UNIVERSIDADE ESTADUAL DO MARANHÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

DANIEL HERRERA DE OLIVEIRA LEMOS

ANÁLISE PREDITIVA NA CATEGORIZAÇÃO DE ACIDENTES DE TRABALHO NA EMPRESA VALE S.A. COM A COMPARAÇÃO DE ALGORITMOS DE CLASSIFICAÇÃO

Projeto de pesquisa apresentado ao Programa de Pós-Graduação em Engenharia de Computação e Sistemas como Dissertação da Universidade Estadual do Maranhão.

Orientador: Prof. Dr. Cícero Costa Quarto

São Luís

2024

Lemos, Daniel Herrera de Oliveira.

Análise preditiva na categorização de acidentes de trabalho na Empresa Vale S.A. com a comparação de algoritmos de classificação. / Daniel Herrera de Oliveira Lemos. – São Luís (MA), 2024.

73p.

Dissertação (Mestrado em Engenharia de Computação e Sistemas – PECS)
Universidade Estadual do Maranhão - UEMA, 2024.

Orientador: Prof. Cícero Costa Quarto.

1.Acidentes de trabalho. 2. Mineração de texto. 3. KNIME. 4. CRISP-DM . 5. Aprendizado de máquina. I. Título.

CDU: 331.46

Elaborado por Luciana de Araújo - CRB 13/445

UNIVERSIDADE ESTADUAL DO MARANHÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

Dissertação de Mestrado do Programa de Pós-Graduação em Engenharia de Computação e Sistemas intitulado **Análise preditiva na categorização de acidentes de trabalho na empresa Vale S.A. Com a comparação de algoritmos de classificação** de autoria de Daniel Herrera de Oliveira Lemos, aprovada pela banca examinadora constituída pelos seguintes professores:

gov.br
Documento assinado digitalmente
CICERO COSTA QUARTO
Data: 30/04/2024 23:29:38-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cícero Costa Quarto
Orientador

gov.br
Documento assinado digitalmente
JODELSON AGUILAR SABINO
Data: 04/05/2024 09:32:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Jodelson Aguilar Sabino
ITV

gov.br
Documento assinado digitalmente
MATHEUS CHAVES MENEZES
Data: 30/04/2024 21:27:48-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Matheus Chaves Menezes
UNDB

gov.br
Documento assinado digitalmente
ANTONIO FERNANDO LAVAREDA JACOB JUNIOI
Data: 29/04/2024 16:05:39-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Antonio Fernando Lavareda Jacob Junior
UEMA

SUMÁRIO

RESUMO

Estatísticas da Organização Internacional do Trabalho mostram que ocorre uma morte a cada 15 segundos devido a acidentes de trabalho ou doenças relacionadas à atividade profissional, totalizando 2,3 milhões de mortes por ano no mundo. No Brasil, uma pessoa morre a cada três horas e meia, colocando o país em quarto lugar no ranking global de acidentes de trabalho com vítimas fatais. (OIT, 2023). Com base na introdução, a seguinte dissertação tem como objetivo realizar análise preditiva para a classificação de acidentes de trabalho na Vale S.A. através dos dados coletados, com uma acurácia acima de 80%. Utilizando a metodologia a CRISP-DM (*Cross-Industry Standard Process for Data Mining*), uma abordagem amplamente aceita para projetos de mineração de dados. A metodologia é dividida em seis fases distintas, cada uma com suas próprias etapas e atividades específicas. Com o resultado de 3 algoritmos treinados e com uma acurácia superior a 80% de acertos, atendendo o objetivo de negócio de atingir uma acurácia acima de 80% dos algoritmos de classificação implementados.

Palavras-chave: Acidentes de trabalho, Mineração de texto, KNIME, CRISP-DM, Aprendizado de máquina.

ABSTRACT

Statistics from the International Labor Organization show that one death occurs every 15 seconds due to workplace accidents or illnesses related to professional activity, totaling 2.3 million deaths per year worldwide. In Brazil, one person dies every three and a half hours, placing the country in fourth place in the global ranking of workplace accidents with fatal victims. (ILO, 2023). Based on the introduction, the following dissertation aims to carry out predictive analysis for the classification of work accidents at Vale S.A. through the data collected, with an accuracy above 80%. Using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely accepted approach for data mining projects. The methodology is divided into six distinct phases, each with its own specific steps and activities. With the result of 3 trained algorithms and with an accuracy greater than 80% of hits, meeting the business objective of achieving an accuracy above 80% of the implemented classification algorithms.

Keywords: Work accidents, Text mining, KNIME, CRISP-DM, Machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DE	*Differential entropy*
DT	*Decision Tree*
EDA	*Exploratory data analysis*
EPI	Equipamento de proteção individual
ETL	*Extract, transform, load*
KNN	*K-Nearest Neighbors*
KNIME	*Konstanz Information Miner*
LGPD	*Lei Geral de Proteção de Dados Pessoais*
LR	*Logistic regression*
NB	*Naive Bayes*
NCR	*National Cash Register*
NLP	*Natural language processing*
OIT	Organização Internacional do Trabalho
POS	*Part-of-speech*
PSD	*Power Spectral Density*
ROC	*Receiver operating characteristic*
SPSS	*Statistical Package for the Social Sciences*
SVM	*Support vector machine*
TF-IDF	*Term frequency-inverse document frequency*
UEMA	Universidade Estadual do Maranhão

LISTA DE APÊNDICES

1 INTRODUÇÃO

1.1 Contextualização

Segundo a Organização Internacional do Trabalho (2023) as estatísticas apresentadas revelam uma realidade alarmante no ambiente de trabalho, com 3 milhões de mortes anuais decorrentes de doenças relacionadas ao trabalho e 330 mil mortes por acidentes ocupacionais no mundo. Esses números evidenciam que a cada 10 segundos, um trabalhador perde a vida em um acidente ou doença relacionada ao trabalho. Essa problemática é agravada em países em desenvolvimento, onde uma parcela significativa da população se dedica a atividades perigosas, como agricultura, construção, pesca e mineração.

Esses números estão relacionados à falta de medidas adequadas de segurança no ambiente de trabalho. Rocha (2021) destaca as duas tragédias ocorridas em 2015 em Mariana e em 2019, com o rompimento da Barragem do Córrego do Feijão pela empresa de mineração Vale S.A., que resultaram em terríveis impactos sociais, incluindo a perda de centenas de vidas, além das questões ambientais associadas. Esses eventos trágicos demonstram a necessidade urgente de medidas eficazes de prevenção de acidentes e segurança ocupacional.

As técnicas de aprendizado de máquina usam o princípio de indução, como o aprendizado supervisionado que infere conclusões a partir de exemplos específicos, o autor Zhang et al. (2018) aborda a técnica de mineração de texto e processamento de linguagem natural em análise de acidentes de trabalho em obras, essa técnica é uma subárea do aprendizado de máquina, a seguinte dissertação é utilizada a técnica de mineração de texto para acidentes de trabalho em uma mineradora.

Dentro dessas técnicas, o aprendizado supervisionado, como descrito por Murphy (2012), envolve aprender a partir de um conjunto de dados rotulados, enquanto o aprendizado não supervisionado busca identificar padrões sem a necessidade de rótulos explícitos. Monard e Baranauskas (2003) complementam, destacando que o aprendizado de máquina é essencialmente sobre desenvolver sistemas capazes de aprender com dados, identificando padrões e *insights* que podem ser aplicados para melhorar a segurança e prevenir acidentes no local de trabalho.

1.2 Motivação

O seguinte projeto tem por motivação principal a categorização de eventos de segurança de maneira mais eficaz. Atualmente esse trabalho é realizado manualmente por um técnico ou engenheiro de segurança do trabalho que leva em média 20 horas por mês para categorizar esses eventos e a seguinte proposta tem por objetivo a utilização de técnicas de aprendizado de máquina para realizar a categorização desses eventos.

Tamers et al. (2020) destacam que as mudanças no ambiente de trabalho, influenciadas pelos avanços tecnológicos, têm levado a um aumento no foco e na priorização dos esforços para lidar com questões de segurança, saúde e bem-estar dos trabalhadores. Essa abordagem está alinhada com a Lei nº 8.213/1991, que enfatiza a importância da prevenção de acidentes de trabalho para proteger tanto os trabalhadores quanto os interesses das empresas. Investir em medidas de segurança e medicina do trabalho, pode ser facilitado pelo uso de técnicas de aprendizado de máquina Fernandes e Filho (2019).

A aplicação dessas técnicas não só permite a análise eficiente de grandes volumes de dados para identificar possíveis riscos ocupacionais, como também contribui para a predição e prevenção de acidentes, no trabalho elaborado por Zhang et al. (2018), foi utilizado algoritmos de aprendizado de máquina para realizar análise textual da pesquisa de entrevistas de saúde (NHIS) durante 1997 e 1998, com isso foi possível identificar os objetos mais comuns que causam acidentes, como ‘escada’, ‘raiz’, ‘caminhão’, 'máquina', 'empilhadeira', 'andaime', 'veículo', 'fogo', 'imprensa', 'árvore'.

A seguinte dissertação tem um foco similar, de realizar análise preditiva nas descrições de eventos de segurança de um período entre 2020 e 2022, e ao invés de mostrar os objetos mais comuns, tem por objetivo realizar análise preditiva de classificação para categorizar os eventos com assertividade de acertos acima de 80%.

Portanto, a implementação de técnicas de aprendizado de máquina é possível de realizar e pode trazer uma melhoria no processo da categorização manual por parte dos profissionais de segurança do trabalho, por observar outros trabalhos correlatos, faz-se viável a implementação da seguinte dissertação utilizando mineração de texto para categorizar os eventos de segurança.

1.3 A questão de pesquisa

Após a identificação das áreas de pesquisa e do problema envolvendo a análise preditiva da classificação de acidentes de trabalho, a referente pesquisa se norteia pela seguinte questão de pesquisa: ***“É possível realizar análise preditiva para a classificação de***

acidentes de trabalho na Vale S.A, considerando os dados de acidentes de trabalho e obter uma acurácia acima de 80%?"

1.4 Objetivos

1.4.1 Objetivo geral

Realizar análise preditiva para a classificação de acidentes de trabalho na Vale S.A. através dos dados coletados com uma acurácia acima de 80%.

1.4.2 Objetivos específicos

- Tratar os dados coletados com a utilização de técnicas de linguagem natural;
- Realizar a comparação da eficácia entre algoritmos de classificação com a utilização da plataforma KNIME;
- Publicar a pesquisa em revista científica com *Qualis* B2 ou acima.

1.5 Visão geral metodológica

A classificação de acidentes de trabalho é uma tarefa crítica para a segurança no ambiente de trabalho e para a prevenção de futuros incidentes. Neste contexto, é fundamental adotar uma metodologia robusta para garantir a eficácia e a precisão do processo de classificação. Neste texto, apresentaremos uma metodologia baseada no CRISP-DM (*Cross-Industry Standard Process for Data Mining*), uma abordagem amplamente aceita para projetos de mineração de dados. A metodologia é dividida em seis fases distintas, cada uma com suas próprias etapas e atividades específicas.

1. Entendimento do Problema

Definição clara do objetivo: O primeiro passo é definir o objetivo da classificação de acidentes de trabalho, como identificar os fatores de risco, avaliar a gravidade dos acidentes ou prever futuros incidentes.

Coleta de dados: É necessário identificar as fontes de dados relevantes, que podem incluir relatórios de acidentes, registros médicos, dados de segurança do trabalho e outros.

2. Entendimento dos Dados

Exploração de dados: Realização de uma análise exploratória dos dados para entender sua estrutura, distribuição e características. Isso envolve estatísticas descritivas e visualizações.

Pré-processamento de dados: Limpeza e tratamento de dados ausentes ou inconsistentes. Isso inclui normalização, codificação de variáveis categóricas e seleção de recursos relevantes.

3. Preparação dos Dados

Engenharia de recursos: Crie recursos adicionais a partir dos dados brutos, como a criação de indicadores de risco ou características específicas dos acidentes.

Divisão de dados: Separe os dados em conjuntos de treinamento, validação e teste para treinar, ajustar e avaliar o modelo.

4. Modelagem

Escolha do algoritmo: Selecione os algoritmos de classificação mais adequados para o problema, como árvores de decisão, redes neurais ou métodos baseados em texto (NLP).

Treinamento do modelo: Use o conjunto de treinamento para treinar o modelo escolhido, ajustando os parâmetros conforme necessário.

Avaliação do modelo: Avalie a eficácia do modelo usando o conjunto de validação, ajustando-o conforme necessário para melhorar o desempenho.

5. Avaliação

Teste do modelo: Avalie o modelo final com o conjunto de teste, medindo métricas de desempenho como precisão, *recall*, *F1-score*, entre outras.

Interpretação dos resultados: Analise os resultados para identificar padrões e *insights* relevantes sobre os acidentes de trabalho.

6. Implantação e Monitoramento

Implantação do modelo: Implemente o modelo em um ambiente de produção, onde ele possa ser usado para classificar novos casos de acidentes de trabalho.

Monitoramento contínuo: Monitore o desempenho do modelo ao longo do tempo e faça ajustes conforme necessário para garantir que ele continue sendo eficaz.

1.6 Estrutura organizacional do texto

A pesquisa se encontra organizada, além do capítulo 1 já descrito, em mais seis capítulos. O capítulo 2 é destinado para a fundamentação teórica, desde a parte da fundamentação do KNIME, os algoritmos de classificação, técnicas de mineração de texto e

trabalhos correlatos. No capítulo 3 é voltado para a metodologia de pesquisa, onde é abordado o CRISP-DM uma técnica utilizada em indústrias para análises preditivas.

No capítulo 4 é apresentado o desenvolvimento e validação do modelo, desde a coleta de dados, preparação, treino e validação do modelo, seguindo a metodologia CRISP-DM. Os resultados são apresentados no capítulo 5, exemplificando os resultados obtidos por meio da análise preditiva. Análise e discussão dos resultados estão no capítulo 6 e as conclusões, considerações finais e perspectivas de trabalhos futuros são apresentadas no capítulo 7.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Software KNIME

As especificações da plataforma de código aberto KNIME se refere à organização e estrutura dos componentes e fluxos de trabalho dentro da plataforma. O KNIME é uma plataforma de análise de dados que permite a criação de pipelines personalizados para processar e analisar dados de forma eficiente.

KNIME é uma plataforma de análise de dados em código aberto que enfatiza a interoperabilidade e integração de uma vasta coleção de ferramentas e bibliotecas existentes (Dietz e Berthold, 2016). Eles destacam que o KNIME é uma plataforma amigável ao usuário, projetada para lidar com grandes volumes de dados heterogêneos, e é amplamente utilizado por profissionais da indústria e da academia desde 2006.

Através de seus nós encapsulados, o KNIME permite a combinação de diferentes ferramentas e domínios para formar fluxos de trabalho, os quais documentam e facilitam a reprodução dos resultados da análise. A plataforma também garante a versatilidade ao manter versões anteriores de módulos mesmo após atualizações, permitindo que fluxos de trabalho criados há anos ainda sejam executados nas versões mais recentes do KNIME.

Essa capacidade de processamento eficiente, inclusive em dispositivos de pequena escala, torna o KNIME adequado para análises de alto rendimento, mesmo em conjuntos de dados volumosos, como os gerados por triagens de alto rendimento.

2.2 Análise Exploratória dos Dados

Também é destacado a importância do processo de Análise Exploratória de Dados (EDA) como um estágio crucial para a descoberta de conhecimento, no qual os cientistas de dados exploram interativamente conjuntos de dados desconhecidos através de uma sequência de operações de análise (como filtragem, agregação e visualização) (Milo e Somech, 2020). Eles apontam que, devido à complexidade conhecida da EDA, que demanda habilidades analíticas profundas, experiência e conhecimento do domínio, sistemas diversos foram desenvolvidos ao longo da última década para facilitar esse processo.

Com base nos autores mencionados, o processo de EDA é algo crucial a se realizar em estudos, projetos e no presente trabalho é possível observarmos alguns pontos importantes, analisando a Tabela 4 mencionada anteriormente, é possível notarmos que foi

realizado um tratamento dos dados para que não houvesse uma exposição dos dados pessoais em acordo com a LGPD.

De acordo com a LGPD, qualquer informação que identifique ou possa tornar uma pessoa identificável está incluída no conceito de dado pessoal (Botelho, 2020). Ele salienta que a LGPD exclui dados anonimizados, os quais, conforme o artigo 5º, inciso III, referem-se a informações sobre o titular que não podem ser identificadas, desde que sejam considerados os meios técnicos razoáveis e disponíveis para utilização pelo controlador durante seu processamento.

Ainda analisando a Tabela 4 sobre os dados coletados, pode-se notar que é realizado um trabalho de mineração de texto para identificar qual a categoria presente em cada descrição de acidente mencionado na primeira coluna, como um exemplo na primeira linha "O motorista conduzindo o ônibus ao fazer uma manobra em marcha ré [...]" se enquadra na categoria de acidente "Direção".

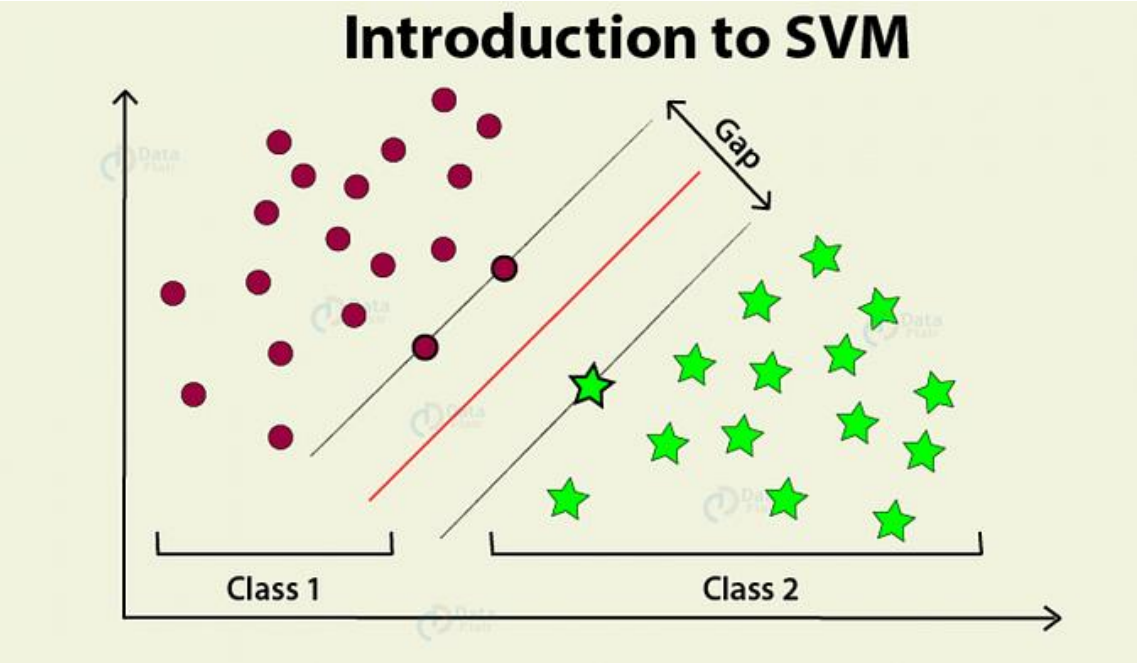
2.3 Algoritmos de Classificação

2.3.1 Support Vector Machine - SVM

O SVM é um modelo de aprendizado de máquina supervisionado. É adequado para conjuntos de dados relativamente pequenos com poucos valores atípicos (Le et al., 2020). O objetivo do SVM é encontrar um hiperplano que maximize a margem entre as duas classes de dados. A margem é a distância entre o hiperplano e os dois pontos de dados mais próximos, correspondentes a duas subclasses. O SVM otimiza o algoritmo ao maximizar o valor da margem, encontrando assim o melhor hiperplano para dividir os dados em duas camadas. Os pontos de dados mais próximos do hiperplano são chamados de vetores de suporte.

A Máquina de Vetores de Suporte (SVM) é um dos algoritmos mais importantes no contexto das técnicas de aprendizado supervisionado (Jain, 2020). Esse algoritmo se baseia na busca pela linha ótima no conjunto de dados, um hiperplano separador. Em outras palavras, dado um conjunto de dados de treinamento rotulados (aprendizado supervisionado), o algoritmo produz um hiperplano ótimo que categoriza novos exemplos. Em um espaço bidimensional, esse hiperplano é uma linha que divide um plano em duas partes, onde cada classe está localizada de um lado, conforme Figura 1.

Figura 1: SVM



Fonte: Jain (2020)

O algoritmo é uma técnica amplamente utilizada em reconhecimento de emoções com base em sinais fisiológicos (Uyanik et al., 2022). O SVM é empregado para classificar os estados emocionais com base nas características extraídas dos sinais, como *Power Spectral Density* (PSD), acoplamento de fase e variabilidade da frequência cardíaca. O SVM tem sido eficaz na classificação de emoções em contextos de realidade virtual, atingindo uma precisão de 85.01% para a classificação de quatro emoções distintas.

Esse algoritmo foi um dos classificadores testados em seu estudo de classificação de dados do espectro autista e que os resultados mostraram que o SVM obteve bom desempenho em termos de precisão, sensibilidade e especificidade ao classificar o conjunto de dados, tanto com quanto sem valores ausentes (Radzi et al., 2022). Além disso, o SVM demonstrou ser eficiente em termos de tempo de execução, levando apenas 0,88 segundos para classificar o conjunto de dados. Este classificador produziu uma curva ROC próxima de 1, indicando sua capacidade de prever com precisão a ocorrência do espectro autista.

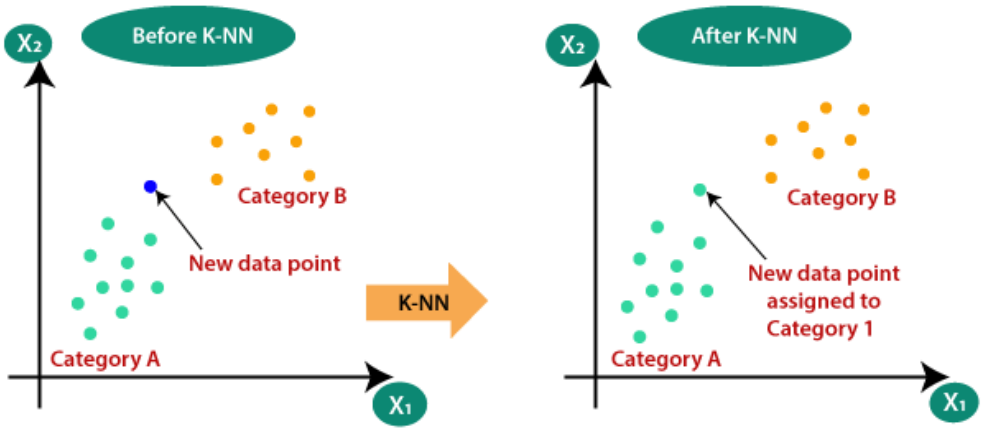
O SVM é um modelo de aprendizado de máquina supervisionado. É adequado para conjuntos de dados relativamente pequenos com poucos valores atípicos (Le et al., 2020). O objetivo do SVM é encontrar um hiperplano que maximize a margem entre as duas classes de dados. A margem é a distância entre o hiperplano e os dois pontos de dados mais próximos, correspondentes a duas subclasses. O SVM otimiza o algoritmo ao maximizar o valor da margem, encontrando assim o melhor hiperplano para dividir os dados em duas camadas. Os pontos de dados mais próximos do hiperplano são chamados de vetores de suporte.

Em conjunto, esses estudos enfatizam a versatilidade e eficácia do SVM em várias aplicações, destacando sua importância no campo de aprendizado de máquina e classificação de dados.

2.3.2 KNN

O algoritmo dos K-Vizinhos Mais Próximos (KNN) é um tipo de algoritmo de aprendizado supervisionado usado tanto para regressão quanto para classificação (Christopher, 2021). O KNN busca prever a classe correta para os dados de teste calculando a distância entre os dados de teste e todos os pontos de treinamento. Em seguida, são selecionados os K pontos mais próximos dos dados de teste. O algoritmo KNN calcula a probabilidade de os dados de teste pertencerem às classes dos 'K' dados de treinamento, e a classe com a maior probabilidade é escolhida. No caso da regressão, o valor é a média dos 'K' pontos de treinamento selecionados. Na Figura 2 abaixo, mostra o detalhamento do funcionamento do algoritmo KNN, a partir de um novo registro, e qual sua probabilidade de categorização, do lado A ou lado B.

Figura 2: KNN



Fonte: Christopher (2021)

O algoritmo KNN é um método não paramétrico simples e eficaz de classificação (Saroj et al., 2022). Ele envolve a coleta dos 'k' vizinhos mais próximos de um determinado registro de dados 't', formando um conjunto de vizinhança 't'. A maioria dos pontos entre os registros de dados na vizinhança é usada para decidir a classificação de 't', com ou sem consideração de ponderação baseada na distância. A escolha de um valor adequado para 'k'

ao aplicar o KNN é crucial, pois o sucesso da classificação depende desse valor. Existem várias maneiras de determinar os valores de 'k', sendo a mais simples a execução do algoritmo várias vezes com valores diferentes de 'k' e a escolha do desempenho ótimo.

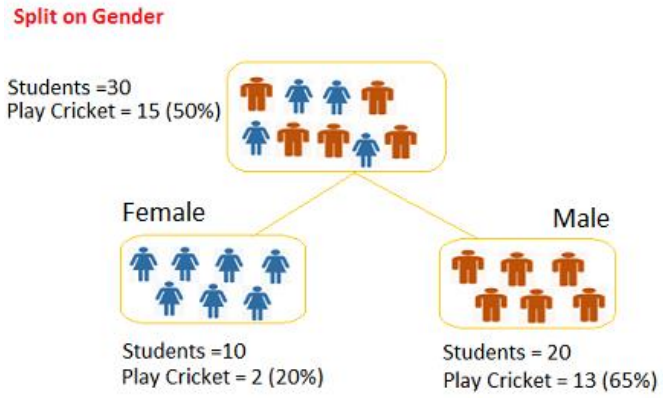
O KNN é baseado em instâncias capaz de classificar vários tipos de conjuntos de dados e realizar ponderação de distâncias, quando 'k' amostras pertencem a um tipo de categoria e se constata que as amostras do conjunto de dados são muito semelhantes a outras 'k' amostras, essas amostras serão classificadas na categoria correspondente (Radzi et al., 2022). O classificador analisa a distância das amostras para a amostra de treinamento vizinha mais próxima no espaço de características.

Em resumo, esses autores realçaram a versatilidade do KNN e sua aplicação em várias áreas, enfatizando a importância da escolha criteriosa de 'K' para obter o melhor desempenho na classificação de dados.

2.3.3 Decision tree

O algoritmo de Árvore de Decisão (DT) são um método de aprendizado supervisionado não paramétrico usado para classificação e regressão (Sehra, 2018). Essas árvores de decisão aprendem com os dados para aproximar uma curva senoidal com um conjunto de regras de decisão do tipo "se-então-senão". Quanto mais profunda a árvore, mais complexas são as regras de decisão e mais ajustado é o modelo. Na Figura 3 mostra uma um exemplo de uso da árvore de decisão, com 30 estudantes se ramificando no gênero de cada um, se for feminino, é alocado na caixa da esquerda, se for masculino na caixa da direita.

Figura 3: Árvore de decisão



Fonte: Sehra (2018)

Este algoritmo foi introduzido como um método em que cada nó da árvore representa propriedades, e os ramos são os valores selecionados para essa propriedade (Le et al., 2020). Seguindo os valores das propriedades na árvore, a DT fornece um valor preditivo. A construção de um algoritmo de árvore começa com a análise dos dados, das características e das variáveis categóricas (valores fictícios), resultando em dados. Em seguida, para encontrar as melhores características para a árvore, trabalha-se com entropia e poder discriminativo usando fórmulas específicas, como entropia (H(X)) e poder discriminativo. Esse processo ajuda a criar uma estrutura de árvore que pode ser usada para prever saídas corretas com base nos dados e características analisadas.

A Árvore de Decisão é uma das técnicas mais intuitivas e diretas em aprendizado de máquina, baseada no paradigma de dividir e conquistar (Saroj et al., 2022). Nesse método, testes (em padrões de entrada) e categorias (de padrões) são utilizados como nós internos e nós folha, respectivamente. Essa técnica também atribui um número de classe a uma matriz de entrada, filtrando a matriz por meio dos testes na árvore. Ela fornece uma estrutura clara para a classificação e tomada de decisão com base nos padrões de entrada e nos testes realizados durante o processo.

A entropia é utilizada para avaliar a impureza ou aleatoriedade de um conjunto de dados (Charbuty e Abdulazeez, 2021). A entropia varia entre 0 e 1, sendo preferível quando seu valor é igual a 0, indicando um estado mais puro. Se o conjunto de dados possui a classe alvo "G" com diferentes valores de atributos, a entropia da classificação do conjunto "S" em relação a "G" pode ser calculada usando a seguinte fórmula conforme Equação 1.

Equação 1: Entropia

$$\text{Entropy}(S) = \sum_{i=1}^{c} P_i \log 2^{P_i}$$

Fonte: Charbuty e Abdulazeez (2021).

O cálculo da entropia ajuda a quantificar o grau de impureza do conjunto de dados, e a ideia é encontrar a divisão que minimize essa entropia, tornando a classificação mais precisa (Charbuty e Abdulazeez, 2021). Quanto mais próxima a entropia estiver de 0, mais homogêneo e puro será o conjunto de dados.

2.3.4 Regressão Logística

A regressão logística é um modelo estatístico comumente utilizado para modelar uma variável dependente binária com a ajuda de uma função logística, que é também conhecida como uma função sigmoide (Rai, 2020). Essa função permite ao modelo de regressão logística restringir os valores de (-k, k) para o intervalo (0, 1). A regressão logística é amplamente empregada em tarefas de classificação binária, embora também possa ser utilizada em problemas de classificação multiclasse.

O algoritmo também é um método estatístico multivariado que exige menos pressuposições (Radzi et al., 2022). Esse método é útil para avaliar a relação entre variáveis independentes e variáveis dependentes, bem como para prever o risco de uma doença com base em variáveis preditoras incorporadas no modelo.

A Regressão Logística (LR) é um algoritmo utilizado para problemas de classificação binária (Le et al., 2020). A regressão logística é semelhante à regressão linear, cujo propósito é encontrar valores para os coeficientes. A logística converterá qualquer valor para o intervalo de 0 a 1. Devido à forma como o modelo é construído, as previsões feitas pela regressão logística também podem ser usadas como probabilidades de uma determinada instância de dados pertencer à classe 0 ou à classe 1. Essa técnica pode ser útil em problemas nos quais é necessário apresentar várias razões para uma previsão.

2.3.5 Naive Bayes

A técnica do Classificador *Naive Bayes* é fundamentada no Teorema de Bayes e é especialmente adequada quando a dimensionalidade das entradas é alta (Le et al., 2020). O algoritmo é frequentemente utilizado para fazer previsões precisas com base em conjuntos de dados coletados, pois é relativamente fácil de entender e altamente preciso. Esse algoritmo pertence ao grupo de algoritmos de aprendizado de máquina supervisionado.

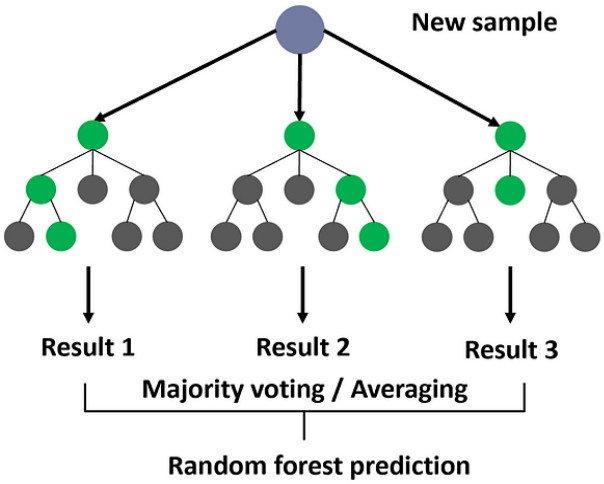
Para entender como o classificador *Naive Bayes* funciona, primeiro é necessário obter um sólido entendimento do Teorema de Bayes (Mall, 2018). P(A|B) representa a probabilidade posterior, ou seja, a probabilidade de A dada a evidência B, enquanto P(B|A) é semelhante, ou seja, a probabilidade de B dado A. Por sua vez, P(A) representa a probabilidade anterior, ou seja, a probabilidade de A sem considerar qualquer evidência, e P(B) é uma constante de normalização usada para obter uma função de densidade de probabilidade com uma soma de 1.

O algoritmo é também um algoritmo simples de aprendizado de máquina baseado no teorema de *Bayes* (Saroj et al., 2022). Ele possui uma suposição fundamental de que os atributos são condicionalmente independentes para a classe fornecida. O *Naive Bayes* oferece uma precisão de classificação competitiva e é amplamente aplicado devido à sua eficiência computacional e às características desejáveis que apresenta.

2.3.6 Random Forest

O algoritmo de floresta aleatória como um modelo de aprendizado de máquina poderoso que se baseia em um conjunto de árvores de decisão, onde cada árvore é desenvolvida usando um subconjunto aleatório do conjunto de dados (Yehoshua, 2023). A previsão final do modelo é obtida por meio de votação majoritária (para classificação) ou média (para regressão) das previsões das árvores na floresta. A média das previsões de várias árvores de decisão torna as previsões mais consistentes, mas pode haver um pequeno aumento no viés do modelo, mas que geralmente melhora o desempenho do modelo final, conforme Figura 4 abaixo.

Figura 4: Floresta aleatória



Fonte: Yehoshua (2023)

O *Random Forest* como uma técnica baseada em árvores de decisão que é usada para classificar e fazer regressão em observações (Radzi et al., 2022). Esta técnica é construída usando uma amostra diferente dos dados originais. O *Random Forest*, que tem raízes no algoritmo de árvore de decisão, consiste em ser uma floresta de árvores de decisão individuais

que funcionam como um conjunto. Aproximadamente um terço dos dados é deixado de fora da amostra e não é utilizado na construção da k-ésima árvore.

O algoritmo é uma técnica que envolve a definição de hiperparâmetros, como o número de árvores e a profundidade máxima de cada árvore (Saroj et al., 2022). O *Random Forest* é uma combinação de abordagens de aprendizado para classificação em aprendizado de máquina e utiliza uma grande coleção de árvores de decisão não correlacionadas.

Portanto, o algoritmo é uma técnica inteligente de aprendizado de máquina baseada em um conjunto de árvores de decisão. O *Random Forest* utiliza uma abordagem de conjunto, onde cada árvore é construída a partir de um subconjunto aleatório do conjunto de dados. A previsão final do modelo é baseada na votação da maioria (para classificação) ou na média (para regressão) das previsões das árvores no conjunto.

2.4 Aplicação de técnicas de Text Mining

2.4.1 Pré-processamento de texto

A importância do pré-processamento de texto como parte da mineração de texto. Eles explicam que, em uma abordagem inicial, o objetivo do pré-processamento é reduzir a dimensionalidade do espaço de representação, o que envolve várias técnicas como remoção de pontuação, palavras minúsculas, *stopwords, tokenization, stemming, lemmatization e POS tagging*.

Remoção de pontuação e palavras minúsculas:

A transformação de palavras minúsculas converte todas as letras no texto para minúsculas, por exemplo, 'Empregado' para 'empregado' (Zhang et al., 2018). Ao converter tudo para minúsculas, reduzimos a variação da mesma palavra e melhoramos a precisão da análise de texto e a remoção de pontuação ocorre retirando pontuações como pontos, vírgulas, pontos de exclamação etc., são removidas do texto. Embora a pontuação seja essencial para a gramática e a legibilidade na linguagem humana, geralmente não contribui significativamente para tarefas de análise de texto, como análise de sentimento, classificação etc. Remover a pontuação ajuda a reduzir o ruído nos dados e diminui o tamanho do conjunto de dados de treinamento.

Stopwords:

As palavras e frases resultantes passam por filtragem, que inclui a eliminação de palavras de parada (*stopwords*) e critérios de frequência (Gupta e Lehal, 2009). O objetivo

desse processo é extrair termos de alta qualidade para indexação e análise. É um filtro eliminação de palavras de parada, como "*the*" e "*a*", que não contribuem significativamente para a tarefa, a padronização do caso das palavras para que todas estejam no mesmo formato de letra.

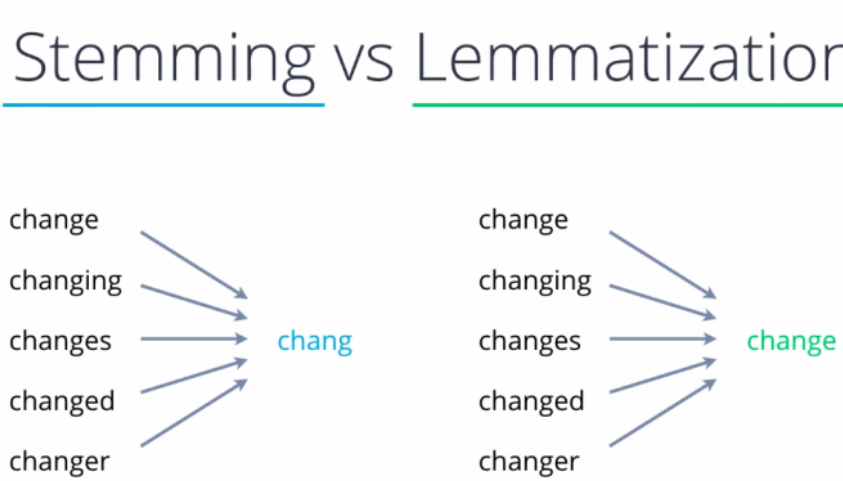
Por exemplo, no domínio médico, palavras como 'pílula' e 'paciente' ocorrem na maioria dos documentos e são consideradas *stopwords*, enquanto no domínio de produtos de computador, uma lista de *stopwords* potenciais pode incluir palavras como 'CPU' e 'memória' (Zhang et al., 2018). Geralmente, as listas de *stopwords* comuns não cobrem esses termos específicos de domínio. Portanto, é recomendado compilar uma lista de *stopwords* específica para o domínio com base no conhecimento adquirido sobre o domínio em questão.

Stemming* e *lemmatization:

O processo de *stemming* é utilizado para tratar palavras que têm formas diferentes, mas são semanticamente semelhantes (Gupta e Lehal, 2009). O resultado desse procedimento é obter o radical de cada palavra, enfatizando sua semântica. O processo de pré-processamento envolve a extração de palavras-chave e frases a partir de um conjunto de resumos de documentos (Tseng, Lin e Lin, 2007). Além disso, idealmente, a indexação inclui o processo de *stemming* (análise morfológica) dos termos e o agrupamento de termos semelhantes que representam conceitos relacionados. Isso não apenas reduz o tamanho do vocabulário para uma análise mais eficiente, mas também resolve o problema de inconsistência de vocabulário, tornando a clusterização mais eficaz.

A técnica de *stemming* tem sido amplamente aplicada em recuperação de informações (Korenius et al., 2004). Ela é vantajosa, pois permite que o usuário não precise se preocupar com o ponto exato de truncamento das palavras-chave de busca. Além disso, o *stemming* reduz o número total de entradas de índice distintas, o que simplifica a indexação e economiza espaço. Essa técnica também expande as consultas ao agrupar variantes de palavras, incluindo derivações.

Figura 5: *Stemming x Lemmatization*



Fonte: Kumar (2022)

O *stemming* e *lematização* são técnicas de pré-processamento de texto usadas para reduzir palavras à sua forma básica ou raiz, facilitando a análise textual (Kumar, 2022). *Stemming* envolve cortar as palavras para encontrar o radical, o que pode resultar em diferentes variações de uma palavra sendo reduzidas à mesma raiz. Por outro lado, a lematização envolve mapear palavras para sua forma básica com base em um dicionário linguístico, o que geralmente resulta em uma forma mais precisa da palavra. Ambas as técnicas são úteis em processamento de texto e recuperação de informações.

A lematização que é outra técnica de normalização na qual, para cada forma de palavra flexionada em um documento ou consulta, é identificada a sua forma básica, ou seja, o lema (Korenius et al., 2004). Os benefícios da lematização são semelhantes aos do *stemming*, incluindo a redução da ambiguidade.

O processo de *stemming* visa traduzir as formas morfológicas de uma palavra para a sua raiz, supondo que todas as variações estejam semanticamente relacionadas (Vijayarani, Ilamathi e Nithya, 2015). Nesse contexto, dois pontos essenciais são considerados: primeiro, palavras que não possuem o mesmo significado devem ser mantidas separadas; segundo, as diferentes formas morfológicas de uma palavra são assumidas ter a mesma base de significado e, portanto, devem ser mapeadas para a mesma raiz. Essas regras são relevantes e suficientes em aplicações de processamento de linguagem ou mineração de texto. Além disso, o *stemming* é geralmente considerado como um mecanismo que melhora o *recall* em sistemas de recuperação de informações, embora sua eficácia possa variar dependendo da complexidade morfológica da língua.

Os algoritmos de *stemming* são classificados em três grupos: métodos de truncamento, métodos estatísticos e métodos mistos (Vijayarani, Ilamathi e Nithya, 2015). Os métodos de truncamento, ou remoção de afixos, envolvem a eliminação de sufixos ou prefixos de uma palavra. Um exemplo simples é o "*Truncate (n) stemmer*," que corta uma palavra no símbolo n, mantendo as n letras iniciais e removendo o restante. Um exemplo de um método de truncamento mais específico é o S-*stemmer*, um algoritmo que converte as formas no singular e plural de substantivos em inglês. Os métodos estatísticos, por outro lado, utilizam análise e técnicas estatísticas em conjunto com a remoção de afixos para identificar as raízes das palavras.

Tokenization

Tokenização é a tarefa de dividir essa sequência em partes, chamadas de tokens, durante o processo, certos caracteres, como pontuação, são filtrados (Zhang et al., 2018).

A tokenização refere-se à divisão de sentenças em palavras, caracteres, pontuações, todos os quais são chamados de tokens (Tabassum, 2020). O critério de divisão ocorre principalmente na ocorrência de um espaço ou uma pontuação. Esta etapa ajuda na filtragem de palavras indesejadas em etapas posteriores de processamento. Por exemplo, a frase "*NLP is the future of Speech Recognition Systems!*" é tokenizada como *"NLP", "is", "the", "future", "of", "Speech", "Recognition", "Systems"* e "!".

POS tagging:

O *Part-of-speech* (POS) *tagging* consiste no procedimento de atribuir a cada palavra em uma sentença sua classe morfossintática apropriada, tal como verbo, substantivo, adjetivo, entre outras (Sousa, 2019). A palavra “O” é um artigo, “cavalo” é um substantivo, “de” uma adposição, “Napoleão” outro substantivo, “é” verbo e “branco” um adjetivo.

Tabela 1: Exemplo de POS Tagging

Palavras	O	cavalo	de	Napoleão	é	branco
Classes	ART	SUBS	ADP	SUBS	V	ADJ

Fonte: Sousa (2019).

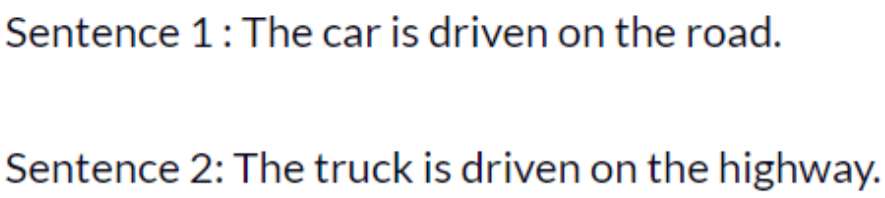
2.4.2 Extração de características textuais (TF e TF-IDF)

O conceito de *term frequency* (TF) e *inverse document frequency* (IDF) no contexto do TF-IDF. TF representa a frequência de um termo em um documento e é usado para medir com que frequência um termo aparece em um documento específico (Qaiser e Ali, 2018).

Para lidar com variações no comprimento de documentos, o TF é calculado dividindo o número de vezes que um termo aparece em um documento pelo número total de termos nesse documento, enquanto o TF representa a frequência de uma palavra em um documento específico, o IDF mede a importância dessa palavra no contexto de todos os documentos. A fórmula para calcular o TF-IDF é a multiplicação desses dois valores, resultando em uma medida que reflete a relevância da palavra-chave pesquisada no documento em questão.

Esse cálculo garante que o valor do TF permaneça consistente em documentos de diferentes comprimentos, permitindo uma avaliação mais precisa da importância do termo em um documento (Qaiser e Ali, 2018). Por exemplo, um documento possui 5000 mil palavras e o termo "*Alpha*" se repete 10 vezes, portanto seu TF é 10/5000 = 0,002. Tripathi (2018) demonstra de maneira prática como é o funcionamento do TF e do TF-IDF com duas frases separadas na Figura 6.

Figura 6: Sentenças



Fonte: Tripathi (2018)

Um exemplo do cálculo TF-IDF que é aplicado a dois documentos representados pelas frases "*The car is driven on the road*" e "*The truck is driven on the highway*" (Tripathi, 2018). Os documentos são tratados como textos separados. O cálculo do TF-IDF revela que as palavras comuns têm um valor de TF-IDF igual a zero, indicando que não são significantes. Em contraste, as palavras *"car," "truck," "road" e "highway"* têm valores de TF-IDF não nulos, o que indica que essas palavras são mais significativas no contexto desses documentos, conforme mostra a Tabela 2.

Tabela 2: TF-IDF

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Fonte: Tripathi (2018).

O uso de TF e IDF como parâmetros para a filtragem de termos. Termos com baixa frequência em documentos (TF) e baixa frequência em toda a coleção de documentos (DF) geralmente são removidos da indexação (Tseng, Lin e Lin, 2007). Além disso, eles mencionam a utilização de uma lista de mais de 250 palavras de parada (*stopwords*) para evitar o cálculo de termos indesejados, incluindo função. Posteriormente, após a análise de documentos de patentes, adicionaram manualmente mais 200 palavras à lista, a maioria delas sendo advérbios. Para melhorar a correspondência entre os termos, eles também realizam o *stemming* das palavras com base no algoritmo de Porter. No entanto, eles observam que o algoritmo padrão é muito agressivo na remoção dos sufixos das palavras, tornando os termos resultantes difíceis de serem lidos. Portanto, eles modificam o algoritmo para remover apenas plurais simples e sufixos gerais, como os do passado regular.

O conceito de IDF que é usado para atribuir maior importância a palavras que são menos frequentes nos documentos (Qaiser e Ali, 2018). Isso ocorre porque, ao calcular apenas o TF, todas as palavras são tratadas igualmente, mesmo *stopwords* com pouca significância. Por exemplo, se a palavra "tecnologia" aparece em 5 de um conjunto de 10 documentos, o IDF é calculado como IDF = log_e (10/5) = 0,3010, atribuindo um peso menor às palavras frequentes e um peso maior às palavras infrequentes nos documentos.

O cálculo do *Term Frequency - Inverse Document Frequency* (TF-IDF) envolve a multiplicação do TF e do IDF, que foram apresentados anteriormente. Em resumo, a combinação dessas métricas, conhecida como TF-IDF, é utilizada para identificar palavras-chave significativas e filtrar termos comuns, melhorando a representação de documentos e facilitando a recuperação de informações relevantes. Essas técnicas são fundamentais para tarefas de processamento de linguagem natural e recuperação de informações.

2.5 Trabalhos correlatos

Nesta seção, são discutidos trabalhos relacionados a dissertação de mestrado, com o objetivo de identificar trabalhos similares, considerar referências teóricas importantes, descrever as metodologias utilizadas e os resultados alcançados. A análise comparativa desses trabalhos é fundamental para destacar as diferenças em relação à presente pesquisa, permitindo identificar pontos fortes e fracos e, assim, contribuir para o avanço do estado da arte.

O autor Zhang et al. (2018) foca em apresentar a análise de acidentes em canteiros de obras por meio de técnicas de mineração de texto e processamento de linguagem natural, visando identificar padrões e *insights* relevantes a partir de grandes conjuntos de dados não estruturados. Além disso, eles destacam a importância da remoção de letras maiúsculas e pontuação para reduzir a variação da mesma palavra e remover informações desnecessárias que não contribuem significativamente para a análise de texto.

A remoção de *stopwords* também é abordada por Zhang et al. (2018), ressaltando a necessidade de eliminar palavras comuns que têm pouco valor na seleção de documentos. Eles mencionam diferentes listas de *stopwords* disponíveis e a importância de adaptá-las de acordo com o domínio específico de cada pesquisa.

Zhang et al. (2018) também descrevem a utilização de cinco classificadores individuais e um modelo de conjunto proposto em seu estudo, destacando o uso do algoritmo SQP para otimizar os pesos do modelo de conjunto e melhorar o mecanismo de votação majoritária. Esses aspectos metodológicos são discutidos detalhadamente nas seções subsequentes do artigo.

Neste contexto, Xu et al. (2021) discutem a importância da análise de relatórios de descrição de acidentes na construção para identificar os fatores de risco de segurança típicos, que são essenciais para orientar a gestão de segurança no futuro. Eles observam que, atualmente, essa prática depende principalmente do julgamento de especialistas do domínio, o que pode ser subjetivo e demorado.

No artigo, uma abordagem aprimorada é desenvolvida para identificar os fatores de risco de segurança a partir de um grande volume de relatórios de acidentes de construção, utilizando tecnologia de mineração de texto (TM). Um framework de TM é concebido e um fluxo de trabalho é estabelecido para a construção de um léxico de domínio personalizado.

Xu et al. (2021) também destacam o uso de abordagens estatísticas, como TF, DF (frequência de documentos) e TF-IDF, para identificar características dos documentos. Eles exemplificam o uso de TF-IDF na priorização de palavras em documentos de solicitação de informações prévias ao envio de propostas, destacando a importância dessas técnicas na identificação de fatores críticos e na tomada de decisões informadas, que é uma continuidade da abordagem explorada por Zhang et al. (2018) na análise de acidentes em canteiros de obras.

Tao, Yang e Feng (2020) ressaltam a capacidade dessa técnica de extrair *insights* valiosos de grandes conjuntos de dados não estruturados. Essas aplicações não se limitam apenas ao campo alimentar, como também são exploradas em outras áreas, como a segurança alimentar e a vigilância de fraudes alimentares. A análise de dados textuais é utilizada para caracterizar padrões dietéticos, minerar opiniões de consumidores, desenvolver novos produtos e gerenciar a cadeia de suprimentos de alimentos e serviços alimentares online.

Adicionalmente, a explanação de Tao, Yang e Feng (2020) sobre a importância do processo de *stemming*, *lematização* e remoção de *stopwords* na análise de texto ressalta a necessidade de métodos precisos e significativos para extrair informações valiosas dos dados. Essas técnicas, que preservam o significado essencial do texto original, são cruciais não apenas na ciência dos alimentos e nutrição, mas também em outras áreas, como na análise de acidentes de trabalho.

Assim como na caracterização de padrões dietéticos, a análise precisa de relatórios de acidentes de trabalho pode beneficiar-se dessas técnicas para identificar fatores de risco, padrões e *insights* relevantes que contribuam para a prevenção de acidentes e a promoção de ambientes de trabalho mais seguros.

Os trabalhos de Zhang et al. (2018), Xu et al. (2021) e Tao, Yang e Feng (2020) abordam técnicas de mineração de texto que são utilizados na seguinte pesquisa, correlacionando diversas áreas em um mesmo núcleo de mineração de texto. Zhang et al. (2018) e Xu et al. (2021) focam na aplicação de técnicas de mineração de texto para análise de acidentes de trabalho em canteiros de obras.

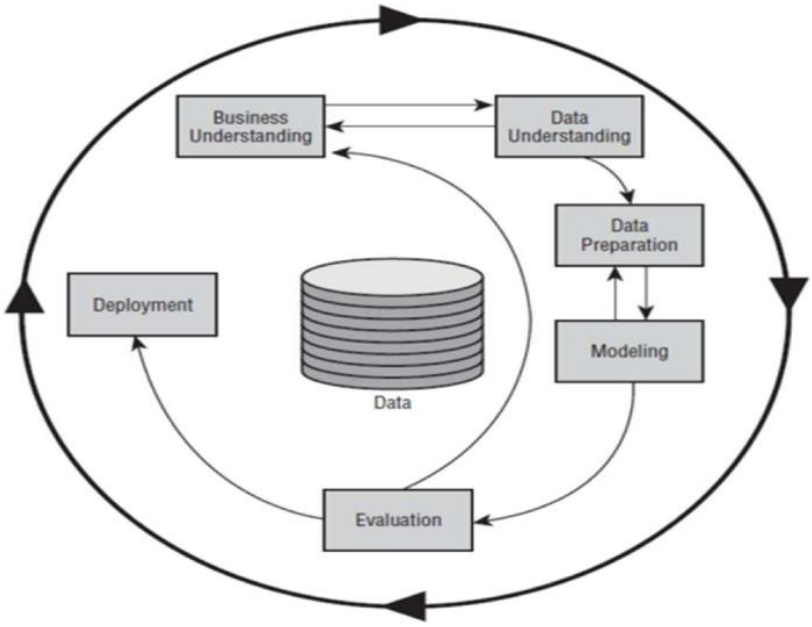
Por sua vez, Tao, Yang e Feng (2020) exploram o potencial da mineração de texto na ciência dos alimentos e nutrição, destacando sua capacidade de extrair *insights* valiosos para aprimorar a produção de alimentos, a segurança alimentar e a nutrição humana. Embora os contextos de aplicação variem, todos os trabalhos demonstram o poder da mineração de texto na extração de conhecimento útil a partir de grandes volumes de dados não estruturados, oferecendo *insights* importantes para diversas áreas de pesquisa e prática.

3 METODOLOGIA DE PESQUISA

3.1 Definição da metodologia CRISP-DM

O *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) foi desenvolvido por Daimler Chrysler (então Daimler-Benz), SPSS e NCR em 1999 (Shafique e Qaiser, 2014). A versão 1.0 do CRISP-DM foi publicada, sendo um framework completo e documentado que fornece um conjunto de diretrizes e uma estrutura uniforme para os profissionais de mineração de dados. Ele é composto por seis fases ou estágios bem estruturados e definidos. Conforme mencionado na Figura 7.

Figura 7: CRISP-DM



Fonte: Shafique e Qaiser (2014).

O processo CRISP-DM foi desenvolvido por meio dos esforços de um consórcio inicialmente composto por DaimlerChrysler, SPSS e NCR. CRISP-DM representa o *Cross-Industry Standard Process for Data Mining* e consiste em um ciclo com seis estágios bem definidos (Azevedo e Santos, 2008).

1. Entendimento do negócio;
2. Entendimento dos dados;
3. Preparação dos dados;
4. Modelagem;

5. Avaliação;
6. Implantação

A sequência desses seis estágios não é rígida, e o CRISP-DM é documentado e bem estruturado, o que facilita a compreensão e revisão de projetos. Além disso, o processo CRISP-DM é independente da ferramenta de mineração de dados escolhida.

3.1.1 Entendimento do Negócio

A primeira fase do processo CRISP-DM se concentra na identificação e na exploração de fatores cruciais, abrangendo critérios de sucesso, objetivos e requisitos de mineração de dados, terminologias comerciais e termos técnicos (Shafique e Qaiser, 2014).

A fase de Entendimento do Negócio envolve a avaliação da situação empresarial para obter uma visão geral dos recursos disponíveis e necessários (Schroer, Kruse e Gómez, 2021). Um dos aspectos mais críticos nesta fase é a definição do objetivo da mineração de dados. Inicialmente, o tipo de mineração de dados deve ser explicado (por exemplo, classificação), juntamente com os critérios de sucesso da mineração de dados (como precisão). Além disso, é fundamental criar um plano de projeto obrigatório.

Com base nos autores acima, referente a essa etapa, o objetivo de negócio da seguinte proposta é reduzir o retrabalho dos técnicos e engenheiros de segurança que realizam a categorização dos eventos manualmente, gastando um tempo médio de 20 horas por mês para fazer este trabalho, propõe-se um cenário em que essa categorização seja de forma automática por meio de algoritmos de classificação. A métrica a ser utilizada para mensurar a eficácia do algoritmo seria a quantidade de eventos categorizados de forma correta, a partir de 80% de acertos, ou seja, uma acurácia acima de 80% seria o mais indicado para o objetivo do algoritmo.

3.1.2 Entendimento dos Dados

A fase de Entendimento dos Dados se inicia com a coleta inicial de dados e, a partir daí, engloba uma série de atividades que têm como objetivo familiarizar-se com os dados, identificar problemas de qualidade dos dados, obter as primeiras percepções sobre os dados e detectar subconjuntos de dados de interesse que possam ser usados para formular hipóteses visando informações ocultas (Azevedo e Santos, 2008). A compreensão dos dados é um passo

crucial no processo de mineração de dados, permitindo a exploração e preparação adequada dos dados para análises posteriores.

Nessa fase, além da coleta de dados de fontes diversas, ocorre também a exploração, descrição e verificação da qualidade dos dados são tarefas essenciais (Schroer, Kruse e Gómez, 2021). Para tornar isso mais concreto, o guia do usuário descreve a tarefa de descrição dos dados por meio de análises estatísticas e a determinação de atributos e sua relação. Essa fase é fundamental para preparar os dados de forma adequada, tornando-os prontos para análises e avaliações subsequentes.

Em resumo, essa fase envolve a busca de ideias iniciais nos dados e a identificação de atributos relevantes. O entendimento dos dados é considerado como uma etapa essencial para o sucesso de qualquer projeto de mineração de dados, permitindo uma análise mais precisa e significativa para a tomada de decisões e descoberta de informações valiosas. Portanto, a fase de Entendimento dos Dados é um pré-requisito para a realização de uma mineração de dados eficaz e informada.

3.1.3 Preparação dos Dados

A fase de Preparação dos Dados compreende várias etapas críticas. Inicialmente, a seleção de dados é conduzida por meio da definição de critérios de inclusão e exclusão, visando à escolha dos dados mais relevantes para a análise (Schroer, Kruse e Gómez, 2021). Problemas relacionados à má qualidade dos dados podem ser abordados por meio da limpeza dos dados. Além disso, dependendo do modelo definido na primeira fase do processo, é necessário construir atributos derivados para a análise. É importante notar que existem diferentes métodos disponíveis para realizar essas etapas, e a escolha dos métodos apropriados pode variar dependendo do modelo de mineração de dados selecionado. Portanto, a Preparação dos Dados é uma etapa crucial que envolve a seleção cuidadosa e a transformação dos dados para garantir que estejam prontos para análises posteriores.

Os dados são compostos por conceitos (representando o que precisa ser aprendido), instâncias (registros independentes relacionados a uma ocorrência) e atributos (que caracterizam um aspecto específico de uma instância dada) (Moro, Laureano e Cortez, 2011). Essa distinção entre conceitos, instâncias e atributos é fundamental para a compreensão da estrutura e natureza dos dados, permitindo uma análise mais precisa e significativa durante o processo de mineração de dados.

Em resumo, a preparação dos dados é considerada uma etapa crítica para garantir que o conjunto de dados final seja de alta qualidade, adequado e pronto para análises posteriores. Consequentemente, a qualidade e a preparação adequada dos dados são pré-requisitos fundamentais para o sucesso de qualquer projeto de mineração de dados.

3.1.4 Modelagem

A fase de Modelagem envolve a seleção e aplicação de várias técnicas de modelagem, bem como a calibração de seus parâmetros para obter os valores ótimos (Wirth e Hipp, 2000). Geralmente, existem várias técnicas disponíveis para o mesmo tipo de problema de mineração de dados, e algumas delas podem exigir formatos de dados específicos. É importante destacar que existe uma estreita relação entre a Preparação dos Dados e a Modelagem, uma vez que muitas vezes são identificados problemas nos dados durante a modelagem, ou ideias para a construção de novos dados podem surgir durante essa fase. Essa interconexão ressalta a importância de ambas as fases no processo de mineração de dados.

Essa etapa tem como objetivo construir o modelo que representa o conhecimento aprendido a partir dos dados (Moro, Laureano e Cortez, 2007). Esse modelo tem a capacidade de, por exemplo, prever o valor alvo que representa o objetivo definido quando fornecido com uma instância. A construção desse modelo é essencial para a aplicação eficaz do conhecimento obtido por meio da mineração de dados em situações práticas e na tomada de decisões informadas.

3.1.5 Avaliação

A quinta etapa do processo CRISP-DM é a Avaliação, que se concentra na análise dos modelos obtidos e na tomada de decisões sobre como usar os resultados (Shafique e Qaiser, 2014). A interpretação dos modelos depende do algoritmo utilizado, e os modelos podem ser avaliados para determinar se atendem adequadamente aos objetivos estabelecidos. Essa fase desempenha um papel fundamental na análise crítica dos resultados da mineração de dados, avaliando a eficácia e relevância dos modelos em relação às metas de negócios.

Após a construção de um ou mais modelos que parecem ter alta qualidade do ponto de vista da análise de dados, é fundamental realizar uma avaliação mais detalhada desses modelos (Azevedo e Santos, 2008). Além disso, é importante revisar os passos executados para construir o modelo para garantir que ele atenda adequadamente aos objetivos de negócios

estabelecidos. Um dos principais objetivos é identificar se existem questões de negócios importantes que não foram devidamente consideradas. No final dessa fase, uma decisão sobre a utilização dos resultados da mineração de dados deve ser tomada, considerando o impacto nos objetivos empresariais.

Uma técnica de avaliação dos modelos é a matriz de confusão, comumente utilizada em aprendizado de máquina, registra os resultados das classificações em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos (Junior et al., 2022). Conforme Tabela 3.

Tabela 3: Matriz de contingência 2x2

	Predição	
Realidade	**Verdadeiro Positivo (VP)**	**Falso Negativo (FN)**
	Falso Positivo (FP)	**Verdadeiro Negativo (VN)**

Fonte: Junior (2022).

A etapa de Avaliação é fundamental para garantir que os modelos atendam aos padrões de qualidade e eficácia, contribuindo para a tomada de decisões informadas. Além disso, enfatizam a importância de avaliar se todas as questões de negócios cruciais foram consideradas, permitindo que as organizações alcancem melhores resultados e direcionem suas ações com base nos *insights* gerados pela mineração de dados.

3.1.6 Implantação e Monitoramento

A fase de Implantação é a sexta e última etapa do processo CRISP-DM, e seu foco principal está na determinação do uso do conhecimento e dos resultados obtidos durante o processo de mineração de dados (Shafique e Qaiser, 2014). Além disso, essa fase concentra-se na organização, relato e apresentação do conhecimento adquirido quando necessário. A implantação é crucial para garantir que as informações e os *insights* gerados sejam aplicados de maneira eficaz nas operações de negócios e que estejam disponíveis para tomada de decisões informadas quando necessário.

Essa fase não marca o encerramento do projeto, geralmente o conhecimento adquirido precisa ser organizado e apresentado de maneira que o cliente possa usá-lo de forma eficaz (Wirth e Hipp, 2000). A complexidade da fase de implantação pode variar de acordo com os requisitos, podendo ser tão simples quanto a geração de um relatório ou tão complexa quanto a implementação de um processo de mineração de dados repetível. Em muitos casos, a

implantação é realizada pelo usuário final, não pelo analista de dados. Portanto, é crucial entender antecipadamente quais ações serão necessárias para realmente aproveitar os modelos criados.

4 DESENVOLVIMENTO E VALIDAÇÃO DO MODELO

O software KNIME como um ambiente modular que facilita a montagem visual e a execução interativa de um pipeline de dados (Berthold et al., 2009). Este ambiente foi projetado com o propósito de ser uma plataforma de ensino, pesquisa e colaboração, permitindo a fácil integração de novos algoritmos e ferramentas, bem como métodos de manipulação ou visualização de dados na forma de novos módulos ou nós. Portanto o desenvolvimento da seguinte dissertação é baseado na plataforma KNIME *Analytics Platform*.

4.1 Coleta de Dados

Conforme citado anteriormente na etapa de entendimento dos dados, esse tópico tem o objetivo de familiarizar com os dados e as fontes de dados utilizadas foram coletadas através de uma base de dados descritiva sobre acidentes de trabalho na empresa Vale, sendo validado juntamente com o Jurídico da empresa em conformidade com a Lei Geral de Proteção de Dados de nº 13.709, de 14 de agosto de 2018. Sendo solicitada também pela UEMA através de ofício para empresa Vale.

Essa é uma amostra da base de dados que possui aproximadamente 11 mil registros com uma série histórica de aproximadamente 3 anos (2020 – 2022). A base tem informação de descrição do acidente de trabalho e a categoria do acidente sendo desdobrado em 5 categorias: estrutural, comportamento, direção, equipamentos e estrutura do veículo.

Tabela 4: Base de dados

DESCRIÇÃO	CATEGORIA
O MOTORISTA CONDUZINDO O ÔNIBUS OYF- AO FAZER UMA MANOBRA EM MACHA RÉ VEIO A	Direção
O MOTORISTA ESTACIONOU ÔNIBUS DE FROTA W/ QEK NA ESTRADA AO LADO DO RESTAURANT	Direção
O MOTORISTA: MAT CONDUZIA O ÔNIBUS DE FROTA W AO FAZER UMA MANOBRA NO TREVO I	Direção
MOTORISTA CONDUZINDO O ÔNIBUS DE ODN COLIDIU NA TRASEIRA DO ÔNIBUS DE OIX DUR	Direção
O MECÂNICO ESTAVA REALIZANDO UMA INSPEÇÃO / VERIFICAÇÃO NO PARA – BRISA DE UM VEICL	Comportamento
NO DIA / POR VOLTA DAS : O CONDUTOR REALIZAVA TRAJETO (RESTAURANTE DE CONCEIÇÃO SEN	Estrutura do veículo
O MOTORISTA CONDUZIA O VEÍCULO / QEE PELA ESTRADA PAULO FONTELES APÓS PASSAR DA VI	Estrutura do veículo
O MOTORISTA CONDUZIA O VEÍCULO PELA ESTRADA PAULO FONTELES APÓS PASSAR DA VILA SA	Estrutura do veículo
O MOTORISTA CONDUZIA O VEÍCULO W/ QVXC PELA ESTRADA PAULO FONTELES APÓS PASSAR DA	Estrutura do veículo
MOTORISTA VIX CONDUZIA O VEÍCULO DE ROTA INICIANDO O REVEZAMENTO DE ALMOÇO DA	Direção
O MOTORISTA UNIMAR REALIZAVA A LINHA PIER DE CARVÃO(CIRCULAR) AO REALIZAR UMA MAN	Direção
O MOTORISTA CONDUZIA O VEÍCULO NA LINHA VV AO REALIZAR SEU PERCURSO NA RUA PIRACIC	Direção
O MOTORISTA DA JSL CONDUZIA O VEÍCULO E AO ACESSAR A OFICINA AVANÇADA PARA DEIXAR C	Direção
O MOTORISTA ESTAVA REALIZANDO SUA ROTA NORMAL CONDUZINDO O VEICULO DE IDENTIFICA	Estrutura do veículo
ÔNIBUS QEW QA JSL TC E ÔNIBUS QDD QA VIX VIX TC PASSARAM SIMULTANEAMENTE PELA PORT	Direção

Fonte: O autor (2023)

Para auxiliar o leitor, o dicionário dos dados é apresentado a seguir, os dados possuem dois campos:

Descrição: Descrição do acidente ocorrido entre o período de 2020 até 2022.

Categoria: Tipo em que o acidente é categorizado, podendo ser: ‘Direção’, ‘Estrutural’, ‘Estrutura do veículo’, ‘Comportamento’ e ‘Equipamentos’, a seguir é detalhado cada categoria.

Categoria direção: Acidente de decorrência de direção, por exemplo, uma manobra de marcha ré que colidiu com um poste.

Categoria estrutural: Acidente que envolve algum problema estrutural, por exemplo, um forro do teto caiu sobre um empregado.

Categoria estrutura do veículo: Acidente que envolve um problema na estrutura do veículo, por exemplo, roda dianteira do veículo se desprendeu do eixo.

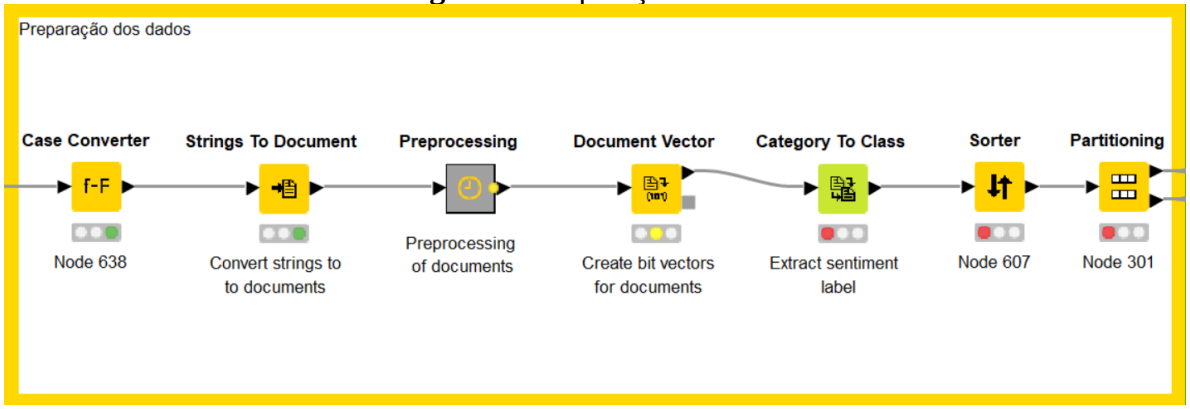
Categoria comportamento: Acidente envolvendo o comportamento do empregado, por exemplo, empregado transitando em área restrita e vindo a lesionar o pé.

Categoria equipamentos: Acidente que envolve a falta de equipamentos de proteção individual (EPI) do empregado, por exemplo, empregado tocou em uma estrutura metálica sem equipamento e veio a sofrer um choque elétrico.

4.2 Preparação dos dados

Conforme citado anteriormente na etapa de preparação dos dados, esse tópico tem o objetivo escolher os dados mais relevantes para a análise. Na etapa atual foi realizado diversos tipos de processamentos de dados textuais para que o algoritmo possa atuar de maneira eficiente. Conforme na Figura 8 abaixo:

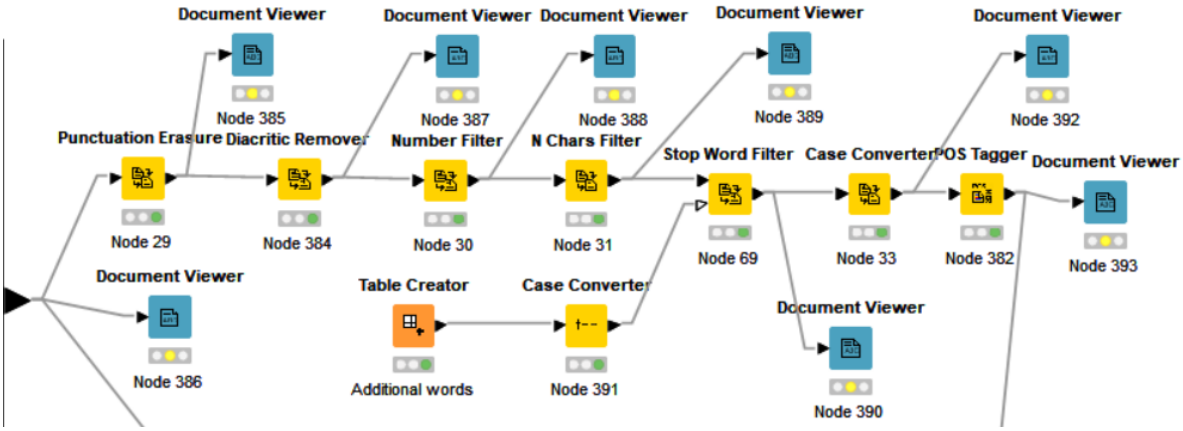
Figura 8: Preparação dos dados



Fonte: O autor (2023)

Na primeira parte da preparação dos dados, foi realizado a conversão de texto para maiúsculo, com intuito de evitar algum erro por parte de ser letras minúsculas, foi convertido também para o tipo de documento, que é o tipo que é realizado toda a preparação textual na ferramenta KNIME. Na Figura 9 é apresentado todas as etapas de pré-processamento dos dados textuais.

Figura 9: ETL dos dados



Fonte: O autor (2023)

Dentro do nó de pré-processamento mencionado acima, é exibido a primeira parte dos nós, onde esses de cor azul têm por objetivo uma análise exploratória, somente para visualização do documento para ter ciência de sua evolução. Os nós de cor amarela são para realizar algum tipo de tratamento, como remoção de acentos, remoção de *stopwords*, o *POS Tagger* descrito na última etapa é um outro tipo de preparação textual.

Destaca-se que o POS *tagging* desempenha um papel essencial na construção de aplicações de processamento de linguagem natural (NLP), uma vez que várias dessas aplicações dependem, em determinado momento, desse tipo de informação linguística para sua execução.

Tabela 5: *Stopwords* manuais

Manually created table - 3:29

File Edit Hilite Navigation

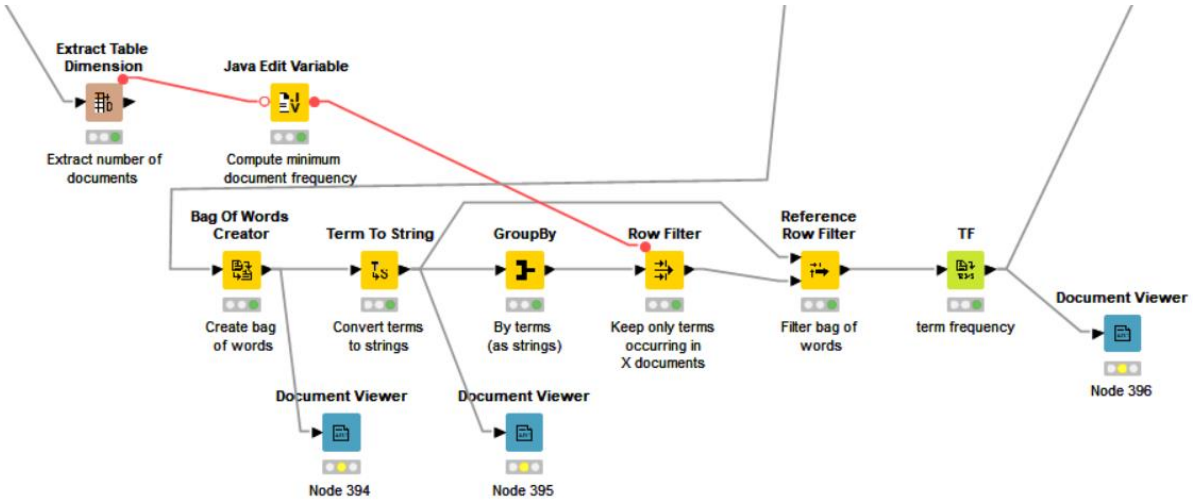
Table "default" - Rows: 221 Spec -

Row ID	S column1
Row196	tera
Row197	teremos
Row198	terao
Row199	teria
Row200	teriamos
Row201	teriam
Row202	hmin
Row203	etc
Row204	empregado
Row205	motorista
Row206	veiculo
Row207	vix
Row208	serra
Row209	verde
Row210	veio
Row211	nao
Row212	nwn
Row213	lyon
Row214	jsl
Row215	manserv
Row216	empresa
Row217	sodexo
Row218	grsa
Row219	segurpro
Row220	vale

Fonte: O autor (2023)

Conforme citado por Zhang et al. (2018) em determinados contextos, é necessário remover *stopwords* manuais conforme a realidade, por exemplo, em domínio médico, palavras como 'pílula' e 'paciente' ocorrem na maioria dos documentos e são consideradas *stopwords*, enquanto no domínio de produtos de computador, uma lista de *stopwords* potenciais pode incluir palavras como 'CPU' e 'memória'. Portanto, no seguinte trabalho as palavras que ocorrem na maioria dos documentos são ‘motorista’, ‘empregado’, ‘veiculo’, entre outras palavras. Conforme Tabela 5, é possível ver uma amostragem de *stopwords* que foi inserido manualmente, pois são comuns nos documentos.

Figura 10: Segunda parte do ETL



Fonte: O autor (2023)

Na segunda parte da preparação dos dados, inicia-se realizado a extração das palavras que contém na descrição, é possível notar que o *POS Tagger* fez a atribuição a cada palavra relevante da descrição como uma palavra exclusivamente como um termo único, conforme mencionado na Tabela 6 abaixo:

Tabela 6: *POS Tagger*

DESCRIÇÃO	GRUPO L1	Document	Term	Term as String
O MOTORISTA CONDUZINDO O ÔNIBUS ...	DIREÇÃO	""	motorista[NN(PO...	motorista
O MOTORISTA CONDUZINDO O ÔNIBUS ...	DIREÇÃO	""	conduzindo[NN(P...	conduzindo
O MOTORISTA CONDUZINDO O ÔNIBUS ...	DIREÇÃO	""	onibus[NN(POS)]	onibus

Fonte: O autor (2023)

Após a extração das palavras, foi realizado um agrupamento que tem por objetivo quantificar quantas palavras aparecem em toda a base de acidentes de trabalho, por exemplo “colaborador” aparece 1726 vezes, “atividade” aparece 1354 vezes, entre outras palavras destacadas no agrupamento realizado, como pode ser visto na Tabela 7 abaixo:

Tabela 7: Contagem de palavras

S Term as String	I Document
colaborador	1726
atividade	1354
veiculo	1285
risco	1243
onibus	1239
empresa	1221
durante	1180
uso	1117
seguranca	1056
mesmo	987
area	982
motorista	947
empregado	942

Fonte: O autor (2023)

Após essa etapa de agrupamento de palavras, é realizado a frequência de termos (TF), que representa a frequência de um termo em um documento e é usado para medir com que frequência um termo aparece em um documento específico.

Conforme citado por Qaiser e Ali (2018) o conceito de IDF que é usado para atribuir maior importância a palavras que são menos frequentes nos documentos. Isso ocorre porque, ao calcular apenas o TF, todas as palavras são tratadas igualmente. Portanto, utilizar o TF-IDF em sua combinação é mais relevante, do que usar somente a frequência dos termos.

Tabela 8: TF-IDF na prática

T Term	S Term as String	D TF rel
motorista[NN(PO...	motorista	0.069
conduzindo[NN(P...	conduzindo	0.034
onibus[NN(POS)]	onibus	0.069
manobra[NN(POS)]	manobra	0.034
veio[NN(POS)]	veio	0.034
veiculo[NN(POS)]	veiculo	0.069
porta[NN(POS)]	porta	0.034
lateral[JJ(POS)]	lateral	0.034

Fonte: O autor (2023)

Quanto menor o TF, menos a palavra se repete no documento, podemos notar no exemplo acima que as palavras que menos se repetem são as “conduzindo”, “manobra”, “veio”, “porta”, “lateral” e as que se repetem com uma maior frequência são as “motorista”, “ônibus” e “veículo”.

Após a etapa de pré-processamento mencionado nas tabelas anteriores, é transformado o documento em texto, mas de maneira que o algoritmo é capaz de realizar seu treinamento, na Tabela 9 abaixo, podemos ver uma amostra de como a preparação de dados realizou todo o tratamento de dados textuais com objetivo de realizar o treinamento dos modelos preditivos.

Tabela 9: Base tratada

D mascara	D luvas	D epis	D oculos	D dss	S Document class
0	0	1	1	0	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO
1	0	0	0	0	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO
0	0	1	0	0	COMPORTAMENTO
0	0	1	0	0	COMPORTAMENTO

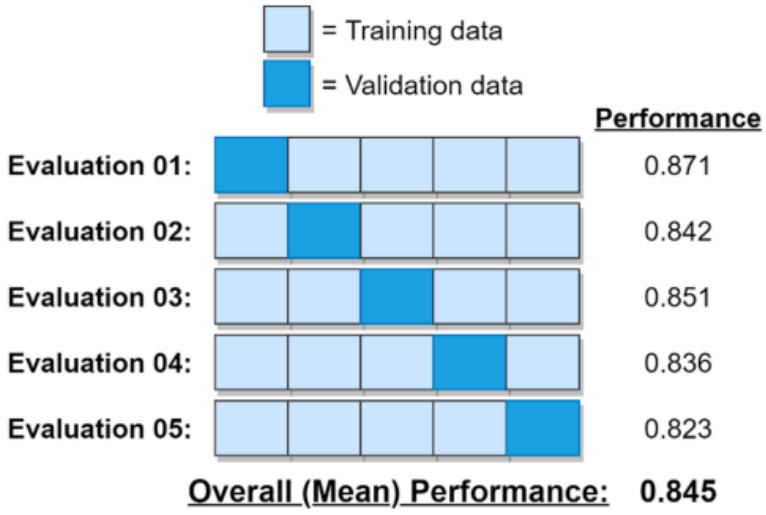
Fonte: O autor (2023)

Na Tabela 9 acima, é possível observar que em uma combinação palavras de um registro se repete a palavra “EPI” e “óculos” a categoria desse acidente é comportamento, com base nos dados históricos, porém é importante ressaltar que essa preparação de dados gerou mais de 200 colunas com outras palavras elencadas, portanto para classificar como uma categoria não é realizado a combinação de uma ou duas palavras para afirmar uma classe de categoria, mas sim um conjunto de palavras.

4.3 Divisão dos dados em treinamento, validação e teste

Soper (2021) aborda sobre a técnica de validação cruzada *k-fold*, onde é um processo que envolve a divisão dos dados em k subconjuntos (chamados de *folds*), sendo cada parte usado iterativamente como conjunto de validação para um modelo candidato treinado com os dados das demais partes. Cada conjunto contém um número igual ou aproximadamente igual de casos, com a atribuição dos dados realizado de forma aleatória.

Figura 11: *K-Fold*



Fonte: Soper (2021)

No seguinte projeto, foi realizada a divisão utilizando um conjunto de treino e teste. Murphy (2012) afirma que o aprendizado supervisionado tem por objetivo aprender a partir de um conjunto de dados categorizados, ou seja, no aprendizado supervisionado faz-se necessário o uso de dados de treinamento para que categorize um aprendizado supervisionado e dados de teste é o conjunto que não foi treinado pelo modelo para realizar o teste da base de treino pelo modelo de aprendizado supervisionado.

No seguinte modelo, foi dividido as bases de treino e teste por 70% treino e 30% para teste. De um total de 10712 linhas, na base de treino foi distribuído aproximadamente 7 mil linhas e para os dados de teste aproximadamente 3 mil linhas.

4.4 Aprendizado de máquina

Decision Tree **-** Conforme citado anteriormente na etapa de modelagem, esse tópico tem o objetivo de realizar a seleção e aplicação de várias técnicas de modelagem. Para o algoritmo de árvore de decisão foram selecionados dois “nós” para treinar e testar os dados, o “nó” de treinamento denominado *Decision Tree Learner* tem como objetivo realizar o treinamento do conjunto de dados, e o “nó” *Decision Tree Predictor* é utilizado para testar os dados que foi treinado a partir do outro "nó" e compare-o com a base de teste de 30% para ver se o algoritmo foi eficiente. As árvores de decisão aprendem com os dados e formam um conjunto de regras de decisão do tipo "se-então-senão". Quanto mais profunda a árvore, mais complexas são as regras de decisão e mais ajustado é o modelo.

Random Forest **-** No algoritmo de floresta aleatória, dois “nós” são selecionados para treinar e testar os dados. O primeiro nó, conhecido como *Random Forest Learner*, é responsável por treinar o conjunto de dados. O segundo nó, o *Random Forest Predictor*, testa os dados treinados pelo nó anterior e os compara com uma base de teste de 30% para avaliar a eficiência do algoritmo.

SVM - No algoritmo SVM foram escolhidos dois componentes para treinamento e teste de dados: o módulo de treinamento, denominado SVM *Learner*, que visa treinar o conjunto de dados, e o módulo SVM *Predictor*, utilizado para testar os dados treinados pelo módulo anterior. Este processo envolve a comparação das previsões com um conjunto de testes separado de 30% para avaliar a eficiência do algoritmo. Conforme citado anteriormente por Jain (2020) ressalta que esse algoritmo se baseia na busca pelo valor/linha previsto ótimo no conjunto de dados. Em sua essência, o SVM é um classificador discriminativo definido formalmente por um hiperplano separador.

KNN - No algoritmo KNN, um único “nó” é designado para fins de treinamento e teste. Este nó treina em um conjunto de dados e posteriormente testa os dados treinados em outro “nó”. Esta comparação envolve avaliar a eficiência do algoritmo comparando-o com uma base de teste de 30%. Em seguida, são selecionados os K pontos mais próximos dos dados de teste.

Naive Bayes **-** No algoritmo *Naive Bayes*, dois "nós" distintos são utilizados para treinar e testar dados. O “nó” de treinamento, conhecido como *Naive Bayes Learner*, é responsável pelo treinamento no conjunto de dados fornecido. Por outro lado, o "nó" *Naive Bayes Predictor* é empregado para testar os dados treinados a partir do "nó" anterior e compará-los com o conjunto de dados de teste de 30%, avaliando a eficiência do algoritmo.

Regressão Logística - O algoritmo de Regressão Logística envolve dois componentes principais para treinar e testar os dados: o *Logistic Regression Learner*, responsável por treinar o conjunto de dados, e o Preditor de Regressão Logística, que testa os dados treinados em um conjunto de testes separado e avalia sua eficiência em relação a uma amostra de 30%.

4.5 Comparação de modelos com avaliação de desempenho

Conforme citado anteriormente na etapa de avaliação, esse tópico se concentra na análise dos modelos obtidos e na tomada de decisões sobre como usar os resultados (Shafique e Qaiser, 2014). Essa fase desempenha um papel fundamental na análise crítica dos resultados

da mineração de dados, avaliando a eficácia e relevância dos modelos em relação às metas de negócios.

4.5.1 Matriz de Contingência

Junior et al. (2022) ressaltam que a matriz de confusão, comumente utilizada em diagnósticos clínicos e em aprendizado de máquina, registra os resultados das classificações em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

Segundo os algoritmos implementados no seguinte projeto, a matriz de confusão possui vários valores a serem preditos, pois a ideia é realizar a previsão das categorias, “estrutural”,”direção”,”estrutura do veículo”,”comportamento”,”equipamentos”. Na Tabela 10 exemplifica a comparação de 3 algoritmos iniciais: árvore de decisão, floresta aleatória e SVM:

Tabela 10: Matriz de confusão dos primeiros algoritmos

Row ID	S Algoritmo	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives
Row0	Árvore de decisão	?	?	?	?
ESTRUTURAL	?	653	194	2210	157
DIREÇÃO	?	226	67	2807	114
ESTRUTURA ...	?	148	70	2904	92
COMPORTAM...	?	1521	271	1235	187
EQUIPAMENTOS	?	28	36	3062	88
Overall	?	?	?	?	?
Row0_dup	Floresta aleatória	?	?	?	?
ESTRUTURAL...	?	721	180	2224	89
DIREÇÃO_dup	?	274	65	2809	66
ESTRUTURA ...	?	156	37	2937	84
COMPORTAM...	?	1574	177	1329	134
EQUIPAMENT...	?	21	9	3089	95
Overall_dup	?	?	?	?	?
Row0_dup_dup	SVM	?	?	?	?
ESTRUTURAL...	?	720	239	2165	90
DIREÇÃO_du...	?	262	90	2784	78
ESTRUTURA ...	?	113	21	2953	127
COMPORTAM...	?	1549	201	1305	159
EQUIPAMENT...	?	8	11	3087	108

Fonte: O Autor (2024)

Realizando a análise entre os três primeiros algoritmos, é possível ver os verdadeiros positivos (*TruePositives*), falsos positivos (*FalsePositives*), verdadeiros negativos (*TrueNegatives*) e os falsos negativos (*FalseNegatives*), de um total de 3.214 registros de teste para cada categoria predita:

- Verdadeiros positivos: Em todas as categorias, exceto a categoria “equipamentos”, o algoritmo que teve maior resultado foi o floresta aleatória, já na categoria “equipamentos” a árvore de decisão obteve mais verdadeiros positivos.
- Falsos positivos: Em todas as categorias, exceto a categoria “estrutura do veículo”, o algoritmo que teve menor resultado foi o floresta aleatória, já nesta categoria, o algoritmo SVM obteve mais falsos positivos.
- Verdadeiros negativos: Em todas as categorias, exceto a categoria “estrutura do veículo”, o algoritmo que teve maior resultado foi o floresta aleatória, já nesta categoria, o algoritmo SVM obteve mais verdadeiros negativos.
- Falsos negativos: Em todas as categorias, exceto a categoria “equipamentos”, o algoritmo que teve menor resultado foi a floresta aleatória, já nesta categoria, a árvore de decisão obteve mais falsos negativos.

Com a comparação dos 3 algoritmos iniciais pode-se perceber que o floresta aleatória obteve melhor desempenho.

Tabela 11: Matriz de confusão dos últimos algoritmos

Row ID	S Algoritmo	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives
Row0_dup_d...	KNN	?	?	?	?
ESTRUTURAL...	?	731	301	2103	79
DIREÇÃO_du...	?	163	21	2853	177
ESTRUTURA ...	?	147	74	2900	93
COMPORTAM...	?	1493	264	1242	215
EQUIPAMENT...	?	11	9	3089	105
Overall_dup_...	?	?	?	?	?
Row0_dup_d...	Naive Bayes	?	?	?	?
EQUIPAMENT...	?	116	3098	0	0
ESTRUTURAL...	?	0	0	2404	810
DIREÇÃO_du...	?	0	0	2874	340
ESTRUTURA ...	?	0	0	2974	240
COMPORTAM...	?	0	0	1506	1708
Overall_dup_...	?	?	?	?	?
Row0_dup_d...	Regressão Logís...	?	?	?	?
ESTRUTURAL...	?	713	185	2219	97
DIREÇÃO_du...	?	265	60	2814	75
ESTRUTURA ...	?	167	51	2923	73
COMPORTAM...	?	1542	177	1329	166
EQUIPAMENT...	?	23	31	3067	93
Overall_dup_...	?	?	?	?	?

Fonte: O Autor (2024)

- Verdadeiros Positivos: Em todas as categorias, exceto "estrutural" e “equipamentos”, o algoritmo Regressão Logística apresentou o maior número de verdadeiros positivos. Na categoria "Estrutural", o KNN obteve mais verdadeiros positivos.

- Falsos Positivos: Em todas as categorias, exceto "equipamentos", o *Naive Bayes* teve o menor número de falsos positivos. Porém, todas as demais categorias o algoritmo *Naive Bayes* apresentou uma quantidade zerada de falsos positivos.
- Verdadeiros Negativos: Em todas as categorias, exceto " equipamentos ", o *Naive Bayes* apresentou o maior número de verdadeiros negativos. Somente essa categoria que estava com uma quantidade zerada.
- Falsos Negativos: Em todas as categorias, exceto "equipamentos", o *Naive Bayes* teve o maior número de falsos positivos. Somente essa categoria que estava com uma quantidade zerada.

Com a comparação dos 3 últimos algoritmos pode-se perceber que o algoritmo *Naive Bayes* se destacou como o melhor em várias métricas, No entanto, esses valores que estão zerados, pode influenciar no algoritmo. No próximo tópico iremos comparar a precisão, *recall*, sensibilidade, especificidade e o *F1-score*.

4.5.2 Precisão, sensibilidade, especificidade, *F1-score* e acurácia.

- Precisão:

Lima (2023) destaca que a precisão é definida como a proporção de verdadeiros positivos em relação a todos os exemplos rotulados como positivos. Em sua explicação, ele ressalta que a precisão representa a exatidão das predições positivas.

Tabela 12: Precisão

Row ID	S Algoritmo	D Precision	Row ID	S Algoritmo	D Precision
Row0	Árvore de decisão	?	Row0_dup_d...	KNN	?
ESTRUTURAL	?	0.771	ESTRUTURAL...	?	0.708
DIREÇÃO	?	0.771	DIREÇÃO_du...	?	0.886
ESTRUTURA ...	?	0.679	ESTRUTURA ...	?	0.665
COMPORTAM...	?	0.849	COMPORTAM...	?	0.85
EQUIPAMENTOS	?	0.438	EQUIPAMENT...	?	0.55
Overall	?	?	Overall_dup_...	?	?
Row0_dup	Floresta aleatória	?	Row0_dup_d...	Naive Bayes	?
ESTRUTURAL...	?	0.8	EQUIPAMENT...	?	0.036
DIREÇÃO_dup	?	0.808	ESTRUTURAL...	?	?
ESTRUTURA ...	?	0.808	DIREÇÃO_du...	?	?
COMPORTAM...	?	0.899	ESTRUTURA ...	?	?
EQUIPAMENT...	?	0.7	COMPORTAM...	?	?
Overall_dup	?	?	Overall_dup_...	?	?
Row0_dup_dup	SVM	?	Row0_dup_d...	Regressão Logís...	?
ESTRUTURAL...	?	0.751	ESTRUTURAL...	?	0.794
DIREÇÃO_du...	?	0.744	DIREÇÃO_du...	?	0.815
ESTRUTURA ...	?	0.843	ESTRUTURA ...	?	0.766
COMPORTAM...	?	0.885	COMPORTAM...	?	0.897
EQUIPAMENT...	?	0.421	EQUIPAMENT...	?	0.426
Overall_dup_...	?	?	Overall_dup_...	?	?

Fonte: O Autor (2024)

Analisando os algoritmos, o algoritmo que se destacou na precisão foi o floresta aleatória em 3 categorias, estrutural, comportamento e equipamentos, com uma taxa de mais de 80%, exceto em equipamentos com 70%. Na categoria de direção, o algoritmo KNN teve melhor precisão com 88% e na categoria estrutura do veículo, o SVM teve melhor precisão com 84%.

- Sensibilidade:

A sensibilidade, também conhecida como *recall*, refere-se à taxa de detecção da classe positiva, expressa como a proporção de verdadeiros positivos em relação ao total de exemplos positivos (Lima, 2023). Um baixo *recall* indica a presença de muitos falsos negativos, ou seja, casos positivos que não foram identificados pelo modelo. O autor ressalta que o *recall* deve ser priorizado em situações em que a ocorrência de falsos negativos seja mais prejudicial do que a de falsos positivos.

Tabela 13: Sensibilidade

Row ID	Algoritmo	Sensitivity	Row ID	Algoritmo	Sensitivity
Row0	Árvore de decisão	?	Row0_dup_d...	KNN	?
ESTRUTURAL	?	0.806	ESTRUTURAL...	?	0.902
DIREÇÃO	?	0.665	DIREÇÃO_du...	?	0.479
ESTRUTURA ...	?	0.617	ESTRUTURA ...	?	0.613
COMPORTAM...	?	0.891	COMPORTAM...	?	0.874
EQUIPAMENTOS	?	0.241	EQUIPAMENT...	?	0.095
Overall	?	?	Overall_dup_...	?	?
Row0_dup	Floresta aleatória	?	Row0_dup_d...	Naive Bayes	?
ESTRUTURAL...	?	0.89	EQUIPAMENT...	?	1
DIREÇÃO_dup	?	0.806	ESTRUTURAL...	?	0
ESTRUTURA ...	?	0.65	DIREÇÃO_du...	?	0
COMPORTAM...	?	0.922	ESTRUTURA ...	?	0
EQUIPAMENT...	?	0.181	COMPORTAM...	?	0
Overall_dup	?	?	Overall_dup_...	?	?
Row0_dup_dup	SVM	?	Row0_dup_d...	Regressão Logís...	?
ESTRUTURAL...	?	0.889	ESTRUTURAL...	?	0.88
DIREÇÃO_du...	?	0.771	DIREÇÃO_du...	?	0.779
ESTRUTURA ...	?	0.471	ESTRUTURA ...	?	0.696
COMPORTAM...	?	0.907	COMPORTAM...	?	0.903
EQUIPAMENT...	?	0.069	EQUIPAMENT...	?	0.198
Overall_dup_...	?	?	Overall_dup_...	?	?

Fonte: O Autor (2024)

Essa análise mostra que diferentes algoritmos apresentam melhor desempenho em diferentes categorias. O *Naive Bayes* obteve a maior sensibilidade na categoria de equipamentos com 100%, a Floresta Aleatória de direção com 80% e comportamento com 92%, o KNN na categoria estrutural com 90%, enquanto a Regressão Logística se destacou na categoria de estrutura do veículo com 69% de sensibilidade (*recall*).

- Especificidade:

A especificidade se refere à taxa de detecção da classe negativa, sendo também conhecida como taxa de verdadeiros negativos (Santos, 2020). Ele aborda que essa medida é calculada pela proporção de verdadeiros negativos em relação ao total de exemplos negativos. Vamos agora analisar a especificidade dos algoritmos mencionados no seguinte projeto:

Tabela 14: Especificidade

Row ID	S Algoritmo	D Specificity
Row0	Árvore de decisão	?
ESTRUTURAL	?	0.919
DIREÇÃO	?	0.977
ESTRUTURA ...	?	0.976
COMPORTAM...	?	0.82
EQUIPAMENTOS	?	0.988
Overall	?	?
Row0_dup	Floresta aleatória	?
ESTRUTURAL...	?	0.925
DIREÇÃO_dup	?	0.977
ESTRUTURA ...	?	0.988
COMPORTAM...	?	0.882
EQUIPAMENT...	?	0.997
Overall_dup	?	?
Row0_dup_dup	SVM	?
ESTRUTURAL...	?	0.901
DIREÇÃO_du...	?	0.969
ESTRUTURA ...	?	0.993
COMPORTAM...	?	0.867
EQUIPAMENT...	?	0.996
Overall_dup_...	?	?

Row ID	S Algoritmo	D Specificity
Row0_dup_d...	KNN	?
ESTRUTURAL...	?	0.875
DIREÇÃO_du...	?	0.993
ESTRUTURA ...	?	0.975
COMPORTAM...	?	0.825
EQUIPAMENT...	?	0.997
Overall_dup_...	?	?
Row0_dup_d...	Naive Bayes	?
EQUIPAMENT...	?	0
ESTRUTURAL...	?	1
DIREÇÃO_du...	?	1
ESTRUTURA ...	?	1
COMPORTAM...	?	1
Overall_dup_...	?	?
Row0_dup_d...	Regressão Logís...	?
ESTRUTURAL...	?	0.923
DIREÇÃO_du...	?	0.979
ESTRUTURA ...	?	0.983
COMPORTAM...	?	0.882
EQUIPAMENT...	?	0.99
Overall_dup_...	?	?

Fonte: O Autor (2024)

A análise da especificidade mostra que o algoritmo *Naive Bayes* fica com 100% em todas as categorias, exceto em equipamentos, que KNN e Floresta Aleatória empatam com 99,7%. Ou seja, existe um acerto majoritário por parte dos algoritmos nos valores que são negativos.

- *F1-score*:

Lima (2023) destaca que o *F1-Score* é uma métrica que equilibra os objetivos de precisão e *recall*, sendo definida como a média harmônica entre essas duas medidas. A fórmula para o cálculo do *F1-Score* é dada por:

Figura 12: Fórmula *F1-score*

$$F1\text{-}Score = 2 * \frac{Precisão * Recall}{Precisão + Recall}$$

Fonte: Lima (2023)

Essa métrica é amplamente utilizada em tarefas de classificação que envolvem classes desbalanceadas. Segue a análise da métrica *F1-score* dos algoritmos mencionados no seguinte projeto:

Tabela 15: *F1-score*

Row ID	Algoritmo	F-measure	Row ID	Algoritmo	F-measure
Row0	Árvore de decisão	?	Row0_dup_d...	KNN	?
ESTRUTURAL	?	0.788	ESTRUTURAL...	?	0.794
DIREÇÃO	?	0.714	DIREÇÃO_du...	?	0.622
ESTRUTURA ...	?	0.646	ESTRUTURA ...	?	0.638
COMPORTAM...	?	0.869	COMPORTAM...	?	0.862
EQUIPAMENTOS	?	0.311	EQUIPAMENT...	?	0.162
Overall	?	?	Overall_dup_...	?	?
Row0_dup	Floresta aleatória	?	Row0_dup_d...	Naive Bayes	?
ESTRUTURAL...	?	0.843	EQUIPAMENT...	?	0.07
DIREÇÃO_dup	?	0.807	ESTRUTURAL...	?	?
ESTRUTURA ...	?	0.721	DIREÇÃO_du...	?	?
COMPORTAM...	?	0.91	ESTRUTURA ...	?	?
EQUIPAMENT...	?	0.288	COMPORTAM...	?	?
Overall_dup	?	?	Overall_dup_...	?	?
Row0_dup_dup	SVM	?	Row0_dup_d...	Regressão Logís...	?
ESTRUTURAL...	?	0.814	ESTRUTURAL...	?	0.835
DIREÇÃO_du...	?	0.757	DIREÇÃO_du...	?	0.797
ESTRUTURA ...	?	0.604	ESTRUTURA ...	?	0.729
COMPORTAM...	?	0.896	COMPORTAM...	?	0.9
EQUIPAMENT...	?	0.119	EQUIPAMENT...	?	0.271
Overall_dup_...	?	?	Overall_dup_...	?	?

Fonte: O Autor (2024)

A análise dos resultados do *F1-Score* para cada classe revela uma variação no desempenho dos algoritmos utilizados. Nas classes como estrutural, comportamento e direção, o algoritmo de Floresta Aleatória demonstrou consistência, alcançando *F1-Scores* de 84%, 91% e 80%, respectivamente, indicando alta precisão e *recall*. Por outro lado, a classe Equipamentos apresentou um desempenho abaixo do esperado, com um *F1-Score* de apenas 31% utilizando o algoritmo de Árvore. A classe Estrutura do veículo também registrou um *F1-Score* um pouco mais baixo, atingindo 72,9% com o algoritmo de Regressão Logística.

- Acurácia:

A acurácia é uma medida que avalia a proporção de acertos em relação ao total de exemplos classificados, sendo calculada como a soma dos verdadeiros positivos e verdadeiros negativos dividida pelo total de exemplos (Lima, 2023). Entretanto, é ressaltado que um alto valor de acurácia nem sempre reflete uma boa performance do modelo, especialmente quando falsos positivos e falsos negativos têm implicações diferentes. Vamos agora analisar a acurácia dos algoritmos mencionados no seguinte projeto:

Tabela 16: Acurácia

Row ID	S Algoritmo	D Accuracy
Row0	Árvore de decisão	?
ESTRUTURAL	?	?
DIREÇÃO	?	?
ESTRUTURA ...	?	?
COMPORTAM...	?	?
EQUIPAMENTOS	?	?
Overall	?	0.801
Row0_dup	Floresta aleatória	?
ESTRUTURAL...	?	?
DIREÇÃO_dup	?	?
ESTRUTURA ...	?	?
COMPORTAM...	?	?
EQUIPAMENT...	?	?
Overall_dup	?	0.854
Row0_dup_dup	SVM	?
ESTRUTURAL...	?	?
DIREÇÃO_du...	?	?
ESTRUTURA ...	?	?
COMPORTAM...	?	?
EQUIPAMENT...	?	?
Overall_dup_...	?	0.825

Row ID	S Algoritmo	D Accuracy
Row0_dup_d...	KNN	?
ESTRUTURAL...	?	?
DIREÇÃO_du...	?	?
ESTRUTURA ...	?	?
COMPORTAM...	?	?
EQUIPAMENT...	?	?
Overall_dup_...	?	0.792
Row0_dup_d...	Naive Bayes	?
EQUIPAMENT...	?	?
ESTRUTURAL...	?	?
DIREÇÃO_du...	?	?
ESTRUTURA ...	?	?
COMPORTAM...	?	?
Overall_dup_...	?	0.036
Row0_dup_d...	Regressão Logís...	?
ESTRUTURAL...	?	?
DIREÇÃO_du...	?	?
ESTRUTURA ...	?	?
COMPORTAM...	?	?
EQUIPAMENT...	?	?
Overall_dup_...	?	0.843

Fonte: O Autor (2024)

Ao analisar os resultados de acurácia dos diferentes algoritmos, observa-se uma variação significativa no desempenho de cada modelo. No entanto, o *Naive Bayes* registrou uma acurácia muito baixa, de apenas 3.6%, o que indica um desempenho insatisfatório. Essa disparidade nos resultados destaca a importância de selecionar cuidadosamente o algoritmo mais adequado para a tarefa específica e ressalta a necessidade de considerar outras métricas além da acurácia, especialmente em casos de desbalanceamento entre as classes (Santos, 2020).

4.5.3 Kappa de Cohen

Henry et al. (2017) enfatizam a importância da estatística Kappa de Cohen na avaliação do desempenho do modelo de regressão logística para classificação de áreas de plantio de cana-de-açúcar. Eles destacam que o coeficiente kappa pode ser interpretado como um coeficiente de correlação interclasse e pode ser utilizado para medir a confiabilidade entre avaliadores. A interpretação do coeficiente kappa permite avaliar a força de concordância entre duas variáveis, conforme descrito na Tabela 17 abaixo:

Tabela 17: Métricas de Kappa

Value of κ	Strength of Agreement
< 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
> 0.80	Very Good

Fonte: Henry et al. (2017)

Henry et al. (2017) apresentam a fórmula do coeficiente de *Cohen's Kappa* que é dada por OA que representa a proporção de concordância observada (ou seja, a proporção de concordância entre os avaliadores ou entre os métodos de medição observados). AC representa a proporção de concordância esperada devida ao acaso.

Equação 2: Fórmula de Kappa

$$\kappa = \frac{OA - AC}{1 - AC}$$

Fonte: Henry et al. (2017)

Essa fórmula leva em consideração a possibilidade de concordância puramente aleatória e, portanto, fornece uma correção para a concordância observada. Conforme Tabela 13 a análise da métrica Kappa de Cohen nos algoritmos utilizados:

Tabela 18: Valores de Kappa

Row ID	S Algoritmo	D Cohen's kappa	Row ID	S Algoritmo	D Cohen's kappa
Row0	Árvore de decisão	?	Row0_dup_d...	KNN	?
ESTRUTURAL	?	?	ESTRUTURAL...	?	?
DIREÇÃO	?	?	DIREÇÃO_du...	?	?
ESTRUTURA ...	?	?	ESTRUTURA ...	?	?
COMPORTAM...	?	?	COMPORTAM...	?	?
EQUIPAMENTOS	?	?	EQUIPAMENT...	?	?
Overall	?	0.681	Overall_dup_...	?	0.663
Row0_dup	Floresta aleatória	?	Row0_dup_d...	Naive Bayes	?
ESTRUTURAL...	?	?	EQUIPAMENT...	?	?
DIREÇÃO_dup	?	?	ESTRUTURAL...	?	?
ESTRUTURA ...	?	?	DIREÇÃO_du...	?	?
COMPORTAM...	?	?	ESTRUTURA ...	?	?
EQUIPAMENT...	?	?	COMPORTAM...	?	?
Overall_dup	?	0.767	Overall_dup_...	?	0
Row0_dup_dup	SVM	?	Row0_dup_d...	Regressão Logís...	?
ESTRUTURAL...	?	?	ESTRUTURAL...	?	?
DIREÇÃO_du...	?	?	DIREÇÃO_du...	?	?
ESTRUTURA ...	?	?	ESTRUTURA ...	?	?
COMPORTAM...	?	?	COMPORTAM...	?	?
EQUIPAMENT...	?	?	EQUIPAMENT...	?	?
Overall_dup_...	?	0.718	Overall_dup_...	?	0.751

Fonte: O Autor (2024)

Henry et al. (2017) reforça que a métrica *Cohen's Kappa* é uma medida que avalia a concordância entre dois avaliadores (ou neste caso, entre os resultados dos diferentes algoritmos) ajustada para concordância aleatória. Geralmente, valores de *Cohen's Kappa* próximos a 1 indicam uma concordância quase perfeita, valores entre 0.61 e 0.80 indicam uma boa concordância, entre 0.41 e 0.60 indicam uma concordância moderada, entre 0.21 e 0.40 indicam uma concordância justa, e valores abaixo de 0.20 indicam uma concordância pobre.

Avaliando os valores fornecidos para os diferentes algoritmos:

- Árvore de decisão: 0.68 (Moderado)
- Floresta aleatória: 0.76 (Bom)
- SVM: 0.71 (Bom)
- KNN: 0.66 (Moderado)
- *Naive Bayes*: 0.0 (Não houve concordância)
- Regressão Logística: 0.75 (Bom)

Baseado nos valores de *Cohen's Kappa* fornecidos, os algoritmos de Floresta Aleatória, SVM e Regressão Logística demonstram uma concordância relativamente boa em relação aos outros algoritmos, enquanto a Árvore de Decisão e KNN mostram uma concordância moderada. No entanto, o *Naive Bayes* parece não ter apresentado concordância em absoluto.

5 APRESENTAÇÃO DE RESULTADOS

Para visualizar o que foi realizado, uma amostra dos valores treinados e aplicados nos dados de teste são apresentadas. Os valores reais de cada combinação de palavras estão na coluna "*document class*", localizada no centro da Tabela 19, enquanto o valor predito está à direita da imagem. Conforme Tabela 16, é possível observar uma acurácia de 80,1% no algoritmo árvore de decisão. Em sua maioria, a classe correta "comportamento" é prevista, enquanto os 20% de erro revelam outras categorias, como "estrutural" e "estrutura do veículo".

Tabela 19: Resultados Árvore de Decisão

ículos	dss	Document cl...	P (Docu...	P (Docu...	P (Docu...	P (Docu...	P (Docu...	Prediction (D...
	0	COMPORTAMENTO	0.009	0.003	0.003	0.985	0	COMPORTAMENTO
	0	COMPORTAMENTO	0.023	0.011	0.011	0.943	0.011	COMPORTAMENTO
	0	COMPORTAMENTO	0	0.105	0.026	0.868	0	COMPORTAMENTO
	0	COMPORTAMENTO	0.049	0	0.024	0.902	0.024	COMPORTAMENTO
	0	COMPORTAMENTO	0.009	0.003	0.003	0.985	0	COMPORTAMENTO
	0	COMPORTAMENTO	0.006	0.008	0.004	0.973	0.008	COMPORTAMENTO
	0	COMPORTAMENTO	0.925	0.002	0.007	0.041	0.024	ESTRUTURAL
	0	COMPORTAMENTO	0.08	0.06	0	0.84	0.02	COMPORTAMENTO
	0	COMPORTAMENTO	0	0	0.857	0.143	0	ESTRUTURA DO ...
	0	COMPORTAMENTO	0.021	0.007	0.003	0.951	0.017	COMPORTAMENTO
	0	COMPORTAMENTO	0.25	0	0.036	0.536	0.179	COMPORTAMENTO
	0	COMPORTAMENTO	0.008	0.005	0.001	0.981	0.004	COMPORTAMENTO
	0	COMPORTAMENTO	0.006	0.024	0.071	0.893	0.006	COMPORTAMENTO
	0	COMPORTAMENTO	0	0	0.026	0.974	0	COMPORTAMENTO
	0	COMPORTAMENTO	0.006	0.008	0.004	0.973	0.008	COMPORTAMENTO
	0	COMPORTAMENTO	0.021	0.007	0.003	0.951	0.017	COMPORTAMENTO
	0	COMPORTAMENTO	0.25	0	0	0.75	0	COMPORTAMENTO
	0	COMPORTAMENTO	0.031	0.01	0	0.958	0	COMPORTAMENTO
	0	COMPORTAMENTO	0.006	0.008	0.004	0.973	0.008	COMPORTAMENTO
	0	COMPORTAMENTO	0.021	0.007	0.003	0.951	0.017	COMPORTAMENTO
	0	COMPORTAMENTO	0.006	0.008	0.004	0.973	0.008	COMPORTAMENTO
	0	COMPORTAMENTO	0.039	0	0.024	0.937	0	COMPORTAMENTO

Fonte: O Autor (2024)

Na Tabela 20 é exibido outra amostra dos dados de treino e aplicado nos dados de teste, conforme algoritmo Floresta Aleatória. Os valores reais de cada combinação de palavras estão na coluna "*document class*", localizada na direita da Tabela 20, enquanto o valor predito está ao seu lado direito. Além disso, também é possível observar uma boa acurácia que foi apresentado anteriormente para este algoritmo. Em sua maioria, a classe correta "comportamento" é prevista.

Tabela 20: Resultados Floresta aleatória

luvas	D epis	D oculos	D dss	S Document cl...	S Prediction (D...	D Predicti...
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.71
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.95
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.59
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	ESTRUTURAL	0.84
	0	1	0	COMPORTAMENTO	COMPORTAMENTO	0.94
	0	0	0	COMPORTAMENTO	ESTRUTURA DO ...	0.53
	1	0	0	COMPORTAMENTO	COMPORTAMENTO	0.99
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.48
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.99
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.78
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.82
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.79
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	1
	1	1	0	COMPORTAMENTO	COMPORTAMENTO	1
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.97
	0	0	0	COMPORTAMENTO	COMPORTAMENTO	0.94

Fonte: O Autor (2024)

Na próxima representação visual, é apresentada uma nova amostra dos dados de treinamento e dos dados de teste, desta vez utilizando o algoritmo SVM. Os valores reais de cada combinação de palavras estão listados na coluna "*document class*", situada à direita na Tabela 21, enquanto as previsões correspondentes estão ao lado. Além disso, é perceptível uma alta acurácia que foi previamente demonstrada para este algoritmo. Na maioria dos casos, a classe correta "comportamento" é identificada com precisão.

Tabela 21: Resultados SVM

luvas	D epis	D oculos	D dss	S Document cl...	S Prediction (D...
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	ESTRUTURAL
	0	1	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	ESTRUTURA DO ...
	1	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	DIREÇÃO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	1	1	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	1	0	COMPORTAMENTO	COMPORTAMENTO

Fonte: O Autor (2024)

Na Tabela 22, é apresentada uma nova exemplificação dos dados de treinamento e aplicação nos dados de teste, utilizando o algoritmo KNN. Os valores reais de cada combinação de palavras estão registrados na coluna "*document class*", situada à direita da figura, enquanto as previsões estão ao lado direito na coluna “Class [KNN]”. Adicionalmente, é perceptível uma precisão notável, conforme indicado anteriormente para este algoritmo, com a predominância de previsões corretas para a classe "comportamento".

Tabela 22: Resultados KNN

D mascara	D luvas	D epis	D oculos	D dss	S Document cl...	S Class [kNN]
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
1	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	ESTRUTURAL
0	0	0	1	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	1	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
0	0	0	0	0	COMPORTAMENTO	ESTRUTURAL

Fonte: O Autor (2024)

Agora na Tabela 23 seguinte, é apresentada uma exemplificação dos dados de treinamento e aplicação nos dados de teste, utilizando o algoritmo *Naive Bayes*. Porém dessa vez os valores não obtiveram uma acurácia satisfatória, todos os valores desta amostra o algoritmo não conseguiu trazer o valor predito corretamente, isso pode acontecer devido alguns problemas

Tabela 23: Resultados *Naive Bayes*

uvas	D epis	D oculos	D dss	S Document cl...	S Prediction...
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	1	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	1	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS
	0	0	0	COMPORTAMENTO	EQUIPAMENTOS

Fonte: O Autor (2024)

Na Tabela 24, é demonstrada uma nova amostra dos dados de treino e teste, agora utilizando o Logistic Regression. Os valores reais de cada combinação de palavras estão listados na coluna "*document class*", posicionada à direita da figura, enquanto as previsões correspondentes estão ao lado. Além disso, é evidente uma alta precisão que foi previamente destacada para este algoritmo. Em grande parte das vezes, a classe correta "comportamento" é corretamente identificada.

Tabela 24: Resultados Regressão Logística

o	D sonolen...	D mascara	D luvas	D epis	D oculos	D dss	S Document cl...	S Prediction (D...
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	1	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	ESTRUTURAL
	0	0	0	0	1	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	ESTRUTURA DO ...
	0	0	0	1	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	1	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	0	0	0	0	0	COMPORTAMENTO	COMPORTAMENTO
	0	1	0	1	1	0	COMPORTAMENTO	COMPORTAMENTO

Fonte: O Autor (2024)

6 ANÁLISE E DISCUSSÃO DE RESULTADOS

Ao realizar a análise do objetivo de negócio, mencionado na primeira etapa do CRISP-DM, a proposta inicial é reduzir o retrabalho dos profissionais de segurança por meio da automatização da categorização de eventos, para que isso ocorra de maneira eficaz, uma acurácia acima de 80% seria um valor inicial para utilização, ao realizar a análise da acurácia dos seis algoritmos, quatro deles obtiveram um valor acima de 80% esperado pelos objetivos de negócio.

Ao analisar os algoritmos, foi percebido que o *Naive Bayes* e o KNN não obtiveram um resultado aceitável, isso acontece devido a vários fatores que possam influenciar a uma boa assertividade do modelo, uma delas é o tamanho dos dados, conforme Le et al. (2020) o algoritmo *Naive Bayes*, por exemplo, é recomendado o uso quando a quantidade de dados de entrada é alta.

Devido as funcionalidades de cada algoritmo, a utilização de uma combinação de algoritmos seria o mais indicado, por exemplo a floresta aleatória que é uma combinação de árvores de decisão, mostrou ser mais eficiente em termos de acurácia do que o algoritmo de árvore de decisão, com 85% comparado a 80% de acurácia. Portanto o algoritmo que se mostrou mais eficaz em acurácia foi a floresta aleatória, com 85% de acurácia.

7 CONCLUSÕES E PERSPECTIVAS DE TRABALHOS FUTUROS

7.1 Objetivos alcançados

Durante esse projeto foi realizado a categorização de eventos de segurança por meio da implementação de algoritmos de classificação. Esse trabalho tem o potencial de reduzir cerca de 20 horas mensais de cada colaborador da área de segurança do trabalho e o projeto foi concluído com uma acurácia superior a 80% de acertos, por quatro algoritmos, árvore de decisão, floresta aleatória, SVM e regressão logística, a proposta de implementação é substituir parte do que é realizado manualmente para o modelo automático, por meio dos algoritmos de classificação.

Além disso, foi possível aplicar seis algoritmos distintos, permitindo uma comparação detalhada e uma análise minuciosa de cada um. O resultado desse esforço conjunto foi a realização de análise preditiva significativa do processo de classificação de acidentes de trabalho. Essa eficácia aprimorada foi documentada e compartilhada através da publicação do trabalho na revista *Safety Science*, uma renomada publicação na área de segurança, com classificação *Qualis* A1.

7.2 Implicações Práticas

O propósito principal deste projeto é auxiliar os profissionais de segurança do trabalho na classificação de acidentes. Além de seu uso nesse campo específico, o projeto também tem potencial para ser aplicado em diversas outras situações que envolvam a mineração de texto, como, por exemplo, análises jurídicas que requerem categorizações semelhantes.

7.3 Trabalhos Futuros

Como perspectiva para trabalhos futuros, será desenvolvido um sistema baseado em modelo preditivo de acidentes de trabalho, com intuito de identificar tendências de prováveis acidentes de trabalho. Isso adicionaria uma variável adicional no processo de previsão. Por exemplo, ao analisar um histórico específico e identificar que certos tipos de acidentes são mais comuns em uma categoria como a de direção, seria possível implementar campanhas de

segurança viária direcionadas a essa categoria, antecipando potenciais ocorrências com base em seu histórico específico.

REFERÊNCIAS

AZEVEDO, Ana.; SANTOS, Manuel Filipe. **KDD, SEMMA and CRISP-DM: a parallel overview**. IADIS European Conference on Data Mining, p. 182-185. 26. Jul. 2008.

BERTHOLD, Michael. et al. **KNIME – The Konstanz Information Miner**. University of Konstanz, Konstanz, Alemanha, Nycomed Chair for Bioinformatics and Information Mining, doi: 10.1145/1656274.1656280, p. 26-31, 16 Nov. 2009.

BOTELHO, Marcos Cesar. **A LGPD e a proteção ao tratamento de dados pessoais de crianças e adolescentes**. Direitos sociais e políticas públicas (UNIFAFIBE), v. 08, n. 02, p. 197-231, 24 Jun. 2020.

BRASIL, **Lei nº 8.213, de 24 de julho de 1991**, 24 Jul. 1991.

CHARBUTY, B.; ABDULAZEEZ, A. **Classification Based on Decision Tree Algorithm for Machine Learning**. Journal of Applied Science and Technology Trends, v. 2, n. 01, p. 20-28, 24 Mar. 2021.

CHRISTOPHER, Antony. **K-Nearest Neighbor.** Medium Article, 02 Fev. 2021.

DIETZ, Christian; BERTHOLD, Michael. **KNIME for open-source bioimage analysis: a tutorial.** Focus on Bio-Image Informatics, v. 219, doi: 10.1007/978-3-319-28549-8_7, p. 179-197, 01 Mai. 2016.

FERNANDES, Fernando Timóteo; FILHO, Alexandre. D. P. C. **Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho**. Revista Brasileira de Saúde Ocupacional, v. 44, n. 13, p. 12, 01 Jan. 2019

GUPTA, Vishal; LEHAL, Gurpreet. **A Survey of Text Mining Techniques and Applications**. Journal of Emerging Technologies in Web Intelligence, v. 01, n. 01, p. 60-76, 01 Ago. 2009.

HENRY, Felix et al. **Sugarcane Land Classification with Satellite Imagery using Logistic Regression Model**. IOP Conference Series: Materials Science and Engineering, v. 185, doi: 10.1088/1757-899X/185/1/012024, p. 012024, 01 Mar. 2017.

JAIN, Rishabh. **Support Vector Machine (SVM),** Medium Article, 17 Jun. 2020.

JUNIOR, G. B. V. et al. **Determinação das métricas usuais a partir da matriz de confusão de classificadores multiclasses em algoritmos inteligentes nas ciências do movimento humano.** Centro de Pesquisas Avançadas em Qualidade de Vida, v. 14, n. 2, p. 1-15, 01 Jan. 2022.

KORENIUS, Tuomo. et al. ***Stemming*** **and Lemmatization in the Clustering of Finnish Text Documents**. International Conference on Information and Knowledge Management, v. CIKM '04, doi: 10.1145/1031171.1031285, p. 625-633. 13 Nov. 2004.

KUMAR, A. ***Stemming*** **and Lemmatization**. LinkedIn Article, p. 1, 23 Abr. 2022.

LE, Tuan Minh. et al. **A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic.** IEEE Engineering in Medicine and Biology Society Section, Ho Chi Minh City, Vietnam National University, v. 9, doi: 10.1109/ACCESS.2020.3047942, p. 7869 – 7884, 22 Dez. 2020.

LIMA, Sarah Coura de. **Detecção de fraudes em pagamentos com cartão de crédito utilizando técnicas de aprendizado de máquina.** Monografia (Graduação em Engenharia de Controle e Automação) - Instituto Federal do Espírito Santo, Coordenadoria do curso Engenharia de Controle e Automação, 01 Dez. 2023.

MALL, Rishabh. ***Naive Bayes*** **Classifier**. Medium Article, 28 Dez. 2018.

MILO, T.; SOMENCH, A. **Automating Exploratory Data Analysis via Machine Learning: An Overview**. International Conference on Management of Data, v. SIGMOD/PODS '20, doi: 10.1145/3318464.3383126, p. 2617-2622, 14 Jun. 2020.

MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos sobre aprendizado de máquina.** Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 89-114, 2003.

MORO, Sérgio; LAUREANO, Raul M. S.; CORTEZ, Paulo. **Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology**. Proceedings of the European Simulation and Modelling Conference, ISBN: 978-90-77381-66-3, p. 3-7, 01 Out. 2011.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. The MIT Press Cambridge, ISBN 978-0-262-01802-9, p. 1-1098, 24 Ago. 2012.

ORGANIZAÇÃO INTERNACIONAL DO TRABALHO. **Quase 3 milhões de pessoas morrem devido a acidentes e doenças relacionados ao trabalho**, 26 Nov. 2023.

QAISER, Shahzad; ALI, Ramsha. **Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents.** International Journal of Computer Applications, v. 181, n. 1, p. 25-29, 16 Jul. 2018.

RADZI, Siti Fairuz et al. **Comparison of classification algorithms for predicting autistic spectrum disorder using WEKA modeler.** BMC Medical Informatics and Decision Making, v. 22, n. 1, 24 Nov. 2022.

RAI, Khushwant. **The math behind logistic regression**. Medium Article, 14 Jun. 2020.

ROCHA, L. C. **As Tragédias de Mariana e Brumadinho: É Prejuízo? Para Quem?** Caderno de Geografia, v. 31, n. 1, p. 184-195, 08 Fev. 2021.

SANTOS, V. M. B. L. **Deteção automática de lesões no intestino delgado por análise de imagens obtidas por cápsula endoscópica.** Biblioteca da Universidade do Minho, Tese de doutoramento, 07. Mai. 2020.

SAROJ, R. K. et al. **Machine Learning Algorithms for understanding the determinants of under-five Mortality**. BioData Mining, v. 15, n. 20, p. 1-22, 24 Set. 2022.

SCHROER, Christoph; KRUSE, Felix; GÓMEZ, Jorge. **A Systematic Literature Review on Applying CRISP-DM Process Model. International Conference on Project Management**. Procedia Computer Science, v. 181, doi: 10.1016/j.procs.2021.01.199, p. 526–534, 01 Jan. 2021.

SEHRA, Chirag. **Decision Trees Explained Easily**. 2018. Medium Article, 19 Jan. 2018.

SHAFIQUE, Umair; QAISER, Haseeb. **A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)**. International Journal of Innovation and Scientific Research, v. 12, n. 1, p. 217-222. 1 Nov. 2014

SOPER, Daniel S. Greed Is Good: **Rapid Hyperparameter Optimization and Model Selection Using Greedy k-Fold Cross Validation**. Electronics, v. 10, n. 16, p. 1973, 16 Ago. 2021.

SOUSA, Rômulo César Costa de. **Part-of-Speech Tagging para Português**. Dissertação de Mestrado Pontifícia Universidade Católica do Rio de Janeiro, 9 Ago. 2019.

TABASSUM, Ayisha; RAJENDRA, R. Patil. **A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing.** International Research Journal of Engineering and Technology, v. 07, n. 06, p. 1-4, 01 Jun. 2020.

TAMERS, S. L. et al. **Envisioning the future of work to safeguard the safety, health, and well-being of the workforce: A perspective from the CDC's National Institute for Occupational Safety and Health**. American Journal of Industrial Medicine, v. 63, n. 10, p. 877-890, 14 Set. 2020.

TAO, Dandan; YANG, Pengkun; FENG, Hao**. Utilization of text mining as a big data analysis tool for food science and nutrition.** Comprehensive Reviews in Food Science and Food Safety, v. 19, n. 2, p. 875-894, 16 Fev. 2020.

TRIPATHI, Mayank. **How to process textual data using TF-IDF in Python**. FreeCodeCamp Article, 06 Jun. 2018.

TSENG, Yuen-Hsien; LIN, Chi-Jen; LIN, Yu-I. **Text Mining Techniques for Patent Analysis.** Information Processing and Management, v. 43, n. 5, p. 1216–1247, 01 Set. 2007

UYANIK, Hakan. et al. **Use of Differential Entropy for Automated Emotion Recognition in a Virtual Reality Environment with EEG Signals**. Diagnostics, v. 12, n. 10, p. 2508, 16 Out. 2022.

VIJAYARANI, S.; ILAMATHI, J.; NITHYA. **Preprocessing Techniques for Text Mining - An Overview**. International Journal of Computer Science & Communication Networks, vol. 5, n. 1, p. 7-16, 27 Fev. 2015.

WIRTH, R; HIPP, J. **CRISP-DM: Towards a Standard Process Model for Data Mining.** Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, p. 29-40, 13 Abr. 2000.

XU, N. et al. **An Improved Text Mining Approach to Extract Safety Risk Factors from Construction Accident Reports.** Safety Science, v. 138, n. 08, p. 105216, 01, Jun. 2021

YEHOSHUA, Roi. **Random Forests**. Medium Article, 24 Mar. 2023.

ZHANG, Fan. et al. **Construction site accident analysis using text mining and natural language processing techniques.** Automation in Construction, v. 99, n. 2019, p. 238-248, 24 Dez. 2018.

APÊNDICES

Apêndice 1: Formalização da concessão dos dados.

RES: SOLICITAÇÃO DE DADOS- OFICIO N. 08/2023-PECS/CCT/UEMA.

Sheila Fialho <sheila.fialho@vale.com>
Sex, 10/11/2023 15:27
Para:PECS <secretaria.pecs@gmail.com>
Cc:daniel301198 <daniel301198@hotmail.com>

1 anexos (323 KB)
08-_2023_%5B1%5D_assinado_%281%29_assinado (1).pdf;

Prezados, boa tarde! Espero que todos estejam bem.
Segue em anexo o documento assinado. Os dados solicitados já estão com o aluno Daniel. Disponibilizamos a base conforme orientação do nosso jurídico, pontos esses que reforço abaixo.

1) Retirar da coluna que contém a descrição dos acidentes as abreviações das iniciais do nome dos motoristas, deixando simplesmente condutor ou motorista, bem como retirar as matrículas e algumas outras referências que possam eventualmente levar a associação de que são as pessoas envolvidas em razão de datas, horários, nome das empresas.
2) Revisar todas as descrições para que de fato não conste o nome de nenhuma pessoa.
3) Não utilizar no curso dos estudos a coluna ID SAP, eis que são referências internas da empresa. Essa referência não permite identificar as pessoas que estão envolvidas na descrição do evento, mas podem levar a identificação da pessoa que notificou/registrou o incidente. Ademais, acima de tudo, estes registros na planilha são irrelevantes para o desenvolvimento do trabalho em questão.

Qualquer dúvida ou ajuda, estou à disposição.
Att,
Sheila Fialho

Apêndice 2: Aprovação da concessão dos dados via ofício da UEMA para empresa Vale

UNIVERSIDADE ESTADUAL DO MARANHÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS - PECS

PECS

Ofício nº08/2023-PECS/CCT /UEMA

São Luís - MA, 25 de outubro de 2023.

Para: Sheila Fialho dos Santos Vinciprova - Coordenação de Perfomance, Compliance e Gestão de Riscos- VALE

Do: Prof.º Dr. Mauro Sérgio Silva Pinto

Coordenador do Programa de Pós-Graduação em Engenharia de Computação e Sistemas- PECS/UEMA.

Prezada Coordenadora, Sheila Fialho dos Santos Vinciprova,

Cumprimentando-a cordialmente, dirijo-me a Vossa Senhoria, solicitando os dados sobre a descrição de acidentes e a categorização dos acidentes de trabalho da Diretoria de Serviços Operacionais e Segurança Empresarial, que já foram tratados pelo setor Jurídico da empresa Vale S.A, com intuito de proteger contra a LGPD. Os dados tem como objetivo, contribuir no desenvolvimento da pesquisa científica do mestrando, Daniel Herrera de Oliveira Lemos - Mat. 20221000547, o mesmo se encontra matriculado no Programade Pós Graduação em Engenharia de Computação e Sistemas, orientado pelo professor Dr. Cícero Costa Quarto.

Prof. Dr. Mauro Sérgio Silva Pinto
Coordenador do Programa de Pós-Graduação em Engenharia de Computação e Sistemas- PECS.

gov.br Documento assinado digitalmente
MAURO SERGIO SILVA PINTO
Data: 31/10/2023 13:28:51-0300
Verifique em https://validar.iti.gov.br

gov.br Documento assinado digitalmente
SHEILA FIALHO DOS SANTOS VINCIPROVA
Data: 09/11/2023 12:14:27-0300
Verifique em https://validar.iti.gov.br

Curso de Mestrado Profissional em Engenharia de Computação e Sistemas

Aprovado pelas Resoluções 911/2010-CEPE/UEMA e 801/2010-CONSUN/UEMA, de 22/04/2010.

Apêndice 3: Amostragem dos dados coletados

EMPREGADOS DE EMPRESA CONTRATADA ACESSANDO SALÃO DE MINA DO LABORATÓRIO FISICO SEM O ÓCULOS DE PROTEÇÃO	Comportamento
COLABORADOR REALIZANDO ATIVIDADE DE FORMA INADEQUADA SEM O USO DO EPI ADEQUADO PARA ATIVIDADE DE LIMPEZA DAS BANDEJAS	Comportamento
COLABORADOR DA TERCEIRIZADA ESTAVA ATENDENDO TELEFONE E CAMINHANDO	Comportamento
SENSOR DE PRESENÇA DO BANHEIRO MASCULINO DO PRÉDIO DO APOIO DO PÁTIO DE MARABÁ ESTÁ QUEIMADO IMPOSSIBILITANDO A LUZ DE ACENDER PODENDO CAUSAR UM TROPEÇO OU QUEDA	Estrutural
DIALOGO COMPORTAMENTAL REALIZADO COM A EQUIPE DE PORTARIA DE TIMBOPEBA SOBRE OS CUIDADOS NO ATENDIMENTO NAS PISTAS RISCO DE DE ATROPELAMENTO	Comportamento
PORTA DE PRINCIPAL DA FAZENDINHA (DUPLA DE VIDRO) QUEBROU E OBSTRUIU A ENTRADA DO PRÉDIO	Estrutural
COLABORADORES DEIXANDO A GARRAFA TÉRMICA DE ÁGUA NO CORREDOR DO MICRO-ÔNIBUS	Comportamento
CINTO DE SEGURANÇA DA POLTRONA DO ÔNIBUS U NÃO ESTÁ DISPONÍVEL ENCONTRA SE TRAVADO	Estrutura do veículo
FOI IDENTIFICADO AO ACESSAR SALA ELÉTRICA LOCALIZADA NA OFICINA DA JSL VÁRIAS CONDIÇÕES DE RISCO : PAINÉIS DE ILUMINAÇÃO E AR CONDICIONADO SEM TAMPAS PORTAS DANIFICADAS INFILTRAÇÃO DE ÁGUA PORTA DA SALA ELÉTRICA SEM TRANCA	Estrutural
AO EMBARCAR NO ÔNIBUS ANTES DE SENTAR AINDA NO CORREDOR (MEIO DO ÔNIBUS) O MOTORISTA DEU PARTIDA CAUSANDO DESEQUILÍBRIO	Estrutura do veículo
FOI OBSERVADO EMPREGADO CAMINHANDO PELA BRITA COM A EXISTÊNCIA DE CAMINHO SEGURO AO LADO	Comportamento
COLABORADORA FAZENDO USO DE ADORNO (ALIANÇA)	Comportamento
ILUMINAÇÃO INSUFICIENTE NOS PATIOS DE FINOS DE TUBARÃO	Estrutural
ENQUANTO O COOPERADO S AGUARDAVA PARA ADENTRAR A ROTATÓRIA DO KM COM O VEÍCULO DE QWWI MODELO SANDERO O MESMO FOI SURPREENDIDO COM COLISÃO NA TRASEIRA NA OCASIÃO HAVIA USUÁRIOS DA ELETRÔNICA NÃO HOUVE DANOS PESSOAIS SÓ DANOS MATERIAIS	Direção
DURANTE ATIVIDADE DE LIMPEZA DO PÁTIO COM EQUIPAMENTO SOPRADOR (SOPRADOR DE FOLHAS) O FUNCIONÁRIO NÃO ESTAVA UTILIZANDO O EPI ADEQUADO	Comportamento
USANDO O CELULAR ATRAVESSANDO A FAIXA DE PEDESTRE	Comportamento
TORNEIRA DO LAVABO DO RESTAURANTE DA USINA QUEBROU OCORRENDO O DERRAMAMENTO DE MUITA ÁGUA NO LOCAL FOI ISOLADO O LOCAL E ABERTO CHAMADO PARA TROCA DESSA TORNEIRA	Estrutural
QUEDA DE CUMEEIRA NO TELHADO DA PORTARIA PRINCIPAL CPBS ÁREA DE ACESSO DE EMPREGADOS PRÓPRIOS E TERCEIROS	Estrutural
ÔNIBUS DA EMPRESA VIX REALIZOU MANOBRA INDEVIDA ENFRENTE AO PRÉDIO DOS BOMBEIROS DA BRITAGEM SECUNDÁRIA PARA RETOMA A OUTRA VIA EM DIREÇÃO A RODOVIÁRIA OU TLEVE	Direção
ACÚMULO DE AGUA	Estrutural
NO CORREDOR DOS FUNDOS DO CCO HÁ UMA DO TETO TRINCADA OBS : NESTA ESTÁ INSTALADO UM DISPOSITIVO DE SEGURANÇA	Estrutural
INSPEÇÃO DE S NO ENTORNO DA OFICINA	Comportamento
FALTA DE SINALIZAÇÃO NO PISO DO ESTACIONAMENTO	Estrutural

Apêndice 4: Submissão de artigo científico na revista *Safety Science*

SAFETY-D-24-00407 - Confirming your submission to Safety Science

SS Safety Science <em@editorialmanager.com>
17/02/2024 20:01

To: Daniel Lemos

This is an automated message.

OPTIMIZATION OF THE CATEGORIZATION OF OCCUPATIONAL ACCIDENTS IN THE COMPANY VALE S.A. WITH THE COMPARISON OF CLASSIFICATION ALGORITHMS

Dear Daniel Herrera Lemos Lemos,

We have received the above referenced manuscript you submitted to Safety Science. It has been assigned the following manuscript number: **SAFETY-D-24-00407**.

To track the status of your manuscript, please log in as an author at https://www.editorialmanager.com/safety/, and navigate to the "Submissions Being Processed" folder.

Thank you for submitting your work to this journal.

Kind regards,
Safety Science

FAQ: How can I reset a forgotten password?
https://service.elsevier.com/app/answers/detail/a_id/28452/supporthub/publishing/
For further assistance, please visit our customer service site: https://service.elsevier.com/app/home/supporthub/publishing/
Here you can search for solutions on a range of topics, find answers to frequently asked questions, and learn more about Editorial Manager via interactive tutorials. You can also talk 24/7 to our customer support team by phone and 24/7 by live chat and email

This journal uses the Elsevier Article Transfer Service. This means that if an editor feels your manuscript is more suitable for an alternative journal, then you might be asked to consider transferring the manuscript to such a journal. The recommendation might be provided by a Journal Editor, a dedicated Scientific Managing Editor, a tool assisted recommendation, or a combination. For more details see the journal guide for authors.

At Elsevier, we want to help all our authors to stay safe when publishing. Please be aware of fraudulent messages requesting money in return for the publication of your paper. If you are publishing open access with Elsevier, bear in mind that we will never request payment before the paper has been accepted. We have prepared some guidelines (https://www.elsevier.com/connect/authors-update/seven-top-tips-on-stopping-apc-scams) that you may find helpful, including a short video on Identifying fake acceptance letters (https://www.youtube.com/watch?v=o5l8thD9XtE). Please remember that you can contact Elsevier s Researcher Support team (https://service.elsevier.com/app/home/supporthub/publishing/) at any time if you have questions about your manuscript, and you can log into Editorial Manager to check the status of your manuscript (https://service.elsevier.com/app/answers/detail/a_id/29155/c/10530/supporthub/publishing/kw/status/).
#AU_SAFETY#

To ensure this email reaches the intended recipient, please do not delete the above code