligência Artificial Explicável, conhecida também como XAI (*eXplainable Artificial Intelligence*). Os modelos de XAI buscam compreender o funcionamento de outros algoritmos e como eles atingiram seus resultados, de forma que apresentem confiabilidade e segurança para gestores e tomadores de decisão (ARRIETA *et al*., 2020). Dessa forma, o uso de XAI aliada aos dados de SST promove subsídios para tomada de decisões nas organizações, além de descrever um avanço na pesquisa em relação às demais publicações que apenas aplicam mineração de dados.

Neste contexto de dados relacionados à SST e explicabilidade da mineração de dados, é desenvolvida esta pesquisa. Com o objetivo geral de prever a ocorrência de óbitos em função de acidentes e doenças ocupacionais com registro no Brasil, este estudo utiliza doze técnicas de mineração de dados, aplicadas ao conjunto de dados da CAT com recorte na indústria de transformação. O resultado da mineração de dados é comparado por meio de métricas, que

demostram quais foram os algoritmos com melhor desempenho. Estes, por sua vez, são submetidos à um algoritmo de XAI, o *SHapley Additive exPlanations* (SHAP), a fim de interpretar os resultados apresentados pelas técnicas com melhor atuação.

2. Referencial Teórico

2.1 Saúde e Segurança do Trabalho

A Saúde e Segurança do Trabalho (SST) tem se tornado tema fundamental na garantia dos direitos dos trabalhadores, atuando para que exerçam suas funções sem exposição a riscos e perigos, em um local de trabalho seguro (ILO, 2020). A SST pode ainda ser definida com o objetivo de identificar e gerenciar riscos à saúde e segurança do trabalhador, de forma que não prejudique este, além de garantir a manutenção dos sistemas de produção e a reputação das organizações perante o mercado (MUTLU; ALTUNTAS, 2019).

Os prejuízos ao trabalhador podem ocorrer por meio de diversos acontecimentos, tais como: incapacidade temporária ou permanente para o trabalho, doenças ou lesões ocupacionais, acidentes de trabalho e de trajeto, ou ainda, a mais grave das ocorrências, o óbito do trabalhador (ILO, 2020). Todos esses incidentes, acumulam por ano ao redor do mundo mais de 2,78 milhões de mortes, além de um impacto econômico de 3,3 trilhões de dólares em decorrência de acidentes e doenças ocupacionais (WANG *et al.*, 2020; ILO, 2019).

Em busca de mapear e mitigar os resultados negativos aos trabalhadores e às organizações privadas e governamentais, os incidentes ligados ao ambiente de trabalhado devem ser registrados após sua ocorrência (BRASIL, 1999). Esses registros levam em consideração fatores de relevância local, por isso são formalizados de maneiras distintas ao redor do mundo, dificultando uma comparação global (ILO, 2020). No Brasil, a maior parte dos dados gerados em relação à saúde estão ligados ao Sistema Único de Saúde (SUS) (SANNI ALI *et al*., 2019).

São exemplos de conjuntos de dados ligados às informações do SUS os dados de óbitos, disponibilizados por meio do Sistema de Informações de Mortalidade (SIM) ou ainda em relação às doenças, pelo Sistema de Informação de Agravos de Notificação (SINAN) (DOS SANTOS *et al*., 2019; SANNI ALI *et al.*, 2019). Com foco na saúde e segurança do trabalhador, o conjunto de dados com abrangência nacional é o de abertura de Comunicação de Acidentes de Trabalho (CAT), que reúne as ocorrências de acidentes, doenças ou óbitos registrados no Brasil, em decorrência da função exercida por um trabalhador.

Os dados ligados à CAT são agrupados e disponibilizados por trimestre, por meio do Portal Brasileiro de Dados Abertos do Governo Federal. Eles abrangem ocorrências relacionadas a todos os estados brasileiros, e apresentam atributos ligados ao empregado, ao empregador, a ocorrência, aos benefícios e a CAT. O conjunto de dados também permite evidenciar quais fatores são mais recorrentes nos registros, como por exemplo o setor de indústria de transformação (REIS *et al.*, 2020).

A Classificação Nacional de Atividades Econômicas (CNAE) apresenta cinco níveis hierárquicos para sua divisão. No primeiro deles são apresentadas 21 seções, onde a terceira é representada pela letra C, e corresponde à indústria de transformação, que atualmente possui o maior número de registros. No segundo nível hierárquico, dentro do setor de transformação, são apresentadas outras 24 categorias que vão desde produtos alimentícios, a metalurgia e artigos têxteis. Essas subdivisões caracterizam todas as indústrias que realizam uma transformação química, física ou biológica, para geração de um novo produto (IBGE, 2021).

Em vista da representatividade deste setor no contexto industrial brasileiro, alguns estudos já foram desenvolvidos com foco em SST e indústria de transformação, mas ainda há horizontes para pesquisa. Cavalcante *et al.* (2013) estudaram a incidência de ruídos na indústria de transformação e concluíram que alguns ramos são mais expostos que outros, chegando em alguns casos a quase 50% dos trabalhadores expostos a ruídos excessivos. Além disso, observaram que ainda são necessários novos estudos, pois há escassez de informações de SST nesse setor. Outra pesquisa foi desenvolvida por Menegon *et al*. (2021), que apresentam uma análise exploratória de dados da indústria têxtil com relação a acidentes de trabalho, representando um decréscimo nos casos registrados.

2.2 Mineração de Dados

O avanço tecnológico associado às diversas esferas da sociedade promove a geração de dados constantemente e em larga escala. A formação de conjuntos de dados está associada ao termo *big data*, utilizado para expressar uma quantidade numerosa de dados (DOS SANTOS *et al*., 2019). Uma solução para lidar com esses grandes conjuntos são as tecnologias associadas à mineração de dados (ZHAO *et al*., 2020). Os conceitos relacionados a aprendizagem de máquina (AM) e mineração de dados (MD), conhecidos como *machine learning* e *data mining*, respectivamente, têm se popularizado a partir do século 21 e ganhado destaque, especialmente na última década (DOS SANTOS *et al*., 2019; LIU *et al*., 2019).

A mineração de dados é apresentada como uma das etapas presentes dentro do processo KDD (*Knowledge Discovery in Databases*), que busca gerar conhecimento por meio do processamento de dados, sendo a MD responsável pela geração de padrões nos dados a partir da aplicação de algoritmos (FAYYAD *et al*., 1996; LIU *et al*., 2019). Embora o processo KDD seja constituído por nove etapas, onde todas são importantes, a maior parte da literatura sobre o KDD apresenta enfoque na etapa de mineração de dados (THOM DE SOUZA, 2013).

O processo de mineração de dados pode ser caracterizado por possuir uma aprendizagem supervisionada ou não supervisionada (ZHAO *et al*., 2020). O desenvolvimento da aprendizagem supervisionada é definido como um método direcionado, enquanto a aprendizagem não supervisionada não possui esse direcionamento (HAJAKBARI; MINAEI-BIDGOLI, 2014). O primeiro método busca prever ou classificar os dados, buscando um resultado específico, que pode estar relacionado inclusive a dados brutos de áudio ou vídeo (JIANG; GRADUS; ROSELLINI, 2020; SCHMIDHUBER, 2015). O segundo método envolve tarefas de descrição, que objetivam encontrar relações entre os dados, mas sem a presença de uma variável resposta supervisionada (JIANG; GRADUS; ROSELLINI, 2020).

Além dos tipos de aprendizagem, a MD também pode ser relacionada quanto à suas tarefas. Na **classificação** busca-se dividir um determinado conjunto de entradas em categorias discretas ou classes. Essas são previamente definidas, e apresentam comumente as informações de entrada de forma binária, que são apontadas sem sobreposição de classes, ou seja, em apenas uma classe por vez (GONZALEZ; FIACCHINI; IAGNEMMA, 2018; HOOD; CRACKNELL; GAZLEY, 2018).

A tarefa de **associação** também se relaciona aos métodos de aprendizagem supervisionada, e busca realizar associações possíveis entre os dados do conjunto selecionado (VOUGAS *et al*., 2019). As primeiras aplicações de associação estão ligadas à observação do comportamento de compra de clientes em mercados, buscando por itens diferentes que eram encontrados concomitantemente na cesta de compras. Esse comportamento pode ser aplicado em outros cenários, como no processo de diagnóstico de doenças (CHOU *et al*., 2020).

A tarefa de **regressão** utiliza das relações entre variáveis de entrada dependentes e independentes, para prever um valor numérico para variáveis de saída (GONZALEZ; FIACCHINI; IAGNEMMA, 2018; JIANG; GRADUS; ROSELLINI, 2020) e também é considerada uma aprendizagem supervisionada. Por outro lado, no aprendizado não supervisionado a tarefa de **clusterização** ou agrupamento é incorporada, e ocorre de maneira semelhante à classificação, com a formação de grupos. No entanto, esses *clusters* não são previamente escolhidos pelo tomador de decisão, mas sim pelo algoritmo, que maximiza as

diferenças entre os indivíduos e assim os separa em grupos (HOOD; CRACKNELL; GAZLEY, 2018).

2.3 Técnicas de Mineração de Dados

A mineração de dados pode também ser expressa por suas técnicas, que são os algoritmos utilizados nas tarefas de mineração. Essas técnicas podem ser subdivididas em *ensemble* e não *ensemble*. Algoritmos não *ensemble* são considerados classificadores básicos, cujos modelos desenvolvem uma técnica de classificação simples para os dados. Por outro lado, modelos *ensemble*, geralmente realizam uma classificação mais robusta e precisa, utilizando um conjunto de classificadores para fazer a previsão do modelo (MARQUÉS; GARCÍA; SÁNCHEZ, 2012; DIETTERICH, 2000). *Bayesian averaging* (ou média bayesiana) é considerado um dos primeiros algoritmos dos métodos *ensemble*, mas em seguida outros passaram a ser utilizados com o mesmo fim (DIETTERICH, 2000).

São exemplos de algoritmos *ensemble: Bagging* (BA), *Voting* (VO), *Stacking* (ST), *Extra Trees* (ET), *Random Forest* (RF) e *XGBoost* (XGB). Algoritmos de *bagging* utilizam as diferentes saídas de um modelo, atribuindo pesos iguais a elas, para gerar uma única saída de decisão. Em contrapartida, os algoritmos de *boosting*, também adotam essa abordagem, mas podem atribuir pesos diferentes para seus fatores (WITTEN; FRANK, 2016). Um exemplo de algoritmo de *boosting* é o *eXtreeme Gradient Boosting* (*XGBoost)*, que tende a superar outros algoritmos devido a sua facilidade de uso e precisão dos resultados, além de ser escalável em diversos cenários e apresentar boa velocidade de execução (CHEN; GUESTRIN, 2016).

Os algoritmos de *votting* funcionam de maneira semelhante ao *bagging*, porém seus classificadores devem apresentar desempenhos que sejam comparáveis. Por outro lado, as técnicas de *stacking* frequentemente combinam modelos com funcionamento diferente para fazer a previsão (WITTEN; FRANK, 2016). A técnica *Extra Trees* utiliza como método um conjunto de árvores, que apresentam nós com pontos de corte aleatórios (GEURTS; ERNST; WEHENKEL, 2006) e *Random Forest*, utilizando métodos de aleatoriedade para construir um conjunto com classificadores individuais (DÉSIR *et al*., 2012).

Considerando as técnicas não *ensemble*, podem ser elencadas: *Support Vector Machine* (SVM), *Logistic Regression* (LR), *K-Nearest Neighbors* (KNN), *Naïve Bayes* (NB), *Decision Trees* (DT) e *Neural Networks* (NN). A técnica SVM se assemelha à algoritmos lineares devido a sua baixa complexidade, mas apresentando bons resultados para regressão, extração ou geração de conhecimento (HEARST *et al.*, 1998). *Logistic Regression* (LR), ou regressão

logística, utiliza de funções lineares para a construção de um modelo baseado em uma variável alvo, que também é conhecida como variável resposta ou independente (WITTEN; FRANK, 2016).

Para o algoritmo KNN, a classificação ocorre por meio da aproximação da amostra à uma das *k* classes existentes no modelo, priorizando aquela que se assemelhar mais às suas características (KELLER; GRAY; GIVENS, 1985). *Decision Trees*, ou árvores de decisão, são algoritmos que constroem seu modelo de classificação por meio de uma estrutura de árvores, onde os nós representam testes que o modelo realiza entre seus atributos e constantes, e nós folhas que destacam as classes ou respostas possíveis (WITTEN; FRANK, 2016).

Outra técnica não *ensamble* é a *Naïve Bayes*, que recebe a característica de algoritmo de ingênuo (*Naïve*) por considerar que as variáveis do modelo não possuem relação, ou seja, são independentes (WITTEN; FRANK, 2016). Por fim, a última técnica não *ensemble*, que também é amplamente conhecida na aprendizagem de máquinas, são as *Neural Networks* (redes neurais), que representam um modelo baseado no funcionamento dos neurônios ao qual aprende e realiza classificações com bom desempenho (ALBER *et al*., 2019).

Em relação às ferramentas utilizadas para aplicação dos métodos de mineração de dados, estas também podem variar, e a natureza dos dados, o problema e os conhecimento de um especialista impactam na escolha da ferramenta (CRACKNELL; READING, 2014). Algumas pesquisas desenvolvem algoritmos em linguagem Python (GOH; UBEYNARAYANA, 2017; MARUCCI-WELLMAN; CORNS; LEHTO, 2017) e outras utilizam *softwares* estatísticos como SAS (NENONEN, 2013; SHIN *et al.*, 2018), SPSS (SHIRALI; NOROOZI; MALEHI, 2018), R (DOS SANTOS *et al*., 2019; HEO *et al*., 2019; LEE; KIM, 2018) e Weka (DOS SANTOS *et al*., 2019; PEKEL *et al.*, 2018; SANMIQUEL *et al*., 2018).

2.4 Mineração de Dados aplicada à SST

Considerando estudos que realizam aplicações de algoritmos de mineração à dados de saúde e segurança do trabalho, são encontrados diferentes conjuntos de dados, técnicas e países de atuação. Há autores que consideram dados de empresas específicas, como Bevilacqua, Ciarapica e Giancchetta (2008), que utilizaram dados coletados de uma refinaria. Outros empregam em seus estudos conjuntos de dados de acidentes totais no país, mas enfatizam um setor industrial em específico, como indústria mineral extrativista (CHENG; LIN; LEU, 2010) ou setor da construção civil (CHENG *et al*., 2012; CHOI *et al*., 2020; LIAO; PERNG, 2008).

Também há autores que utilizam conjuntos de dados públicos, sem especificar um setor industrial, mas fazendo a análise de todas as ocorrências do país, como Nenonen (2013), que relacionou os acidentes de escorregões, tropeços e quedas com suas causas. Cheng, Yao e Wu (2013) e Liao e Perng (2008) manipularam dados de ocorrências de maior gravidade. Os primeiros utilizaram métodos estatísticos e mineração de dados para avaliar a relação entre os fatores na ocorrência de grandes acidentes na indústria petroquímica. Grandes acidentes são aqueles que ferem no mínimo três pessoas ou causam uma ou mais mortes (CHENG; YAO; WU, 2013).

Na pesquisa de Liao e Perng (2008) o foco do estudo é apenas em casos de óbitos em consequência do trabalho, investigando os fatores a fim de ponderar quais são os mais significativos nesse cenário. Outros estudos aplicam técnicas de mineração de dados em conjuntos que registram tanto acidentes quanto óbitos em virtude do trabalho (CHENG; LIN; LEU, 2010; CHOI *et al.*, 2020; CIARAPICA; GIACCHETTA, 2009; JOCELYN; OUALI; CHINNIAH, 2018; MISTIKOGLU *et al*., 2015; SHIN *et al.*, 2018; SHIRALI; NOROOZI; MALEHI, 2018).

A quantidade de dados utilizados na mineração também é variável entre os trabalhos encontrados. Jocelyn, Ouali e Chinniah (2018) utilizam um conjunto de 23 relatórios de acidentes no setor de transportes, buscando entender o comportamento de risco em situações de acidentes. Outros estudos utilizam bancos de dados entre 200 e 350 registros, (BEVILACQUA; CIARAPICA; GIACCHETTA, 2008; CHENG; YAO; WU, 2013; LIAO; PERNG, 2008), mas a maioria das pesquisas analisadas utilizam conjuntos com milhares de dados. Ciarapica e Giancchetta (2009) aplicam modelos para classificação de mais de 190.000 registros, Shin *et al*. (2018) utilizam 98.189 registros de acidentes e óbitos para entender as relações entre as variáveis que impactam na ocorrência dos incidentes.

Alguns estudos descrevem o setor industrial de atuação (CIARAPICA; GIACCHETTA, 2009; HAJAKBARI; MINAEI-BIDGOLI, 2014), outros representam o tamanho da organização por seu número de funcionários (CHENG *et al*., 2012; CHENG; LIN; LEU, 2010; CHOI et al., 2020; SANMIQUEL *et al*., 2018; SANMIQUEL; ROSSELL; VINTRÓ, 2015), outros ainda tem o foco nos projetos que são executados (CHENG *et al*., 2012; CHENG; LIN; LEU, 2010; CHOI *et al*., 2020; LIAO; PERNG, 2008; MISTIKOGLU *et al*., 2015) ou em medidas preventivas adotadas pela organização (MISTIKOGLU *et al*., 2015; SANMIQUEL *et al.*, 2018; SANMIQUEL; ROSSELL; VINTRÓ, 2015).

As variáveis associadas as ocorrências de acidentes e lesões estão descritas em maior número e variedade. Elas podem destacar informações temporais, como data (BEVILACQUA;

CIARAPICA; GIACCHETTA, 2008; CHENG; YAO; WU, 2013; CHOI *et al*., 2020; LIAO; PERNG, 2008; SANMIQUEL; ROSSELL; VINTRÓ, 2015) e horário de ocorrência do acidente (LIAO; PERNG, 2008; SANMIQUEL *et al*., 2018; SANMIQUEL; ROSSELL; VINTRÓ, 2015), localização (CHENG *et al*., 2012; CHENG; LIN; LEU, 2010; HAJAKBARI; MINAEI-BIDGOLI, 2014; MISTIKOGLU *et al*., 2015; SANMIQUEL *et al*., 2018; SANMIQUEL; ROSSELL; VINTRÓ, 2015) ou agente causador do acidentes, variável essa ligada a quase todos os trabalhos citados.

Os resultados encontrados em cada pesquisa estão diretamente associados aos algoritmos escolhidos para a mineração de dados. Eles podem apresentar associações entre as variáveis (CHENG; LIN; LEU, 2010) e descrever quais fatores são mais relevantes para a ocorrência de um acidente (CHENG *et al*., 2012), tais como dias e horários (CHENG; YAO; WU, 2013). Também podem apresentar resultados que atuem de forma preventiva na ocorrência de acidentes, como descrição de ações que mitiguem as ocorrências (BEVILACQUA; CIARAPICA; GIACCHETTA, 2008), conclusões estratégicas para o desenvolvimento de novas políticas de SST (LIAO; PERNG, 2008) e análises quanto a força, condições de trabalho e o ambiente laboral ao qual o trabalhador pertence (HAJAKBARI; MINAEI-BIDGOLI, 2014)

2.5 Inteligência Artificial Explicável

Em contraste às preocupações do mercado com quedas de produção, vendas e lucratividade, a inteligência artificial (AI) tem crescido substancialmente nos últimos anos. Além disso, segundo pesquisas da Accenture (PURDY; DAUGHERTY, 2017), a AI pode ser utilizada como suporte para reverter a queda nos lucros das organizações, com potencial de promover um aumento de 14 trilhões de dólares até 2035, considerando apenas 12 países como foco. Estes ganhos representam 38% a mais de rentabilidade para as empresas (PURDY; DAUGHERTY, 2017).

O conceito de inteligência artificial ou a*rtificial intelligence* está diretamente ligado ao aprendizado computacional (MITCHELL; MICHALSKI; CARBONELL, 2013) e pode ser definido como um conjunto de tecnologias, que combinadas, resultam em formas diferentes para compreender, agir, sentir e aprender (PURDY; DAUGHERTY, 2017). As aplicações de IA são variadas, desde controle do processo produtivo das organizações (ALEXOPOULOS; NIKOLAKIS; CHRYSSOLOURIS, 2020) ao apoio de diagnósticos e tratamento de doenças, como a COVID-19 (BULLOCK *et al*., 2020).

Acompanhando as evoluções tecnológicas, nos últimos anos tem evoluído também o conceito de *eXplainable Artificial Intelligence* (XAI) ou IA explicável, com o objetivo de apresentar modelos de AM que mantenham seu alto desempenho e garantam o entendimento e confiabilidade para os tomadores de decisão (ARRIETA *et al*., 2020). Este alto nível de confiabilidade se faz necessário principalmente em aplicações que se relacionam à vida dos seres humanos, como diagnóstico médico e decisões legais (GOODMAN; FLAXMAN, 2017).

Diante disso, entender os motivos pelos quais um modelo apresentou uma previsão como resultado, é tão, ou mais importante do que apenas a previsão. Para evidenciar a explicabilidade do modelo alguns algoritmos podem ser utilizados, como por exemplo o *SHapley Additive exPlanations* (SHAP) (LUNDBERG; LEE, 2017). O SHAP, utilizando os princípios da teoria dos jogos, atribui um valor específico para cada característica dentro de uma previsão (LUNDBERG; LEE, 2017) e entre suas possibilidades, pode ser aplicado para interpretação de previsões sobre a qualidade do ar (GARCÍA; ASNARTE, 2020) ou para detecção de anomalias em sistemas de aquecimento urbano (PARK; MOON; HWANG, 2020).

Outro algoritmo amplamente utilizado para explicação de previsões é o *Local Interpretable Model-agnostic Explanations* (LIME), que também auxilia o entendimento dos resultados de modelos de AM considerados como caixa preta (RIBEIRO; SINGH; GUESTRIN, 2016). Sendo aplicável em qualquer classificador, o LIME pode atuar na interpretação de modelos complexos, como resultados clínicos (ZHANG *et al*., 2018) ou ainda na classificação de candidatos no processo de recrutamento de uma empresa (BRAMHALL *et al*., 2020).

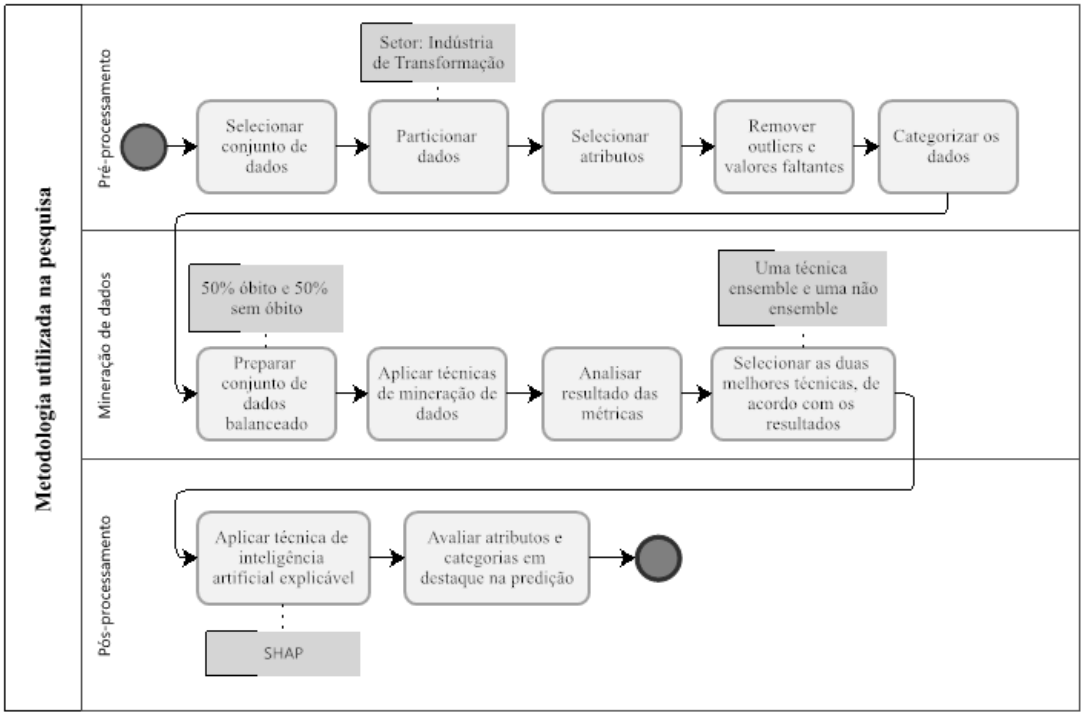
3. Metodologia

A metodologia utilizada se aproxima do processo KKD, como apresentada por Fayyad *et al.* (1996), mas com foco nas etapas de mineração e interpretação dos dados. A etapa inicial do estudo é caracterizada pela seleção dos dados, sendo o escolhido o conjunto de dados com abertura de CAT por sua abrangência nacional e apresentação de dados de óbito, doenças e acidentes de trabalho na mesma base. Em seguida os dados passaram por uma limpeza, onde foi realizado um recorte, optando pelo uso dos dados referente ao setor de indústria de transformação, que acumula o maior número de registros.

Nessa etapa também foram removidos valores faltantes e *outliers* do conjunto de dados, e a partir disso os dados foram adaptados, transformando-os em categóricos em busca de facilitar a aplicação dos algoritmos. A etapa de pré-processamento dos dados, que compõe toda

seleção, limpeza e preparação do conjunto de dados, assim como as demais etapas, são evidenciadas pela Figura 1.

Figura 1 – Metodologia da pesquisa



Fonte: A autora (2021)

Encerrando o pré-processamento dos dados, a etapa posterior foi a realização da mineração, onde, inicialmente foi preparado um conjunto balanceado, composto por 50% de ocorrências com óbitos e 50% sem óbito, totalizando 1500 registros. Em seguida, esse conjunto balanceado foi submetido à 12 técnicas de mineração de dados, buscando prever a ocorrência de óbitos a partir de atributos registrados com a abertura da CAT.

Para aplicação da mineração de dados foi utilizado o ambiente do *Jupyter Notebook*, na distribuição *Anaconda*, desenvolvendo os modelos a partir de funções do *scikit-learn*[1] na linguagem *Python*. Como sistema computacional foi utilizado um Intel Core i3-5005U, com memória RAM de 4,00 GB e uma unidade de armazenamento (SSD) de 120 GB.

Na etapa de execução os algoritmos utilizados foram seis caracterizados como *ensemble* (*Bagging, Extra Trees, Random Forest, Stacking, Voting* e *XGBoost*) e outros seis não *ensemble* (*Decision Trees, K-Nearest Neighbors*, *Logistic Regression*, *Naïve Bayes*, *Neural Networks* e *Support Vector Machine*,). Os resultados das técnicas foram avaliados a partir de cinco métricas, sendo elas: acurácia, precisão, *recall*, F1 score e curva ROC/AUC (*Receiver Operating Characteristic/Area Under Curve*). As métricas utilizadas e suas respectivas fórmulas são apresentadas no Quadro 1.

Quadro 1 – Métricas utilizadas e suas respectivas fórmulas

Métrica	**Fórmula**
Acurácia	$A = \frac{TP + TN}{TP + FP + TN + FN}$
Precisão	$P = \frac{TP}{TP + FP}$
Recall	$R = \frac{TP}{TP + FN}$
F1 score	$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$
ROC/AUC	$AUC = \sum_i \{(1 - \beta_i \times \Delta\alpha) + \frac{1}{2}[\Delta(1 - \beta) \times \Delta\alpha]\}$

Fonte: Adaptado de Callahan e Shah (2017); Davis e Goadrich (2006); Bradley (1997)

O objetivo da avaliação por meio de métricas é quantificar a assertividade da classificação dos algoritmos utilizados. Os índices *True Positive* (TP - Verdadeiro positivo) e *False Positive* (FP - Falso positivo) apresentam uma relação com a classificação positiva de acordo com os dados originais, no primeiro caso uma classificação correta e no segundo, incorreta. Já o *True Negative* (TN – Verdadeiro negativo) e *False Negative* (FN – Falso negativo) se referem à uma classificação negativa da classe de interesse, onde o primeiro é identificado de maneira correta e o segundo de maneira errônea (CALLAHAN; SHAH, 2017; DAVIS; GOADRICH, 2006; BRADLEY, 1997).

Em busca de entender os resultados das técnicas aplicadas, a última etapa da pesquisa foi o pós-processamento, com a execução de um algoritmo de explicabilidade, para entender as previsões geradas na etapa anterior. Para isso foram selecionadas duas técnicas, aquelas que obtiveram melhores resultados no grupo *ensemble* e não *ensemble*, e para elas foi aplicado o *SHAP*, um algoritmo comumente utilizado no contexto de modelos explicáveis. A partir dos resultados encontrados, algumas conclusões e ponderações foram realizadas ao final do artigo.

4. Resultados e discussões

4.1 Seleção, pré-processamento e transformação dos dados

O conjunto de dados selecionado para esta pesquisa foi extraído dos dados abertos disponibilizados pela Empresa de Tecnologia e Informações da Previdência (DATAPREV), ligada ao Ministério da Economia do Brasil. Os dados abertos apresentam alguns conjuntos

disponibilizados publicamente, dentre eles os dados de acidentes com abertura de Comunicação de Acidentes do Trabalho (CAT). Estes conjuntos em questão foram selecionados para a presente pesquisa.

Os dados com CAT são agrupados trimestralmente e disponibilizados no formato *.CSV* (*Comma-Separated Values*). Até o momento desta pesquisa, os dados apresentados equivalem ao horizonte temporal de julho de 2018 a setembro de 2020, e para tanto, foi selecionada a totalidade destes dados para aplicação. Compilando todos os trimestres, a fim de formar um único conjunto de dados, o resultado foi de 990.870 instâncias relacionadas a 25 atributos.

Em busca de refinar a pesquisa e possibilitar observações pontuais, a partir de uma análise exploratória o conjunto de dados foi particionado, tanto em função de suas instâncias como de seus atributos. O estudo exploratório realizado proporcionou o destaque de valores faltantes, *outliers*, atributos não confiáveis, atributos com informações repetidas e outros que poderiam ser agrupados. Além disso, a fase exploratória do estudo também proporcionou a avaliação de quais categorias dos atributos eram mais preponderantes no conjunto de dados total, expondo quais seriam foco na pesquisa.

Como resultado da análise exploratória foi selecionado um recorte nos dados, correspondente ao setor de atuação da indústria de transformação. Essa escolha se justifica, pois, dentre os 21 grupos da Classificação Nacional de Atividades Econômicas (CNAE) presentes no conjunto de dados, a indústria de transformação compreende o maior número de casos, ocupando 27% do total de dados. Este recorte corresponde a 267.616 instâncias no geral, mas com a remoção de *outliers* e valores faltantes resulta em 263.629.

Outro particionamento dos dados foi em função de seus atributos, onde foram removidos aqueles com informações duplicadas, grande parte dos valores faltantes ou tendenciosos para apenas uma categoria. Um resumo dos dez atributos selecionados para a pesquisa é apresentado no Quadro 2, assim como sua descrição, valores e tipo.

Quadro 2 – Caracterização dos atributos do conjunto de dados

Atributo	**Descrição**	**Valores**	**Tipo**
Idade	Idade inteira do trabalhador até a data do acidente	{16 a 75}	Numérico
Sexo	Gênero do trabalhador	{Feminino, Masculino}	*String*
CBO	Código Brasileiro da Ocupação associada ao trabalhador	{3 a 999.999}	Categórico
CNAE subgrupo	Classificação Nacional de Atividades Econômicas do empregador associado ao acidente	{1.000 a 3.399}	Categórico
UF empregador	Unidade Federativa de inscrição do empregador	{26 estados e Distrito Federal}	*String*
Tipo de acidente	Classificação do acidente	{Típico, Trajeto ou Doença}	*String*
Agente causador	Agente (podendo ser coisa, substância ou ambiente) que provocou o acidente	{273 categorias}	*String*
Parte de corpo atingida	Região no corpo do trabalhador atingida pelo acidente	{41 categorias}	*String*
Natureza da lesão	Identificação da lesão segundo suas características principais	{28 categorias}	*String*
Óbito	Indica morte do trabalhador relacionado ao acidente ou doença	{Sim, Não}	*String*

Fonte: A autora (2021)

A idade do trabalhador foi o único atributo que precisou ser calculado, pois não era apresentado dessa maneira no conjunto de dados inicial. Seu cálculo foi realizado a partir da subtração da data do acidente e da data de nascimento, ambas informações apresentadas no conjunto disponibilizado. Todos os demais atributos já eram apresentados no conjunto de dados conforme a descrição no Quadro1. No entanto, como pode ser observado no quadro, alguns deles apresentavam muitas possibilidades de categorização, e fazer uso dos dados dessa forma aumentaria a complexidade e tempo dos métodos utilizados.

Buscando agilidade na aplicação das técnicas, todos os atributos no formato de *string* foram adaptados para categóricos. O Quadro 3 apresenta os atributos selecionados, as adaptações realizadas na categorização, os novos valores e tipos dos dados.

Quadro 3 – Categorização dos atributos do conjunto de dados

Atributo	Adaptação realizada	Valores	Tipo
Idade	Inclusão de faixa etárias com intervalo de 10 anos	{0 a 5}	Categórico
Sexo	Transformação de *string* para numérico	{0 e 1}	Categórico
CBO	Categorização de acordo com o primeiro nível de grandes grupos da CBO	{0 a 9}	Categórico
CNAE	Categorização de acordo com o segundo nível de grupos da CNAE	{0 a 23}	Categórico
UF empregador	Representação numérica para cada unidade federativa	{0 a 26}	Categórico
Tipo de acidente	Representação numérica para cada tipo	{0 a 2}	Categórico
Agente causador	Agrupamento de agentes causadores e categorização	{0 a 31}	Categórico
Parte de corpo atingida	Agrupamento de partes de corpo atingidas e categorização	{0 a 5}	Categórico
Natureza da lesão	Agrupamento de naturezas da lesão e categorização	{0 a 22}	Categórico
Óbito	Transformação de *string* para numérico	{0 e 1}	Categórico

Fonte: A autora (2021)

As faixas etárias foram divididas em intervalos de dez anos, com a primeira iniciando em 16 e indo até 25 anos, e a última de 66 a 75 anos. O sexo do trabalhador foi representado por uma escala binária, onde zero (0) representa trabalhadoras do sexo feminino e um (1) do sexo masculino. O Código Brasileiro de Ocupação (CBO) possui 2.511 ocupações considerando a última divisão da hierarquia, no entanto, para esta pesquisa foi utilizado apenas seu primeiro nível, com dez grandes grupos, dispostos conforme a ordenação do Ministério do Trabalho e Emprego (MTE).

Em relação à CNAE do empregador, devido ao recorte nos dados para a indústria de transformação, o primeiro nível da hierarquia já foi fixado. Para este estudo foram consideradas as categorias do segundo nível, apenas correspondentes ao setor industrial selecionado, resultando em 24 possibilidades de categorização deste atributo, descritas conforme a ordenação proposta pela Comissão Nacional de Classificação (CONCLA). Para a unidade federativa, foram considerados os 26 estados brasileiros e o Distrito Federal, categorizados de acordo com a ordem alfabética deles, em ordem crescente. Considerando os tipos de acidentes, os típicos receberam a categoria zero, acidentes de trajeto elencados ao número um e doenças com o número dois.

Quanto ao atributo de parte do corpo atingida, as 41 categorias foram agrupadas em seis, sendo elas: cabeça, membros inferiores, membros superiores, partes múltiplas, sistemas e

aparelhos e tronco, com essa respectiva ordenação. Os atributos de agente causador e natureza da lesão também passaram por junção de categorias, o primeiro passando de 273 para 32 possibilidades e o segundo passando de 28 para 23. A categorização destes atributos é apresentada nos Apêndices A e B deste trabalho, respectivamente. Por fim, o último atributo adaptado foi o óbito, também transformado em variável binária, onde zero (0) indica acidentes sem óbito e um (1) indica o óbito do trabalhador.

4.2 Descrição e avaliação do desempenho das técnicas

O conjunto de dados da CAT com recorte da indústria de transformação foi submetido às técnicas de mineração de dados, com o objetivo de prever a ocorrência de óbitos baseado nos acidentes e doenças já registrados. Para isso foram considerados os dez atributos apresentados na seção anterior, onde a variável dependente (ou variável resposta) do modelo é o óbito do trabalhador, e as outras nove variáveis elencadas são consideradas independentes.

Como técnicas para aplicação do modelo foram selecionadas doze, sendo seis delas caracterizadas como *ensemble* e seis como não *ensemble*. Esta divisão na seleção das técnicas foi realizada a fim de possibilitar uma comparação do desempenho de cada uma delas e sua complexidade. As técnicas *ensemble* selecionadas foram: *Extra Trees* (ET), *Random Forest* (RF), *XGBoost* (XGB), *Bagging* (BA), *Voting* (VO) e *Stacking* (ST). Já as técnicas não *ensemble* foram: *Support Vector Machine* (SVM), *Logistic Regression* (LR), *K-Nearest Neighbors* (KNN), *Naïve Bayes* (NB), *Decision Trees* (DT) e *Neural Networks* (NN).

Os modelos construídos em cada uma das técnicas utilizaram funções de classificação provenientes do *scikit-learn*. As funções utilizadas e suas respectivas técnicas estão apresentadas no Quadro 4.

Quadro 4 – Técnicas utilizadas e suas respectivas funções do s*cikit-learn*

	Técnica	**Função do s*cikit-learn***
ENSEMBLE	*Extra Trees* (ET)	ExtraTreesClassifier
	Random Forest (RF)	RandomForestClassifier
	XGBoost (XGB)	HistGradientBoostingClassifier
	Bagging (BA)	BaggingClassifier
	Voting (VO)	VotingClassifier
	Stacking (ST)	StackingClassifier
NÃO ENSEMBLE	*Support Vector Machine* (SVM)	SVC
	Logistic Regression (LR)	LogisticRegression
	K-Nearest Neighbors (KNN)	KNeighborsClassifier
	Naïve Bayes (NB)	CategoricalNB
	Decision Trees (DT)	tree.DecisionTreeClassifier
	Neural Networks (NN)	MLPClassifier

Fonte: A autora (2021)

Cada técnica foi aplicada utilizando o mesmo conjunto de dados, construído de maneira balanceada, de forma que 50% dos dados estavam associados ao óbito do trabalhador e 50% não associados. Como, para o setor de indústria de transformação, havia 750 instâncias com óbito, foram elencadas mais 750 instâncias sem associação ao óbito. A escolha dos dados sem óbito foi aleatória, utilizando o Microsoft Excel® para geração de números aleatórios entre zero e um para cada linha da tabela. Com os números aleatórios gerados, foram selecionados os 750 maiores, e essa foi a seleção para compor o conjunto de dados balanceado, com 1.500 ocorrências.

Com o objetivo de possibilitar a comparação entre as técnicas utilizadas, cinco métricas foram escolhidas, sendo elas: acurácia, precisão, recall, F1 score e ROC/AUC. Estas métricas também foram aplicadas por meio de funções do *scikit-learn*, respectivamente descritas por: *accuracy_score*, *precision_score*, *recall_score*, *f1_score*, *roc_auc_score*. A partir da definição do conjunto de dados, atributos, técnicas e métricas que seriam utilizadas, a etapa de execução pôde ser iniciada, sendo esta realizada por meio de três experimentos, que serão detalhados na sequência. Esses experimentos se diferenciam quanto à divisão dos conjuntos de treino e teste utilizados nos métodos e a declaração de parâmetros do modelo.

4.2.1 Experimento 1

O primeiro experimento foi realizado utilizando o conjunto de dados com um particionamento de 70% treino e 30% teste. Esta divisão foi realizada por meio da função *train_test_split* do *scikit-learn*, que divide aleatoriamente o conjunto de dados em duas partes, segundo especificações do modelo. Neste modelo, foi definido que uma das partes seria composta por 70% dos dados de treinamento do modelo e a outra com os 30% restantes, compondo o conjunto de teste. Assim, para este experimento, o conjunto de treino se caracterizava por 1.050 instâncias, enquanto o conjunto de teste apresentava apenas 450 instâncias.

A partir da caracterização do conjunto de dados de treino e teste, que foi submetido às 12 técnicas escolhidas, utilizando para todas o padrão de seus parâmetros, foi possível avaliar os resultados das métricas e tempos de execução para cada uma delas. Esta comparação é apresentada na Tabela 1.

Tabela 1 – Resultados das métricas e tempo computacional das técnicas (experimento 1)

Técnica	**Acurácia**	**Precisão**	***Recall***	**F1 score**	**ROC/AUC**	**Tempo (s)**
			Ensemble			
ET	0,9022	0,9031	0,9031	0,9031	**0,9599**	**3,3981**
RF	0,8978	0,8841	0,9156	0,8996	0,9482	3,8558
XGB	0,8800	0,8618	0,8863	0,8738	0,9439	4,8366
BA	0,8889	**0,9050**	0,8734	0,8889	0,9494	4,4417
VO	0,8133	0,8319	0,8034	0,8174	0,9101	3,5404
ST	**0,9044**	0,8945	**0,9217**	**0,9079**	0,9529	13,4521
			Não *ensemble*			
SVM	0,7267	0,7333	0,6379	0,6822	0,8106	3,7272
LR	0,7067	0,7333	0,6962	0,7143	0,7654	5,5397
KNN	0,7467	0,7727	0,6892	0,7286	0,8067	3,5548
NB	**0,8822**	**0,8943**	0,8750	**0,8845**	**0,9448**	**3,1591**
DT	0,8667	0,8703	**0,8776**	0,8740	0,8661	3,9163
NN	0,8133	0,7759	0,8491	0,8108	0,8996	4,1987

Fonte: A autora (2021)

De maneira geral é possível destacar que as técnicas categorizadas como *ensemble* obtiveram resultados melhores que aquelas elencadas como não *ensemble*. Esta característica é esperada, devido à complexidade dos algoritmos *ensemble* ser maior. No entanto, tiveram exceções nas duas subdivisões, pois a técnica de *Voting* (VO) atingiu valores nas métricas

próximos aos algoritmos não *ensemble*, assim como *Naïve Bayes* (NB) obteve resultados semelhantes às técnicas caracterizadas como *ensemble*, mesmo não estando neste grupo.

Além disso, na Tabela 1 estão em destaque quais foram os melhores valores para cada métrica, assim como qual foi o menor tempo dentre as técnicas utilizadas. Para o grupo de algoritmos *ensembles*, o menor tempo foi para o *Extra Tree* (ET), que também obteve o melhor valor de ROC/AUC. Em contrapartida, o melhor resultado de acurácia, *recall* e F1 score esteve associado à *Stacking* (ST), enquanto a melhor precisão foi do algoritmo de *Bagging* (BA). Já em relação às técnicas não *ensemble*, *Naïve Bayes* (NB) acumulou os melhores resultados de métricas, assim como o menor tempo computacional, com exceção do *recall*, que o melhor valor foi em função de *Decision Trees* (DT).

4.2.2 Experimento 2

Para o segundo experimento a mudança foi em relação ao seu conjunto de treino e teste, ainda utilizando o padrão para definição dos parâmetros. Enquanto o experimento anterior fez uso do método de *train_test_split*, aqui a função utilizada foi a *cross_validate*. Esta função tem o objetivo de realizar a validação cruzada dos dados, a partir da divisão do conjunto em n partes (*folds*), onde uma das partes é separada para testes e as outras $(n-1)$ são subconjuntos de treino. Para este experimento foi utilizado o valor de $n = 10$, ou seja, dez subconjuntos de dados separados aleatoriamente, onde nove são para treino e apenas um deles para teste.

Após a nova caracterização do conjunto de dados de treino e teste, as 12 técnicas de mineração de dados escolhidas foram aplicadas. As métricas relacionadas as técnicas, assim como os tempos computacionais para sua aplicação, são apresentadas na Tabela 2.

Tabela 2 - Resultados das métricas e tempo computacional das técnicas (experimento 2)

Técnica	Acurácia	Precisão	*Recall*	F1 score	ROC/AUC	Tempo (s)
			Ensemble			
ET	0,8940	0,8924	0,8973	0,8945	0,9574	7,5466
RF	**0,9067**	0,8983	**0,9187**	**0,9080**	0,9607	6,4333
XGB	0,8973	0,8933	0,9040	0,8981	0,9503	17,9580
BA	0,8880	0,9008	0,8733	0,8863	0,9456	**3,5365**
VO	0,8233	0,8212	0,8280	0,8237	0,9145	7,1474
ST	0,9060	**0,9043**	0,9093	0,9065	**0,9608**	23,9468
			Não *ensemble*			
SVM	0,7433	0,7816	0,6747	0,7229	0,8239	7,4659
LR	0,7100	0,7118	0,7067	0,7082	0,7662	3,8888
KNN	0,7593	0,7886	0,7093	0,7457	0,8124	3,8752
NB	**0,8860**	**0,8907**	**0,8800**	**0,8850**	**0,9541**	3,1593
DT	0,8613	0,8609	0,8627	0,8615	0,8613	**3,1094**
NN	0,8267	0,8313	0,8187	0,8234	0,9054	21,4574

Fonte: A autora (2021)

Similar ao experimento 1, os resultados das técnicas *ensemble* superaram as não *ensemble*, como era esperado. Neste caso, *Random Forest* (RF) obteve melhor resultado em três métricas: acurácia, *recall* e F1 score, enquanto *Stacking* (ST) superou as demais em precisão e ROC/AUC. No entanto, esta apresentou o maior tempo computacional, chegando a quase 24 segundos para execução e o melhor tempo computacional foi apresentado pela técnica de *Bagging* (BA). Para os algoritmos não *ensemble* o melhor resultado de todas as métricas foi vinculado a *Naïve Bayes* (NB), que também apresentou um tempo de execução baixo, ficando atrás apenas de *Decision Tree* (DT).

4.2.3 Experimento 3

O terceiro experimento também foi realizado utilizando a validação cruzada para os dados, por meio de dez subconjuntos, onde nove são para treino e apenas um para teste, assim como no experimento anterior. No entanto, para este caso, não foi utilizado apenas o padrão de seus parâmetros, mas sim a função *GridSearchCV*. Com o objetivo de buscar o melhor ajuste de parâmetros para o modelo, essa função realiza combinações dos parâmetros a fim de avaliá-los. O resultado das métricas para cada técnica e seu tempo de execução são apresentados na Tabela 3, assim como os melhores parâmetros para cada uma delas é descrito na Tabela 3.

Tabela 3 - Resultados das métricas e tempo computacional das técnicas (experimento 3)

Técnica	**Acurácia**	**Precisão**	***Recall***	**F1 score**	**ROC/AUC**	**Tempo (s)**
			Ensemble			
ET	0,8987	0,9010	0,9053	0,8983	0,9586	86,5549
RF	**0,9067**	0,8990	**0,9160**	**0,9100**	0,9604	89,7024
XGB	0,8980	0,8918	0,9067	0,8988	0,9561	70,6722
BA	0,9000	0,8792	**0,9160**	0,9042	0,9568	**21,0993**
VO	0,8280	0,8230	0,8333	0,8267	0,9146	96,7537
ST	0,9053	**0,9064**	0,9133	0,9082	**0,9611**	400,5376
			Não *ensemble*			
SVM	0,7673	0,8043	0,7107	0,7518	0,8558	65,0319
LR	0,7100	0,7118	0,7080	0,7085	0,7664	24,5548
KNN	0,7740	0,8611	0,7200	0,7605	0,8328	187,655
NB	**0,8860**	**0,8947**	**0,8800**	0,8850	**0,9541**	5,7400
DT	0,8660	0,8825	0,8627	**0,8669**	0,8693	**4,8724**
NN	0,8527	0,8575	0,8667	0,8613	0,9255	11.444,64

Fonte: A autora (2021)

Acompanhando os experimentos anteriores, também são observados resultados superiores nas técnicas *ensemble* em comparação com as não *ensemble*. Em relação às técnicas *ensemble*, continua se destacando *Random Forest* (RF) em relação às métricas: acurácia, *recall* e F1 score, no entanto *Bagging* (BA) atingiu o mesmo valor de *recall*, assim como o menor tempo computacional. Similarmente, *Stacking* (ST) se manteve com a melhor precisão e ROC/AUC como no experimento anterior.

Consideração as técnicas não *ensemble*, as melhores métricas se mantiveram com *Naïve Bayes* (NB), com exceção de F1 score que agora está relacionada a *Decision Tree* (DT), que também apresentou menor tempo de execução. Além dos resultados das métricas, outra saída da aplicação dos métodos usando *GridSearchCV* foram os melhores parâmetros elencados pelos modelos, como é observado no Quadro 5.

Quadro 5 – Técnicas e melhores parâmetros elencados pela função *GridSearchCV*

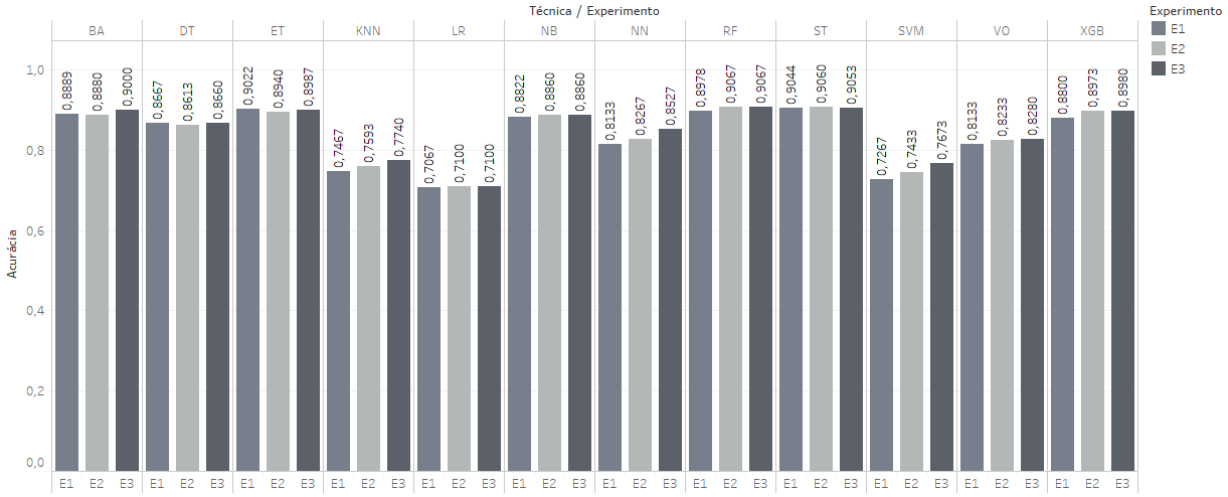
Técnica	**Melhores parâmetros**
ET	criterion: 'gini'; max_features: 'sqrt'
RF	criterion: 'entropy'; max_features: 'sqrt'
XGB	loss: 'auto'; max_leaf_nodes: 10
BA	max_samples: 200; n_estimators: 100
VO	flatten_transform: False; n_jobs:
ST	n_jobs: 10; stack_method: 'predict'
SVM	gamma: 'auto'; kernel: 'rbf'
LR	multi_class: 'auto'; penalty: 'l2'
KNN	algorithm: 'auto'; n_neighbors: 4, weights: 'distance'
NB	alpha: 1; fit_prior: True
DT	criterion: 'entropy'; splitter: 'best'
NN	activation: 'tanh'; solver: 'adam'

Fonte: A autora (2021)

4.2.4 Discussão dos resultados

A utilização de três experimentos comparativos é significativa nesse contexto, pois permite a observação das técnicas utilizadas quando se mantém o conjunto de dados, mas se modifica seu particionamento. Além disso, também é possível observar os resultados quando seus parâmetros são utilizados no *default* do modelo ou são delimitados e comparados. A fim de ilustrar os resultados dos classificadores e suas métricas, a Figura 2 apresenta um resumo das acurácias para os experimentos 1, 2 e 3.

Figura 2 – Relação entre as acurácias dos classificadores e os experimentos



Fonte: A autora (2021)

Observando a Figura 2, as técnicas *Bagging* (BA), *Extra Trees* (ET), *Random Forest* (RF) e *Stacking* (ST) apresentaram acurácia superior a 90% em pelo menos um dos experimentos, sendo aquelas que descrevem melhor performance geral do modelo. Em contrapartida, as técnicas *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR) e *Support Vector Machine* (SVM) apresentam os menores valores de acurácia, ficando abaixo de 80% em todos os experimentos. As demais técnicas apresentaram desempenho mediano em comparação as outras, pois suas acurácias foram entre 81% e 89%.

Além disso, também é possível observar que na maior parte das técnicas os resultados da acurácia foram iguais ou superiores nos experimentos 2 e 3, em relação ao experimento 1. No entanto, a melhoria dos resultados não foi significativa em nenhum dos casos, pois o percentual de mudança foi baixo de um experimento para outro. Isto representa que a forma de particionar os dados e a limitação dos parâmetros do modelo, não alteram de forma significativa seu desempenho.

Em contrapartida, o tempo de execução apresentou um aumento nos últimos experimentos, principalmente com a execução de comparação de parâmetros (experimento 3). Considerando que acurácia é utilizada para descrever o desempenho do modelo de modo geral, as técnicas que apresentam acurácia em torno de 90% apontam que o modelo classifica corretamente 90% dos casos, tanto para óbito quando para não óbito do trabalhador.

Os resultados observados para a acurácia, também foram considerados nas demais métricas, que apresentaram uma melhoria em seus resultados nos outros experimentos, mas nada significativo. Para precisão, os algoritmos que obtiveram melhores resultados foram: *Bagging* (BA), *Extra Trees* (ET) e *Stacking* (ST) que apresentaram resultados superiores a 90%

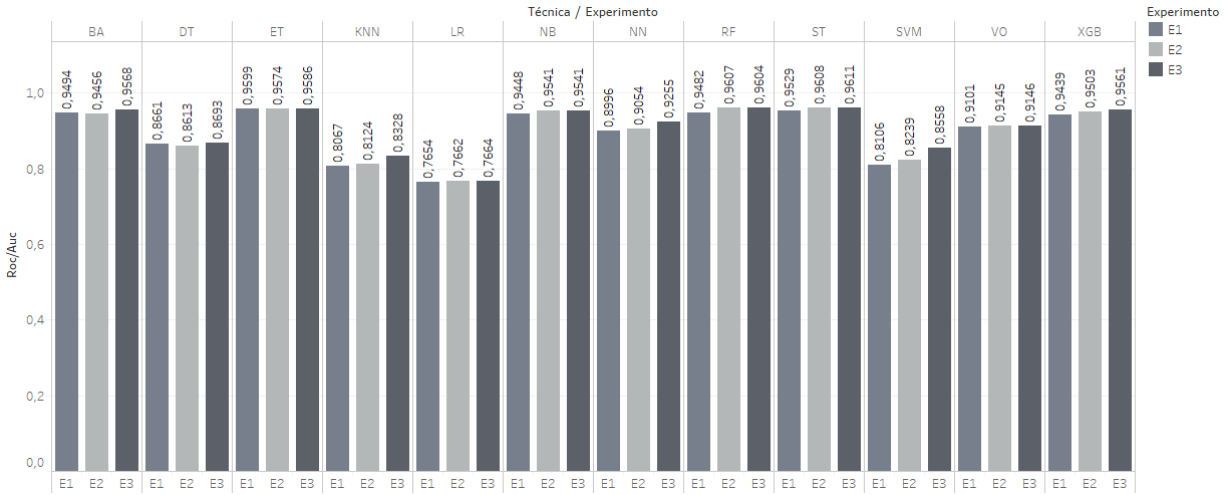
em pelo menos um dos experimentos. Por outro lado, *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR), *Neural Networks* (NN) e *Support Vector Machine* (SVM) apresentaram precisão próxima a 70% em pelo menos um experimento, sendo *Logistic Regression* (LR) o pior desempenho com todos os resultados abaixo de 74%. As demais técnicas apresentaram resultados de precisão dentro deste intervalo.

Em relação a precisão, seus resultados indicam a relação das previsões de positivas que foram classificadas corretamente em comparação com todas as classes positivas do modelo, mesmo aquelas que são consideradas como falso positivas. Isto representa que as técnicas que obtiveram resultados superiores a 90% de precisão classificaram 90% dos resultados positivos de maneira correta.

Recall apresentou os resultados mais baixos em comparação com as demais métricas, onde *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR) e *Support Vector Machine* (SVM) atingiram percentuais abaixo de 70%. Por outro lado, as técnicas *Bagging* (BA), *Extra Trees* (ET), *Random Forest* (RF) e *Stacking* (ST) mantiveram seus resultados acima de 90% como para as demais métricas analisadas. Estes algoritmos com melhor desempenho para *recall* indicam que em 90% das ocorrências, os casos positivos são classificados corretamente.

F1 score, por ser uma relação entre as métricas precisão e *recall*, acompanhou o desempenho das anteriores, em relação às técnicas que se destacaram e aquelas que obtiveram desempenho inferior. Essa métrica descreve a média harmônica entre as duas outras métricas citadas, por isso está diretamente relacionada com as discussões apresentadas anteriormente. Considerando a última métrica avaliada, a curva ROC e AUC, sete técnicas apresentaram resultados superiores a 95%, como é descrito pela Figura 3.

Figura 3 - Relação entre as ROC/AUC dos classificadores e os experimentos

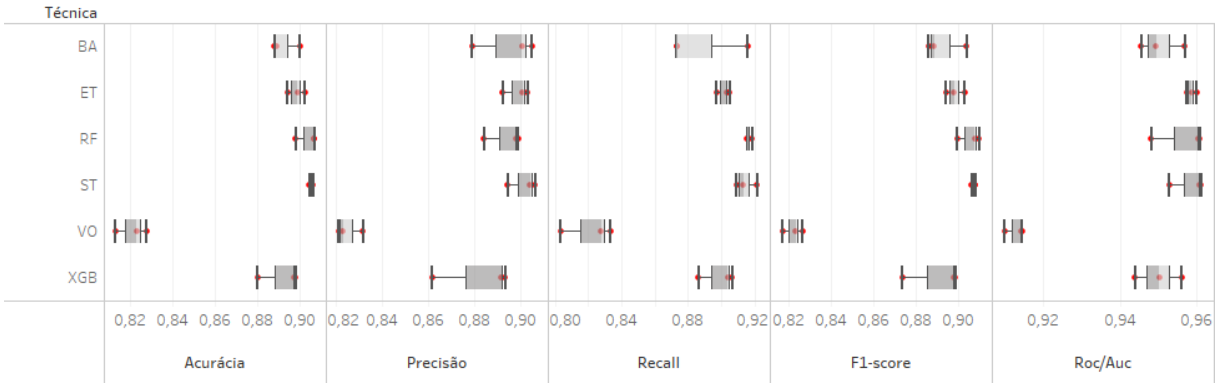


Fonte: A autora (2021)

Bagging (BA), *Extra Trees* (ET), *Naïve Bayes* (NB), *Random Forest* (RF), *Stacking* (ST) e *XGBoost* (XGB) atingiram os melhores resultados, dentre elas com destaque para *Extra Trees* (ET) com média de 95,86% entre os três experimentos. Em contrapartida, *Logistic Regression* (LR) apresentou os valores mais baixos para ROC/AUC, pois, para todos os experimentos os resultados foram em torno de 76%. A curva ROC e AUC auxiliam no entendimento da separabilidade das classes do modelo e assim, descrevendo quais das técnicas apresentam melhor classificação para os dados.

Além disso, os resultados das métricas também foram analisados a partir da configuração de gráficos *bloxpot* para cada uma das métricas em relação às técnicas aplicadas. A Figura 4 apresenta o gráfico para as técnicas *ensemble* e a Figura 5 para as técnicas não *ensemble*, para cada uma das cinco métricas utilizadas.

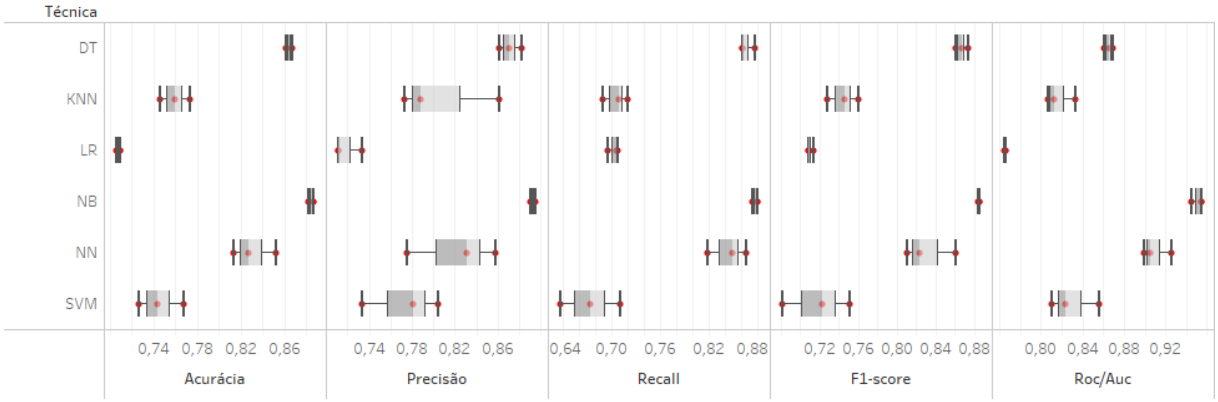
Figura 4 – *Boxplot* da relação entre métricas e técnicas *ensemble*



Fonte: A autora (2021)

Considerando a acurácia, a técnica que apresentou melhor resultado foi *Stacking*, pois contempla o melhor resultado e menor intervalo entre os experimentos avaliados. Para precisão, recall e ROC/AUC também foi *Stacking* que se destacou, atingindo os maiores percentuais. Apenas para a métrica F1 score houve um resultado diferente, onde *Random Forest* apresentou melhores resultados, mas *Stacking* esteve logo atrás. Mesmo que uma técnica tenho sido destaque na maior parte das métricas, os resultados foram bem próximos para todas, com exceção de *Votting*, que apresentou valores inferiores em todos os casos.

Figura 5 – *Boxplot* da relação entre métricas e técnicas não *ensemble*



Fonte: A autora (2021)

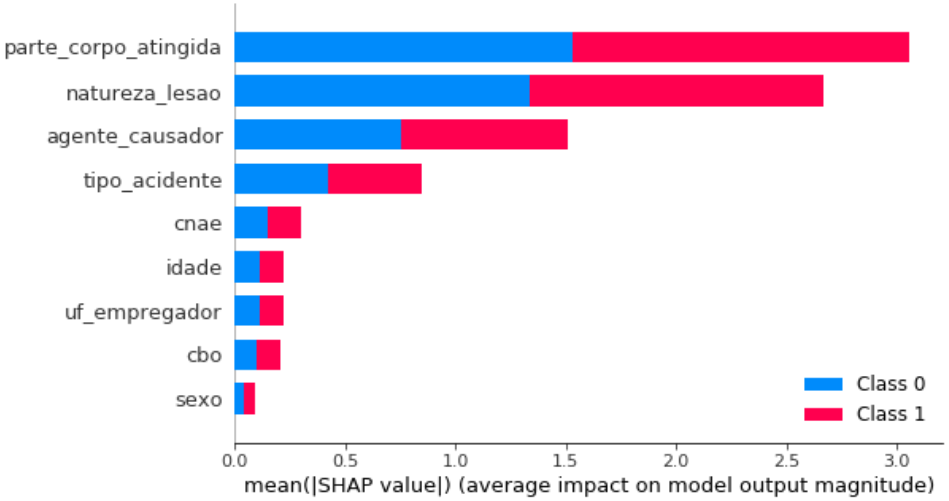
Similarmente às interpretações da Figura 4, com os algoritmos *ensemble*, para os não *ensemble* (Figura 5) também houve destaque de uma técnica. *Naïve Bayes* apresentou melhores resultados em todas as métricas, assim como descreveu baixa variação entre os três experimentos avaliados. Em contrapartida, as técnicas *Logistic Regression* e *SVM* ocuparam os resultados inferiores dentre aquelas analisadas.

4.3 Explicabilidade das técnicas

Avaliar o desempenho das técnicas apenas por meio das métricas não é suficiente para entender seu funcionamento. Em busca da interpretação dos classificadores utilizados, foram selecionados aqueles com melhores resultados, tanto para o grupo de técnicas *ensemble* como não *ensemble*. Para a escolha foram considerados os resultados do último experimento, por ser considerado o mais robusto dentre eles. Assim, as técnicas escolhidas para análise por meio de explicabilidade foram *Random Forest* (RF) e *Naïve Bayes* (NB), pois acumularam os melhores resultados na maior parte das métricas analisadas.

Além disso, para compor a etapa de explicação foi selecionado o *SHapley Additive exPlanations* (SHAP) como algoritmo de inteligência artificial explicável, cujo objetivo é entender o funcionamento e as saídas de um modelo de aprendizado de máquinas. Para este fim, o SHAP realiza explicações locais usando princípios da teoria de jogos para explicar as previsões do algoritmo classificador. Iniciando as explicações pela técnica não *ensemble*, *Naïve Bayes* (NB), a Figura 6 apresenta um dos resultados desta etapa.

Figura 6 – Relação entre os atributos e sua relevância nos resultados de *Naïve Bayes*



Fonte: A autora (2021)

A Figura 6 representa um resumo dos impactos de cada atributo no resultado do modelo, ou seja, na previsão do óbito. Dessa forma, a variável de parte do corpo atingida descreve a maior relevância para previsão do óbito, seguida pela natureza de lesão, destacando que o óbito do trabalhador é diretamente impactado pela região do seu corpo que é atingida pelo acidente ou doença, assim como o tipo da lesão sofrida.

Em contrapartida, o sexo do trabalhador não apresenta relevância para a previsão de óbitos na indústria de transformação, assim como o CBO e idade do empregado, e o estado e CNAE do empregador. Assim como os atributos mais influentes são destacados pela aplicação do SHAP, as categorias de atributos mais significativas também são apresentadas (Figura 7). A figura apresenta o valor base previsto pelo modelo, assim como as variáveis de entrada que empurram a base para valores mais altos, representada pelo lado esquerdo (rosa) e as variáveis que empurram a base para valores mais baixos, no lado direito da figura (azul).

Figura 7 – Categorias de atributos e sua relevância nos resultados de *Naïve Bayes*

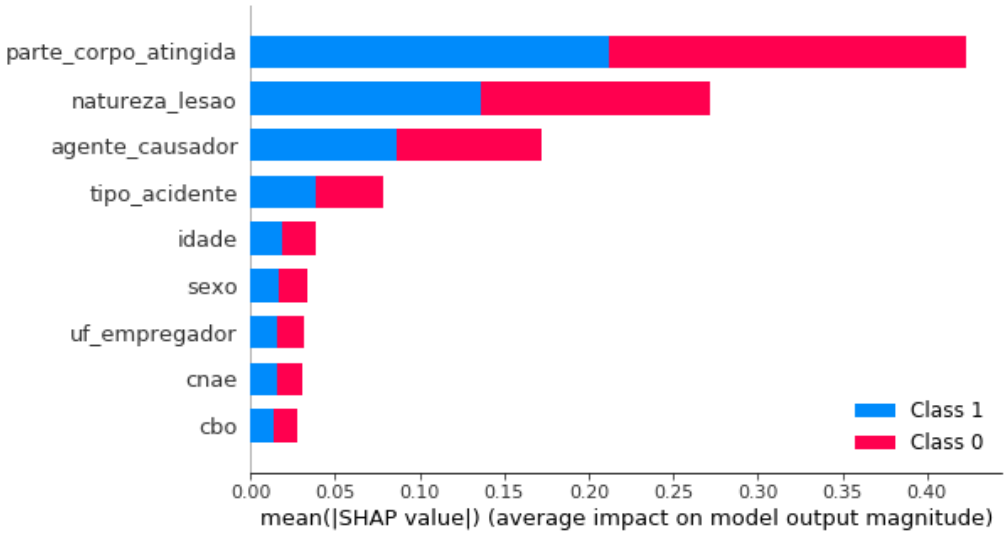
Fonte: A autora (2021)

Conforme apresentado pela Figura 7, as categorias que mais impactam na incidência de óbitos do trabalhador são: categoria de número cinco para natureza da lesão, de número cinco para parte do corpo atingida e zero para agente causador. Em relação à natureza da lesão, a classe cinco representa contusões e esmagamento, enquanto a categoria de parte do corpo

atingida referente ao grupo cinco é descrita por acidentes e doenças que afetam o tronco do trabalhador. Considerando o agente causador, a classe designada pelo número zero corresponde à veículos, meios de transporte e equipamentos de transporte.

Os mesmos processos e interpretações realizados para a técnica *Naïve Bayes*, em busca de explicação por meio do algoritmo SHAP, foram aplicados para *Random Forest*, a técnica *ensemble* com melhores resultados. Dessa forma, as Figuras 8 e 9 descrevem a explicabilidade do algoritmo utilizando SHAP, evidenciando os atributos e categorias mais influentes na previsão do óbito do trabalhador.

Figura 8 - Relação entre os atributos e sua relevância nos resultados de *Random Forest*



Fonte: A autora (2021)

De maneira análoga ao algoritmo anterior, para *Random Forest* os atributos mais significativos para incidência de óbitos do trabalhador também foram parte do corpo atingida, natureza da lesão e agente causador do acidente. A Figura 9 apresenta o valor base previsto pelo modelo de 0.04, assim como as variáveis de entrada que empurram a base para valores altos (rosa) ou baixos (azul).

Figura 9 - Categorias de atributos e sua relevância nos resultados de *Random Forest*

Fonte: A autora (2021)

Nesta técnica, o grupo de lesões (classe 18), que constituem: lesões imediatas, múltiplas ou outras lesões foi a natureza de lesão com maior relação na previsão do óbito. Já em relação

a parte do corpo atingida, o grupo três representa partes múltiplas afetadas pela doença ou acidente, o que caracteriza uma alta gravidade à saúde do trabalhador.

Considerando o agente causador, a classe com maior impacto do resultado do algoritmo são os veículos, meios de transporte e equipamentos de transporte (classe 0). Esta informação somada ao tipo de acidente, que na Figura 9 é descrito pelo número um (1) como a maior relação, que por sua vez representa acidentes de trajeto, implica que os acidentes que ocorrem no trajeto de casa para o trabalho ou do trabalho para casa têm grande impacto nos óbitos da indústria de transformação.

5. Considerações finais

Com o intuito de prever os óbitos na indústria de transformação brasileira, a partir dos dados com abertura de CAT, este trabalho utilizou técnicas de mineração de dados para este fim. Por meio de três experimentos vinculados à 12 algoritmos diferentes, foi possível identificar as primeiras conclusões da pesquisa. Os experimentos se diferenciavam de acordo com seu particionamento de dados e apontamento de parâmetros dos modelos.

O primeiro experimento contava com uma divisão de 70% de dados para treino e 30% para teste, com os parâmetros dos algoritmos utilizados no padrão do modelo. Já o segundo o experimento, também com os parâmetros no default, se diferenciou quanto ao particionamento dos dados, pois utilizou a validação cruzada, onde nove subconjuntos formaram os dados de treino e um outro subconjunto foi utilizado para teste. O terceiro e último experimento, também utilizando a validação cruzada também aplicou o método de *Grid Search* para comparação e definição dos melhores parâmetros do modelo.

No entanto, mesmo com a evolução dos algoritmos, tornando-os mais robustos conforme avançavam os experimentos, eles não apresentaram evolução significativa em seus resultados. Utilizando cinco métricas comparativas (acurácia, precisão, *recall*, F1 *score* e ROC/AUC) foi possível observar que as técnicas mantiveram seu padrão de desempenho, independente da mudança dos experimentos, e ainda, elevaram seu tempo de execução.

Nesta etapa da pesquisa também foram identificados os algoritmos com melhores resultados, sendo elencados *Naïve Bayes* para o grupo de técnicas não *ensemble* e *Random Forest* para as técnicas *ensemble*. Estas indicações se basearam nos resultados das métricas dos experimentos realizados, e embasaram a etapa seguinte da pesquisa, que foi a aplicação de IA explicável. Para este fim foi utilizado o algoritmo SHAP, que buscou explicar os fatores que impactaram nas previsões de cada um dos modelos.

A aplicação de explicabilidade nas técnicas demonstrou quais atributos foram mais significativos na previsão dos óbitos dos trabalhadores da indústria de transformação. Estes atributos foram: parte de corpo atingida, natureza da lesão e agente causador do acidente. Os resultados do SHAP também demostraram dentro de cada atributo, quais são as categorias mais expressivas na previsão dos modelos.

O conhecimento gerado por meio da aplicação da mineração de dados e de AI explicável em dados de saúde e segurança do trabalho, repercute em avanços tanto na academia como na gestão de organizações públicas e privadas. Na academia, pois apresenta o envolvimento de várias áreas tornando-se um estudo multidisciplinar, envolvendo saúde, exatas, engenharia e tecnologia. Na gestão de organizações, apresenta quais são os fatores que mais impactam na letalidade dos acidentes e doenças na indústria de transformação, auxiliando gestores na indicação de fatores de risco e com isso, atuando na redução de ocorrências para esse setor industrial.

Estes fatores, diretamente associados à parte do corpo atingida e natureza da lesão, evidenciam a importância da utilização, controle e fiscalização do uso de equipamentos de proteção pelos trabalhadores. Para isso, existem algumas normas regulamentadoras (NRs) que amparam as especificações para proteção de máquinas e equipamentos durante seu manuseio, limpeza e manutenção, como são as NRs 12, 18 e 22. Seguir as orientações normativas, assim como oferecer e fiscalizar a utilização de EPIs e EPCs são práticas necessárias no contexto da indústria de transformação.

Este estudo apresentou limitações em relação aos dados utilizados, pois alguns atributos do conjunto de dados selecionado apresentavam valores faltantes ou categorias não identificadas e foi necessário que fossem retirados do modelo. Por consequência, um modelo que considera um maior número de variáveis poderia apresentar classificações mais assertivas e uma melhor previsão dos óbitos. Como sugestão para pesquisas futuras, a aplicação de técnicas de mineração de dados em outros setores, tais como: comércio, reparação de veículos automotores e motocicletas, e saúde humana e social, que também apresentam significativos percentuais de registro de CAT.

Referências

ALBER, M.; LAPUSCHKIN, S.; SEEGERER, P.; HÄGELE; M.; SCHÜTT, K. T.; MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R.; DÄHNE, S.; KINDERMANS, P.-J. iNNvestigate neural networks!. **Journal of Machine Learning Research**, v. 20, n. 93, p. 1-8, 2019.

ALEXOPOULOS, K.; NIKOLAKIS, N.; CHRYSSOLOURIS, G. Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing. International Journal of Computer Integrated Manufacturing, v. 33, n. 5, p. 429-439, 2020.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCIA, S.; GIL-LOPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, v. 58, p. 82-115, 2020.

BEVILACQUA, M.; CIARAPICA, F. E.; GIACCHETTA, G. Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach. **Accident Analysis & Prevention**, v. 40, no. 4, p. 1468–1479, 2008.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, v. 30, n. 7, p. 1145-1159, 1997.

BRAMHALL, S.; HORN, H.; TIEU, M.; LOHIA, N. Qlime-a quadratic local interpretable model-agnostic explanation approach. SMU **Data Science Review**, v. 3, n. 1, p. 4, 2020.

BRASIL. Decreto nº 3.048 de 06 de Maio de 1999. **Regulamento da Previdência Social**. Brasília, DF, 6 mai. 1999. 178º da Independência e 111º da República. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/d3048.htm>. Acesso em: 30 mai. 2021.

BULLOCK, J.; LUCCIONI, A.; PHAN, K. H.; LAM, C. S. N.; LUENGO-OROZ, M. Mapping the landscape of artificial intelligence applications against COVID-19. **Journal of Artificial Intelligence Research**, v. 69, p. 807-845, 2020.

CALLAHAN, A.; SHAH, N. H. Machine learning in healthcare. In: **Key Advances in Clinical Informatics**. Academic Press, 2017. p. 279-291.

CAVALCANTE, F.; FERRITE, S.; MEIRA, T. C. Exposição ao ruído na indústria de transformação no Brasil. **Revista CEFAC**, v. 15, n. 5, p. 1364-1370, 2013.

CHEN, T.; GUESTRIN, C. **Xgboost: A scalable tree boosting system**. *In*: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

CHENG, C.-W.; LEU, S.-S.; CHENG, Y.-M.; WU, T.-C.; LIN, C.-C. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. **Accident Analysis & Prevention**, v. 48, p. 214–222, 2012.

CHENG, C. W.; LIN, C. C.; LEU, S. SEN. Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. **Safety Science**, v. 48, no. 4, p. 436–444, 2010.

CHENG, C. W.; YAO, H. Q.; WU, T. C. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. **Journal of Loss Prevention in the Process Industries**, v. 26, no. 6, p. 1269–1278, 2013.

CHOI, J.; GU, B.; CHIN, S.; LEE, J.-C. Machine learning predictive model based on national data for fatal accidents of construction workers. **Automation in Construction**, v. 110, p. 102974, 2020.

CHOU, C.-A.; CAO, Q.; WENG, S.-J.; TSAI, C.-H. Mixed-integer optimization approach to learning association rules for unplanned ICU transfer. **Artificial Intelligence in Medicine**, v. 103, p. 101806, 2020.

CIARAPICA, F. E.; GIACCHETTA, G. Classification and prediction of occupational injury risk using soft computing techniques: An Italian study. **Safety Science**, v. 47, no. 1, p. 36–49, 2009.

CRACKNELL, M. J.; READING, A. M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. **Computers and Geosciences**, v. 63, p. 22–33, 2014.

DAVIS, J.; GOADRICH, M. The relationship between Precision-Recall and ROC curves. *In*: **Proceedings of the 23rd international conference on Machine Learning**. 2006. p. 233-240.

DÉSIR, C.; PETITJEAN, C.; HEUTTE, L.; SALAÜN, M.; THIBERVILLE, L. Classification of endomicroscopic images of the lung based on random subwindows and extra-trees. **IEEE Transactions on Biomedical Engineering**, v. 59, n. 9, p. 2677-2683, 2012.

DIETTERICH T. G. **Ensemble Methods in Machine Learning**, 2000. *In*: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, v. 1857. Springer, Berlin, Heidelberg.

DOS SANTOS, B. S.; STEINER, M. T. A.; FENERICH, A. T.; LIMA, R. H. P. 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, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in databases. **AI Magazine**. v. 17, no. 3, p. 37–54, 1996.

GARCÍA, M. V.; AZNARTE, J. L. Shapley additive explanations for NO2 forecasting. **Ecological Informatics**, v. 56, p. 101039, 2020.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, v. 63, n. 1, p. 3-42, 2006.

GOH, Y. M.; UBEYNARAYANA, C. U. Construction accident narrative classification: An evaluation of text mining techniques. **Accident Analysis & Prevention**, v. 108, p. 122-130, 2017.

GONZALEZ, R.; FIACCHINI, M.; IAGNEMMA, K. Slippage prediction for off-road mobile robots via machine learning regression and proprioceptive sensing. **Robotics and Autonomous Systems**, v. 105, p. 85–93, 2018.

GOODMAN, B.; FLAXMAN, S. European Union Regulations on Algorithmic Decision-

Making and a “Right to Explanation”. **AI Magazine**, [S. l.], v. 38, n. 3, p. 50-57, 2017.

HAJAKBARI, M. S.; MINAEI-BIDGOLI, B. A new scoring system for assessing the risk of occupational accidents: A case study using data mining techniques with Iran’s Ministry of Labor data. **Journal of Loss Prevention in the Process Industries**, v. 32, p. 443–453, 2014.

HEARST, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. **IEEE Intelligent Systems and Their Applications**, v. 13, n. 4, p. 18-28, 1998.

HEO, S.-J.; KIM, Y.; YUN, S.; LIM, S.-S.; KIM, J.; NAM, C.-M.; PARK, E.-C.; JUNG, I.; YOON, J.-H. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers’ Health Examination Data. **International Journal of Environmental Research and Public Health**, v. 16, n. 2, p. 250, 2019.

HOOD, S. B.; CRACKNELL, M. J.; GAZLEY, M. F. Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning. **Journal of Geochemical Exploration**, v. 186, p. 270–280, 2018.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Estrutura das atividades econômicas na CNAE**, 2021. Disponível em: <https://cnae.ibge.gov.br/>. Acesso em: 27 Abr 2021.

ILO – INTERNATIONAL LABOUR ORGANIZATION. **Quick guide on sources and uses of statistics on occupational safety and health**, 2020. Disponível em: <https://www.ilo.org/global/statistics-and-databases/publications/WCMS_759401/lang--en/index.htm>. Acesso em: 26 Abr 2021.

ILO - INTERNATIONAL LABOUR ORGANIZATION. **Rules of the game: An introduction to the standards-related work of the International Labour Organization**, 2019. Disponível em: <https://www.ilo.org/global/standards/information-resources-and-publications/publications/WCMS_672549/lang--en/index.htm>. Acesso em: 26 Abr 2021.

JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. Supervised machine learning: A brief primer. **Behavior Therapy**, v. 51, n. 5, p. 675-687, 2020.

JOCELYN, S.; OUALI, M.-S.; CHINNIAH, Y. Estimation of probability of harm in safety of machinery using an investigation systemic approach and Logical Analysis of Data. **Safety Science**, v. 105, p. 32, 2018.

KELLER, J. M.; GRAY, M. R.; GIVENS, J. A. A fuzzy k-nearest neighbor algorithm. **IEEE Transactions on Systems, Man, and Cybernetics**, n. 4, p. 580-585, 1985.

LEE, J.; KIM, H.-R. Prediction of return-to-original-work after an industrial accident using machine learning and comparison of techniques. **Journal of Korean Medical Science**, v. 33, no. 19, 2018.

LIAO, C.-W.; PERNG, Y.-H. Data mining for occupational injuries in the Taiwan construction industry. **Safety Science**, v. 46, no. 7, p. 1091–1102, 2008.

LIU, J.; KONG, X.; ZHOU, X.; WANG, L.; ZHANG, D.; LEE, I.; XU, B.; XIA, F. Data Mining and Information Retrieval in the 21st century: A bibliographic review. **Computer Science Review**, v. 34, p. 100193, 2019.

LUNDBERG, S.; LEE, S. A unified approach to interpreting model predictions. **arXiv** preprint arXiv:1705.07874, 2017.

MARQUÉS, A. I.; GARCÍA, V.; SÁNCHEZ, J. S. Exploring the behaviour of base classifiers in credit scoring ensembles. **Expert Systems with Applications**, v. 39, n. 11, p. 10244-10250, 2012.

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

MENEGON, L. S.; MENEGON, F. A.; MAENO, M.; KUPEK, E. Incidência e tendência temporal de acidentes de trabalho na indústria têxtil e de confecção: análise de Santa Catarina, Brasil, entre 2008 e 2017. **Revista Brasileira de Epidemiologia**, v. 24, 2021.

MITCHELL, R.; MICHALSKI, J.; CARBONELL, T. **An artificial intelligence approach**. Berlin: Springer, 2013.

MISTIKOGLU, G.; GEREK, I. H.; ERDIS, E.; USMEN, P. E. M.; CAKAN, H.; KAZAN, E. E. Decision tree analysis of construction fall accidents involving roofers. **Expert Systems with Applications**, v. 42, no. 4, p. 2256–2263, 2015.

MUTLU, N. G.; ALTUNTAS, S. Risk analysis for occupational safety and health in the textile industry: Integration of FMEA, FTA, and BIFPET methods. **International Journal of Industrial Ergonomics**, v. 72, p. 222–240, 2019.

NENONEN, N. 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, no. 2, p. 215–224, 2013.

PARK, S.; MOON, J.; HWANG, E. **Explainable Anomaly Detection for District Heating Based on Shapley Additive Explanations**. In: 2020 International Conference on Data Mining Workshops (ICDMW). IEEE, 2020. p. 762-765.

PEKEL, E.; AKŞCHIR, Z. D.; METO, B.; AKLEYLEK, S.; KILIÇ, E. A Bayesian Network Application in Occupational Health and Safety. In: 2018 3rd International Conference on Computer Science and Engineering (UBMK). **IEEE**, 2018. p. 239-243.

PURDY, M.; DAUGHERTY, P. **How AI Boosts Industry Profits and Innovation**. Accenture Research, 2017. Disponível em: <https://www.accenture.com/pt-pt/insight-ai-industry-growth>. Acesso em: 30 Mai 2021.

REIS, B. L.; RAFAEL, C.; LEAL, G. C. L.; THOM DE SOUZA, R. C.; GALDAMEZ, E. V. C. **Doenças e acidentes de trabalho no Brasil: uma análise exploratória de dados**. *In*: X Congresso Brasileiro de Engenharia de Produção, 10, 2020. Anais eletrônicos do ConBRepro

2020, p. 1-12.

SANNI-ANIBIRE, M. O.; MAHMOUD, A. S.; HASSANAIN, M. A.; SALAMI, B. A. A risk assessment approach for enhancing construction safety performance. **Safety Science**, v. 121, p. 15–29, 2020.

SANNI ALI, M.; ICHIARA, M. Y.; LOPES, L. C.; BARBOSA, G. C. G.; PITA, R.; CARREIRO, R. P.; SANTOS, D. B.; RAMOS, D.; BISPO, N.; RAYNAL, F.; CANUTO, V.; ALMEIDA, B. A.; FIACCONE, R. L.; BARRETO, M. E.; SMEETH, L.; BARRETO, M. T. Administrative data linkage in Brazil: Potentials for health technology assessment. **Frontiers in Pharmacology**, v. 10, 2019.

SANMIQUEL, L.; BASCOMPTA, M.; ROSSEL, J. M.; ANTICOI, H. F.; GUASH, E. Analysis of Occupational Accidents in Underground and Surface Mining in Spain Using Data-Mining Techniques. **International Journal of Environmental Research & Public Health**, v. 15, no. 3, p. 462, 2018.

SANMIQUEL, L.; ROSSELL, J. M.; VINTRÓ, C. Study of Spanish mining accidents using data mining techniques. **Safety Science**, v. 75, p. 49–55, 2015.

SCHMIDHUBER, J. Deep Learning in neural networks: An overview. **Neural networks**, v. 61, p. 85-117, 2015.

SHIN, D.-P.; PARK, Y.-J.; SEO, J.; LEE, D.-E. Association Rules Mined from Construction Accident Data. **KSCE Journal of Civil Engineering**, v. 22, no. 4, p. 1027-1039, 2018.

SHIRALI, G. A.; NOROOZI, M. V; MALEHI, A. S. Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran. **Journal of Public Health Research**, v. 7, no. 2, 2018.

THOM DE SOUZA, R. C. **Uma metodologia para classificação de dados nominais baseada no processo KDD: ênfase aos algoritmos culturais, estimação de distribuição e análise de correspondência múltipla**. Tese (Doutorado em Métodos Numéricos em Engenharia) – Universidade Federal do Paraná, Curitiba, Paraná.

VOUGAS, K.; SAKELLAROPOULOS, T.; KOTSINAS, A.; FOUKAS, G.-R. P.; NTARGARAS, A.; KOINIS, F.; POLYZOS, A.; MYRIANTHOPOULOS, V.; ZHOU, H.; NARANG, S.; GEORGOULIAS, V.; ALEXOPOULOS, L.; AIFANTIS, I.; TOWNSEND, P. A.; SFIKAKIS, P.; FITZGERALD, R.; THANOS, D.; BARTEK, J.; ... GORGOULIS, V. G. Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining. **Pharmacology & Therapeutics**, p. 107395, 2019.

WANG, Y.; HONG, C.; BEI, L.; MENGHUA, Y.; QIANYI, L. A Systematic Review on the research progress and evolving trends of occupational health and safety management: a bibliometric analysis of mapping knowledge domains. **Frontiers in Public Health**, v.8, p. 81, 2020.

WITTEN, I., FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. 4 ed. São Francisco: Morgan Kaufmann, 2016.

ZHANG, Z.; BECK, M. W.; WINKLER, D. A.; HUANG, B.; SIBANDA, W.; GOYAL, H. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. **Annals of Translational Medicine, v. 6, n. 11, 2018.**

ZHAO, Y.; ZHANG, C.; ZHANG, Y.; WANG, Z.; LI, J. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. **Energy and Built Environment**, v. 1, no. 2, p. 149–164, 2020.

Apêndices

Apêndice A – Categorização do atributo agente causador.

Código (atributo)	Descrição	Nº de atributos agrupados
0	Veículos, meios de transporte e equipamentos de transporte	19
1	Produtos alimentícios e/ou de origem animal	13
2	Ferramentas manuais sem força motriz	21
3	Ferramentas portáteis com força motriz ou aquecimento	15
4	Mobiliário e acessórios	8
5	Superfícies e equipamentos utilizados para sustentar pessoas	13
6	Edifício ou estrutura	16
7	Embalagem ou recipiente (vazio ou cheio)	5
8	Dispositivo de transmissão de energia mecânica	6
9	Condições ambientais ou do ambiente	20
10	Compostos e/ou substâncias químicas	16
11	Equipamentos de guindar	11
12	Equipamentos elétricos	11
13	Equipamentos ou substâncias emissoras de radiação (ionizante e não ionizante)	8
14	Petróleo, combustíveis e derivados	10
15	Equipamentos para trabalho em ambiente de pressão anormal	4
16	Máquinas	21
17	Cerâmica, utensílios e materiais derivados	7
18	Fornos e caldeiras	3
19	Vestuário e têxteis	2
20	Bombas, motores e turbinas	5
21	Agente infeccioso, produto biológico e medicamentos	3
22	Metais e minerais	3
23	Equipamentos sob pressão	2
24	Ser vivo	4
25	Aprisionamento, atrito, abrasão, impacto ou queda	8
26	Madeira	1
27	Água e líquidos	3
28	Vidro	1
29	Fogo e materiais inflamáveis	1
30	Esforço excessivo e movimentos involuntários	3
31	Outros	8

Apêndice B – Categorização do atributo natureza da lesão.

Código (atributo)	**Descrição**	**Nº de atributos agrupados**
0	Amputação ou enucleação	1
1	Asfixia, estrangulamento e afogamento	1
2	Choque elétrico e eletroplessão	1
3	Concussão cerebral	1
4	Congelamento e geladura	1
5	Contusão e esmagamento	1
6	Corte, laceração, ferida contusa e punctura	1
7	Dermatose	1
8	Distensão e torção	1
9	Doença	2
10	Efeito de radiação (Imediato e mediato)	1
11	Envenenamento sistêmico	1
12	Escoriação e abrasão	1
13	Esforço excessivo	1
14	Fratura	1
15	Hérnia de qualquer natureza e ruptura	1
16	Inflamação de articulação, tendão ou músculo	1
17	Intermação, insolação, cãibra e exaustão	1
18	Lesões	4
19	Luxação	1
20	Perda ou diminuição de sentido (Imediato e mediato)	1
21	Pneumoconiose	1
22	Queimaduras	2

5

CONSIDERAÇÕES FINAIS

Este trabalho buscou lançar luz sobre a questão de pesquisa: Como a mineração de dados pode contribuir para a saúde e segurança do trabalho no Brasil? Para isso, foi traçado o objetivo principal da pesquisa, voltado a prever a ocorrência de óbitos no Brasil, a partir de registros de doenças e acidentes de trabalho. Mas seu desenvolvimento não foi apenas focado na previsão de ocorrências fatais ao trabalhador, de forma que esta pesquisa se iniciou com estudo exploratório, tanto da literatura, como dos dados selecionados para tal.

Um mapeamento sistemático da literatura, realizado pelo grupo de pesquisa nas áreas de mineração de dados e saúde e segurança do trabalho foi o passo inicial para identificação das pesquisas desenvolvidas no tema e de lacunas na literatura. Este estudo serviu de subsídio para identificação de potenciais pesquisas, como é o caso da presente dissertação, pois foi possível identificar a escassez de estudos ligados a mineração de dados e SST desenvolvidos no Brasil. Menor ainda foi o número de pesquisas utilizando conjuntos de dados públicos e de abrangência nacional. Essa carência de pesquisas na área foi um motivador para seleção do conjunto de dados da CAT, que expressa ocorrências registradas, disponibilizadas publicamente e englobando todo território nacional.

Após a seleção do conjunto de dados, com a realização da análise exploratória, foi possível entender os atributos representados nos dados selecionados. As primeiras conclusões com relação a essa análise foram as variáveis não confiáveis do conjunto, que apresentavam grande percentual de informações incompletas, como o estado em que ocorreu o acidente ou as variáveis relacionadas ao benefício. Também foram considerados os *outliers* presentes no

conjunto de dados, como os casos de idade discrepantes com a faixa etária legal para trabalho ou as ocorrências com registro *"n class"* para algumas categorias.

Dentre as variáveis com registros válidos, foi possível constatar que o gênero mais representativo é o mais masculino nas ocorrências de acidentes e que entre 26 e 35 anos é a faixa etária com maior número de casos. A ocupação profissional que mais registra acidentes também foi avaliada e está relacionada ao grupo de trabalhadores da produção de bens e serviços industriais, assim como o setor industrial com maior número de registros é a indústria de transformação. Esses empregadores estão localizados em sua grande maioria no estado de São Paulo e considerando as cidades com maior destaque para registros de acidentes, as mais representativas são capitais.

Considerando as variáveis relativas à ocorrência, aproximadamente 75% registravam acidentes típicos. Para os acidentes o principal agente causador foram as motocicletas, enquanto para as doenças foi o esforço excessivo. Em relação à classificação das doenças, o capítulo XIX descrito por "lesões, envenenamentos e algumas outras consequências de causas externas", é o destaque entre os registros, apresentando o maior índice, enquanto a parte do corpo atingida com maior número de ocorrências foram os membros superiores. Para a natureza da lesão, a classe de "corte, laceração, ferida contusa e punctura (ferida aberta)" foi frequentemente registrada. Também foi possível observar através dos dados o período para abertura da CAT e quem foram seus responsáveis, concluindo que na maioria das vezes os comunicados são registrados no mesmo mês de ocorrência e pelo próprio empregador.

Com o estudo exploratório inicial foi possível observar os dados mais atentamente e tomar decisões iniciais, como a retirada de algumas variáveis do conjunto. Além disso, nesta etapa, optou-se por particionar o conjunto de dados, selecionando apenas as ocorrências da indústria de transformação. Essa escolha foi realizada pois este setor representa aproximadamente 27% da totalidade de ocorrências, sendo o mais significativo nesse contexto. Uma análise exploratória para este setor industrial também foi detalhada, no entanto, a maior parte das categorias de destaque acompanhou o que já havia sido observado no conjunto de dados geral.

O sexo predominante em números de registros foi o masculino, assim como a faixa etária em destaque permaneceu entre 26 e 35 anos. A grande maioria dos trabalhadores são considerados empregados pela Previdência Social (98,5%), com destaque no CBO 7 que são trabalhadores da produção de bens e serviços industriais. Quanto à localização do empregador, o estado de São Paulo continua acumulando o maior número de ocorrências e pelo subgrupo de CNAE do empregador os destaques são para fabricação de produtos alimentícios. O capítulo

XIX do CID-10 permaneceu como predominante nos dados, assim como a natureza da lesão (corte, laceração, ferida contusa e punctura), o agente causador (veículos, meios de transporte e equipamentos de transporte), e a parte do corpo atingida (membros superiores).

Com a preparação dos dados a etapa posterior da pesquisa, descrita no segundo artigo, teve início. Este, por sua vez, utilizou doze técnicas de mineração de dados, aplicadas ao conjunto de dados da CAT com recorte na indústria de transformação, buscando prever a ocorrência de óbitos dos trabalhadores. As técnicas foram submetidas à três experimentos e os resultados foram avaliados por meio de cinco métricas: acurácia, precisão, *recall*, F1 score e ROC/AUC. Os experimentos se diferenciavam quanto aos seus subconjuntos de treino e teste e quantos aos parâmetros.

O primeiro experimento utilizou um conjunto de 70% de dados para treino e 30% para teste, com seus parâmetros no padrão do modelo. O segundo experimento, com os parâmetros ainda no padrão, utilizou a validação cruzada para divisão dos seus subconjuntos de treino e teste, onde decompôs em dez partes, sendo apenas uma utilizada para teste. O último experimento, também utilizando a validação cruzada com dez subconjuntos, realizou uma comparação entre parâmetros para avaliar a melhor combinação, utilizando para isso a função *Grid Search*. A partir da etapa de mineração foram elencadas as duas melhores técnicas, para as *ensemble* o melhor desempenho foi de *Naïve Bayes* e para as não *ensemble* foi *Random Forest*.

Após a seleção das duas técnicas de destaque, estas foram submetidas à um algoritmo de inteligência artificial explicável, em busca de uma explicação para as previsões de óbito. O algoritmo escolhido foi o SHAP, e como resultado de sua aplicação foi possível levantar os atributos e categorias mais relevantes no momento da decisão do modelo. Tanto para *Naïve Bayes* quanto para *Random Forest*, os atributos de parte do corpo atingida, natureza da lesão e agente causador, foram os mais influentes na decisão. A etapa de pós-processamento com a aplicação do SHAP foi importante para entender a previsão realizada pelos modelos de mineração de dados.

A pesquisa realizada apontou contribuições não apenas para o ambiente científico ao qual foi desenvolvida, mas também para demais organizações e sociedade. No contexto científico, o levantamento na literatura e identificação de lacunas na pesquisa, assim como o desenvolvimento deste trabalho na condição de multidisciplinariedade, envolvendo áreas de tecnologia e saúde. Para a sociedade, o estudo da saúde e segurança do trabalhador pode contribuir para a prevenção e mitigação de acidentes, prezando pela vida, saúde e segurança do ser humano em seu ambiente de trabalho.

Relacionados aos benefícios para a sociedade, estão as contribuições para organizações públicas e privadas. O conhecimento gerado pelo estudo exploratório do conjunto de dados de doenças e acidentes de trabalho, assim como os resultados da mineração de dados e aplicação de XAI, podem subsidiar novas medidas tomadas pelas organizações. Estas ações podem envolver a definição de normas de proteção ao trabalhador ou ao ambiente de trabalho, voltadas principalmente para as categorias de maior incidência de acidentes, ou ainda, aquelas que mais impactam na mortalidade.

Os resultados da pesquisa também auxiliam gestores e equipe de SST na tomada de decisão relacionada aos riscos, principalmente voltados para a indústria de transformação que foi o objeto de estudo. Além disso, os levantamentos desta pesquisa podem ser utilizados em consonância com outras medidas já existentes, como por exemplo o uso de equipamentos de proteção (EPIs e EPCs), normas regulamentadoras e políticas internas, destacando quais os pontos requerem maior atenção a partir dos resultados da pesquisa. Destacando as contribuições desta pesquisa, a utilização de dados da indústria de transformação, pois é um setor com grande número de registros de acidentes no Brasil, além de que estudos semelhantes já foram conduzidos em outros países (LIAO; PERNG, 2008), demostrando a relevância da pesquisa na área.

Em relação às limitações da pesquisa, pode ser destacada a restrição de alguns fatores, tais como a tarefa de mineração utilizada e o recorte no conjunto de dados. Foram utilizadas apenas técnicas de classificação e dados da indústria de transformação, limitando os resultados encontrados. Outra limitação da pesquisa está relacionada à disponibilidade dos dados, que são disponibilizados publicamente, mas não apresentam uma periodicidade padrão para sua publicação. Também é uma restrição da pesquisa a utilização de dados apenas no contexto brasileiro, sem comparar com outros contextos, pois a realidade de outros países pode ser diferente dos acidentes registrados no Brasil.

Como sugestão para pesquisa futuras, podem ser elencadas a utilização de novos conjuntos de dados, o teste de outras técnicas e métricas, tanto para mineração de dados como para o pós-processamento e avaliação de outros setores industriais. Para a sugestão de novos conjuntos de dados, a comparação com dados de outros países, outros conjuntos de dados nacionais ou ainda a aplicação das mesmas técnicas para previsão de óbitos utilizando os últimos dados disponibilizados pela CAT.

Quanto à sugestão de utilizar outras técnicas, podem ser avaliadas outras técnicas de mineração de dados, buscando verificar seus resultados ou ainda, avaliar as mesmas técnicas por meio de outras métricas, como Kappa, sensitividade e especificidade. Além de outras

técnicas, podem ser utilizadas outras tarefas de mineração de dados, como regressão ou agrupamento (*clustering*), avaliando o mesmo conjunto de dados. Em relação aos setores industriais, analisar não apenas a indústria de transformação, mas também outros setores com alto percentual de ocorrências, como: comércio, reparação de veículos automotores e motocicletas; saúde humana e serviços sociais; e atividades administrativas e serviços complementares.

REFERÊNCIAS

ALDOWAH, H.; AL-SAMARRAIE, H.; FAUZY, W. M. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. **Telematics and Informatics**, v. 37, p. 13-49, 2019.

ARJI, G.; SAFDARI, R.; REZAEIZADEH, H.; ABBASIAN, A.; MOKHTARAN, M.; AYATI, M. H. A systematic literature review and classification of knowledge discovery in traditional medicine. **Computer Methods and Programs in Biomedicine**, v. 168, p. 39–57, 2019.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCIA, S.; GIL-LOPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, v. 58, p. 82-115, 2020.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14280**: Cadastro de acidente do trabalho - Procedimento e classificação. Rio de Janeiro, p. 1-94, 2001.

BAUER, E.; KOHAVI, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. **Machine Learning**, v. 36, n. 1, p. 105-139, 1999.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory**. 1992, p. 144-152.

BRASIL. Decreto nº 3.048 de 06 de Maio de 1999. **Regulamento da Previdência Social**. Brasília, DF, 6 mai. 1999. 178º da Independência e 111º da República. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/d3048.htm>. Acesso em: 6 jun 2021.

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

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BULLOCK, J.; LUCCIONI, A.; PHAN, K. H.; LAM, C. S. N.; LUENGO-OROZ, M. Mapping the landscape of artificial intelligence applications against COVID-19. **Journal of Artificial Intelligence Research**, v. 69, p. 807-845, 2020.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. 2016. p. 785-794.

CHOI, J.; GU, B.; CHIN, S.; LEE, J.-C. Machine learning predictive model based on national data for fatal accidents of construction workers. **Automation in Construction**, v. 110, p. 102974, 2020.

CHOU, C.-A.; CAO, Q.; WENG, S.-J.; TSAI, C.-H. Mixed-integer optimization approach to learning association rules for unplanned ICU transfer. **Artificial Intelligence in Medicine**, v. 103, p. 101806, 2020.

DEL POZO-ANTÚNEZ, J. J.; ARIZA-MONTES, A.; FERNANDÉZ-NAVARRO, F.; MOLINA-SANCHÉZ, H. Effect of a job demand-control-social support model on accounting professionals' health perception. **International Journal of Environmental Research and Public Health**, v. 15, no. 11, 2018.

DOS SANTOS, B. S.; STEINER, M. T. A.; FENERICH, A. T.; LIMA, R. H. P. 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, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in databases. **AI Magazine**. v. 17, no. 3, p. 37–54, 1996.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3-42, 2006.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4ª ed. São Paulo: Atlas, 2017.

HOOD, S. B.; CRACKNELL, M. J.; GAZLEY, M. F. Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning. **Journal of Geochemical Exploration**, v. 186, p. 270–280, 2018.

ILO – INTERNATIONAL LABOUR ORGANIZATION. **Quick guide on sources and uses of statistics on occupational safety and health**, 2020. Disponível em: <https://www.ilo.org/global/statistics-and-databases/publications/WCMS_759401/lang--en/index.htm>. Acesso em: 06 Jun 2021.

LIAO, C.-W.; PERNG, Y.-H. Data mining for occupational injuries in the Taiwan construction industry. **Safety Science**, v. 46, no. 7, p. 1091–1102, 2008.

LIAO, X.; XUE, Y.; CARIN, L. Logistic regression with an auxiliary data source. In: **Proceedings of the 22nd international conference on Machine learning**. 2005. p. 505-512.

LUNDBERG, S. M.; LEE, S.-I. GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H. (eds.). A Unified Approach to Interpreting Model Predictions (PDF), **Advances in Neural Information Processing Systems 30**, Curran Associates, Inc., p. 4765–4774, 2017.

MINISTÉRIO DA ECONOMIA. **Relação Anual de Informações Sociais – RAIS 2019**,

2020. Disponível em: <https://static.poder360.com.br/2020/10/Sumario-Executivo_RAIS-2019.pdf>. Acesso em: 6 jun 2021.

MINISTÉRIO DA FAZENDA. **Anuário Estatístico de Acidentes de Trabalho - AEAT 2018**, 2018. Disponível em: <https://www.gov.br/previdencia/pt-br/assuntos/previdencia-social/saude-e-seguranca-do-trabalhador/dados-de-acidentes-do-trabalho/arquivos/aeat-2018.pdf>. Acesso em: 6 de jun 2021.

MINISTÉRIO DO TRABALHO E EMPREGO. **Classificação Brasileira de Ocupações**: CBO - 2010 – 3ª ed. Brasília: MTE, SPPE, 2010.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81-106, 1986.
SANNI-ANIBIRE, M. O.; MAHMOUD, A. S.; HASSANAIN, M. A.; SALAMI, B. A. A risk assessment approach for enhancing construction safety performance. **Safety Science**, v. 121, p. 15–29, 2020.

SARKAR, M.; LEONG, T.-Y. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. In: Proceedings of the AMIA Symposium. **American Medical Informatics Association**, 2000. p. 759.

SCHUH, G.; REINHART, G.; PROTE, J.-P.; SAUERMANN, F.; HORSTHOFER, J.; OPPOLZER, F.; KNOLL, D. Data mining definitions and applications for the management of production complexity. In: 52nd **CIRP** conference on manufacturing systems. 2019. p. 874-879.

TORRECILLA, J. L.; ROMO, J. Data learning from big data. **Statistics and Probability Letters**, v. 136, p. 15–19, 2018.

WOLPERT, D. H. Stacked generalization. **Neural networks**, v. 5, n. 2, p. 241-259, 1992.

ZHANG, G. P. Neural networks for classification: a survey. **IEEE Transactions on Systems, Man, and Cybernetics**, Part C (Applications and Reviews), v. 30, n. 4, p. 451-462, 2000.

ZHANG, H. The Optimality of Naive Bayes,". In: Seventeenth International Florida Artificial Intelligence Research Society Conference (**FLAIRS 2004**). 2004. p. 1-6.

ZHAO, Y.; ZHANG, C.; ZHANG, Y.; WANG, Z.; LI, J. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. **Energy and Built Environment**, v. 1, no. 2, p. 149–164, 2020.