UNIVERSIDADE FEDERAL DE ALAGOAS
UNIDADE ACADÊMICA CENTRO DE TECNOLOGIA
CURSO DE ENGENHARIA QUÍMICA

LARISSA SILVA COSTA

USO DA ENGENHARIA 4.0 PARA MODELAGEM DE RISCOS: TEXT MINING E MACHINE LEARNING PARA CLASSIFICAÇÃO AUTOMÁTICA DE ACIDENTES DE TRABALHO

MACEIÓ – AL

2024

LARISSA SILVA COSTA

USO DA ENGENHARIA 4.0 PARA MODELAGEM DE RISCOS: TEXT MINING E MACHINE LEARNING PARA CLASSIFICAÇÃO AUTOMÁTICA DE ACIDENTES DE TRABALHO

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia Química da Universidade Federal de Alagoas como requisito parcial para obtenção do título de Bacharel em Engenharia Química.

Orientador/a: Prof. Dr. Frede de Oliveira Carvalho.

MACEIÓ – AL

2024

Catalogação na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Janis Christine Angelina Cavalcante – CRB-4 –1667

C837u Costa, Larissa Silva

Uso da engenharia 4.0 para modelagem de risco : *Text Mining* e *Machine Learning* para classificação automática de acidentes de trabalho / Larissa Silva Costa. – 2024.

35 f. il. : figs. ; tabs. color.

Orientador: Frede de Oliveira Carvalho.

Monografia (Trabalho de Conclusão de Curso em Engenharia Química). Universidade Federal de Alagoas. Centro de Tecnologia. Maceió, 2024.

Bibliografia: f. 33-35.

1. Indústria 4.0. 2. Acidentes de Trabalho. 3. Segurança Ocupacional. 4. Redes Neurais Artificiais. 5. Máquinas de Vetores de Suporte. I. Título.

CDU: 66.03:331.4

RESUMO

A evolução da indústria ao longo dos séculos tem sido marcada por revoluções que transformaram a forma como os produtos são fabricados e os processos são gerenciados. A Quarta Revolução Industrial, também conhecida como Indústria 4.0, tem desempenhado um papel fundamental na transformação dos processos industriais, impulsionando avanços significativos na automação, digitalização e segurança no ambiente de trabalho. Este trabalho propõe o desenvolvimento de algoritmos computacionais em Python para identificar gravidade de lesões originadas por acidentes de trabalho, com ênfase no uso de mineração de texto com aprendizado de máquina. Utilizando técnicas avançadas de inteligência artificial, como Redes Neurais Artificiais (RNA) e Máquinas de Vetores de Suporte (SVM), a pesquisa busca compreender a ocorrência de acidentes no ambiente de trabalho, destacando a importância do controle desses incidentes para a segurança e bem-estar dos trabalhadores. Os resultados obtidos demonstram que os modelos de RNA apresentam alta precisão na estimação das condições de trabalho e na classificação de acidentes, enquanto os modelos de SVM oferecem uma abordagem menos eficaz na identificação específica das classificações. Recomenda-se a integração dessas técnicas para aprimorar o monitoramento contínuo e a gestão de riscos ocupacionais, contribuindo para a construção de ambientes de trabalho mais seguros, saudáveis e produtivos.

Palavras-chave: Indústria 4.0; Acidentes de Trabalho; Segurança Ocupacional; Redes Neurais Artificiais; Máquinas de Vetores de Suporte; Classificação de Acidentes; Python.

ABSTRACT

The evolution of industry throughout the centuries has been marked by revolutions that have transformed the way products are manufactured and processes are managed. The Fourth Industrial Revolution, also known as Industry 4.0, has played a fundamental role in transforming industrial processes, driving significant advances in automation, digitalization, and workplace safety. This work proposes the development of computational algorithms in Python to identify the severity of injuries resulting from workplace accidents, with an emphasis on the use of text mining with machine learning. Using advanced artificial intelligence techniques such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM), the research seeks to understand the occurrence of accidents in the workplace, highlighting the importance of controlling these incidents for the safety and well-being of workers. The results obtained demonstrate that ANN models exhibit high accuracy in estimating working conditions and classifying accidents, while SVM models offer an effective approach in identifying specific classifications. Integration of these techniques is recommended to enhance continuous monitoring and management of occupational risks, contributing to the development of safer, healthier, and more productive work environments.

Keywords: Industry 4.0; Workplace Accidents; Occupational Safety; Artificial Neural Networks; Support Vector Machines; Accident Classification; Python.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RNA	Redes Neurais Artificiais
SVM	Máquinas de Vetores de Suporte
CAT	Comunicação de Acidente de Trabalho
BoW	*Bag of Words*
CNAE	Classificação Nacional de Atividades Econômicas
RBF	*Radial Basis Function*
RELU	*Rectified Linear Unit*
SGD	*Stochastic Gradient Descent*
RMSProp	*Root Mean Square Propagation*
Adagrad	*Adaptive Gradient*
Adam	*Adaptive Moment Estimation*

SUMÁRIO

1. INTRODUÇÃO

Acidentes no local de trabalho são uma preocupação séria no Brasil e no mundo. Segundo a Organização Internacional do Trabalho (OIT), quase três milhões de trabalhadores e trabalhadoras morrem a cada ano devido a acidentes e doenças relacionados com o trabalho. Além disso, em 2022, cerca de 395 milhões de trabalhadores e trabalhadoras em todo o mundo sofreram lesões ocupacionais não fatais. Especificando o Brasil, em 2022, houve 648.366 acidentes de trabalho, incluindo os típicos, de trajeto e doenças do trabalho, de acordo com o AEPS (Anuário Estatístico da Previdência Social) (BRASIL, 2022).

A dificuldade em reduzir o número de acidentes do trabalho está relacionado com alguns fatores do sistema de gestão, como o modo com que o mesmo é implementado e mantido (RICHERS, 2009). Existem muitas áreas que possibilitam a investigação desses acidentes, tendo o monitoramento de perigos e a avaliação de riscos entre os tópicos de investigação mais populares. Neste ambiente de estudo, estão disponíveis vários resultados de trabalhos relacionados à identificação de fatores de risco de acidentes ocupacionais, simulação de cenários de perigo, análise da frequência e gravidade de eventos de lesões e contramedidas de prevenção baseadas em regras de decisão (SARKAR, 2020). Todos esses esforços estão relacionados pelos documentos registrados de acidentes de trabalho. No Brasil, uma fonte para esses dados é a Comunicação de Acidente de Trabalho (CAT), que apresenta a particularidade de que são registros feitos em linguagem natural.

A identificação da gravidade das lesões visa melhorar o desempenho da segurança e sugerir medidas corretivas proativas para possíveis acidentes observados, por exemplo (DI NOIA et al., 2020) que utilizaram estratégias de aprendizado de máquina no contexto de acidentes de trabalho, prevendo riscos de doenças ocupacionais por meio de inteligência computacional e técnicas de reconhecimento de padrões. Eles usaram dados reais sobre o trabalhador e o local de trabalho a partir de dados da Autoridade Sanitária Italiana (ASL).

Como as mortes e lesões em massa são uma característica comum dos acidentes, a distribuição da gravidade dos acidentes nos dados textuais é comparativamente uniforme e adequada como um caso experimental para investigar a relação implícita entre os fatores de atributos e a gravidade das lesões. Embora a identificação da gravidade das lesões seja essencial para a prevenção de acidentes, na literatura não foram encontradas muito trabalhos analisando dados textuais para avaliação de risco e otimização de modelos para acidentes de trabalho. Tendo isso em vista, a aplicação do método de máquinas de aprendizado como Máquinas de Vetores de Suporte (SVM) e Redes Neurais Artificiais (RNA) pode oferecer resultados valiosos

para esse tipo de análise, pois agrupa incidentes semelhantes em *clusters*, facilitando a identificação de padrões intrínsecos nos dados de CATs (BEIER, 2020).

Neste contexto, o presente trabalho tem o objetivo de implementar, em linguagem *Python*, uma classificação automática de gravidade de acidentes de trabalho, a partir do CAT. Foram utilizadas técnicas de pré-processamento de linguagem natural em conjunto com máquinas de aprendizado, mais especificamente SVM e RNA, permitindo, dessa forma classificar automaticamente novos acidentes relatados.

2. OBJETIVOS

2.1 Geral

O objetivo deste trabalho de conclusão de curso é explorar o potencial das técnicas da engenharia 4.0, trazendo a mineração de texto em conjunto com as máquinas de aprendizado, mais especificamente Redes Neurais Artificiais e o SVM, como abordagens eficientes para a classificação e padronização automática de acidentes no contexto da segurança do trabalho.

2.2 Específicos

- Estudo do processo de Diagnóstico de CAT utilizando Mineração de texto para pré-processamento de texto;
- Estudo do processo de Diagnóstico de CAT utilizando o modelo SVM;
- Estudo do processo de Diagnóstico de CAT utilizando o modelo RNA;
- Avaliar o desempenho dos modelos na etapa de Diagnóstico de CAT para determinar o que obteve a melhor performance;
- Avaliar as classificações obtidas com a mineração do texto das CATs e determinar o melhor modelo para classificação automática da gravidade da lesão.

3. REVISÃO BIBLIOGRÁFICA

3.1 Saúde Ocupacional e Segurança do Trabalho

A Lei n. 8.213, de 24 de julho de 1991, conceitua acidente de trabalho típico como aquele "que ocorre pelo exercício do trabalho a serviço da empresa, provocando lesão corporal ou perturbação funcional que cause a morte ou a perda ou redução, permanente ou temporária, da capacidade para o trabalho (BRASIL, 1991). Equiparam-se aos acidentes de trabalho as doenças profissionais, aquelas produzidas ou desencadeadas "pelo exercício do trabalho peculiar a determinada atividade", e as doenças de trabalho, adquiridas ou desencadeadas "em função de condições especiais em que o trabalho é realizado e com ele se relacione diretamente (IBID, 2020).

Quando ocorre acidente ou doença ocupacional, o empregador é obrigado a emitir a Comunicação de Acidente de Trabalho – CAT até o primeiro dia útil da ocorrência (IBID, 2020), nela estão todos os dados referentes ao acidente, onde são utilizados para proteger os direitos do trabalhador, bem como para garantir que tanto o empregador quanto os órgãos competentes estejam formalmente cientes do incidente. No entanto, com o passar dos eventos, as CATs muitas vezes são consideradas insignificantes, sem ações de acompanhamento, perdendo assim sua importância. Apesar de as organizações terem experimentado eventos fatais, sérios e com prejuízos, ainda é desafiador reverter essa situação, pois o AEPS (Anuário Estatístico da Previdência Social), divulgado pelo governo, indica um aumento de 11,6% nos dados relativos aos acidentes de trabalho no Brasil em 2022 em relação ao ano de 2021. Já no caso dos óbitos, o crescimento foi de 4,6% em relação ao ano anterior. Com tantos dados históricos disponíveis, juntamente com estratégia e conhecimento técnico adequados, é coerente utilizar técnicas advindas da indústria 4.0, que traz revolução tecnológica para integração de sistemas ciberfísicos, Internet das Coisas (IoT) e computação em nuvem para otimizar processos industriais e criar fábricas inteligentes e conectadas. Este poderia ser o novo valor agregado ao ecossistema e indústrias relacionados, considerando que as classificações automáticas de lesões ocupacionais podem auxiliar a identificar áreas de maior risco, priorizar medidas preventivas e melhorar a segurança no ambiente de trabalho.

Assim podemos perceber que, diante do aumento dos índices de acidentes de trabalho e óbitos, é crucial que as empresas adotem uma abordagem tecnológica e proativa. Integrar técnicas da indústria 4.0 com a análise eficiente dos dados das CATs possibilita identificar e

prevenir riscos, promovendo uma cultura de segurança e inovação, para alcançar ambientes de trabalho mais seguros e produtivos.

3.2 *Text Mining*

Hoje em dia, há uma maior simplicidade na procura da informação pretendida, daí surgem as ferramentas de auxílio neste sentido, que permitiram aumentá-la de forma significativa, agilizando o processo com inteligência, tal como a mineração de texto (REPORT, 2005). Inspirada no *Data Mining*, o *Text Mining* é definido como a ciência responsável pelo tratamento do processamento de informação/texto. Une diferentes áreas tais como a estatística, linguística, informática e ciência cognitiva. Em outras palavras, a mineração de texto consiste, através da identificação de diferentes tendências ou padrões, na extração de dados regulares de textos estruturados ou semi-estruturados (ARANHA, 2006).

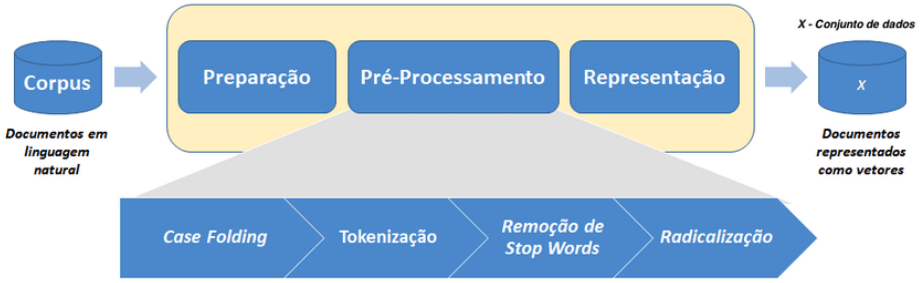
O *Text Mining* envolve a análise de um conjunto de documentos chamado *corpus*, em que ocorre primeiramente a preparação desses dados, geralmente em formato *MS Excel*, em que se retiram colunas indesejadas para a análise e células vazias.

Após isso, é realizado o processo chamado de *Case Folding*, que transforma todos os dados de texto em letras minúsculas, já que letras maiúsculas e minúsculas podem ser interpretadas de forma diferente por programas de computador se não forem tratadas de maneira consistente. Em seguida, cada documento é examinado para criar uma lista de termos, também chamados de *tokens*, onde são excluídos os termos considerados irrelevantes para a análise, como artigos, preposições e caracteres especiais e os termos definidos pelo usuário em uma lista especial chamada *stopwords* são removidos.

Depois disso, os termos são simplificados para suas formas mais básicas, eliminando prefixos e sufixos para evitar diferentes variações da mesma palavra, como em "interessante" e "interessantíssimo", que expressam ideias semelhantes com diferentes intensidades. Esse processo de simplificação é conhecido como *stemming* ou radicalização.

Por fim, a presença dos termos é contabilizada em uma representação vetorial binária ou em uma representação baseada nas frequências dos termos (BOSCARIOLI, 2016). Esse processo completo foi mostrado na Figura 1.

Figura 1 – Esquema típico de um pré-processamento de dados



Fonte: Diaz et. al (2018).

Em conjuntos de dados do mundo real, valores errôneos podem ser registrados por uma variedade de razões, incluindo erros de edição, julgamentos subjetivos e mau funcionamento, ou uso indevido do equipamento de correção ortográfica automática. Essas questões são levantadas e podem ser tratadas também na fase de pré-processamento.

Por fim, os textos são representados em um formato estruturado na forma de números que preserve as principais características dos dados (REZENDE et al., 2011).

3.2.1 Representação de documentos textuais

Existem várias maneiras de realizar a conversão de documentos de texto simples para instâncias com um número fixo de atributos. É possível contar o número de vezes que as frases especificadas ocorrem, ou tomar uma combinação de duas palavras consecutivas, ou ainda contar a ocorrência de duas ou três combinações de caracteres (conhecidas como bigramas e trigramas, respectivamente). Uma representação simples baseada em palavras é a mais comumente utilizada, conhecida como representação de saco de palavras (BoW, do inglês *bag-of-words*) (BRAMER, 2016).

Em processamento de linguagem natural, denominamos corpus (ou documentos) toda uma coleção de textos que temos. Um *corpus* pode ser decomposto em fragmentos. Os fragmentos podem ser frases simples, parágrafos ou documentos de várias páginas, sendo algo que desejamos tratar como uma amostra. Por exemplo, se estivermos analisando documentos clínicos, cada documento de admissão de paciente pode ser um fragmento; se estivermos analisando o sentimento nas mídias sociais, cada comentário do usuário é um fragmento; e assim por diante.

Um modelo de saco de palavras é feito transformando cada palavra do *corpus* em um recurso e em cada linha, sob aquela palavra, conta-se quantas vezes a palavra ocorre naquele fragmento. Nesta abordagem, a ordem das palavras é perdida, porém é uma das principais formas de converter a linguagem em recursos a serem alimentados em um algoritmo de aprendizado de máquina.

A literatura apresenta geralmente duas possibilidades de se trabalhar, ambas envolvem a construção de um dicionário de palavras para uma coleção de documentos. A primeira é a abordagem do dicionário local, usando apenas as palavras que aparecem nos documentos classificados como pertencentes a determinada categoria. Isso permite que cada dicionário seja relativamente pequeno ao custo de precisar construir N deles, onde há N categorias. A segunda abordagem é construir um dicionário global, que inclui todas as palavras que ocorrem pelo menos uma vez em qualquer um dos documentos (SKANSI, 2018).

A seguir, na Tabela 1 apresenta um exemplo de saco de palavras para um conjunto de mídia social simples.

Tabela 1 – Exemplo de entrada de saco de palavras

Usuário	Comentário
S. A	produto bom
A. V	prazo de entrega bom
E. F	produto bom, prazo de entrega ruim
P. M	prazo de entrega bom, produto ruim

Fonte: Autora (2024).

Para criar um saco de palavras a partir dos comentários, é necessário realizar duas etapas. A primeira apenas coleta todas as palavras que ocorrem e as transforma em recursos, ou seja, coleta as palavras e cria colunas a partir delas, e a segunda, escreve nos valores reais, conforme a Tabela 2:

Tabela 2 – Exemplo de saco de palavras

Usuário	produto	bom	prazo	de	entrega	ruim
S. A	1	1	0	0	0	0
A. V	0	0	1	1	1	1
E. F	1	1	1	1	1	1
P. M	1	1	1	1	1	1

Fonte: Autora (2024).

De qualquer forma, é importante considerar também a redução do tamanho do espaço de recursos (ou seja, o conjunto de palavras incluídas no dicionário) (BRAMER, 2016). Neste sentido, deve-se remover a pontuação que também facilita o caminho para uma segmentação de palavras adequada. O processo em que um texto é segmentado em sentenças e palavras, formando pedaços é chamado de *Tokenização*, normalmente é o primeiro passo da etapa de pré-processamento, reservando unidades mínimas de texto para análise.

3.3 *Machine Learning*

Machine Learning é uma área da inteligência artificial que se baseia em técnicas estatísticas e computacionais para dar aos computadores a capacidade de "aprender" com base em dados, sem serem explicitamente programados (RUSSELL, 2020).

Existem dois tipos de aprendizado de máquina: a aprendizagem supervisionada e a não supervisionada. Na aprendizagem não supervisionada, os algoritmos procuram por padrões em registros que possuem características semelhantes, comparando os valores de seus atributos. É frequentemente utilizada em situações de agrupamento (também conhecido como *clusterização*) ou para reduzir a complexidade de conjuntos de dados com várias variáveis. Quando aplicada ao agrupamento, os dados são avaliados com base em suas semelhanças ou diferenças, sendo organizados em grupos por meio de métodos de medição de distância, como a distância quadrática euclidiana (JAMES et al., 2013).

Já a aprendizagem supervisionada envolve usar dados que já têm resultados conhecidos para treinar um modelo, geralmente rotulados por especialistas ou obtidos experimentalmente. Para criar esse tipo de modelo, primeiro deve-se escolher parte dos dados para treinamento e ajuste, enquanto outra parte é reservada para testar o quão bem o modelo funciona.

Em resumo, com relação ao uso de algoritmos de *machine learning*, a maioria dos problemas de análise pode ser inserido em duas categorias principais: aprendizado supervisionado, em que o desfecho de um conjunto de dados é conhecido, ou seja, existe um valor da variável resposta a ser predito; e o aprendizado não supervisionado, em que não existe uma variável resposta específica, por exemplo no caso de identificar populações parecidas de acordo com suas similaridades ou reduzir a dimensionalidade de um conjunto de dados (JAMES et al., 2013).

3.3.1 Máquinas de Vetores de Suporte

As técnicas de SVM são métodos de aprendizado de máquina amplamente aplicados em tarefas de classificação e regressão. Esses métodos são fundamentados em hiperplanos que separam as diferentes classes de decisão nos espaços de características (VAPNIK, 1998).

Em situações em que os conjuntos de dados são linearmente separáveis no espaço de característica, o SVM constrói um hiperplano que idealmente divide ambas as classes em duas regiões distintas. Quando os dados são não-lineares, o SVM recorre a funções do tipo kernel, responsáveis por mapear os dados em uma dimensão superior (MOHAMMADI et al., 2021). Adicionalmente, os algoritmos de SVM podem ser aplicados tanto em tarefas de classificação binárias quanto em multiclasses (GÉRON, 2019).

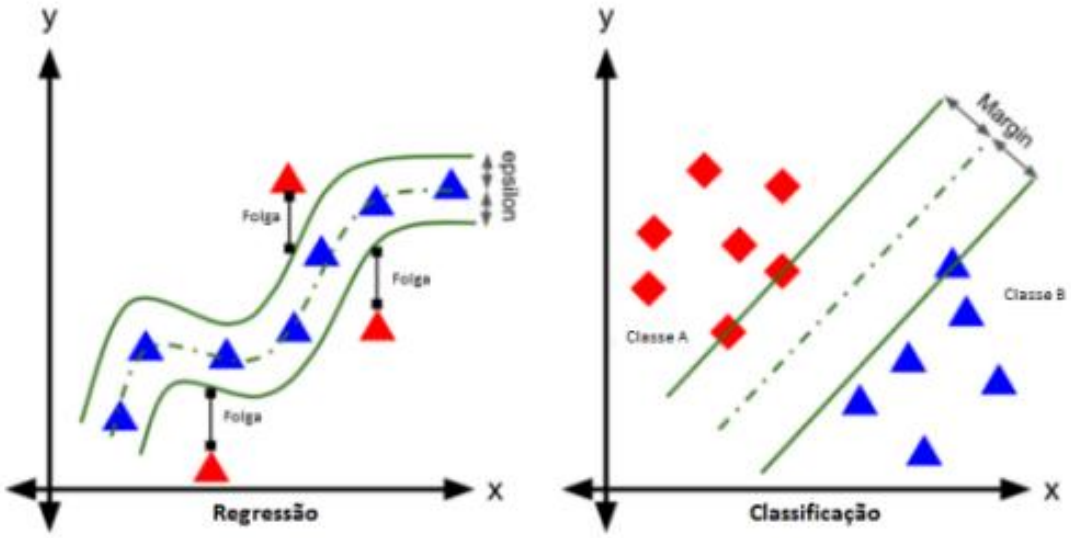
Na construção da SVM, tanto para problemas de regressão quanto para problemas de classificação, são definidos os parâmetros ‘kernel’, ‘C’ e ‘gamma’. O kernel é uma função que permite transformar os dados em um espaço de características de maior dimensão, facilitando a separação entre as classes. Existem diferentes tipos de função kernel: linear, polinomial, sigmóide e RBF. A escolha do kernel adequado depende da natureza dos dados e do problema a ser resolvido (GÉRON, 2019).

O parâmetro C controla a compensação entre a maximização da margem entre as classes e a minimização do erro de classificação. Um valor pequeno de C resulta em uma margem mais ampla, o que pode levar a uma melhor generalização. No entanto, um valor muito pequeno pode resultar em uma margem muito ampla, permitindo muitos erros de classificação. Um valor de C grande resulta em uma margem mais estreita, forçando o modelo a se ajustar aos dados de treinamento, o que pode levar ao *overfitting* (GÉRON, 2019).

Por fim, o parâmetro gamma é usado especificamente no kernel RBF e determina a forma da função de base radial. Um valor pequeno de gamma resulta em uma função RBF mais flexível, permitindo que a fronteira de decisão se adapte melhor aos dados. No entanto, um valor muito pequeno de gamma pode levar ao *overfitting*. Por outro lado, um valor grande de 8 gamma resulta em uma função RBF mais rígida, o que pode levar a um modelo menos flexível e com menor capacidade de adaptação aos dados (GÉRON, 2019).

Uma representação do funcionamento da técnica de SVM, tanto em tarefas de classificação quanto em tarefas de regressão, é mostrada na Figura 2.

Figura 2 - SVM para regressão e classificação.



Fonte: Adaptado de ACHSAN (2019).

Conforme Smola et al. (2000), as SVM são técnicas bastante atrativas devido à sua notável capacidade de generalização, obtendo resultados satisfatórios em tarefas de classificação e regressão para conjuntos de dados que não fazem parte do conjunto utilizado no treinamento. Além disso, apresentam robustez em situações com alta dimensionalidade e possuem fundamentação teórica sólida nos campos da matemática e estatística.

No contexto de aprendizado de máquina supervisionado, o conceito de generalização se refere à habilidade de um modelo em adaptar-se e efetuar previsões precisas em dados não vistos durante o treinamento. Em outras palavras, um modelo com boa generalização é capaz de aplicar o conhecimento obtido durante o treinamento a novos exemplos, apresentando um bom desempenho (GÉRON, 2019).

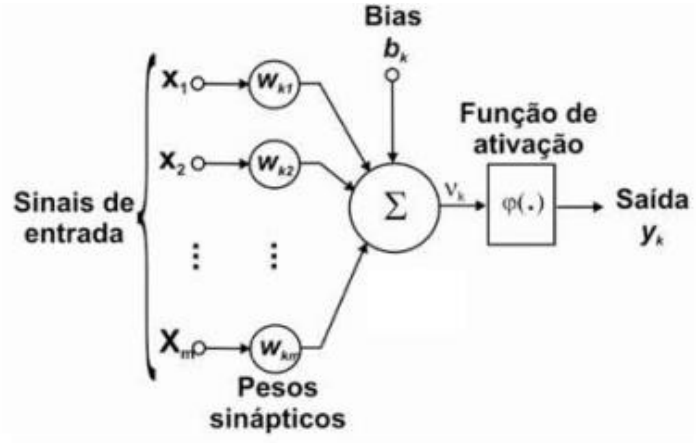
3.3.2 Redes Neurais Artificiais

Na literatura, as redes neurais artificiais têm sido amplamente estudadas e aplicadas em diversos campos de pesquisa. Essas redes foram desenvolvidas com base no conhecimento do funcionamento do cérebro humano, buscando simular o processo de aprendizagem e processamento de informações realizados pelo cérebro (HAYKIN, 2001). As redes neurais são compostas por unidades de processamento simples, denominadas neurônios, que se comunicam entre si através de conexões sinápticas. Essas conexões sinápticas são caracterizadas por pesos sinápticos que têm a capacidade de armazenar o conhecimento adquirido durante o processo de

aprendizagem supervisionada. Ajustando-se os pesos sinápticos, é possível alcançar objetivos pré-estabelecidos, permitindo que as redes neurais aprendam padrões e realizem generalizações. Isso resulta na capacidade de produzir saídas satisfatórias para entradas que não fazem parte do conjunto de treinamento na etapa de aprendizagem (HAYKIN, 2001).

As redes neurais podem ser compostas por um ou mais neurônios, responsáveis pelo processamento de informações. O modelo neuronal é constituído por três componentes fundamentais: um conjunto de sinapses, associadas a pesos; um somador, responsável por somar os sinais de entrada ponderados pelos pesos sinápticos; e uma função de ativação que restringe a amplitude do sinal de saída do neurônio (HAYKIN, 2001). A Figura 3 ilustra um modelo de neurônio com os sinais de entrada ($x1$, $x2$, $x3$, …, xm), as sinapses ($wk1$, $wk2$, $wk3$, …, wkm) a função de ativação $f(vk\ )$ e o sinal de saída do neurônio.

Figura 3 – Modelo de um neurônio artificial



Fonte: Extraído de Haykin (2001)

Matematicamente, um neurônio k pode ser representado pela equação (1):

$$y_k = f\left(\sum_{j=1}^{m} w_{kj} x_j\right) \tag{1}$$

em que xj é o sinal de entrada j, wkj são os pesos sinápticos associados ao sinal de entrada j do neurônio k, e yk é o sinal de saída do neurônio k (HAYKIN, 2001).

Alguns hiperparâmetros importantes das RNA incluem a quantidade de camadas, o número de neurônios em cada camada, a função de ativação, o otimizador e o tamanho do lote. A quantidade de neurônios e o número de camadas definem as diferentes arquiteturas da RNA.

A função de ativação é aplicada aos neurônios da rede para introduzir não-linearidades no modelo. As funções de ativação mais comuns são: ‘tanh’ (tangente hiperbólica), ‘sigmoid’ (função sigmoide), ‘linear’ (função de identidade) e ‘relu’ (*Rectified Linear Unit*) (GÉRON, 2019).

O otimizador é o algoritmo utilizado para atualizar os pesos da rede neural durante o treinamento, minimizado a função de custo. Alguns dos otimizadores mais comuns são: ‘SGD’ (*Stochastic Gradient Descent*), ‘RMSProp’ (*Root Mean Square Propagation*), ‘Adagrad’ (*Adaptive Gradient*) e ‘Adam’ (*Adaptive Moment Estimation*). Já o tamanho do lote (*batch size*) especifica a quantidade de amostras de treinamento empregadas para atualizar os pesos em cada iteração do algoritmo de otimização (GÉRON, 2019).

3.3.3 Redes Neurais Artificiais x Máquinas de Vetores de Suporte

Uma das principais diferenças entre as RNA e as SVM está na forma como elas abordam a aprendizagem e a generalização a partir dos dados. As RNA são modelos flexíveis e adaptáveis que podem aprender a partir de dados não lineares e complexos, ajustando-se aos padrões intrínsecos nos dados. Por outro lado, as SVM procuram encontrar o hiperplano de separação ótimo que maximiza a margem entre as diferentes classes, buscando uma solução que seja mais robusta e generalizável para novos dados (VAPNIK, 2013).

Em termos de implementação e aplicação prática, as RNA e as SVM têm vantagens e desvantagens distintas. As RNA são mais flexíveis e podem lidar com uma variedade de problemas de aprendizado de máquina, mas podem ser computacionalmente intensivas e exigir grandes conjuntos de dados para treinamento eficaz. As SVM, por outro lado, são mais eficientes em termos computacionais e geralmente fornecem uma solução mais simples e interpretável para problemas de classificação, mas podem não ser tão eficazes em problemas de alta dimensionalidade ou com conjuntos de dados não lineares (BISHOP, 2016).

É importante ressaltar que tanto a Rede Neural Artificial (RNA) quanto o Máquina de Vetores de Suporte (SVM) demonstram eficácia que está intrinsecamente ligada à qualidade e quantidade dos dados disponíveis. Portanto, é imperativo validar sua performance por meio da aplicação em conjuntos de dados relevantes.

3.4 Métricas de desempenho

Métricas de desempenho são utilizadas para avaliar a qualidade dos modelos de aprendizado de máquina, permitindo compreender o desempenho deles na realização de 9 previsões (GÉRON, 2019). Em problemas de classificação utilizando máquinas de aprendizado, as principais métricas incluem acurácia, precisão, recall e F1 score (HARRISON, 2019).

A acurácia mede a proporção de previsões corretas em relação ao total de previsões. É uma métrica comum para avaliar o desempenho geral de um modelo, porém pode ser enganosa em situações com classes desbalanceadas (GÉRON, 2019).

Precisão é a proporção de verdadeiros positivos (TP) em relação à soma dos verdadeiros positivos e falsos positivos (FP). Em outras palavras, a precisão indica a fração de previsões corretas da classe positiva em relação a todas as previsões feitas para essa classe (GÉRON, 2019).

Recall, também conhecido como sensibilidade, é a proporção de verdadeiros positivos (TP) em relação à soma dos verdadeiros positivos e falsos negativos (FN). O recall indica a fração de exemplos positivos corretamente identificados pelo modelo em relação a todos os exemplos positivo (HARRISON, 2019).

F1 score é uma métrica que combina precisão e recall em um único valor, dando igual importância a ambas. O F1 score é calculado como a média harmônica entre precisão e recall, e é uma medida útil quando se deseja equilibrar os dois aspectos, especialmente em casos de classes desbalanceadas (HARRISON, 2019).

Por último, há também a tabela de confusão, que mostra os resultados da classificação. As linhas indicam as previsões feitas para uma classe, enquanto as colunas representam as instâncias reais dessa classe. Essas tabelas são simples de interpretar quando o sistema tem apenas duas classes. Se houver mais classes, elas podem ser simplificadas para duas: a classe de interesse (positiva) e todas as outras podem ser agrupadas em uma única classe (negativa). O modelo da matriz de confusão é representado conforme a Figura 4.

Figura 4 – Matriz de confusão

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: (Nogare, 2020)

Resumindo, e como se pode verificar na figura anterior, podemos considerar a matriz de confusão como uma tabela com duas linhas e 2 colunas que regista o número de Verdadeiro Negativo (TN), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Positivo (TP) (POWERS, 2007).

- TN – Instâncias negativas classificadas como negativas;
- FP – Instâncias negativas classificadas como positivas;
- FN – Instâncias positivas classificadas como negativas;
- TP – Instâncias positivas classificadas como positivas.

3.5 Linguagem *Python*

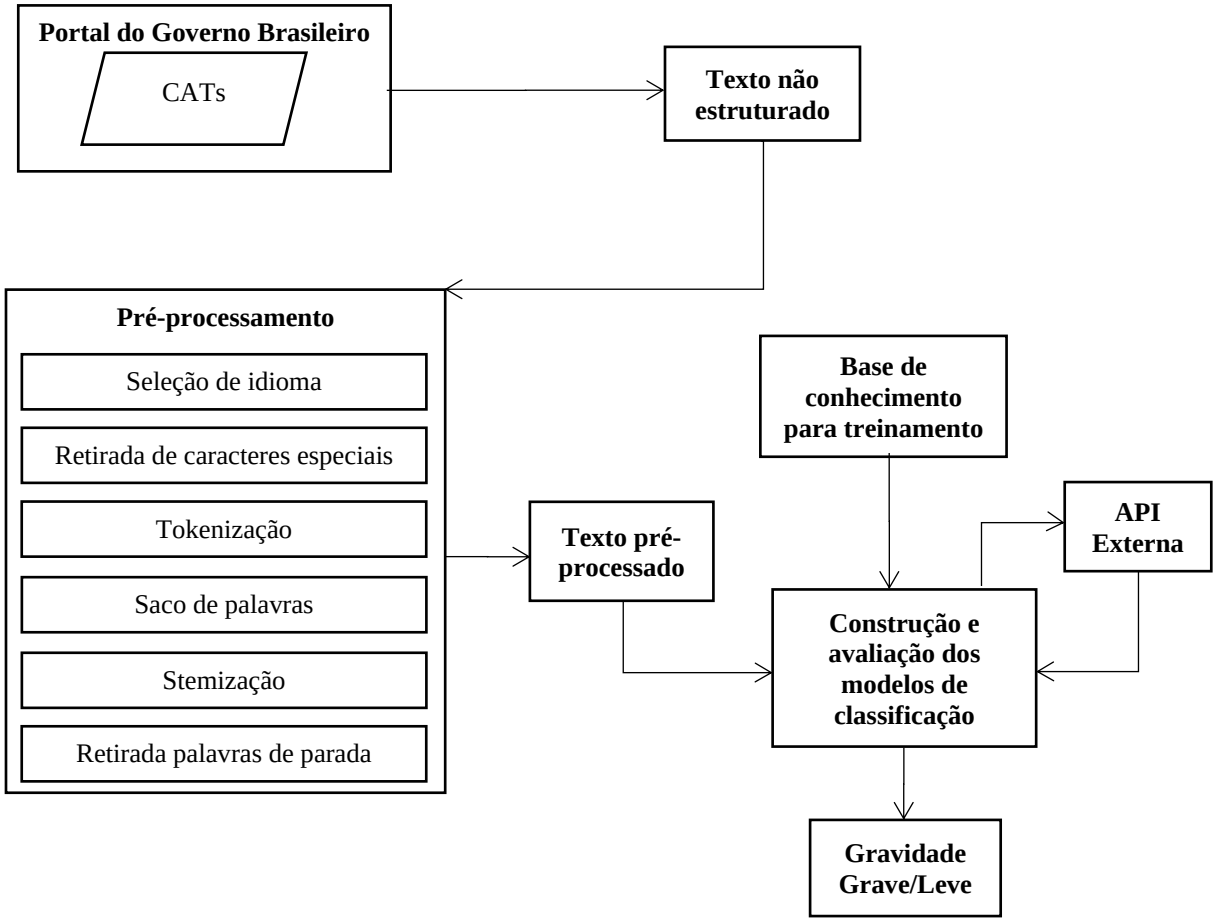
Por conta da linguagem de programação *Python* se tornar cada vez mais popular, principalmente na área da engenharia, e ser gratuita, aliado a um desenvolvimento crescente de novas bibliotecas, graças a uma ampla comunidade, esse trabalho busca implementar diversas bibliotecas em *Python* para a construção dos modelos de classificação usando RNA e SVM.

Algumas bibliotecas foram utilizadas para realização deste trabalho, incluindo a *NumPy*, para computação científica com cálculos de álgebra linear, Pandas, para leitura e manipulação de dados, NLTK para realizar o pré-processamento do texto e *Scikit-learn*, para aplicação da RNA e SVM e das métricas de avaliação.

4. METODOLOGIA

A metodologia utilizada para este trabalho pode ser resumida conforme a Figura 5, primeiramente foi construída a rotina computacional do *Text Mining* para etapa de pré-processamento e seguidamente, a construção das máquinas de aprendizado em RNA e SVM para a etapa de classificação para identificação da gravidade da lesão e, logo após uma análise comparativa entre às duas técnicas de máquinas de aprendizado.

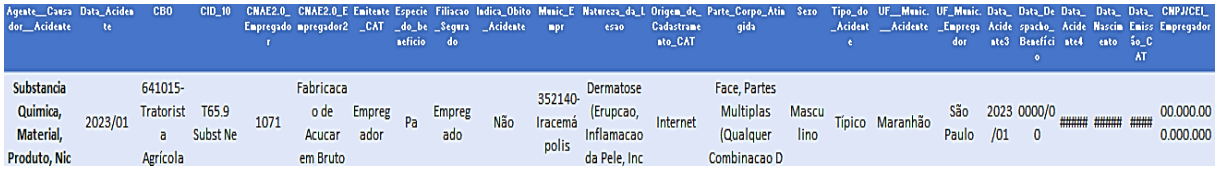
Figura 5 - Fluxo de classificação



Fonte: Autora (2024).

Os dados utilizados foram retirados do Portal de Dados Abertos do governo brasileiro, onde foram utilizados para treinamento da máquina Comunicações de Acidentes de Trabalho (CATs) de janeiro de 2023 de diferentes setores industriais. O arquivo com todas as CATs do mês estava em formato MS Excel, seguindo o modelo da Figura 6.

Figura 6 – Exemplo de uma CAT



Agente__Causador__Acidente	Data_Acidente	CBO	CID_10	CNAE2.0_Empregador	CNAE2.0_Empregador2	Emitente_CAT	Especie_do_beneficio	Filiacao_Segurado	Indica_Obito_Acidente	Munic_Empr	Natureza_da_Lesao	Origem_de_Cadastramento_CAT	Parte_Corpo_Atingida	Sexo	Tipo_do_Acidente	UF__Munic._Acidente	UF_Munic._Empregador	Data_Acidente3	Data_Despacho_Benefício	Data_Acidente4	Data_Nascimento	Data_Emissão_CAT	CNPJ/CEI_Empregador
Substancia Quimica, Material, Produto, Nic	2023/01	641015-Tratorista Agrícola	T65.9 Subst Ne	1071	Fabricacao de Acucar em Bruto	Empregador	Pa	Empregado	Não	352140-Iracemápolis	Dermatose (Erupcao, Inflamacao da Pele, Inc	Internet	Face, Partes Multiplas (Qualquer Combinacao D	Masculino	Típico	Maranhão	São Paulo	2023/01	0000/00	#####	#####	####	00.000.000.000.000

Fonte: Autora (2024).

Foi analisada para teste uma amostra de 400 casos de acidentes de trabalho registrados que tinham a CNAE do empregador em extração, fabricação, manutenção e produção dos diferentes tipos de produtos, onde a intenção é investigar o impacto do pré-processamento em dois tipos de aprendizado de máquina, a RNA e o SVM.

Para os experimentos I e II foram implementados algoritmos para fazer a classificação de gravidade da lesão e os modelos avaliados conforme as métricas descritas na Seção 3.4. Os classificadores utilizados foram construídos usando duas técnicas de *Machine Learning*: RNA e SVM. Será analisado o impacto da classificação com o pré-processamento realizado com a técnica *Bag Of Words*.

Os dados foram divididos em 30% para teste e 70% para treinamento. A ferramenta utilizada nestes experimentos foi o *Google Colab*, para implementação em *Python*, com bibliotecas tais como NLTK, *Pandas data analysis* e *Scikit-learn machine learning.*

4.1 Tratamento dos dados

Nessa etapa, os dados foram importados por meio da biblioteca pandas e foram armazenados em um objeto chamado *DataFrame*, que é bastante similar a uma planilha do *MS Excel* com suas linhas e colunas.

Para evitar erros durante a modelagem, foi verificada a existência de dados duplicados, nulos ou faltantes. E uma vez que eles existam, estes foram apagados, pois não possuem informações relevantes para a análise do modelo.

Inicialmente, na etapa de pré-processamento, são removidos todos os caracteres que não são letras do alfabeto, substituindo-os por espaços em branco. Isso visa a limpar o texto de quaisquer caracteres especiais que possam interferir na análise. Posteriormente, o texto é convertido para minúsculas, garantindo que não haja distinção entre palavras escritas em maiúsculas e minúsculas. Em seguida, o texto é dividido em palavras individuais, facilitando a manipulação e análise subsequente. Para garantir que apenas palavras relevantes sejam

consideradas, são removidas as *stopwords* da língua portuguesa, ou seja, palavras comuns que geralmente não contêm informações úteis para a análise. Por fim, é aplicada a *lematização* em cada palavra, reduzindo-as à sua forma base, o que ajuda a consolidar termos semelhantes, facilitando a análise posterior. As palavras processadas são então reagrupadas em uma única *string* e adicionadas ao *corpus* de texto, que servirá como base para análises posteriores.

Na Tabela 3, encontramos um exemplo do resultado de textos selecionados com as informações do agente causador do acidente, a natureza da lesão e a parte do corpo atingida de uma CAT após as técnicas de pré-processamento utilizadas no decorrer deste trabalho, aqui conseguimos obter maior percepção de como atuam os filtros. Com estas técnicas, a redução de um texto livre à sua essência é fundamental para a construção do modelo de classificação.

Tabela 3 - Exemplo de aplicação pré-processamento em CAT

CAT sem Filtros de Pré-Processamento de texto	CAT após os Filtros de Pré-Processamento de texto
Substancia Quimica, Material, Produto, Nic - Queimadura Quimica (Lesao de Tecido Provoca - Olho (Inclusive Nervo Otico e Visao)	substância química material produto queimadura química lesão olho

Fonte: Autora (2024).

4.2 Construção do modelo de classificação

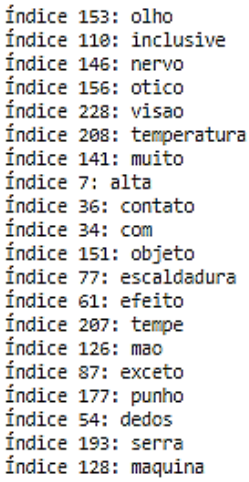
Foi realizado o processo de construção e avaliação dos modelos de classificação utilizando RNA e SVM, com o objetivo de determinar qual técnica de *machine learning* melhor se adequa à classificação da gravidade de lesões em acidentes de trabalho, considerando as características dos dados obtidos das CAT.

4.2.1 Preparação dos dados

Antes da construção dos modelos, os dados foram preparados para o treinamento e teste. Foram selecionadas as colunas "Agente Causador do Acidente", "Natureza da Lesão" e "Parte do Corpo Atingida" como características relevantes para a classificação da gravidade da lesão, que foi categorizada como "Leve" ou "Grave". Além disso, foi utilizada a técnica *Bag of Words*

para representar o texto, transformando-o em um formato numérico compreensível pelos algoritmos de *machine learning*, conforme apresentado na Figura 7.

Figura 7 – Exemplos de índices obtidos após utilizar a técnica *Bag of Words*



Fonte: Autora (2024).

4.2.2 Construção dos modelos

- ***Support Vector Machine* (SVM)**: O primeiro modelo construído foi baseado em SVM, uma técnica de aprendizado supervisionado que busca encontrar o hiperplano ótimo de separação entre as classes. Para isso, utilizamos a implementação SVC disponível na biblioteca *scikit-learn* em Python.
- **Rede Neural Artificial (RNA)**: Em seguida, implementamos um modelo de classificação baseado em Rede Neural Artificial. Essa abordagem é conhecida por sua capacidade de capturar padrões complexos nos dados, adaptando-se bem a diferentes tipos de problemas. Utilizamos a implementação MLPClassifier disponível na biblioteca scikit-learn para construir e treinar a rede neural.

4.2.3 Avaliação dos modelos

Para avaliar o desempenho dos modelos, utilizamos a matriz de confusão. Esta matriz permite visualizar a performance dos modelos na classificação das instâncias em suas respectivas classes ("Leve" ou "Grave"). Comparamos os resultados obtidos pelos modelos SVM e RNA, analisando métricas como precisão, recall e F1-score.

5. RESULTADOS E DISCUSSÃO

Os resultados apresentados aqui, que dependem de máquina local para execução, foram realizados em um notebook Lenovo Ideapad S145, com 8Gb de memória e 1TB de disco.

5.1 Experimento 1 – SVM

O conjunto de dados de CATs retiradas do Portal de Dados Abertos foi submetido ao algoritmo escrito em Python, com as devidas parametrizações de idioma em português. Os resultados são apresentados conforme a Tabela 4, onde modelo apresentou uma acurácia de 80%, indicando um alto grau de precisão das previsões realizadas. Além disso, o recall e o F1-Score, que são medidas de sensibilidade e equilíbrio, respectivamente, o primeiro apresentou um desempenho moderadamente bom na recuperação de exemplos positivos e o segundo entregou também um valor de (80%). Esses resultados sugerem que o modelo é eficiente na minimização de falsos positivos e falsos negativos.

Tabela 4 – Dados de métricas de desempenho com SVM

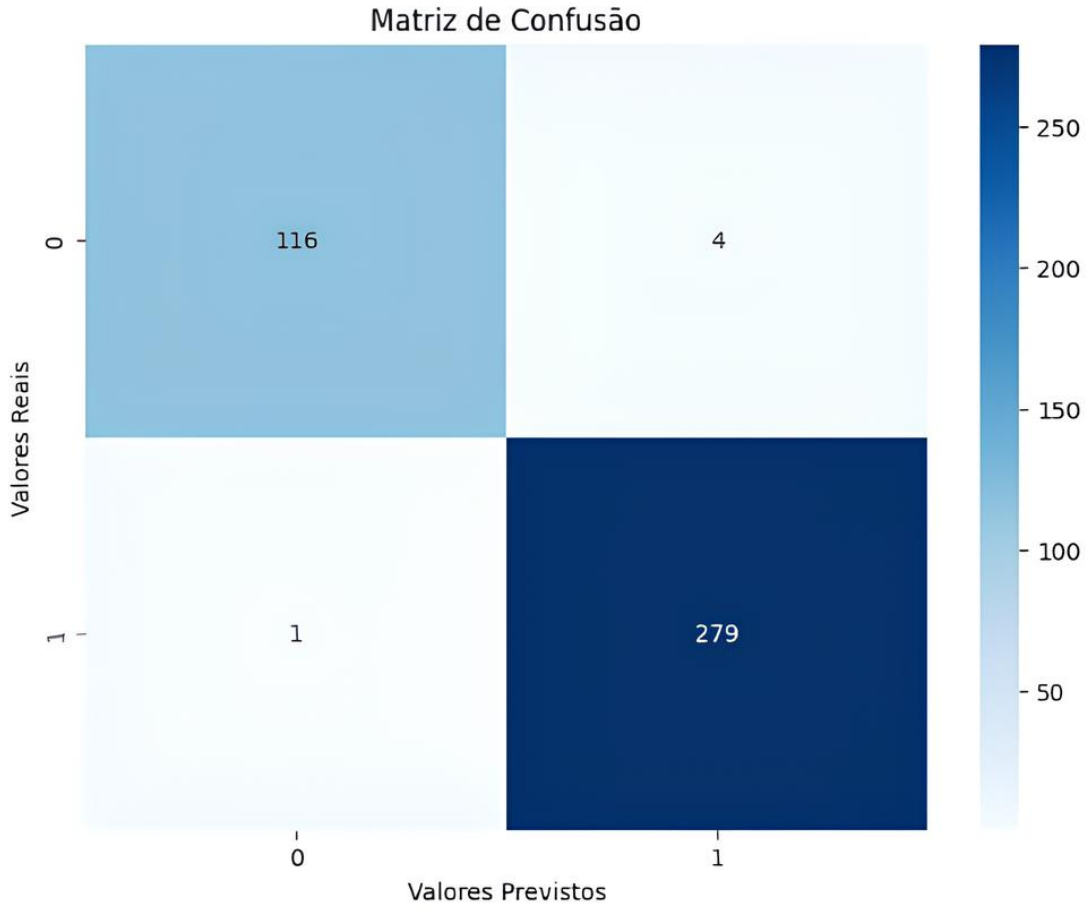
Métrica	Valores obtidos
Acurácia	0.80000000
Precisão	1.00000000
Recall	0.66666667
F1_score	0.80000000

Fonte: Autora (2024).

A Figura 8 apresenta a matriz de confusão obtida ao avaliar o desempenho do modelo no conjunto de dados de teste. A matriz de confusão fornece um resumo das previsões corretas e incorretas realizadas pelo modelo durante a etapa de validação, permitindo verificar a capacidade de generalização do modelo para conjuntos de dados não vistos na etapa de treinamento.

Na análise de sua matriz confusão, pode-se observar que o uso do SVM gerou um resultado de 95,83% de classificação correta, conforme a Figura 8, onde se esperava 120 classificações de gravidade da lesão como “Grave”, foram classificadas 116 corretamente e onde era esperado 280 classificações de gravidade da lesão como “Leve” foram classificadas 279 corretamente.

Figura 8 – Matriz confusão da análise com uso de SVM – CAT



Fonte: Autora (2024).

Os resultados obtidos na avaliação do modelo de classificação SVM para a detecção de gravidade da lesão são promissores. O modelo alcançou uma porcentagem de classificação correta de 95,83%, indicando que em 95,83% dos casos o modelo classificou corretamente os documentos. Isso sugere uma alta precisão na identificação das categorias ou classes alvo.

Já o recall do modelo, que mede a proporção de exemplos positivos corretamente identificados em relação ao número total de exemplos positivos no conjunto de dados, foi observado em 66,67%. Embora o modelo tenha uma certa capacidade de recuperar as instâncias positivas e não existam trabalhos anteriores na literatura para comparar os dados obtidos, pode ser um valor considerado relativamente baixo.

Primeiramente, a natureza intrinsecamente complexa do problema de classificação pode ter dificultado a distinção entre as classes, levando a uma identificação imprecisa das instâncias positivas. Além disso, um possível desbalanceamento entre as classes, com uma classe sendo significativamente mais prevalente do que a outra, pode ter influenciado a capacidade do

modelo em identificar corretamente as instâncias da classe minoritária. O impacto de ruído nos dados também não pode ser descartado, pois a presença de ruído pode levar a confusão entre as classes e, consequentemente, reduzir o desempenho do modelo. Além disso, a complexidade inerente do problema pode exigir abordagens mais sofisticadas ou ajustes adicionais para melhorar o recall. Considerando esses fatores, uma análise mais aprofundada pode ser necessária para identificar estratégias de melhoria do desempenho do modelo, visando alcançar um recall mais satisfatório.

5.2 Experimento 2 – RNA

Os testes iniciais utilizando as CATs retiradas do Portal de Dados Abertos demonstraram bons resultados dos valores obtidos, tendo apenas o Recall com um valor abaixo da média dentre as métricas utilizadas, dando 66,67%, como apresentado na Tabela 5.

Tabela 5 – Dados de métricas de desempenho com RNA

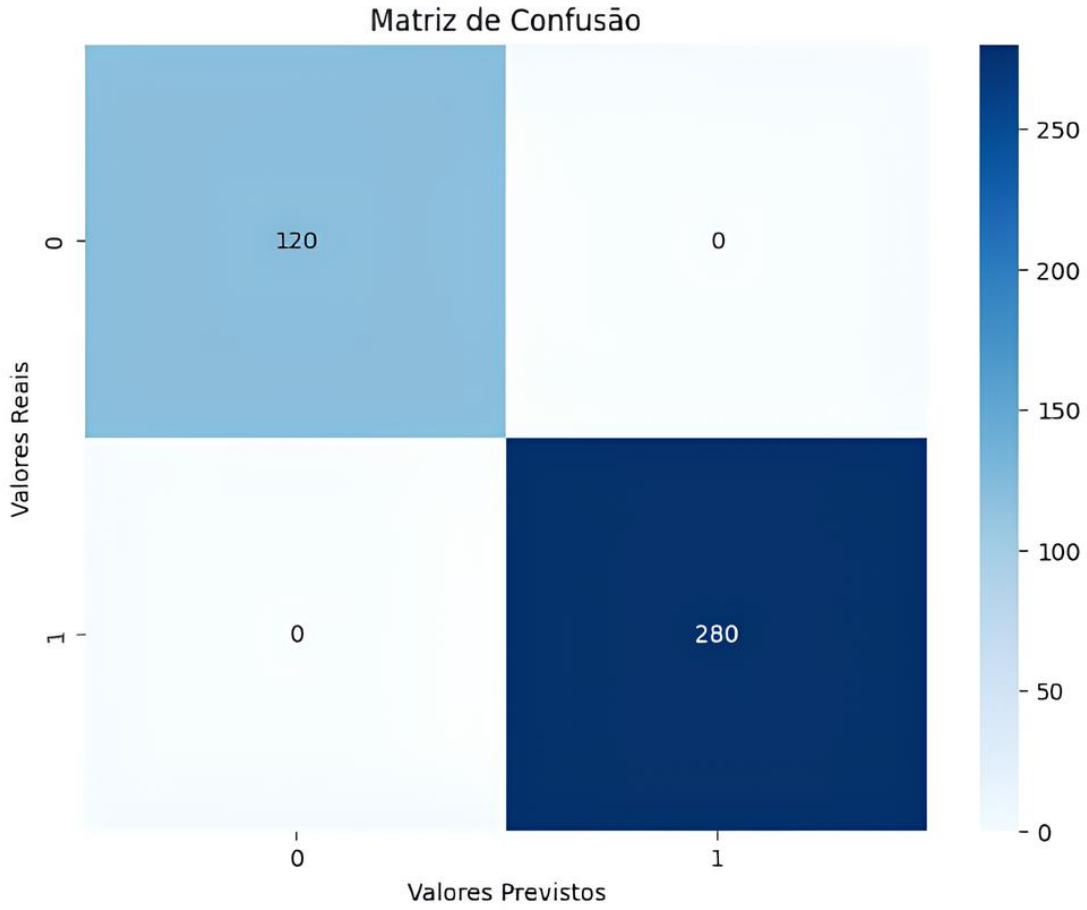
Métrica	Valores obtidos
Acurácia	0.80000000
Precisão	1.00000000
Recall	0.66666667
F1_score	0.80000000

Fonte: Autora (2024).

A Figura 9 mostra a matriz de confusão utilizada para avaliar o desempenho do modelo no conjunto de dados de teste. Essa matriz fornece um resumo das previsões corretas e incorretas realizadas pelo modelo de aprendizado de máquina durante a etapa de validação, comparando os valores previstos com os valores reais das classes. Isso permite verificar a capacidade de generalização do modelo para conjuntos de dados não vistos na etapa de treinamento.

Quando analisada a matriz de confusão desta classificação, conforme apresentado na Figura 9, observa-se que o algoritmo teve um bom desempenho na identificação das opiniões positivas e negativas. Dos 120 casos de gravidade da lesão “Grave”, o algoritmo conseguiu classificar corretamente 120 casos. Da mesma forma, para os 280 casos esperados de gravidade da lesão “Leve”, o algoritmo classificou corretamente 280 casos. Isso resulta em um índice de acerto de 100%, demonstrando a capacidade do modelo em classificar com precisão uma grande proporção das instâncias graves e leves.

Figura 9 – Matriz confusão da análise com uso de RNA – CAT



Fonte: Autora (2024).

A análise da matriz de confusão revelou que o modelo de RNA alcançou um desempenho notável na classificação dos dados. Com uma taxa de acerto de 100%, o modelo foi capaz de classificar corretamente todas as instâncias do conjunto de dados. Esse resultado é altamente encorajador, indicando uma capacidade excepcional da RNA em distinguir entre as diferentes classes ou categorias.

O sucesso do modelo pode ser atribuído a vários fatores, incluindo a robustez da arquitetura da RNA, a qualidade e quantidade dos dados utilizados para o treinamento, e a eficácia dos hiperparâmetros selecionados. A precisão excepcional alcançada pelo modelo demonstra sua utilidade potencial em aplicações práticas, onde a precisão da classificação é de extrema importância. No entanto, é importante notar que, embora o modelo tenha alcançado uma taxa de acerto perfeita neste conjunto de dados específico, é essencial realizar uma

validação adicional em conjuntos de dados independentes para avaliar a generalização do modelo e sua aplicabilidade em diferentes contextos.

Embora a matriz de confusão revele um desempenho perfeito do modelo de Rede Neural Artificial (RNA) em termos de taxa de acerto, com um índice de 100%, o recall do modelo foi observado em 66,67%. Este valor mais baixo indica que o modelo teve dificuldade em recuperar corretamente todas as instâncias positivas no conjunto de dados. A discrepância entre a taxa de acerto e o recall sugere que o modelo pode ter tido uma tendência a classificar incorretamente algumas instâncias positivas como negativas.

5.3 Comparação dos modelos de *Machine Learning*

A comparação entre os modelos de Máquinas de Vetores de Suporte (SVM) e Rede Neural Artificial (RNA) revelou resultados notáveis e similares em termos de métricas de desempenho. Ambas as técnicas demonstraram um desempenho matemático equivalente, com acurácia, precisão, recall e F1_score atingindo valores idênticos de 0.80000000. Esse resultado indica que, em termos gerais, as duas abordagens foram igualmente eficazes na classificação dos dados de Comunicação de Acidente de Trabalho.

No entanto, ao analisar as matrizes de confusão, surgem algumas distinções significativas entre os dois métodos. O SVM apresentou uma porcentagem de classificação correta de 95,83%, enquanto a RNA alcançou uma taxa de 100%. Isso sugere que, embora ambos os modelos tenham sido capazes de classificar corretamente a maioria das instâncias, a RNA obteve uma precisão um pouco mais elevada em suas previsões, não cometendo nenhum erro de classificação no conjunto de dados analisado, o que pode ser atribuído a adaptabilidade da técnica aos diferentes tipos de dados, como os textuais.

É importante ressaltar que, apesar das diferenças na precisão das classificações, ambas as técnicas foram aplicadas à mesma base de dados pré-processada de Comunicação de Acidente de Trabalho. Isso indica que as diferenças de desempenho observadas são provavelmente atribuíveis às características específicas dos algoritmos de SVM e RNA, bem como à sua capacidade de lidar com o conjunto de dados em questão.

Em resumo, enquanto o SVM e a RNA demonstraram resultados semelhantes em termos de métricas de desempenho global, as análises detalhadas das matrizes de confusão destacam nuances distintas no desempenho de cada técnica. A escolha entre os dois métodos pode depender das necessidades específicas do problema e das características dos dados em questão.

6. CONCLUSÃO

A aplicação de técnicas de *machine learning* na automação da classificação de Comunicações de Acidentes de Trabalho (CATs) em ambientes industriais revelou-se uma abordagem altamente eficaz e promissora. Os resultados obtidos demonstraram que tanto o modelo de Máquinas de Vetores de Suporte (SVM) quanto a Rede Neural Artificial (RNA) foram capazes de atingir altos níveis de acurácia, precisão e F1_score, indicando a sua viabilidade para automatizar tarefas cruciais de classificação em ambientes industriais. Esses resultados corroboram a importância e a eficácia do uso de técnicas de *machine learning* para lidar com questões de segurança ocupacional e saúde do trabalhador.

Ao comparar o desempenho dos modelos SVM e RNA, observou-se uma diferença distinta na precisão das classificações, destacada pelas matrizes de confusão. Enquanto o SVM apresentou uma porcentagem de classificação correta de 95,83%, a RNA alcançou uma taxa de 100%. Essa divergência destaca nuances importantes no desempenho dos modelos e destaca a necessidade de uma análise cuidadosa ao selecionar a abordagem mais adequada para a automação da classificação de CATs em ambientes industriais.

Além de melhorar a eficiência da classificação de acidentes de trabalho, a automação por meio de técnicas de *machine learning* oferece oportunidades significativas para otimizar o ambiente de trabalho industrial. Ao identificar e classificar rapidamente os acidentes de trabalho, as organizações podem implementar medidas preventivas e corretivas de forma mais ágil, reduzindo os riscos de acidentes futuros e promovendo um ambiente de trabalho mais seguro e saudável para todos os colaboradores.

No entanto, apesar dos resultados positivos obtidos neste estudo, é importante reconhecer que ainda existem desafios e oportunidades futuras a serem explorados. Refinar e aprimorar continuamente os modelos de *machine learning* para lidar com casos mais complexos e cenários de acidentes de trabalho menos frequentes é essencial. Além disso, a integração de outras fontes de dados e o desenvolvimento de abordagens mais avançadas podem enriquecer ainda mais o processo de automação da classificação de CATs em ambientes industriais.

7. REFERÊNCIAS

Aranha, C., Passos, E. **A Tecnologia de Mineração de Textos**. RESI-Revista Eletrônica de Sistemas de Informação, 2006.

BEIER, G. et al. **Industry 4.0: How it is defined from a sociotechnical perspective and how much sustainability it includes – A literature review**. Journal of Cleaner Production, v. 259, p. 120856, jun. 2020.

BISHOP, C. M. **Pattern Recogniton and Machine Learning**. [S.l.: s.n.], 2006.

BRAMER, M. **Principles of Data Mining.** London: Springer, 2016.

BRASIL. **Lei nº 8.213, de 24 de julho de 1991**. Dispõe sobre os planos de benefícios da previdência social e dá outras providências. Brasília, DF: Presidência da República, 1991.

BRASIL. Ministério da Fazenda. Instituto Nacional do Seguro Social. Empresa de Tecnologia e Informações da Previdência. **Anuário estatístico de acidentes do trabalho**, Brasília, v. 1, p. 1-996, 2022

CARDOSO, Marla. **Aumento de acidentes de trabalho no Brasil em 2022**. Revista Proteção, [S.l.], 2023.

CASTRO, G. B. **Modelo de rede neural bioinspirada para o controle de trânsito urbano**. PUC-SP, 2017.

CHAVES, R. R. **Redes neurais deep learning aplicadas ao reconhecimento facial**. UCS, 2018

di Noia, Antonio & Martino, Alessio & Montanari, Paolo & Rizzi, Antonello. (2020). **Supervised Machine Learning Techniques and Genetic Optimization for Occupational Diseases Risk Prediction**. Soft Computing. 10.1007/s00500-019-04200-2.

DUTRA, V. P. **Redes neurais e o reconhecimento de padrões de texto**. USF, 2011.

ERTEL, W. **Introduction to Artificil Intelligence**. Weingarten: Springer, 2017.

Fernandes, F. T. **Data mining e aprendizado de máquina na área da saúde e segurança no trabalho: uma revisão da literatura.** Revista Brasileira de Saúde Ocupacional, 44 e13, 2019.

Figueiredo, Ernanny. **Classificação de sentimentos em textos de e-commerce utilizando redes neurais artificiais.** Dissertação (Trabalho de Pós-Graduação em Ciência da Computação) - Universidade Estadual do Oeste do Paraná – Unioeste, campus de Cascavel, 2022.

Santos, Cedric Michael d. **Classificação de Documentos com Processamento de Linguagem Natural.** Monografia (Trabalho para Mestrado em Informática e Sistemas) - Instituto Superior de Engenharia de Coimbra, 2015.

SARKAR, S.; MAITI, J. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. Safety Science, v. 131, p. 104900,

nov. 2020. GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow: conceitos, ferramentas e técnicas para construção de sistemas inteligentes.** Rio de Janeiro: Alta Books Editora, 2019.

HARRISON, Matt. **Machine Learning Guia de Referência Rápido: trabalhando com dados estruturados em python.** São Paulo: Novatec Editora Ltda, 2019.

HAYKIN, S. **Redes neurais: princípios e prática.** São Paulo: Bookman, 2001.

Huang, C.; Davis, L. S.; Townshend, J. R. G. **An assessment of support vector machines for land cover classification.** International Journal of Remote Sensing, v. 23, n. 4, p. 725-749, 2002.

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

JUVÊNCIO, Mario Henrique Cosme. **Redes neurais e SVM aplicadas no desenvolvimento de sensores virtuais e detecção e diagnóstico de falhas em processos reacionais complexos**. 51 f. Monografia (Trabalho de Conclusão de Curso em Engenharia Química) - Universidade Federal de Alagoas, Centro de Tecnologia, Maceió, 2023.

Diaz, et al. (2018). **Uma análise comparativa das ferramentas de pré-processamento de dados textuais: NLTK, PreTexT e R**. USP, 2018.

KINGMA, J. L. B. D. P. **Adam: A method for stochastic optimization.** ICLR, 2015.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. **A Review of Machine Learning Algorithms for Text-Documents Classification**. JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, vol. 1, no. 1. 2010.

Li, Chang, et al. **"Machine learning for text mining based on prediction of occupational accidents and safety risk calculation."** Australian Journal of Engineering and Applied Science 13.6 (2020)

McKinney, Wes. **Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython**. Novatec, 2018.

Melgani, F.; Bruzzone, L. **Classification of Hyperspectral Remote Sensing Images with Support Vector Machines**. IEEE Transactions on Geoscience and Remote Sensing, vol. 42, No. 8, August 2004.

MELLO, A. J. T. S. **Uso de técnicas de redes neurais em instrumentação para astronomia.** UFSC, 2014.

Nogare, D. **Medir a performance da descoberta de padrões**. 2020.

Powers, D. M. **Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.** Buchner, 2007.

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

Report, T., & Vincent, K. P. **Mineração de texto Methods for Event Recognition in Stories**. UK, 2005.

Richers, Rosane Schmalz. **Cultura de segurança: estudo exploratório em organização com sistema OHSAS de gestão da saúde e segurança do trabalho.** Catálogo da USP, 2009.

Russel S, Norvig P. **Inteligência artificial**. Rio de Janeiro: Elsevier; 2013.

Silva L, Peres S, Boscarioli C. **Introdução à mineração de dados com aplicações** em R. Rio de Janeiro: Elsevier; 2016.

SKANSI, S. **Introduction to Deep Learning**. Zagreb, Croatia: Springer, 2018.

SOUZA, C. G. e; SOUZA, L. de Mello Tostes e. **O Uso de Redes Neurais Artificiais no Diagnóstico de Doenças Reumatológicas**. UFSC, 2004.

Suthakar U, Magnoni L, Smith DR, Khan A, Andreeva J. **Uma estratégia eficiente para a coleta e armazenamento de grandes volumes de dados para computação**.J Grandes Dados. (2016)

THOMAZ, C. E.; VELLASCO, M. M. **Análise de Tendências de Mercado por Redes Neurais Artificiais.** PUC-RJ, 2016.

VAPNIK, Vladimir N. **Statistical Learning Theory.** New York: Wiley, 1998.

Vapnik, V. N. **The nature of statistical learning theory**. Springer Science & Business Media, 2013.