Universidade de Brasília - UnB
Faculdade UnB Gama - FGA
Curso de Engenharia Eletrônica

Uso de aprendizado de máquinas para análise de acidentes de trabalho

Autor: Fábio Barbosa Pinto
Orientador: Prof. Dr. John Lenon C. Gardenghi

Brasília, DF
2022

Fábio Barbosa Pinto

Uso de aprendizado de máquinas para análise de acidentes de trabalho

Monografia submetida ao curso de graduação em Engenharia Eletrônica da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia Eletrônica.

Orientador: Prof. Dr. John Lenon Cardoso Gardenghi

Brasília, DF
2022

CIP – Catalogação Internacional da Publicação*

Pinto, Fábio Barbosa.

Um estudo sobre pré-processamento de dados e modelagem de classificação para análise de acidentes de trabalho / Fábio Barbosa Pinto Brasília: UnB, 2022. 103 p. : il. ; 29,5 cm.

Monografia De Engenharia Eletrônica – Universidade de Brasília

Faculdade do Gama, Brasília, 2022. Orientação: John Lenon Cardoso Gardenghi

1.Aprendizado de máquinas. 2. Acidentes de trabalho. 3. Análise de correlação. Gardenghi, John Lenon Cardoso. II. Uso de técnicas de aprendizado de máquinas para análise de acidentes de trabalho.

CDU Classificação

Uso de aprendizado de máquinas para análise de acidentes de trabalho

Fábio Barbosa Pinto

Monografia submetida como requisito parcial para obtenção do Título de Bacharel em Engenharia Eletrônica da Faculdade UnB Gama - FGA, da Universidade de Brasília, em 12 de maio de 2022, apresentada e aprovada pela banca examinadora abaixo assinada:

Prof. Dr. John Lenon Cardoso Gardenghi UnB/FGA
Orientador

Prof. Dr. Daniel Sundfeld Lima UnB/FGA
Membro Convidado

Prof. Dr. Glauco Vitor Pedrosa UnB/FGA
Membro Convidado

Brasília, DF
2022

Não há exemplo maior de dedicação do que o da nossa família. À minha querida família, que tanto admiro, dedico o resultado do esforço realizado ao longo deste percurso.

AGRADECIMENTOS

Gostaria de agradecer, especialmente, aos meus pais, por me ensinarem a ser forte, honesto, solidário e a nunca deixar de sonhar.

Agradeço aos meus irmãos, sobrinhos, madrinha e tias por todo amor, compreensão e apoio, durante essa longa jornada.

Agradeço ainda aos professores e a todos os profissionais que trabalham na FGA que possibilitaram a realização do curso nesses últimos anos. Em especial ao Prof. John Lenon Cardoso Gardenghi por me acompanhar e orientar nessa reta final.

Por fim, gostaria de agradecer a todos os meus amigos e companheiros, que inúmeras vezes ergueram a minha cabeça e não me deixaram desistir. O mérito desse trabalho divido com vocês: Alan, Andréia, Augusto, Guilherme, Igor, João, Kelly, Maisa, Matheus e Pedro.

RESUMO

Algoritmos de aprendizado de máquina (ML) tem sido cada vez mais utilizados para classificação em diferentes aplicações. Entretanto, sua exploração é recente, na classificação de acidentes, no campo da segurança do trabalho. Apenas a utilização de classificadores otimizados pode não ser suficiente para avaliar a importância das variáveis presentes no banco de dados. Neste trabalho, o algoritmo C4.5 foi utilizado para classificar a gravidade de acidentes de trabalho, usando uma base de dados abertos da Comunicação de Acidentes de Trabalho (CAT). Em particular, as variáveis mais influentes são analisadas por meio de métodos como qui-quadrado e ganho de informação. Foram realizados dois experimentos para comparar a seleção de variáveis e a eficiência do classificador. O primeiro experimento foi realizado com a base desbalanceada, sendo que a classe minoritária, representa 40% da amostra e o segundo experimento foi realizado com a variável alvo balanceada, utilizando a técnica SMOTE. Todas as variáveis foram selecionadas a partir do qui-quadrado. O ganho de informação ordenou as variáveis mais importantes. O nó raiz da árvore criada foi a variável natureza da lesão e as variáveis emitente e filiação não foram relevantes para o modelo. O algoritmo C4.5 teve melhor desempenho utilizando o SMOTE, com 89,23% de acurácia, sendo 1% maior do que a base sem balanceamento.

Palavras-chave: Aprendizado de máquinas. Acidentes de trabalho. SMOTE.

LISTA DE TABELAS

Tabela 1. Gerações das teorias de acidente 16
Tabela 2. Bases de dados relacionadas à Saúde do Trabalhador 19
Tabela 3. Resultados da aplicação do qui-quadrado, para a base sem balanceamento 44
Tabela 4. Resultados da aplicação do ganho de informação, para a base sem balanceamento 44
Tabela 5. Resultado das métricas da árvore de decisão, com a base sem balanceamento 45
Tabela 6. Resultados da aplicação do qui-quadrado, para a base com balanceamento 47
Tabela 7. Resultados da aplicação do ganho de informação, para a base com balanceamento 48
Tabela 8. Resultado das métricas da árvore de decisão, com a base com balanceamento 49
Tabela 9. Ranque de variáveis para Etapas 1 e 2 50
Tabela 10. Comparação de métricas para a árvore de decisão com e sem SMOTE 51

LISTA DE FIGURAS

Figura 1. Série histórica de acidentes de trabalho no Brasil 13
Figura 2. Fluxograma de emissão e registro da CAT 18
Figura 3. Representação do valor-P 26
Figura 4. Validação cruzada com 10 iterações 27
Figura 5. Árvore de decisão genérica 29
Figura 6. Matriz de confusão genérica 32
Figura 7. Representação de uma curva ROC 33
Figura 8. Quantidade de acidentes de trabalho, por mês, no ano de 2019 38
Figura 9. Distribuição de idade dos acidentados 38
Figura 10. Distribuição das classes da variável alvo 39
Figura 11. Fluxograma de métodos experimentais 42
Figura 12. Curva ROC gerada através do Weka, com a base desbalanceada 46
Figura 13. Curva ROC gerada através do Weka, com a base balanceada 49

LISTA DE SIGLAS

AUC - *Area Under the Curve*

C4.5 – Classificador 4.5

CAT – Comunicação de Acidente de Trabalho

CBO – Classificação Brasileira de Ocupações

CID – Código Internacional de Doenças

CLT – Consolidação das Leis do Trabalho

CNAE – Classificação Nacional da Atividade Econômica

GL – Graus de Liberdade

IBGE – Instituto Brasileiro de Geografia e Estatística

ID3 – *Iterative Dichotomiser* 3

IG – *Information Gain*

IGR – *Information Gain Ratio*

INSS – Instituto Nacional do Seguro Social

ML – *Machine Learning*

PIB – Produto Interno Bruto

ROC – *Receiver Operator Characteristic Curve*

SMOTE – *Synthetic Minority Oversampling Technique*

Sumário

1. INTRODUÇÃO 12

2. FUNDAMENTAÇÃO TEÓRICA 15

2.1 ACIDENTES DE TRABALHO 15

2.2 COMUNICAÇÃO DE ACIDENTE DE TRABALHO 17

2.3 ANÁLISE DE TRABALHOS CORRELATOS 19

2.4 APRENDIZADO DE MÁQUINAS 20

2.4.1 Pré-processamento de dados 23

2.4.2 Ganho de informação 23

2.4.3 Teste do qui-quadrado 24

2.4.4 Validação Cruzada 26

2.5 ÁRVORES DE DECISÃO 28

2.6 SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE 30

2.7 MÉTRICAS DE AVALIAÇÃO 31

3. METODOLOGIA 34

3.1 CLASSIFICAÇÃO DA METODOLOGIA 34

3.2 BASE DE DADOS 35

3.2.1 Agente causador 35

3.2.2 Classificação Brasileira de Ocupações (CBO) 35

3.2.3 Código Internacional de Doenças (CID) 36

3.2.4 Classificação Nacional de Atividades Econômicas (CNAE) 36

3.2.5 Emitente de CAT 36

3.2.6 Filiação do segurado 37

3.2.7 Natureza da lesão 37

3.2.8 Sexo 37

3.2.9 Tipo de acidente 37

3.2.10 Estado 37

3.2.11 Mês do acidente 38

3.2.12 Idade 38

3.2.13 Gravidade 39

3.3 MÉTODOS E TÉCNICAS 39

3.3.1 Seleção de variáveis 40

3.3.2 Treinamento da árvore de decisão com validação cruzada 40

3.3.3 Balanceamento da base de dados 41

3.3.4 Fluxograma de métodos experimentais 41

4. RESULTADOS E ANÁLISES 43

4.1 ETAPA 1: EXPERIMENTO SEM BALANCEAMENTO 43
4.1.1 Resultado da seleção de variáveis 43
4.1.2 Resultado do treinamento da árvore de decisão 45
4.2 ETAPA 2: EXPERIMENTO COM BALANCEAMENTO 46
4.2.1 Resultado da seleção de variáveis 46
4.2.2 Resultado do treinamento da árvore de decisão 48
4.3 DISCUSSÕES 49
5. CONCLUSÕES 52
REFERÊNCIAS BIBLIOGRÁFICAS 54

1. INTRODUÇÃO

Nos últimos anos, tem crescido o número de pesquisas sobre prevenção de acidentes de trabalho, tendo como objetivo fortalecer a gestão da segurança e a cultura de segurança nas empresas. Dentre os vários fatores necessários à prevenção de acidentes, destaca-se a necessidade de um amplo esforço voltado para a mudança de comportamento e atitudes durante as atividades laborais (LUND & AARØ, 2004).

O que é característico e comum para todos os acidentes é que eles ocorrem de repente, são inesperados ou não, e causam danos imediatamente. No entanto, as causas diretas podem ser semelhantes e geralmente estão relacionadas com o tipo de perigo e evento acidental, enquanto as causas raízes estão relacionadas a uma série de condições, como gestão, organização, planejamento, treinamento e competências (JØRGENSEN, 2015).

Depois que um acidente acontece, pode ser simples perceber o que deveria ter sido feito para preveni-lo. As situações em que ocorrem acidentes são semelhantes e se repetem em outras ocorrências.

Apesar da implementação de diversos sistemas de gestão em Saúde e Segurança do Trabalho, a segurança ocupacional é baixa. Essa gestão é complexa devido à grande quantidade de entidades envolvidas no processo produtivo (AYHAN; TOKDEMIR, 2020).

Além disso, a maior parte do trabalho é realizada por humanos, portanto, técnicas para classificar acidentes de trabalho por meio de correlações simples são limitadas. Esse cenário está mudando com a recente evolução dos estudos de predição de acidentes, aplicando técnica de análise de dados e aprendizado de máquinas (ML – *Machine Learning*).

Na Figura 1 é apresentada a série histórica de acidentes de trabalho no Brasil. Os dados expostos permitem identificar a evolução quantitativa dos registros em seus números absolutos.

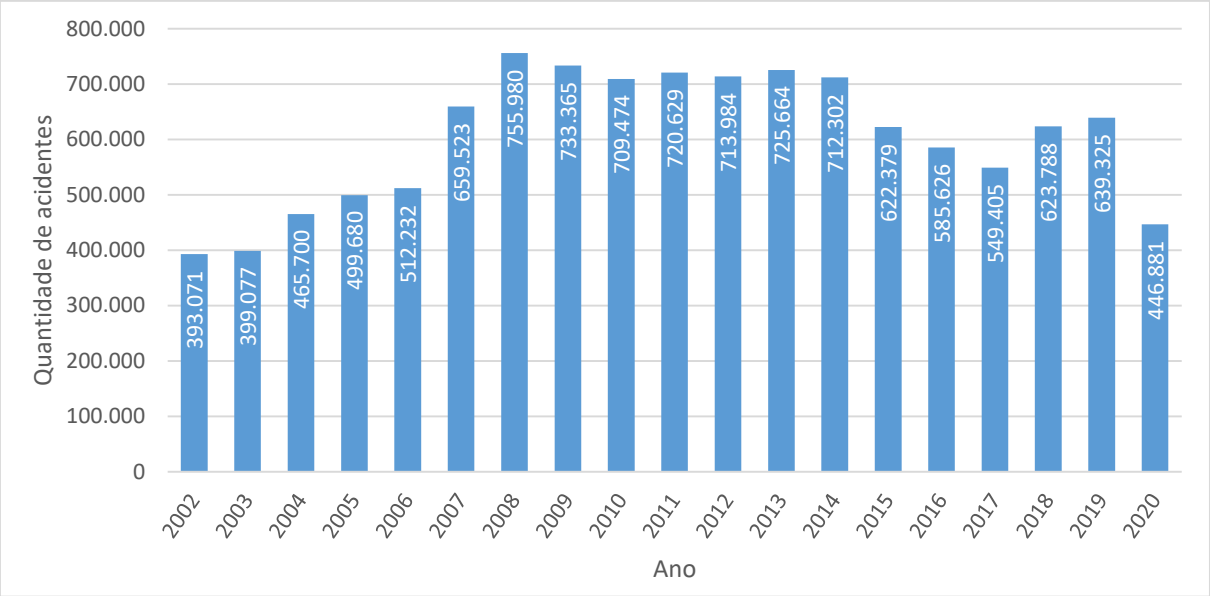

Figura 1. Série histórica de acidentes de trabalho no Brasil (Fonte: Brasil, 2017)

Cerca de 639 mil acidentes aconteceram no Brasil no ano de 2019. Segundo a Organização Internacional do Trabalho (2018), 4% do produto interno bruto (PIB) são perdidos com os efeitos desses acidentes.

Mesmo com métodos estatísticos tradicionais, análises avançadas podem ser realizadas usando grandes conjuntos de dados, para identificar padrões ocultos nos dados e obter informações valiosas (CHEN; LUO, 2016).

Entretanto, estudos têm mostrado que, ao comparar modelos estatísticos tradicionais com métodos de ML, o último é superior na predição de eventos futuro (SARKAR *et al.*, 2019).

Os benefícios potenciais do ML não só podem ser percebidos através da capacidade de processar grandes quantidade de dados, mas também do(a):

- capacidade de lidar com grandes problemas dimensionais,
- flexibilidade em reproduzir a estrutura de geração de dados, independentemente da complexidade e
- poder preditivo e interpretativo através da extração das regras.

O presente trabalho tem como objetivo analisar as variáveis mais importantes, presentes no banco de dados da Comunicação de Acidente de Trabalho (CAT), para

a classificação da variável alvo gravidade, fornecendo assim, insumos para análises de acidentes de trabalho. Como objetivos específicos, tem-se:

- Extrair e tratar dados contidos na base de dados do Instituto Nacional do Seguro Social (INSS), relativos a acidentes de trabalho, em empresas do território brasileiro, legalizadas, com registro no INSS, de todos os ramos econômicos;
- Comparar diferentes técnicas de seleção de variáveis;
- Aplicar a técnica *Synthetic Minority Oversampling Technique* (SMOTE) para o balanceamento da variável alvo, na base de dados;
- Analisar as variáveis que foram selecionadas pela árvore de decisão com e sem SMOTE e confrontá-las com as mais importantes indicadas pelas técnicas de seleção de variáveis.

A presente monografia está estruturada da seguinte forma:

- Introdução: Apresenta o tema, contexto, problema, justificativas e objetivos do trabalho;
- Fundamentação teórica: Introduz os conceitos necessários para a familiarização com o tema abordado e explora conceitos, críticas e hipóteses de outros autores;
- Metodologia: Indica a classificação do tipo de pesquisa e as ferramentas utilizadas para alcançar os objetivos.
- Resultados e análises: Explora os resultados obtidos desenvolvendo análises pertinentes ao objetivo da monografia;
- Conclusões: Discorre sobre as principais inferências, após a análise de resultados.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 ACIDENTES DE TRABALHO

Um acidente de trabalho pode ser definido como uma ocorrência imprevista e indesejável, instantânea ou não, no decorrer da atividade laboral podendo, eventualmente, resultar em lesão pessoal (ABNT, 2001).

Para Jørgensen (2015), o conceito de acidente pode ser definido como o resultado ocorrido de uma série de eventos que levam a um evento inesperado no qual uma pessoa é ferida pela exposição a um perigo.

Lesão e acidente são termos intimamente relacionados. Frequentemente, os termos são usados como sinônimo (HALE & HALE, 1970), porém, eles não são sinônimos.

Acidentes não precisam necessariamente resultar em lesões, mas todas as lesões são resultado de um incidente que pode ser denominado como acidente. De acordo com a definição, transferência de energia para o corpo do homem em excesso do limiar fisiológico é um pré-requisito necessário para que exista a lesão (ABNT, 2001).

Acidentes são resultado de sequências de eventos ocorrendo em sistemas de trabalho. A presença do perigo é a condição primária para a ocorrência de um acidente.

Os acidentes de trabalho podem ser divididos em três tipos (INSS, 1998):

- Acidente típico: ocorre na execução do trabalho;
- Acidente de trajeto: ocorre no percurso da residência para o trabalho ou vice-versa;
- Acidente devido a doença do trabalho: desencadeada ou produzida pelo exercício do trabalho ou em função de condições especiais em que o trabalho é realizado e com ele se relacione diretamente.

Uma pessoa que trabalha nas proximidades de um perigo está exposta ao risco de acidente de trabalho. Fatores causais são responsáveis pela transformação do perigo em um acidente. Existem muitas teorias disponíveis na literatura que explicam a causa de acidentes.

Khanzode, *et al.* (2012) explica as várias teorias por trás dos acidentes em sua pesquisa, como a teoria de propensão a acidentes (KUNCE, 1967), teoria do dominó (HEINRICH *et al*., 1980), epidemiologia de lesões (HADDON *et al*., 1964), teoria de sistemas (HALE; HALE, 1970), teoria de sistemas sociotécnicos (ROBINSON, 1982) e teoria da macro ergonomia (HENDRICK, 1986). Um resumo dessas teorias pode ser visto na Tabela 1.

Tabela 1. Gerações das teorias de acidente

Geração	Teoria	Características importantes
1ª geração	Teoria da propensão	Traços de personalidade são responsáveis pelo acidente, Intervenções comportamentais
2ª geração	Teorias do dominó	Ato e condição inseguros são predecessores imediatos do acidente, intervenções focadas em atos inseguros
3ª geração	Teoria epidemiológica	Transferência descontrolada de energia, controle nos estágios pré-lesão, lesão e pós-lesão
4ª geração	Teoria de sistemas	Abordagem holística, segurança integrada sistemas
	Teoria de sistema sociotécnico	Interagindo com subsistemas sociais e técnicos
	Teoria macro ergonômica	Abordagem holística, como modelos de sistema, abordagem centrada na organização

Fonte: KHANDOZE, et al. (2012)

Um acidente de trabalho ocorre devido à presença de uma cadeia de eventos ou fatores causais. Se as causas forem conhecidas, os resultados (ou seja, acidentes) podem ser prevenidos.

A experiência profissional desempenha um papel importante na identificação fatores causais, ela é necessária para identificar perigos em um determinado sistema de trabalho, porém, a análise humana é limitada, pois o ambiente laboral possui muitas variáveis.

Diferentes técnicas de modelagem são propostas por pesquisadores para avaliar o risco de acidentes/lesões. Primeiro, os dados categóricos - dados que possuem categorias naturais ordenadas e as distâncias entre as categorias não são conhecidas - são usados para avaliar consequências de acidentes/lesões (KEJRIWAL, 2002).

Rao-Tummala & Leung (1996) propõem uma matriz de prioridade bidirecional para avaliação de risco com base na probabilidade de ocorrência e gravidade das consequências. Índices de danos é uma ferramenta útil para comparar o risco ocupacional (SOLOMON & ALESCH, 1989).

O risco de acidente/lesão é modelado por meio de distribuições estatísticas, por ajuste de distribuição de probabilidade de ocorrência ou consequências (CHUNG *et al.* 1986; FREIVALDS & JOHNSON, 1990; MALLICK & MUKHERJEE, 1996). Modelo baseado em distribuição beta com dias de trabalho perdidos (modelo de consequência) como indicador de risco de lesão é um dos métodos mais comuns (COLEMAN & KERKERING, 2007).

2.2 COMUNICAÇÃO DE ACIDENTE DE TRABALHO

A CAT é um documento direcionado ao INSS, com informações sobre um acidente de trabalho, de trajeto ou doença ocupacional. A Figura 2 apresenta um fluxograma com o roteiro de emissão e registro da CAT.

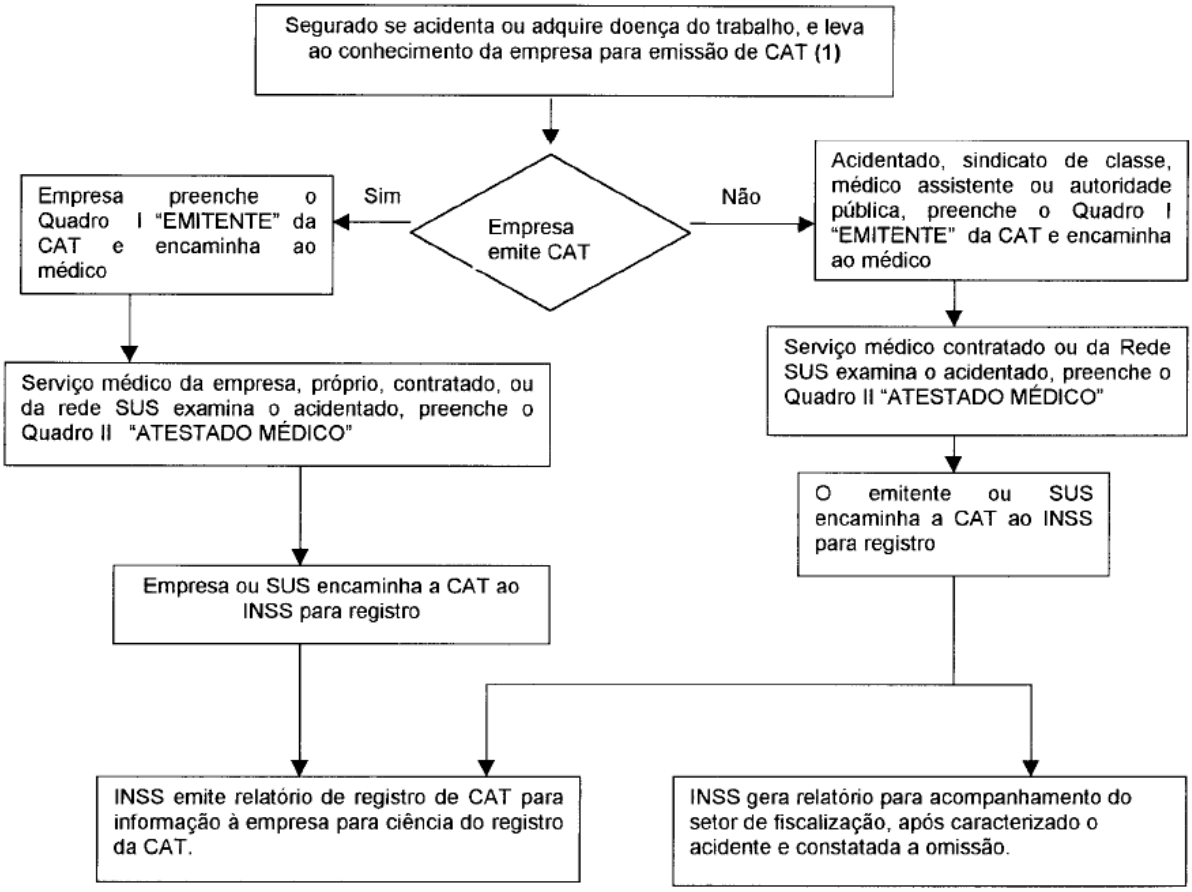

Figura 2. Fluxograma de emissão e registro da CAT (Fonte: INSS, 1999)

Após ocorrido o acidente ou doença do trabalho, sem necessariamente o empregado ter sido afastado da atividade laboral, a empresa tem até um dia útil para relatar o evento à Previdência Social, através do preenchimento da CAT. Caso isso não seja feito, será caracterizado como crime (BRASIL, 1943).

Todos os anos, milhares de acidentes de trabalho acontecem no Brasil, e o que permite a contabilização destes dados é a CAT. As informações contidas nesse banco de dados, serão apresentadas e descritas na subseção 3.2

A Tabela 2 apresenta outros bancos de dados que abordam essa temática, expondo suas características, fonte e o tipo de informação encontrada.

Tabela 2. Bases de dados relacionadas à Saúde do Trabalhador

Tipo de dado	Fonte / Base	Informação	Abrangência
População trabalhadora	IBGE	População geral PEA – Pop. Economicamente Ativa PEA Ocupada	Censo populacional decenal Pesquisas amostrais anuais
	Dataprev Suibe	Segurados da Previdência Social: RGPS e Segurados Especiais Auxílio Acidente de Trabalho	Censo populacional decenal Pesquisas amostrais anuais
	Rais Caged	Nº de empregos celetistas: segundo gênero, idade, raça/cor, escolaridade, remuneração, setor de atividade econômica e tamanho da empresa	Informadas pelas empresas ao MTE: • Rais anualmente • Caged mensalmente
Atividades produtivas, estabelecimentos	IBGE	Censos e pesquisas econômicas específicas: Censo Agropecuário etc.	População em geral, inclusive todos os trabalhadores, independente de tipo de vínculo empregatício
	Rais	Estabelecimentos: porte (nº trabalhadores), atividade econômica	Trabalhadores empregados de empresas públicas e privadas, vínculos CLT
Mortalidade	Datasus: SIM Declaração Óbito	Causas de óbito por grupos de causa, sexo, idade, ocupação, escolaridade, raça/cor etc. Causas externas de óbito, incluindo acidentes de trabalho	População em geral, inclusive todos os trabalhadores, independente de tipo de vínculo empregatício
	Dataprev: Suibe, CAT com óbito	Óbitos registrados em CAT Pensões por mortes acidentárias concedidas aos familiares	Inclui casos de acidentes de trabalho e doenças do trabalho com óbito de trabalhadores segurados
Morbidade	Dataprev: CAT, NTEP, Suibe	Benefícios concedidos em casos de acidentes de trabalho, típico e de trajeto, e doenças do trabalho registrados por meio de CAT ou de NTEP	Trabalhadores empregados de empresas, públicas e privadas, vínculos CLT, segurados pelo Seguro Acidente de Trabalho; Segurados Especiais
	Datasus: Sinan	Agravos, acidentes e doenças de notificação compulsória, incluindo os relacionados ao trabalho	População geral e todos os trabalhadores, independente de tipo de vínculo empregatício
	Datasus: SIH	Internações hospitalares por grupos de causas, incluindo causas externas	População geral e todos os trabalhadores, independente de tipo de vínculo empregatício
	Datasus: SIA	Procedimentos assistenciais e de vigilância em saúde registrados pelos serviços de saúde cadastrados no CNES	População geral e todos os trabalhadores, independente de tipo de vínculo empregatício

Fonte: BRASIL (2014)

É importante reconhecer a importância dos bancos de dados relacionados à saúde do trabalhador, porém, cabe destacar que os dados, em sua maioria, são inconsistentes ou estão desatualizados.

2.3 ANÁLISE DE TRABALHOS CORRELATOS

A análise tradicional de acidentes de trabalho é feita através de estatística descritiva, entretanto, por haver muitas variáveis, não é uma análise precisa. Nos últimos anos, tem crescido o número de estudos envolvendo ML e acidentes de trabalho.

Em SARKAR et al. (2019) são utilizados os métodos de análise de classes latentes, métodos de vetor de suporte, redes neurais e árvores de decisão para analisar o resultado do acidente (lesão, quase lesão e dano material), tendo 16 variáveis de entrada. Os autores destacam a importância do pré-processamento de dados, para reduzir variáveis através do método do qui-quadrado.

A pesquisa de BARTOLOMEU (2002) utiliza os softwares Access, para gerenciamento do banco de dados, e o See5 (possui método semelhante a árvore de decisão). São analisados os resultados para duas variáveis alvo diferentes, sexo e situação geradora do acidente, para essa tarefa, são utilizados os dados da CAT do estado de Santa Catarina, do ano 2000.

Já em LEE et al. (2020) são utilizados os algoritmos de máquina de vetor de suporte, árvores de decisão, análise de componentes principais e *ensemble.* A partir de 15 variáveis de entrada, a gravidade (leve, grave e fatal) é classificada. Cabe destacar que, a variável com maior correlação com a variável alvo foi retirada da análise, pois, segundo os autores, a classificação é determinada com base no tipo de lesão, e a presença dela iria diminuir muito a influência dos outros recursos.

Kubat e Matwin (1997) subamostraram seletivamente a classe majoritária, mantendo a população original da classe minoritária. Eles usaram a média geométrica como medida de desempenho do classificador, que pode ser relacionada a um único ponto na curva Curva Característica de Operação do Recepto (ROC - *Receiver Operator Characteristic Curve*) e observaram uma melhora com as classes balanceadas.

Ling et al. (1998) combinaram a superamostragem da classe minoritária com a subamostragem da classe majoritária. Os autores utilizaram o índice de sustentação (métrica semelhante a uma curva ROC). Em um experimento, eles superamostraram a classe minoritária e notaram que o melhor índice de sustentação é obtido quando as classes são igualmente representadas.

Diante dos trabalhos apresentados nessa seção, é possível perceber que: há uma necessidade em utilizar métodos de seleção de atributos; os algoritmos baseados em árvores de decisão são comuns em análise de acidentes de trabalho; e bases desbalanceadas podem gerar prejuízos para a precisão de algoritmos de aprendizado de máquinas.

2.4 APRENDIZADO DE MÁQUINAS

O aprendizado de máquinas pode ser entendido como a capacidade de melhorar a realização de determinada tarefa em função da experiência (MITCHELL,

1997). Portanto, um algoritmo é capaz de aprimorar a sua função através da aquisição automática de conhecimento a partir de novos dados.

Todos os algoritmos e métodos de aprendizagem usados na mineração de dados são baseados em vários modelos de teste lógico que possuem uma base estatística sólida (MEASE; WYNER, 2008). Existem três tipos de técnicas de mineração de dados: aprendizagem não supervisionada, semisupervisionada e supervisionada.

Na aprendizagem supervisionada, o algoritmo trabalha com um conjunto de exemplos cujos rótulos são conhecidos. Os rótulos podem ser valores nominais, no caso da tarefa de classificação, ou valores numéricos, no caso da tarefa de regressão. Já na aprendizagem não supervisionada, os rótulos dos exemplos no conjunto de dados são desconhecidos, e o algoritmo normalmente visa agrupar os exemplos de acordo com a similaridade de seus valores de atributos, caracterizando uma tarefa de agrupamento. Finalmente, a aprendizagem semisupervisionada é geralmente usada quando um pequeno subconjunto de exemplos rotulados está disponível, junto com um grande número de exemplos não rotulados (NEELAMEGAM; RAMARAJ, 2013).

Os algoritmos mais comuns de classificação baseado em aprendizado de máquinas incluem árvores de decisão, redes neurais artificiais, máquinas de aprendizado extremo, K vizinhos mais próximos, redes bayesianas e máquinas de vetores de suporte (WITTEN; FRANK; HALL, 2011).

Árvores de decisão ganharam popularidade como um algoritmo de classificação poderoso que é transparente e facilmente interpretável (OLSON; DELEN; MENG, 2012). São usadas principalmente por sua capacidade de analisar padrões quantitativos e qualitativos de dados para encontrar informações ocultas (SARKAR *et al.*, 2019).

Com essas vantagens, as árvores de decisão foram aplicadas com sucesso em uma variedade de campos de pesquisa, incluindo medicina (OZTEKIN *et al.*, 2018), ciências sociais (OLSON; DELEN; MENG, 2012), gestão de negócios (AVIAD; GELBARD, 2011), engenharia e gestão de construção (LEU; CHANG, 2013), e indústria de processo (BEVILACQUA, M.; CIARAPICA, F.E.; GIACCHETTA, 2008).

Redes neurais artificiais geralmente fornecem melhores resultados do que as técnicas convencionais de classificação simples. No entanto, essa abordagem é uma caixa preta tecnológica, sendo difícil para os humanos interpretarem. Portanto, é um processo complexo identificar as correlações entre variáveis nos dados de acidentes (AN *et al.*, 2007).

As máquinas de vetor de suporte atraíram atenção considerável devido à sua capacidade para autoaprendizagem e sua alta capacidade de generalização (CHENG *et al.*, 2010). No entanto, esse algoritmo usa um longo processo de tentativa e erro (AN *et al.*, 2007), têm um alto nível de complexidade algorítmica e requer memória extensa (RAVI KUMAR; RAVI, 2007).

Muitos pesquisadores realizaram estudos de análise em vários campos usando vários dados de acidentes; no entanto, várias limitações foram encontradas na análise dos dados de acidentes de trabalho. Experimentalmente, tem sido mostrado que modelos combinados apresentam melhor desempenho do que um sistema decisório único.

Em primeiro lugar, as opiniões individuais e subjetivas de quem elabora o relatório de acidente de trabalho estão refletidas nos dados; portanto, é difícil processar e refletir as características dos dados de acidentes de trabalho, que são criados sem um procedimento de composição e incluem muitos tipos de variáveis e valores (AYHAN; TOKDEMIR, 2020).

Em segundo lugar, a estrutura de dados de acidentes de trabalho inclui variáveis mistas (por exemplo, representações de texto numéricas e categóricas) e informações ausentes.

Esses numerosos tipos de variáveis, junto com a composição de muitas categorias, criam dificuldades e ambiguidades na interpretação dos resultados com elementos de dados; como resultado, apenas correlações limitadas podem ser derivadas entre as variáveis (SARKAR *et al.*, 2019).

A seguinte conclusão pode ser tirada dos resultados das pesquisas existentes: vários tipos de variáveis e valores estão presentes nos dados de acidentes de trabalho, o que torna difícil processar dados, refletir características e interpretar

correlações. No entanto, se as variáveis são excessivamente reduzidas, suas características são perdidas e não podem ser tiradas conclusões significativas. Portanto, os tipos e intervalos de valores para as variáveis devem ser padronizados.

2.4.1 Pré-processamento de dados

O pré-processamento de dados é uma tarefa essencial na mineração de dados e consome, em média, mais de 60% do esforço total de todo o processo (HOUARI *et al.*, 2016). Em particular, porque acidentes de trabalho incluem inúmeras variáveis e tipos de valores, portanto, o conjunto de dados deve ser pré-processado ou padronizado antes da análise.

Caso contrário, a presença de *outliers*, omissões e inconsistências de termos nos dados dificultariam a interpretação dos resultados da análise, o que tornaria as tendências incompreensíveis e poderiam, assim, produzir resultados de análise enganosos.

Além disso, reduzir variáveis e elementos facilita a obtenção de interpretações significativas. O pré-processamento de dados deve ser executado comparando vários métodos em vez de um.

Os conjuntos de dados para análise podem conter muitos atributos, dos quais, alguns podem ser irrelevantes para a tarefa de mineração ou redundantes. Embora possa ser possível para um especialista selecionar alguns dos atributos úteis, isso pode ser uma tarefa difícil e demorada, especialmente quando o comportamento dos dados não é bem conhecido.

Manter atributos irrelevantes pode ser prejudicial, causando confusão para o algoritmo de mineração utilizado. Assim, a diminuição da dimensionalidade reduz o tamanho dos dados removendo tais atributos deles (YANG; PEDERSEN, 1997).

2.4.2 Ganho de informação

Uma abordagem comumente utilizada para selecionar atributos considera cada atributo individualmente, ordenando-os de acordo com suas capacidades preditivas e selecionando os melhores atributos para compor o subconjunto que será utilizado pelo

algoritmo de mineração de dados, a técnica do ganho de informação (IG - *Information Gain*) é um exemplo desse tipo de abordagem (CSISZÁR; SHIELDS, 2004).

IG é frequentemente utilizada em aplicações onde a dimensionalidade dos dados proíbe o uso de técnicas mais sofisticadas. Seja *A* um atributo de um banco de dados e *C* seu conjunto de classes. O cálculo da entropia do atributo de classe antes e depois da observação do atributo *A* é dado pelas Equações (1) e (2), respectivamente.

$$\mathrm{E(C)} = -\sum_{c \in C} p(c) \log_2 p(c) \tag{1}$$

Onde p(c) é a probabilidade da classe *c* ocorrer na base de dados

$$\mathrm{E(C \mid A)} = -\sum_{a \in A} p(a) \sum_{c \in C} p(c \mid a) \log_2 p(c \mid a) \tag{2}$$

Onde p(a) é a probabilidade do valor a ocorrer na base de dados e p(c|a) a probabilidade da classe c ocorrer, dado que o valor de atributo a ocorreu.

A redução causada na entropia de *C* devido à informação adicional fornecida pelo atributo *A* é denominada Ganho de Informação. Na técnica *Information Gain Attribute Ranking*, presente no Weka, cada atributo A_i da base de dados é associado a um valor correspondente ao IG, calculado da seguinte maneira:

$$IG_i = E(C) - E(C \mid A_i) \tag{3}$$

Desse modo, os atributos associados aos maiores ganhos de informação serão selecionados pelo algoritmo, durante o processo de treinamento do modelo.

2.4.3 Teste do qui-quadrado

O teste do qui-quadrado analisa os dados categóricos por meio da distribuição do qui-quadrado para verificar a significância das frequências observadas e esperadas. É usado principalmente para verificar a qualidade do ajuste, homogeneidade e independência dos dados (SARKAR *et al.*, 2019). Há cinco etapas para realizar este teste:

- Etapa 1: Formulação de hipóteses.

Hipótese nula (H0): Não há correlação significativa entre duas variáveis.

Hipótese alternativa (Ha): Existe uma correlação significativa entre duas variáveis.

- Etapa 2: Especificar os valores esperados para cada célula da tabela (quando a hipótese nula for verdadeira).

Os valores esperados especificam quais seriam os valores de cada célula da tabela se não houvesse associação entre as duas variáveis. A fórmula para calcular os valores esperados requer o tamanho da amostra, os totais das linhas e os totais das colunas, sendo representada na equação (4).

$$\text{Contagem esperada} = \frac{\text{Total da Linha} * \text{Total da Coluna}}{\text{Total da Tabela}} \tag{4}$$

- Etapa 3: Para ver se os dados fornecem evidências convincentes contra a hipótese nula, deve-se comparar as contagens observadas da amostra com as contagens esperadas, supondo que H0 seja verdadeira.

Os valores observados são as contagens reais calculadas a partir da amostra.

- Etapa 4: Calcular o qui-quadrado.

A estatística qui-quadrado compara os valores observados com os valores esperados. Esse teste é usado para determinar se a diferença entre os valores observados e esperados é estatisticamente significativa. A fórmula está apresentada na equação (5):

$$X^2 = \sum \frac{(ValorObservado - ValorEsperado)^2}{ValorEsperado} \tag{5}$$

- Etapa 5: Analisar se o qui-quadrado é estatisticamente significativo.

A etapa final do teste qui-quadrado de significância é determinar se o valor da estatística qui-quadrado é grande o suficiente para rejeitar a hipótese nula. As distribuições do qui-quadrado assumem apenas valores positivos e são assimétricas à direita, sua função de densidade é dada pela equação (6):

$$f(x,v) = \frac{1}{\tau\left(\frac{v}{2}\right) * 2^{\frac{v}{2}}} * x^{\frac{v}{2-1}} * e^{\frac{-x}{2}} \quad (6)$$

Sendo v a quantidade de graus de liberdade, determinados pela equação (7):

$$v = (L - 1) * (C - 1) \quad (7)$$

L representa a quantidade de linhas e C a quantidade de colunas da tabela.

Por convenção, ao utilizar essa estatística de teste para estimar a probabilidade de que a hipótese esteja errada, rejeita-se a hipótese se essa probabilidade for 95% ou maior. Em outras palavras, rejeita-se a hipótese se houver 5% (valor-P = 0,05) ou menos de probabilidade de haver um erro ao utilizá-la (MOORE; NOTZ; FLINGER, 2013). O valor-P é a área sob a curva de densidade desta distribuição qui-quadrado à direita do valor da estatística de teste, como pode ser visto na Figura 3.

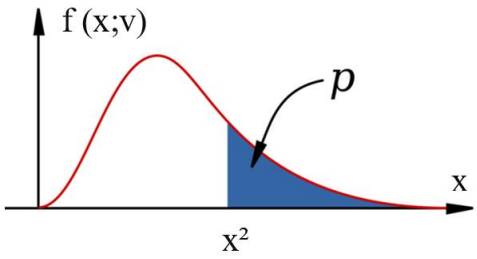

Figura 3. Representação do valor-P

Quanto maior for x^2, maior será a rejeição a hipótese nula (H0), ou seja, as duas variáveis possuem correlação significativa. A distribuição do qui-quadrado possui diversas aplicações importantes, sendo que está tabulada para diferentes valores de graus de liberdade (MEYER, 1983).

2.4.4 Validação Cruzada

Quando a quantidade de dados é grande, a validação cruzada k-fold deve ser empregada para estimar a precisão do modelo induzido a partir de um algoritmo de classificação, pois a precisão resultante dos dados de treinamento do modelo geralmente possui viés (WITTEN; FRANK; HALL, 2011).

Nesse método, os dados são divididos em subconjuntos. Um dos subconjuntos é usado como conjunto de teste e os outros subconjuntos são colocados juntos para

formar um conjunto de treinamento. A estimativa de erro é calculada como a média de todas as tentativas para obter a eficácia total do modelo. Uma representação da validação cruzada pode ser vista na Figura 4.

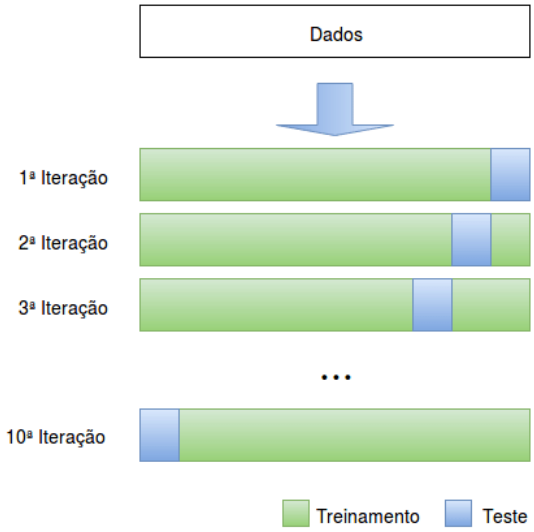

Figura 4. Validação cruzada com 10 iterações (Fonte: LIMA; BATISTA, 2018)

A validação cruzada reduz significativamente o viés, pois utiliza a maioria dos dados para treino e reduz significativamente a variância, pois a maioria dos dados também está sendo usada no conjunto de teste.

Um k grande é aparentemente desejável, pois, com um k maior há mais estimativas de desempenho, e o tamanho do conjunto de treinamento está mais próximo do tamanho total dos dados, aumentando assim a possibilidade de que qualquer conclusão feita sobre o aprendizado dos algoritmos em teste seja generalizada para o caso em que todos os dados são usados para treinar o modelo de aprendizado.

No entanto, à medida que k aumenta, a sobreposição entre os conjuntos de treinamento também aumenta. Por exemplo, com validação cruzada de 5 vezes, cada conjunto de treinamento compartilha apenas 3/4 de suas instâncias com cada um dos outros quatro conjuntos de treinamento, enquanto que com validação cruzada de 10 vezes, cada conjunto de treinamento compartilha 8/9 de suas instâncias com cada um dos outros nove conjuntos de treinamento.

Além disso, aumentar k reduz o tamanho do conjunto de teste, levando a medições menos precisas e menos refinadas da métrica de desempenho. Por

exemplo, com um tamanho de conjunto de teste de 10 instâncias, só é possível medir a precisão com 10% de certeza, enquanto com 20 instâncias a precisão pode ser medida com 5% de certeza.

Todos esses fatores concorrentes foram considerados e o consenso geral na comunidade de mineração de dados é que k igual a 10 é um valor atraente, pois faz previsões usando 90% dos dados, tornando mais provável que seja generalizável para os dados completos (REFAEILZADEH; TANG; LIU, 2009).

2.5 ÁRVORES DE DECISÃO

Um classificador de árvore de decisão consiste em nós que formam uma árvore enraizada, o que significa que é uma árvore direcionada, com um nó denominado “raiz” que não possui arestas de entrada. Todos os outros nós têm exatamente uma aresta de entrada. Um nó com arestas de saída é chamado de nó interno ou de teste. Todos os outros nós são chamados de folhas (também conhecidos como nós terminais) (WITTEN; FRANK, 2005).

As árvores de decisão começaram a desempenhar um papel importante no aprendizado de máquina com a publicação do Quinlan's ID3 (*Iterative Dichotomiser 3*) (QUINLAN, 1986). Posteriormente, Quinlan também apresentou o algoritmo C4.5 (Classificador 4.5) (QUINLAN, 1993), que é uma versão avançada do ID3. Desde então, o C4.5 tem sido considerado um modelo padrão na classificação supervisionada (MANTAS; ABELLÁN, 2014).

Árvores de decisão são modelos baseados em um método de partição recursiva, cujo objetivo é dividir o conjunto de dados usando uma única variável em cada nível. Esta variável é selecionada com um determinado critério. Idealmente, eles definem um conjunto de casos em que todos os casos pertencem à mesma classe.

Sua representação tem uma estrutura de árvore simples, como ilustrada na Figura 5. Pode ser interpretada como um conjunto compacto de regras em que cada nó da árvore é rotulado com uma variável de atributo que produz ramos para cada valor. Os nós folha são rotulados com um rótulo de classe.

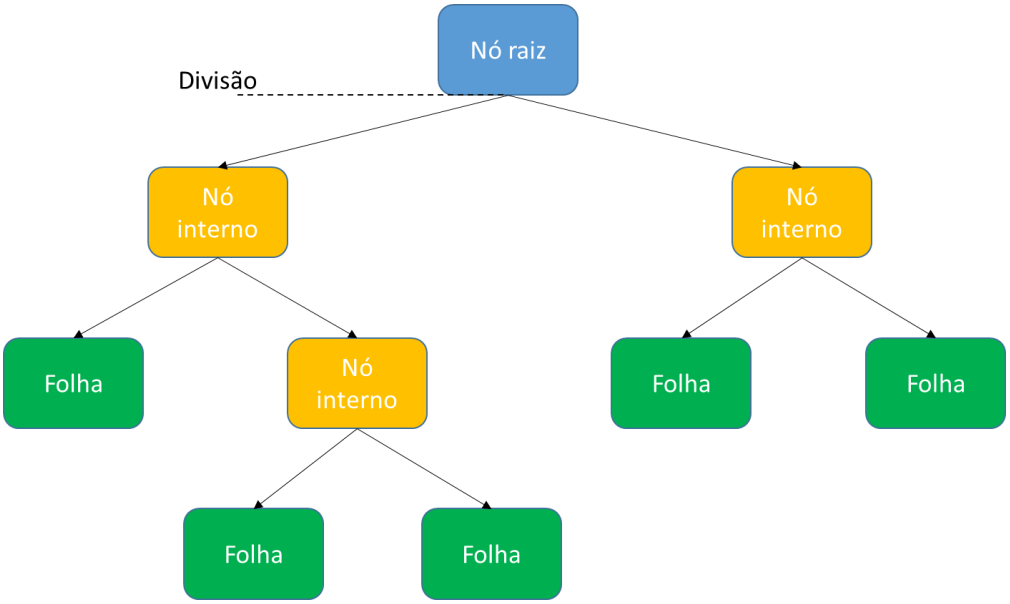

Figura 5. Árvore de decisão genérica (Fonte: De autoria própria)

O processo para construir uma árvore de decisão é determinado principalmente pelos seguintes aspectos:

- Os critérios usados para selecionar o atributo a ser inserido em um nó e ramificação (critérios de divisão);
- Os critérios para impedir a ramificação da árvore;
- O método para atribuir um rótulo de classe ou uma distribuição de probabilidade nos nós folha;
- O processo de pós-poda usado para simplificar a estrutura da árvore.

As árvores de decisão são construídas usando um conjunto de dados de treinamento. Um conjunto diferente, chamado de conjunto de dados de teste, é usado para verificar o modelo. Quando se obtém uma nova amostra ou instância do conjunto de dados de teste, pode-se tomar uma decisão ou previsão sobre o estado da variável de classe seguindo o caminho na árvore da raiz até um nó folha, usando os valores da amostra e a estrutura da árvore.

Para determinar a classe para uma nova instância usando uma árvore de decisão, começando pela raiz, nós internos sucessivos são gerados até que um nó folha seja alcançado. No nó raiz e em cada nó interno, um teste é aplicado. O resultado

do teste determina o ramo atravessado e o próximo nó gerado. A classe da instância é a classe do nó folha.

O critério de estimação no algoritmo de árvore de decisão é a seleção de um atributo a ser testado em cada nó de decisão na árvore. O objetivo é selecionar o atributo mais útil para classificar exemplos. Uma boa medida quantitativa do valor de um atributo é a entropia - equação (1) e (2) -, ela é usada para estimar a aleatoriedade da variável dependente. A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia, ou seja, a aleatoriedade - dificuldade de previsão - da variável que define as classes (HAMILTON *et al.*, 2012).

A métrica do Ganho de Informação, representada pela Equação (3), foi introduzida por Quinlan como base para seu modelo ID3. O modelo tem as seguintes principais características: foi definido para obter árvores de decisão com variáveis discretas, não trabalha com valores omissos, não possui processo de poda e é baseado na entropia.

Para melhorar o ID3, Quinlan desenvolveu o modelo C4.5 (QUINLAN, 1993), onde o critério de divisão (ganho de informação) é substituído por um critério de razão do ganho de informação (IGR – *Information Gain Ratio*) que penaliza variáveis com muitos valores. O IGR é definido pela equação (8), sendo a divisão do ganho de informação do atributo por sua entropia:

$$\mathrm{IGR} = \frac{\mathrm{IG}}{\mathrm{E}} \tag{8}$$

Por esses motivos, o C4.5 possui um procedimento mais completo, definido para trabalhar com variáveis contínuas e dados ausentes. Tem uma poda posterior robusta que é introduzida para melhorar os resultados e obter estruturas menos complexas (MANTAS; ABELLÁN, 2014).

2.6 SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

Utilizando a SMOTE a classe minoritária é superamostrada criando exemplos sintéticos. Essa abordagem é inspirada em uma técnica que provou ser bem-sucedida no reconhecimento de caracteres manuscritos (HA; BUNKE, 1997). Nesse estudo, foram criados dados extras de treinamento realizando certas operações em dados

reais, operações como rotação e inclinação eram formas naturais de perturbar os dados de treinamento.

A classe minoritária é superamostrada pegando cada amostra de classe minoritária e introduzindo exemplos sintéticos ao longo dos segmentos de linha que unem os k vizinhos mais próximos da classe minoritária. Dependendo da quantidade de sobreamostragem necessária, os vizinhos dos k vizinhos mais próximos são escolhidos aleatoriamente.

Amostras sintéticas são geradas da seguinte maneira: Pega-se a diferença entre o vetor de recursos (amostra) em consideração e seu vizinho mais próximo. Multiplica-se essa diferença por um número aleatório entre 0 e 1 e adiciona o vetor de recursos em consideração. Isso provoca a seleção de um ponto aleatório ao longo do segmento de linha entre dois recursos específicos. Essa abordagem força a região de decisão da classe minoritária a se tornar mais geral.

Os exemplos sintéticos fazem com que o classificador crie regiões de decisão maiores e menos específicas, em vez de regiões menores e mais específicas. Esse método permite que as árvores de decisão possam generalizar melhor (CHAWLA *et al.*, 2002).

2.7 MÉTRICAS DE AVALIAÇÃO

Uma maneira comum de avaliar os resultados de experimentos de com ML é através das métricas: precisão, acurácia, revocação (*recall*), F1-*Score* (F-Measure) e a área sob a curva ROC (AUC - Area Under the Curve). Para entender esses conceitos, primeiramente, é necessário entender a matriz de confusão.

A matriz de confusão é uma tabela que indica os erros e acertos do modelo, comparando com o resultado esperado, a Figura 6 apresenta um exemplo de matriz de confusão.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 6. Matriz de confusão genérica

- Verdadeiros Positivos: classificação correta da classe Positivo;
- Falsos Positivos (Erro Tipo I): erro em que o modelo previu a classe Positivo quando o valor real era classe Negativo;
- Falsos Negativos (Erro Tipo II): erro em que o modelo previu a classe Negativo quando o valor real era classe Positivo;
- Verdadeiros Negativos: classificação correta da classe Negativo.

A acurácia pode ser entendida como: Dentre todas as classificações, quantas o modelo classificou corretamente. Seu cálculo é apresentado na equação (9):

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{9}$$

A revocação (*recall*) é: a quantidade correta, dentre todas as situações de classe Positivo como valor esperado. Seu cálculo é apresentado na equação (10):

$$revocação = \frac{VP}{VP + FN} \tag{10}$$

A especificidade é: a capacidade do método de detectar resultados negativos. Seu cálculo é apresentado na equação (11):

$$especificidade = \frac{VN}{VN + FP} \tag{11}$$

A precisão é: uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos. Seu cálculo é apresentado na equação (12):

$$precisão = \frac{VP}{VP + FP} \tag{12}$$

O F1-Score é a média harmônica entre precisão e recall. Seu cálculo é apresentado na equação (13):

$$F1Score = 2 * \frac{precisão * revocação}{precisão + revocação} \tag{13}$$

A curva ROC, representada na Figura 7 é uma ferramenta estatística visual para resumir a relação entre sensibilidade e especificidade, uma abordagem comumente utilizada é calcular a métrica AUC (FAWCETT, 2006).

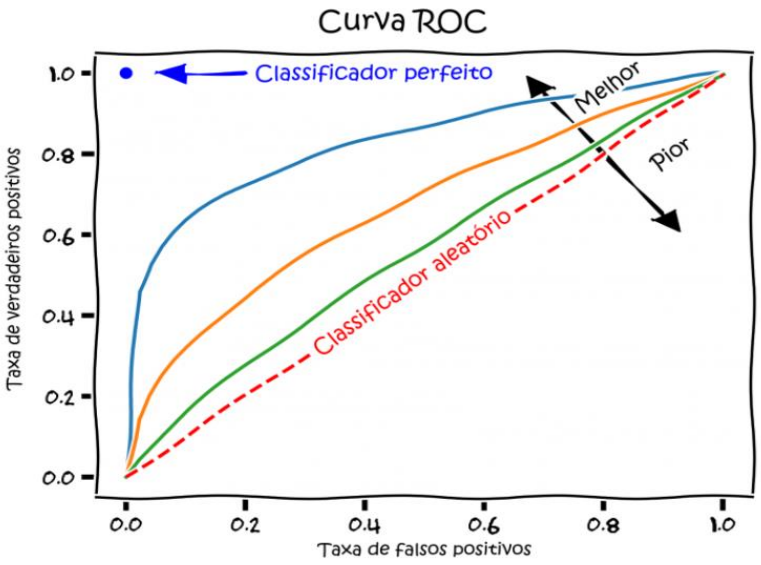

Figura 7. Representação de uma curva ROC (Fonte: THOMA, 2018)

A AUC é a área abaixo da curva ROC e varia de 0 a 1, sendo que 1 representa uma melhor capacidade em rotular as classes.

3. METODOLOGIA

A análise de acidentes por humanos tem a desvantagem de avaliar uma única variável dependente por uma única variável independente, e está inferindo apenas uma variável dependente com dados escritos qualitativa e subjetivamente. Neste contexto, faz-se necessário um método de classificação capaz de aprender com os acidentes anteriores para minimizar o risco dos futuros.

Para cumprir com os objetivos descritos, primeiramente, foi realizada uma pesquisa bibliográfica, para obter embasamento teórico e levantar hipóteses acerca do tema proposto.

A seguinte conclusão pode ser tirada dos resultados das pesquisas existentes: muitos tipos de variáveis e valores estão presentes nos dados de acidentes de trabalho, o que torna difícil processar dados, refletir características e interpretar correlações.

Serão utilizados dados abertos da CAT, registradas no ano de 2019, disponibilizados pelo INSS. Os dados são referentes às empresas no território brasileiro, legalizadas, com registro no INSS, de todos os ramos econômicos.

Esse capítulo descreve a metologia aplicada nesse trabalho, e está dividido da seguinte maneira: na seção 3.1 é apresentada a classificação metodológica desta pesquisa; na seção 3.2 é descrita a base de dados incluindo a descrição dos atributos, bem como, uma breve análise estatística dos dados; na seção 3.3 é apresentado o planejamento da execução dos testes experimentais.

3.1 CLASSIFICAÇÃO DA METODOLOGIA

Quanto à abordagem, a pesquisa é quantitativa, pois enfatiza o raciocínio dedutivo, as regras da lógica e os atributos mensuráveis da experiência humana, portanto, a realidade pode ser compreendida através da análise de dados, recorrendo à linguagem matemática e computacional para descrever as causas de acidentes de trabalho, utilizando correlações de variáveis.

Quanto à natureza, a pesquisa é classificada como aplicada, o seu objetivo é a solução de um problema específico, ou seja, atenuar a quantidade de acidentes de

trabalho e seus desdobramentos. Como consequência última, há a possibilidade de aplicação prática em empresas no território brasileiro, legalizadas, com registro no INSS, de todos os ramos econômicos

Quanto aos objetivos, a pesquisa é caracterizada como explicativa, ou seja, este tipo de pesquisa explica o porquê, determina os fatores que contribuem para a ocorrência de um AT.

Quanto aos procedimentos, a pesquisa é bibliográfica (é feita a partir de referências teóricas já analisadas e publicadas) e documental (utilização de bases de dados sem tratamento analítico).

3.2 BASE DE DADOS

3.2.1 Agente causador

Esse atributo refere-se ao agente causador diretamente relacionado ao acidente, ou a situação em que gerou o acidente.

Essa variável é do tipo nominal e possui 272 valores distintos, sendo os três mais frequentes, com suas respectivas proporções em relação a amostra: motocicleta ou motoneta (8%); metal - inclui liga ferrosa e não ferrosa (5%); chão - superfície utilizada para sustentar pessoas (5%).

3.2.2 Classificação Brasileira de Ocupações (CBO)

O Código Brasileiro de Ocupações (CBO) tem como objetivo a identificação das ocupações no mercado de trabalho, para fins classificatórios junto aos registros administrativos e domiciliares. Os efeitos de uniformização pretendida pela Classificação Brasileira de Ocupações são de ordem administrativa e não se estendem as relações de trabalho (BRASIL, 2002).

A estrutura do CBO é o conjunto de códigos e títulos que é utilizada na sua função enumerativa que contém 6 dígitos. A base de dados possui 2130 ocupações distintas, sendo as três mais frequentes, com suas respectivas proporções em relação a amostra: 322205 - técnico de enfermagem (6%); 784205 - alimentador de linha de produção (6%); 514320 – faxineiro (4%).

3.2.3 Código Internacional de Doenças (CID)

Os estudos para criação do CID começaram em 1983 e foram afirmados na 43ª Assembleia Mundial de Saúde. Os estados-membros utilizaram a classificação, pela primeira vez, em 1994.

O código Internacional de Doenças (CID) é dividido em 22 capítulos, que agrupam doenças com características semelhantes. A catalogação é feita representada por uma letra e números.

A base de dados possui 4877 CIDs distintos, sendo os três mais frequentes, com sua respectiva proporção da amostra total: S610 - ferimento de dedo(s) sem lesão da unha (6%); S934 - entorse e distensão do tornozelo (4%); S626 - Fratura de outros dedos (3%).

3.2.4 Classificação Nacional de Atividades Econômicas (CNAE)

Todas as atividades econômicas possuem seu código na CNAE, desde empresas públicas ou privadas e até mesmo atividades sem fins lucrativos ou de pessoas físicas em atividades autônomas, e estão relacionados com o objeto social da entidade (BRASIL, 2007).

O modelo de codificação adotado na CNAE é um código numérico, que está relacionado a uma atividade econômica. A base de dados possui 865 valores distintos para essa variável, sendo os três mais frequentes, com sua respectiva proporção da amostra total: 8610 – atividades de atendimento hospitalar (11%); 4711 - comércio varejista de mercadorias em geral, com predominância de produtos alimentícios (4%); transporte rodoviário de carga (3%).

3.2.5 Emitente de CAT

Esse atributo possui cinco valores distinto, ele especifica o responsável pela emissão da CAT, são eles: empregador (98,6%); sindicato (0,8%); segurado ou dependente (0,3%); médico (0,2%); autoridade pública (0,1%).

3.2.6 Filiação do segurado

Esse atributo possui dois valores distintos e representa o tipo de filiação à previdência Social do segurado da CAT, são eles: empregado (99,8%) e trabalhador avulso (0,2%).

3.2.7 Natureza da lesão

Esse atributo possui 28 valores distintos, e representa o tipo de lesão e/ou quadro clínico da doença, sendo os três mais frequentes, com sua respectiva proporção da amostra total: corte, laceração ou ferida contusa (21%); fratura (17%); contusão ou esmagamento (14%).

3.2.8 Sexo

Esse atributo possui dois valores distintos e representa o sexo do trabalhador acidentado, sendo eles: masculino (66,6%) e feminino (33,4%).

3.2.9 Tipo de acidente

Os acidentes de trabalho podem ser divididos em três tipos (INSS, 1998):

- Acidente típico: ocorre na execução do trabalho;
- Acidente de trajeto: ocorre no percurso da residência para o trabalho ou vice-versa;
- Acidente devido a doença do trabalho: desencadeada ou produzida pelo exercício do trabalho ou em função de condições especiais em que o trabalho é realizado e com ele se relacione diretamente.

Essa variável é do tipo nominal e possui a seguinte distribuição dentro da amostra: típico (77,0%); trajeto (21,3%); doença (1,7%).

3.2.10 Estado

Unidade federativa em que ocorreu o acidente, com 27 valores distintos, sendo os três mais frequentes: Maranhão (36,8%); São Paulo (30,9%); Rondônia (10,3%).

3.2.11 Mês do acidente

A variável possui 12 valores distintos e representa o mês do acidente de trabalho registrado na CAT, no ano de 2019. Na Figura 8 é apresentada a distribuição da quantidade de acidentes por mês.

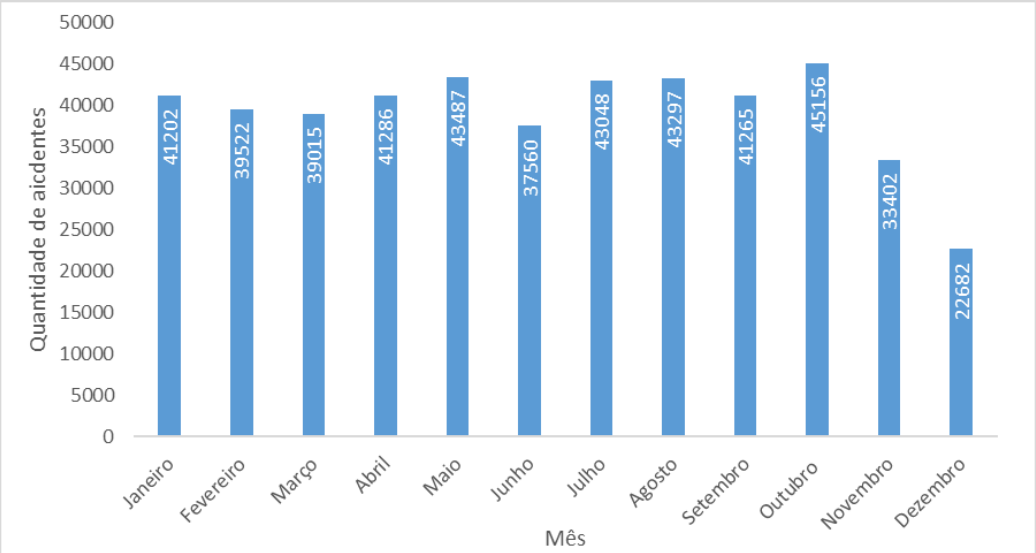

Figura 8. Quantidade de acidentes de trabalho, por mês, no ano de 2019 (Fonte: De autoria própria)

3.2.12 Idade

Essa variável é do tipo numérica e representa a idade do acidentado, sendo a média de idades igual a 35 anos, sua distribuição pode ser vista na Figura 9.

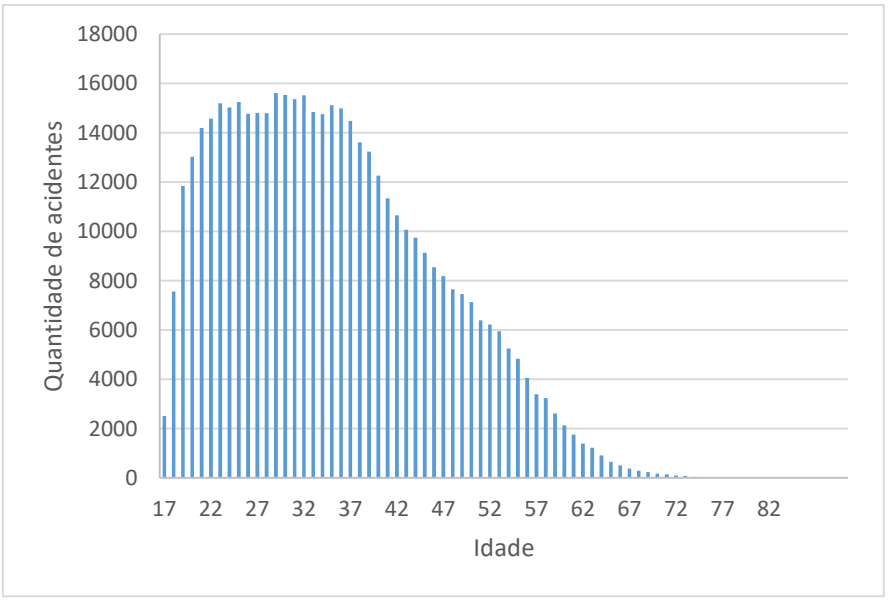

Figura 9. Distribuição de idade dos acidentados (Fonte: De autoria própria)

3.2.13 Gravidade

Vários estudos anteriores utilizaram o aprendizado de máquina para analisar as correlações em dados complexos de acidentes de construção, com foco na classificação da gravidade (LEE *et al.*, 2020).

No banco de dados utilizado nesta monografia, não há a variável gravidade, porém, é possível combinar as variáveis que indicam óbito e espécie de benefício, para parametrizar essa variável. Os seguintes parâmetros foram adotados, atribuindo dois valores para a gravidade:

- Leve: a espécie do benefício indica afastamento menor do que 15 dias (não há pagamento de auxílio por parte do INSS);
- Grave: a espécie do benefício indica algum tipo de auxílio (afastamento igual ou superior a 15 dias);

A variável é do tipo nominal e sua distribuição é apresentada na Figura 10:

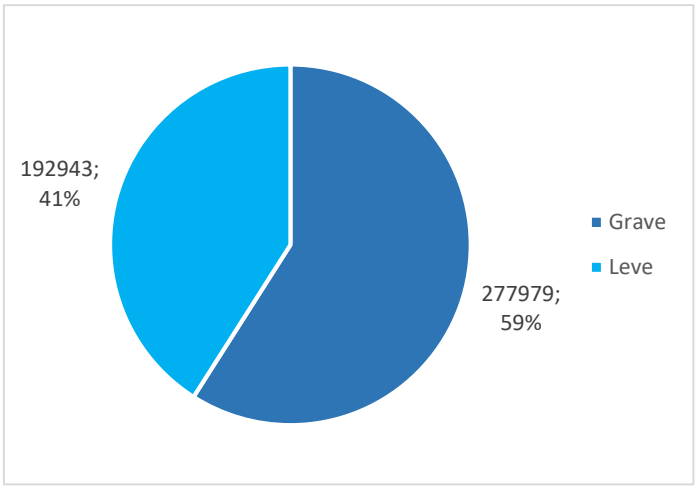

Figura 10. Distribuição das classes da variável alvo (Fonte: De autoria própria)

3.3 MÉTODOS E TÉCNICAS

Essa seção apresenta o planejamento da execução dos testes experimentais, que está dividido em duas etapas: experimento sem o balanceamento da base de dados; e experimento com o balanceamento de dados. Nessa segunda etapa, será aplicada a técnica SMOTE para balanceamento da base de dados.

No experimento executado em ambas as etapas, foram aplicadas as técnicas de seleção de variáveis e execução da árvore de decisão com aplicação da validação cruzada. Os tópicos a seguir, detalham as etapas a cima.

3.3.1 Seleção de variáveis

Na análise de dados, foi utilizado o software Weka (Waikato Environment for Knowledge Analysis), baseado em java. O software Weka foi desenvolvido pela Universidade de Waikato, Nova Zelândia (TIWARI; JHA; YADAV, 2012). As principais razões para utilizar o software Weka na análise de dados são: a ampla utilização do software e seu sistema de código aberto.

A finalidade dessa etapa é reduzir as variáveis e ruídos presentes no banco de dados, para facilitar o treinamento dos modelos, para isso, dois métodos foram avaliados, qui-quadrado e IG.

Algumas variáveis foram removidas da base de dados, essas variáveis traziam apenas informações descritivas de códigos presentes na base de dados, ou seja, são informações redundantes, como: descrição do CBO, descrição do CNAE e descrição do CID.

Instâncias com algum valor não classificado foram removidas (representando menos de 0,1% do banco de dados). Por fim, para esse banco de dados, foram utilizadas/comparadas as técnicas de qui-quadrado e IG, para a seleção de atributos mais importantes.

3.3.2 Treinamento da árvore de decisão com validação cruzada

Através do software Weka, o algoritmo de árvore de decisão será executado. Os experimentos serão realizados utilizando parâmetros default do Weka. As métricas de avaliação de modelo utilizadas nessa pesquisa serão: precisão, acurácia, revocação (*recall*), F1-*Score* (F-Measure) e AUC.

A validação cruzada k-fold será empregada para estimar a precisão do modelo induzido apartir da árvore de decisão, sendo o valor de k igual a 10.

3.3.3 Balanceamento da base de dados

Na base de dados utilizada nessa pesquisa, há desbalanceamento da variável alvo, sendo que a classe leve contém 40% das amostras e a classe grave possui 60% das amostras.

Nessa pesquisa, será utilizada a SMOTE para superamostrar a classe leve, criando exemplos sintéticos. Os exemplos sintéticos fazem com que o classificador crie regiões de decisão maiores e menos específicas, em vez de regiões menores e mais específicas. Esse método permite que as árvores de decisão possam generalizar melhor.

No capítulo 4 serão comparadas a seleção de atributos da árvore de decisão com os métodos do qui-quadrado e IG utilizando a base de dados com e sem a técnica SMOTE.

3.3.4 Fluxograma de métodos experimentais

Na Figura 11 são apresentados os métodos e técnicas experimentais, mencionados nessa seção, em forma de fluxograma.

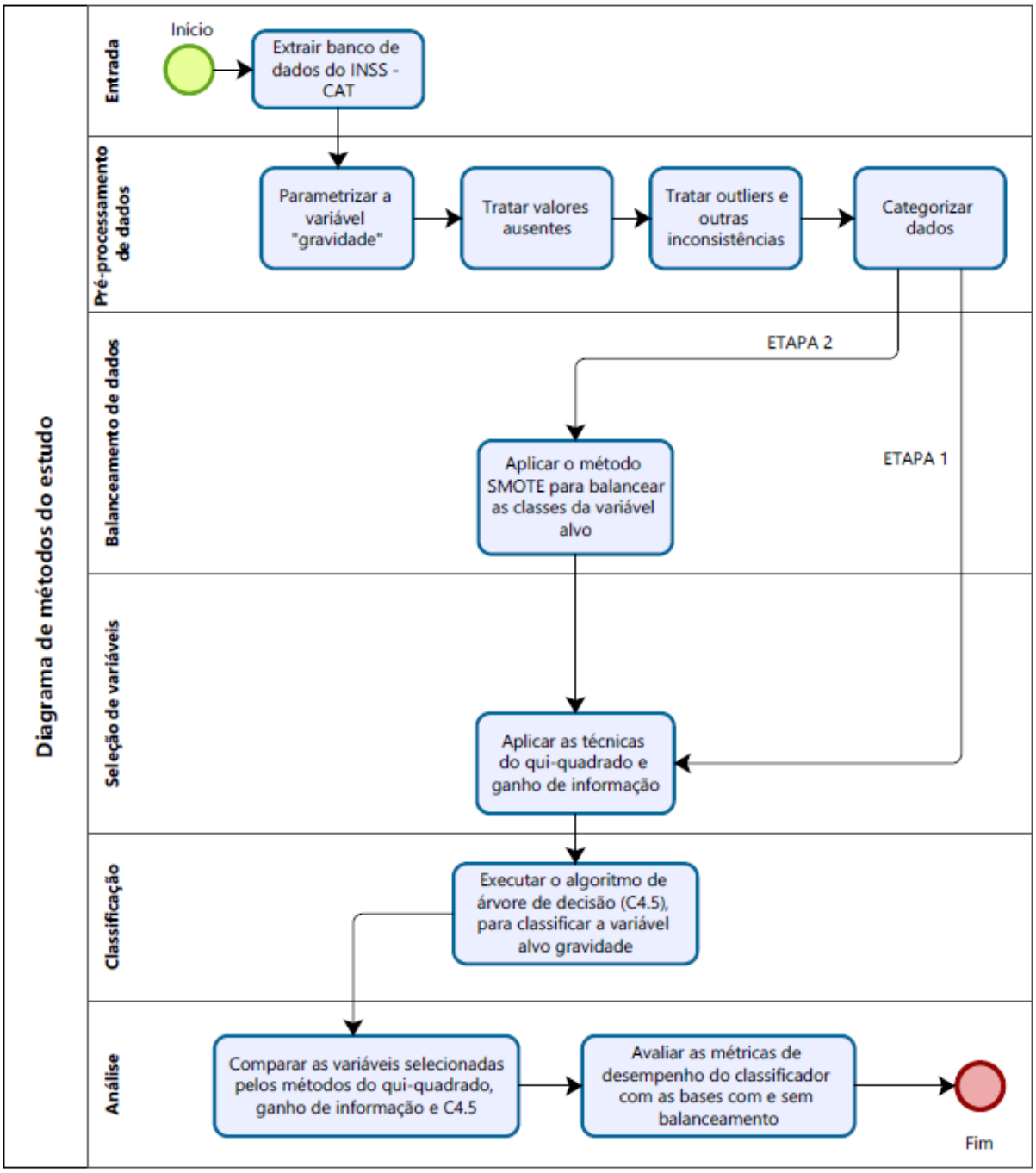

Figura 11. Fluxograma de métodos experimentais (Fonte: De autoria própria)

4. RESULTADOS E ANÁLISES

Neste capítulo, serão apresentados os resultados dos experimentos realizados seguindo as quatro questões que embasaram a realização dos experimentos:

1. Quais são as variáveis mais importantes para a classificação da gravidade, na base de dados?
2. Os atributos da seleção de variáveis são os mesmos selecionados pela árvore de decisão?
3. Os atributos da seleção de variáveis são os mesmos utilizando a técnica SMOTE para o balanceamento da gravidade?
4. As métricas de desempenho da classificação melhoram utilizando a técnica SMOTE?

Os resultados e discussões serão apresentados nos tópicos a seguir, com base nas questões apresentadas acima.

4.1 ETAPA 1: EXPERIMENTO SEM BALANCEAMENTO

Conforme abordado na seção 0, foram realizadas duas etapas de experimentação. Na primeira etapa, foi realizada a seleção de variáveis e, posteriormente, executado a árvore de decisão, ambos com a base de dados sem balanceamento, ou seja, não foi utilizada a técnica SMOTE. Os resultados dessa etapa são descritos nas subseções abaixo.

4.1.1 Resultado da seleção de variáveis

O conjunto de dados de entrada utilizado para essa análise possui 470922 instâncias. Na Tabela 3, são apresentados os valores do qui-quadrado, graus de liberdade e valor-p.

Tabela 3. Resultados da aplicação do qui-quadrado, para a base sem balanceamento

Variável	Qui-quadrado	Graus de Liberdade	Valor-P
cid_10	230508,643	4876	0,000
natureza_lesao	173822,957	27	0,000
agente_causador	134490,089	271	0,000
tipo_acidente	34714,347	2	0,000
cbo	32877,376	2129	0,000
cnae_empregador	21647,719	864	0,000
emitente	2502,782	4	0,000
Idade	1701,125	73	0,000
sexo	829,157	1	0,000
estado_acidente	478,859	26	0,000
mes_acidente	296,389	11	0,000
filiacao_segurado	5,991	1	0,014

Fonte: De autoria própria

Conforme pode ser observado na tabela acima, todas as variáveis assumem um valor-P bem menor que 0,05, neste caso, rejeita-se a hipótese de que as variáveis não possuem correlação. Ou seja, o método do qui-quadrado indica que todas as variáveis são importantes para o algoritmo que será executado.

Na Tabela 4 é apresentada o ranqueamento das variáveis, segundo o método do ganho de informação.

Tabela 4. Resultados da aplicação do ganho de informação, para a base sem balanceamento

Variável	Ganho de Informação
cid_10	0,41832184
natureza_lesao	0,30894751
agente_causador	0,22288011
tipo_acidente	0,05854099
cbo	0,05264550
cnae_empregador	0,03499210
emitente	0,00467569
Idade	0,00261050
sexo	0,00127533
estado_acidente	0,00073878
mes_acidente	0,00045071
filiacao_segurado	0,00000929

Fonte: De autoria própria

As variáveis que tiveram um desempenho maior que 0,1 foram as três primeiras, sendo: cid 10, natureza da lesão e agente causador. Contudo, as três

variáveis com menor influência na variável alvo foram: estado do acidente, mês do acidente e filiação do segurado.

Além disso, é possível observar que a ordem da classificação das variáveis geradas no IG coincide com o ranque gerado a partir dos valores do qui-quadrado, gerado no Weka.

4.1.2 Resultado do treinamento da árvore de decisão

A árvore de decisão utilizada nessa pesquisa é o algoritmo C4.5, ele possui uma técnica própria para selecionar as variáveis, baseado no IGR (conforme descrito na seção 2.5). Esse algoritmo selecionou as seguintes variáveis, para a construção de sua árvore: CID 10, natureza da lesão, agente causador, tipo de acidente, CBO, CNAE do empregador, idade, sexo, estado do acidente e mês do acidente. O algoritmo não utilizou duas variáveis presentes no banco de dados: emitente e filiação do segurado.

Analisando o arquivo da árvore de decisão, em texto, é possível perceber que as variáveis mais importantes são: natureza da lesão; CID 10; e agente causador. Sendo, a primeira, o nó raiz.

Na Tabela 5, são apresentados os resultados das métricas (descritas na subseção 2.7) definidas para a avaliação do modelo de árvore de decisão, na base de dados sem a aplicação da técnica de balanceamento (SMOTE).

Tabela 5. Resultado das métricas da árvore de decisão, com a base sem balanceamento

	Métricas de Avaliação				
Algoritmo	Acurácia	Precisão	Revocação	F1-Score	AUC
Árvore de decisão sem SMOTE	88,0719	0,881	0,881	0,880	0,933

Fonte: De autoria própria

De modo geral, o algoritmo atingiu um percentual de 88% de acurácia, precisão, revocação e F1-Score. Outros estudos, que avaliaram acidentes de trabalho, entretanto, utilizando base de dados diferentes, obtiverem em média uma precisão, utilizando validação cruzada, de 70,19% (LEE *et al.*, 2020) e 90,67% (SARKAR *et al.*, 2019).

A validação cruzada foi utilizada nessa pesquisa para dar mais segurança aos resultados obtidos, pois, quando não utilizado, o resultado pode gerar viés.

Outra métrica avaliada foi a AUC com valor de 0,93, conforme apresentado na Tabela 5Tabela 3. Sua curva pode ser vista na Figura 12.

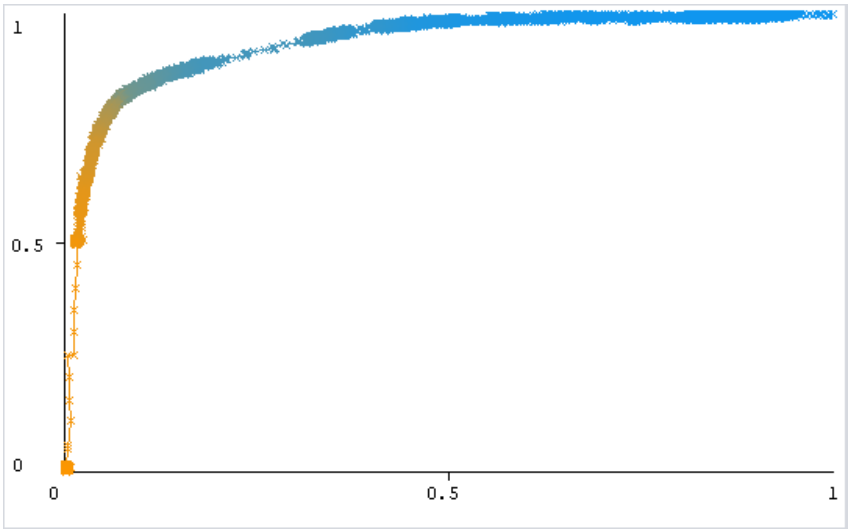

Figura 12. Curva ROC gerada através do Weka, com a base desbalanceada (Fonte: Próprio autor**)**

A curva ROC é uma métrica invariante em escala, ou seja, ela avalia a precisão das classificações ao invés de seu valor absoluto. Ela pode variar de 0 a 1, sendo 0 uma baixa qualidade nas classificações do modelo e 1 indica uma alta qualidade.

4.2 ETAPA 2: EXPERIMENTO COM BALANCEAMENTO

Na segunda etapa, primeiramente, foi aplicada a técnica SMOTE, para realizar o balanceamento da base de dados. Posteriormente, os mesmos procedimentos da primeira etapa foram reaplicados. Os resultados dessa etapa são descritos nas subseções abaixo.

4.2.1 Resultado da seleção de variáveis

O conjunto de dados de entrada, após a aplicação da técnica de balanceamento, teve um aumento no número de amostras, totalizando em 555816 instâncias. Na Tabela 6, são apresentados os valores do qui-quadrado, graus de liberdade e valor-p.

Tabela 6. Resultados da aplicação do qui-quadrado, para a base com balanceamento

Variável	Qui-quadrado	Graus de Liberdade	Valor-P
cid_10	299927,544	4876	0,000
natureza_lesao	215709,667	27	0,000
agente_causador	171293,879	271	0,000
tipo_acidente	55224,399	2	0,000
idade	52067,461	73	0,000
cbo	42094,512	2129	0,000
cnae_empregador	32262,489	864	0,000
emitente	3952,576	4	0,000
estado_acidente	3643,302	26	0,000
mes_acidente	2512,675	11	0,000
sexo	703,968	1	0,000
filiacao_segurado	24,062	1	0,000

Fonte: De autoria própria

Conforme pode ser observado na tabela acima, o valor-p para todas as variáveis é zero, isto é, mesmo após o SMOTE, todas as variáveis mantiveram-se relevantes para o modelo. Sobretudo, algumas tiveram sua ordem alterada, sendo elas: idade e sexo.

Ressalta-se que os valores de qui-quadrado nessa etapa tiveram seu valor aumentado para todas as variáveis.

Na Tabela 7 é apresentada o ranqueamento das variáveis, segundo o método de ganho de informação.

Tabela 7. Resultados da aplicação do ganho de informação, para a base com balanceamento

Variável	Ganho de Informação
cid_10	0,4742135
natureza_lesao	0,3443714
agente_causador	0,2472588
idade	0,0904309
tipo_acidente	0,0770366
cbo	0,0573682
cnae_empregador	0,0456623
emitente	0,0058971
estado_acidente	0,0047528
mes_acidente	0,0032703
sexo	0,0009139
filiacao_segurado	0,0000314

Fonte: De autoria própria

Após a aplicação do SMOTE, as variáveis que tiveram um desempenho maior que 0,1 mantiveram as mesmas, sendo: cid 10, natureza da lesão e agente causador. Contudo, a variável idade subiu quatro posições, sendo que antes, ela possuía valor de 0,002 e passou a ter 0,090. A variável sexo caiu duas posições, antes, possuía 0,0012 e foi para 0,0009. As três variáveis com menor influência na variável alvo foram: mês do acidente, sexo e filiação do segurado.

Além disso, é possível observar que a ordem da classificação das variáveis geradas no IG coincide com o ranque gerado a partir dos valores do qui-quadrado, exceto as variáveis idade e tipo de acidente, que tiveram suas posições trocadas.

4.2.2 Resultado do treinamento da árvore de decisão

Na segunda etapa, utilizando o SMOTE, a árvore de decisão selecionou as mesmas variáveis da primeira etapa, sendo elas: CID 10, natureza da lesão, agente causador, tipo de acidente, CBO, CNAE do empregador, idade, sexo, estado do acidente e mês do acidente. Igualmente à primeira etapa, o algoritmo não utilizou duas variáveis presentes no banco de dados: emitente e filiação do segurado.

As três variáveis mais importantes foram definidas como: natureza da lesão; CID 10; e agente causador. Sendo, a primeira, o nó raiz, da mesma forma que ocorreu na primeira etapa.

Na Tabela 8, são apresentados os resultados das métricas (descritas na subseção 2.7) definidas para a avaliação do modelo de árvore de decisão, na base de dados com a aplicação da técnica de balanceamento (SMOTE).

Tabela 8. Resultado das métricas da árvore de decisão, com a base com balanceamento

	Métricas de Avaliação				
Algoritmo	Acurácia	Precisão	Revocação	F1-Score	AUC
Árvore de decisão com SMOTE	89,2391	0,893	0,892	0,892	0,947

Fonte: De autoria própria

De modo geral, o algoritmo atingiu um percentual de 89% de acurácia, precisão, revocação e F1-Score. Sendo 1% maior do que o resultado da etapa 1. A validação cruzada também foi utilizada nessa etapa.

Quanto a métrica AUC o valor foi de 0,94, conforme apresentado na Tabela 8, obteve aumento de 0,01Tabela 3. Sua curva pode ser vista na Figura 13.

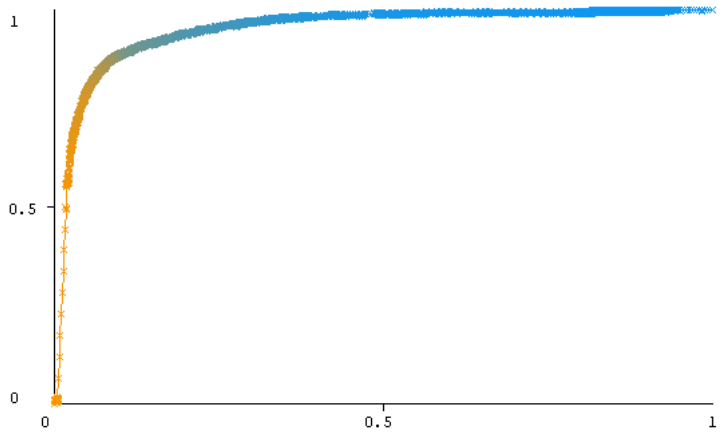

Figura 13. Curva ROC gerada através do Weka, com a base balanceada (Fonte: Próprio autor)

A curva ROC manteve a mesma forma da curva da Figura 12, da base de dados sem balanceamento.

4.3 DISCUSSÕES

Foram observadas pequenas diferenças na classificação da gravidade com e sem utilização do SMOTE. O IG apontou, na primeira etapa, que idade e sexo tiveram ranque 8 e 9, respectivamente, e passaram a ter, na segunda etapa, após o balanceamento, o ranque 4 e 11, respectivamente, como pode ser observado na Tabela 9.

Tabela 9. Ranque de variáveis para Etapas 1 e 2

	Variável	
Ranque	Etapa 1	Etapa 2
1	cid_10	cid_10
2	natureza_lesao	natureza_lesao
3	agente_causador	agente_causador
4	tipo_acidente	idade
5	cbo	tipo_acidente
6	cnae_empregador	cbo
7	emitente	cnae_empregador
8	Idade	emitente
9	sexo	estado_acidente
10	estado_acidente	mes_acidente
11	mes_acidente	sexo
12	filiacao_segurado	filiacao_segurado

Fonte: De autoria própria

Além disso, foi possível observar um aumento nos valores de qui-quadrado e IG, após utilizar o SMOTE.

Portanto, as variáveis mais importantes para a classificação da gravidade, foram definidas conforme aplicação do IG para as duas etapas: sem aplicação do SMOTE, definido como etapa 1; e com aplicação do SMOTE, definido como etapa 2.

Em ambas as etapas, a variável mais importante indicada pela árvore de decisão, ou seja, o nó raiz, foi a variável natureza da lesão. Entretando, conforme apresentado na Tabela 9, a variável mais importante, obtida através do IG é o CID 10. Uma justificativa para essa diferença é que o algoritmo C4.5 aplica o IGR para definir a raiz e os nós da árvore, que é diferente do método IG, conforme descrito nas subseções 2.4.2 e 2.5.

Com o método qui-quadrado foi possível observar que todas as variáveis são importantes, nas duas etapas, para a classificação da gravidade, pois o seu valor-p é menor que o limite, adotado na literatura, de 0,05 (MOORE; NOTZ; FLINGER, 2013). No entanto, notou-se, em ambas as etapas, que a árvore de decisão não utilizou duas variáveis, sendo elas: emitente e filiação.

Com relação às métricas de desempenho da árvore de decisão, que estão apresentadas na Tabela 10, é possível observar que na etapa 2, com a utilização do SMOTE, todas as métricas tiveram aumento de 1%.

Tabela 10. Comparação de métricas para a árvore de decisão com e sem SMOTE

	Métricas de Avaliação				
Algoritmo	Acurácia	Precisão	Revocação	F1-Score	AUC
Árvore de decisão sem SMOTE	88,0719	0,881	0,881	0,880	0,933
Árvore de decisão com SMOTE	**89,2391**	**0,893**	**0,892**	**0,892**	**0,947**

Fonte: De autoria própria

Vale ressaltar que a métrica AUC é comum na análise de experimentos com bases desbalanceadas, porque ela consegue avaliar sensibilidade e especificidade. Seu valor varia 0 a 1, quanto maior esse valor, melhor a capacidade de classificação, portanto, o C4.5 atingiu um bom resultado, conforme essa métrica.

5. CONCLUSÕES

Apesar da implementação de diversos sistemas de gestão em Saúde e Segurança do Trabalho, a segurança ocupacional é baixa. Essa gestão é complexa devido à grande quantidade de entidades envolvidas no processo produtivo.

Além disso, a maior parte do trabalho é realizada por humanos, portanto, técnicas para classificar acidentes de trabalho por meio de correlações simples são limitadas. Esse cenário está mudando com a recente evolução dos estudos de predição de acidentes, aplicando técnica de análise de dados e aprendizado de máquinas

Este estudo apresentou a seleção de variáveis no contexto de acidentes de trabalho, tendo como variável alvo a gravidade, bem como experimentos com o algoritmo de classificação C4.5, disponível na ferramenta Weka.

Este trabalho serviu para consolidar os conceitos relacionados a aprendizados de máquinas, no qual, foi adquirido uma habilidade em trabalhar com o software Weka. Essa ferramenta tem potencial para facilitar a realização de diferentes experimentos rapidamente. Contudo, a ferramenta perde performance quando há análise com grande volume de dados.

Com o objetivo do trabalho de analisar as variáveis mais importantes, presentes no banco de dados da CAT, para a classificação da variável alvo gravidade, foi feita a seleção de variáveis por meio das técnicas do qui-quadrado e IG, aplicadas em duas etapas, com e sem balanceamento de dados.

Foram comparados e discutidos os experimentos realizados nas duas etapas. Observou-se que a técnica de balanceamento aumentou em 1% as métricas de avaliação do modelo.

Em geral, com base nessa pesquisa, podemos observar que: vários tipos de variáveis e valores estão presentes na base de dados de acidentes de trabalho, o que torna difícil processar dados, refletir características e interpretar correlações sem a utilização de ferramentas computacionais.

Neste contexto, os principais benefícios do aprendizado de máquinas são: o aumento da produtividade, uma vez que pode-se automatizar processos internos e a redução de custos.

Como trabalhos futuros, é sugerido investigar quais são as variáveis mais importantes, a partir da árvore de decisão resultante da ferramenta Weka, com um grande volume de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

AN, S.-H. *et al.* Application of Support Vector Machines in Assessing Conceptual Cost Estimates. **Journal of Computing in Civil Engineering**, [*s. l.*], v. 21, n. 4, p. 259–264, 2007. Available at: https://doi.org/10.1061/(ASCE)0887-3801(2007)21:4(259)

AVIAD, B.; GELBARD, R. Classification by clustering decision tree-like classifier based on adjusted clusters. **Expert Systems with Applications: An International Journal**, [*s. l.*], v. 38, n. 7, p. 8220–8228, 2011.

AYHAN, B. U.; TOKDEMIR, O. B. Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial Neural Networks. **Journal of Construction Engineering and Management**, [*s. l.*], v. 146, n. 3, p. 04019114, 2020. Available at: https://doi.org/10.1061/(ASCE)CO.1943-7862.0001762

BARTOLOMEU, T. A. Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento. [*s. l.*], p. 302, 2002.

BEVILACQUA, M.; CIARAPICA, F.E.; GIACCHETTA, G. Industrial and occupational ergonomics in the petrochemical process industry: A regression tree approach. **Accid. Anal. Prev.**, [*s. l.*], v. 40, p. 122–133, 2008.

BRASIL. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Introdução à Classificação Nacional de Atividades Econômicas - CNAE versão 2.0**2007.

BRASIL. MINISTÉRIO DA ECONOMIA. **Código Brasileiro de Ocupações**Brasília, 2002.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. **Journal Of Artificial Intelligence Research**, [*s. l.*], v. 16, p. 321–357, 2002.

CHEN, H.; LUO, X. Severity Prediction Models of Falling Risk for Workers at Height. **Procedia Engineering**, [*s. l.*], v. 164, p. 439–445, 2016. Available at: https://doi.org/10.1016/j.proeng.2016.11.642

CHENG, M.-Y. *et al.* Estimate at Completion for construction projects using Evolutionary Support Vector Machine Inference Model. **Automation in Construction**, [*s. l.*], v. 19, n. 5, p. 619–629, 2010. Available at: https://doi.org/10.1016/j.autcon.2010.02.008

CHUNG, M. K.; WU, S.-C. H.; HERRIN, G. D. The use of a mixed Weibull model in occupational injury analysis. **Journal of Occupational Accidents**, [*s. l.*], v. 7, n. 4, p. 239–250, 1986. Available at: https://doi.org/10.1016/0376-6349(86)90016-7

COLEMAN, P. J.; KERKERING, J. C. Measuring mining safety with injury statistics: Lost workdays as indicators of risk. **Journal of Safety Research**, [*s. l.*], v. 38, n. 5, p. 523–533, 2007. Available at: https://doi.org/10.1016/j.jsr.2007.06.005

CSISZÁR, I.; SHIELDS, P. **Information Theory and Statistics: A Tutorial**. [*S. l.: s. n.*], 2004.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recogn. Lett.**, [*s. l.*], v. 27, p. 861–874, 2006.

FREIVALDS, A.; JOHNSON, A. B. Time-series analysis of industrial accident data. **Journal of Occupational Accidents**, [*s. l.*], v. 13, n. 3, p. 179–193, 1990. Available at: https://doi.org/10.1016/0376-6349(90)90020-V

HA, T. M.; BUNKE, H. Off-line, handwritten numeral recognition by perturbation method. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [*s. l.*], v. 19, n. 5, p. 535–539, 1997.

HAMILTON, H. *et al.* **Overview of Decision Trees**. University of California: [*s. n.*], 2012.

HOUARI, R. *et al.* Dimensionality reduction in data mining: A Copula approach. **Expert Systems with Applications**, [*s. l.*], v. 64, p. 247–260, 2016. Available at: https://doi.org/10.1016/j.eswa.2016.07.041

INSS. **Manual de instruções para preenchimento da comunicação de acidente do trabalho - CAT**. Brasília: [*s. n.*], 1999.

JØRGENSEN, K. **Serious work accidents and their causes - An analysis of data from Eurostat**. [*S. l.*]: Safety Science Monitor, 2015.

KHANZODE, V. V.; MAITI, J.; RAY, P. K. Occupational injury and accident research: A comprehensive review. **Safety Science**, [*s. l.*], v. 50, n. 5, p. 1355–1367, 2012. Available at: https://doi.org/10.1016/j.ssci.2011.12.015

KUBAT, M.; MATWIN, S. **Addressing the curse of imbalanced training sets: one-sided selection**. *In*: PROC. 14TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING. [*S. l.: s. n.*], 1997. p. 179–186.

LEE, J. Y. *et al.* A Study on Data Pre-Processing and Accident Prediction Modelling for Occupational Accident Analysis in the Construction Industry. **Applied Sciences**, [*s. l.*], v. 10, n. 21, p. 7949, 2020. Available at: https://doi.org/10.3390/app10217949

LEU, S. S.; CHANG, C. M. Bayesian-network-based safety risk assessment for steel construction projects. **Accident; Analysis and Prevention**, [*s. l.*], v. 54, p. 122–133, 2013.

LIMA, A.; BATISTA, E. **Uma Analise de Ambientes de Programação em Blocos com Base em Recomendações de Interação Criança-Computador**. 1ª edição ed. Porto Alegre: Sociedade Brasileira de Computação - SBC, 2018.

LING, C.; LING, C. X.; LI, C. **Data Mining for Direct Marketing: Problems and Solutions**. *In*: IN PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. [*S. l.: s. n.*], 1998. p. 73–79.

MALLICK, S.; MUKHERJEE, K. An empirical study for mines safety management through analysis on potential for accident reduction. **Omega**, [*s. l.*], v. 24, n. 5, p. 539–550, 1996. Available at: https://doi.org/10.1016/0305-0483(96)00020-5

MANTAS, C. J.; ABELLÁN, J. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. **Expert Systems with Applications**, [*s. l.*], v. 41, n. 10, p. 4625–4637, 2014. Available at: https://doi.org/10.1016/j.eswa.2014.01.017

MEASE, D.; WYNER, A. Evidence contrary to the statistical view of boosting: A rejoinder to responses. **Journal of Machine Learning Research**, [*s. l.*], v. 9, p. 195–201, 2008.

MEYER, P. L. **Probabilidade: aplicações à estatística**. 2. ed. São Paulo: [*s. n.*], 1983.

MITCHELL, T. M. **Machine learning**. New York: [*s. n.*], 1997.

MOORE, D. S.; NOTZ, W. I.; FLINGER, M. A. **The basic practice of statistics**. 6. ed. New York: [*s. n.*], 2013.

NEELAMEGAM, S.; RAMARAJ, E. Classification algorithm in data mining: An Overview. **International Journal of P2P Network Trends and Technology (IJPTT)**, [*s. l.*], v. 3 (5), p. 1–5, 2013.

OLSON, D. L.; DELEN, D.; MENG, Y. Comparative analysis of data mining methods for bankruptcy prediction. **Decision Support Systems**, [*s. l.*], v. 52, n. 2, p. 464–473, 2012. Available at: https://doi.org/10.1016/j.dss.2011.10.007

OZTEKIN, A. *et al.* A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. **European Journal of Operational Research**, [*s. l.*], v. 266, n. 2, p. 639–651, 2018. Available at: https://doi.org/10.1016/j.ejor.2017.09.034

QUINLAN, J. R. C4.5: programs for machine learning. **Morgan Kaufmann Publishers Inc.**, [*s. l.*], v. 340, p. 302, 1993.

QUINLAN, J. R. Introduction of Decision Trees. **Machine Learning**, [*s. l.*], v. 1, p. 81–106, 1986.

RAVI KUMAR, P.; RAVI, V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. **European Journal of Operational Research**, [*s. l.*], v. 180, n. 1, p. 1–28, 2007. Available at: https://doi.org/10.1016/j.ejor.2006.08.043

REFAEILZADEH, P.; TANG, L.; LIU, H. **Cross-Validation**. *In*: ENCYCLOPEDIA OF DATABASE SYSTEMS. [*S. l.*]: Springer, 2009.

ROBINSON, G. H. Accidents and sociotechnical systems: principles for design. **Accident Analysis & Prevention**, [*s. l.*], v. 14, n. 2, p. 121–130, 1982. Available at: https://doi.org/10.1016/0001-4575(82)90078-1

SARKAR, S. *et al.* Application of optimized machine learning techniques for prediction of occupational accidents. **Computers and Operations Research**, [*s. l.*], v. 106, p. 210–224, 2019. Available at: https://doi.org/10.1016/j.cor.2018.02.021

SOLOMON, K. A.; ALESCH, K. A. The index of harm: A measure for comparing occupational risk across industries. **Journal of Occupational Accidents**, [*s. l.*], v. 11, n. 1, p. 19–35, 1989. Available at: https://doi.org/10.1016/0376-6349(89)90003-5

TIWARI, M.; JHA, M. B.; YADAV, O. Performance analysis of Data Mining algorithms in Weka. **IOSR Journal of Computer Engineering**, [*s. l.*], v. 6, n. 3, p. 32–41, 2012. Available at: https://doi.org/10.9790/0661-0633241

WITTEN, I. H.; FRANK, E. **Data mining: Practical machine learning tools and techniques.** Burlington: Morgan Kaufmann Publisher, 2005.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Massachusetts: Morgan Kaufmann, 2011.

YANG, Y.; PEDERSEN, J. O. **A Comparative Study on Feature Selection in Text Categorization**. USA: [*s. n.*], 1997.