

Protótipo de um sistema inteligente para a prevenção de acidentes graves em empresas

Davi A. Leão* Carlos Wilker N. Santana** Iago S. Rodrigues**
Carlos Henrique L. Cavalcante**** Asley A. Batista****

* Instituto Federal de Educação, Ciência e Tecnologia, CE,
(e-mail: davi.alves.leao61@aluno.ifce.edu.br).

** Instituto Federal de Educação, Ciência e Tecnologia, CE,
(e-mail: carlos.wilker.nogueira07@aluno.ifce.edu.br,
iago.sousa.rodrigues05@aluno.ifce.edu.br).

*** Instituto Federal de Educação, Ciência e Tecnologia, CE,
(e-mail: henriqueleitao@ifce.edu.br).

**** Farias Brito, CE, (e-mail: asley.batista@fariasbrito.com.br).

Abstract: In the work routine, accidents cause several losses, both for the contractor and the people hired. At the same time, profit and materials are lost in the companies, the work capacity and, in the worst cases, the worker's life is affected. From 2002 to 2020, a rate of 6 deaths was recorded for every 100 thousand formal jobs in Brazil. The present work resulted in the evaluation of a computational model using machine learning to prevent serious accidents. Through the use of sliding window algorithms for data standardization, satisfactory results were achieved, in which it was possible to predict the occurrence of severe accidents for periods of up to seven days. Additionally, a prototype of an intelligent web system was proposed, using *cloud* microservices, using the prediction model created.

Resumo: Na rotina de trabalho, os acidentes acarretam diversos prejuízos, tanto para o contratante, quanto para o contratado, enquanto lucro e os materiais são perdidos nas empresas, a capacidade de trabalho e, nos piores casos, a vida do trabalhador é afetada. De 2002 a 2020, foi registrada uma taxa de 6 óbitos para cada 100 mil empregos formais no Brasil. O presente trabalho resultou na avaliação de um modelo computacional usando aprendizado de máquina para prevenção de acidentes graves. Através do uso de algoritmos de janela deslizante para padronização dos dados, foram alcançados resultados satisfatórios, em que se conseguiu prever a ocorrência de acidentes graves por períodos de até sete dias. Adicionalmente, foi proposto um protótipo de um sistema web inteligente, usando micro serviços de *cloud*, utilizando o modelo de predição criado.

Keywords: Occupational injuries; Computational Intelligence; Generic Model; Machine Learning

Palavras-chaves: Acidentes de trabalho; Inteligência Computacional; Modelo genérico; Aprendizado de máquina

1. INTRODUÇÃO

Acidentes no ambiente de trabalho podem ser leves ou graves, uma lesão permanente ou até mesmo ter sua vida ceifada e tendem a gerar prejuízos financeiros para empresas ou danos para a saúde do trabalhador. Os acidentes mais graves são os que mais impactam profundamente a vida das organizações e seus trabalhadores. Segundo estatísticas, o número de Acidentes de Trabalho Fatais (ATF) são expressivos e preocupantes. De acordo com Takala et al. (2014), morrem aproximadamente dois milhões de trabalhadores por acidentes de trabalho, anualmente, no mundo. No Brasil, de 2002 a 2020, foi registrado uma taxa de 6 óbitos a cada 100 mil empregos formais, de acordo com o relatório do Ministério Público do Trabalho e da Organização Internacional do Trabalho (Basilio, 2021).

Para Cordeiro et al. (2017), um acidente é um dos maiores agravadores da saúde dos trabalhadores no Brasil. Em 2013, o Ministério da Previdência Social (MPS) registrou a ocorrência de 702.685 acidentes de trabalho e trajeto, dentre eles 2.797 vieram a falecer (Cordeiro et al., 2017). Já em 2018, foram registrados 576.951 acidentes (Ramos, 2020), havendo uma diminuição de 17% em 4 anos. Todavia, de acordo com Ramos (2020), estes números do ano de 2018 equivalem apenas aos trabalhadores com carteira assinada, o que leva a refletir que a taxa de diminuição foi efetivamente menor.

Estes eventos acabam gerando um alto custo para o governo e, em alguns casos, para a empresa na qual o funcionário exerce sua profissão. Apenas em 2014, foram registradas 176.007 internações hospitalares pelo Sistema Único de Saúde (SUS), implicando em um gasto total de R\$ 244 milhões (Brasil, 2018). Para as empresas, pode

ou não haver gastos, dependendo apenas das medidas preventivas utilizadas, porém ao provar que a empresa possui culpa deste acontecimento, esta, de acordo com o inciso XXVIII do Artigo 7 da Constituição Federal de 1988: “seguro contra acidentes de trabalho, a cargo do empregador, sem excluir a indenização a que este está obrigado, quando incorrer em dolo ou culpa” (Brasil, 1988), deve realizar uma indenização ao seu funcionário (Cardoso and Batista, 2019).

Dessa forma, esta pesquisa tem o objetivo principal de encontrar padrões que possibilitem inferir, de forma antecipada e com um grau de confiança satisfatória, a ocorrência de acidentes de trabalhos graves, através do uso de algoritmos de inteligência computacional. Adicionalmente, pretende-se fornecer informações que possibilitem a antecipação de políticas de prevenção de acidentes e criar um protótipo de um sistema web inteligente para a aplicação do modelo de dados proposto, no dia a dia das empresas.

O presente trabalho encontra-se dividido em 5 Seções, sendo a primeira a introdução. A segunda Seção atenta-se a mostrar a fundamentação teórica sobre os acidentes de trabalho, inteligência artificial e aprendizado de máquina. A terceira expõe os aspectos metodológicos adotados na pesquisa, o modelo proposto e a arquitetura do sistema. A quarta Seção discorre sobre os resultados obtidos e explica a interface criada. A última contém considerações a respeito do modelo criado e sugestões de trabalhos futuros.

2. TRABALHOS RELACIONADOS

Alguns trabalhos semelhantes que, da mesma forma, propõem uma solução para a contenção dos acidentes, são estes: Scott et al. (2021) que propôs um algoritmo para identificação de possíveis acidentes em agricultura usando dados hospitalares, Alawad et al. (2020) modela algoritmos de aprendizado de máquina para a análise de acidentes em estações ferroviárias, Wang et al. (2020) desenvolveu sensores para um capacete e sapatos inteligentes para a prevenção de acidentes em escadas, Mosquera et al. (2021) apresenta uma proposta de sistema para a identificação e prevenção de acidentes ocupacionais em armazéns aplicando alguns algoritmos de aprendizado de máquina como as árvores de decisão e redes Bayesianas, estas também utilizadas por Abad et al. (2018) para prevenir doenças relacionadas a exposição à pesquisa nas indústrias de energia e construção, e por Gerassis et al. (2017) para prever acidentes ocupacionais durante a construção de aterros.

3. INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA

Inteligência Artificial (IA) é uma área da ciência da computação, essa disponibiliza diversos algoritmos para a construção de sistemas inteligentes, no qual conseguem agir da melhor forma possível, dependendo das entradas que lhe forem fornecidas. Atualmente, diversos tipos de sistemas inteligentes estão sendo criados como sistemas na área médica (Haq et al., 2018; Jin et al., 2020), ambos aplicam métodos inteligentes para o diagnóstico prematuro de doenças, Haq et al. (2018) propôs um sistema inteligente para o diagnóstico de doença do coração e Jin et al. (2020) para o COVID-19.

Aprendizagem de máquina ou Machine Learning (ML), para Alzubi et al. (2018) pode ser definida como uma categoria da IA que possibilita as máquinas pensarem e aprenderem por si mesmas. Esse ramo da IA, o aprendizado de máquina, possui diversos tipos de algoritmos dentre os quais existem dois ramos principais: os algoritmos supervisionados e não supervisionados. Na pesquisa, preocupou-se apenas com modelos supervisionados, mais especificamente aqueles que resolvem tarefas de classificação, visto que foram os únicos aplicados no sistema proposto.

A tarefa de classificação, é resolvida por modelos supervisionados no qual é um processo indutivo onde o modelo aprenderá a partir de exemplos com pares de entrada e saída $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)$, este par (\vec{x}_i, y_i) é representado por dois componentes, um vetor de características \vec{x}_i e um rótulo discreto y_i que indica a qual grupo ou classe aquele conjunto de características pertence (Norvig and Russell, 2013; Abu-Mostafa et al., 2012).

4. METODOLOGIA

4.1 Coleta de dados

Os dados aplicados nesta pesquisa foram fornecidos por uma empresa nacional prestadora de serviço público. A base de dados foi composta por acidentes ocorridos em um período compreendido entre os anos de 2016 e 2019, possuindo um total de 4487 acidentes. Por questões de confidencialidade, o ramo e nome da empresa foram ocultados.

4.2 Protótipo de um Sistema de Predição de Acidentes - SISPREV

Foi desenvolvido um protótipo do sistema, denominado SISPREV, que permitirá o envio de dados por empresas com incidência sobre acidentes de trabalhos ocorridos entre seus funcionários. Primeiramente, uma base de dados é obtida a partir de seus registros e será enviada para o sistema de acordo com o fluxo descrito na Figura 1. Em seguida, o sistema irá criar o modelo de dados conforme descrito na seção 4.2.1, que por sua vez é utilizado como entrada para gerar o modelo de predição criado a partir do uso dos classificadores de aprendizagem de máquina. Em um próximo passo, são retornadas as métricas (acurácia, precisão, etc) do modelo. Após o modelo de predição criado, o usuário pode enviar novos dados para verificar a ocorrência futura de um acidente grave ou não. Além do envio da base completa, também é possível realizar o *upload* de incrementos, em que o sistema irá adicionar novos dados aos existentes e realizar as etapas de treinamento novamente. O fluxo da figura 1 reflete o comportamento do protótipo do software em um possível ambiente real, os resultados evidenciados na seção 5.3 são do modelo de predição proposto com base no conjunto de dados que equivale a primeira carga no sistema (criação do modelo).

Modelo de dados proposto O protótipo desenvolvido neste trabalho, aplica um processamento de dados genérico criado para a previsão dos acidentes graves. Este baseia-se no conceito de janela deslizante, dispondo de dois componentes principais: a janela e o passo.

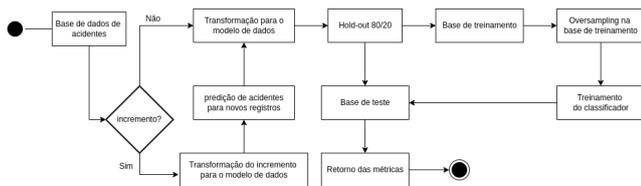


Figura 1. Fluxo de atividade do protótipo SISPREV.

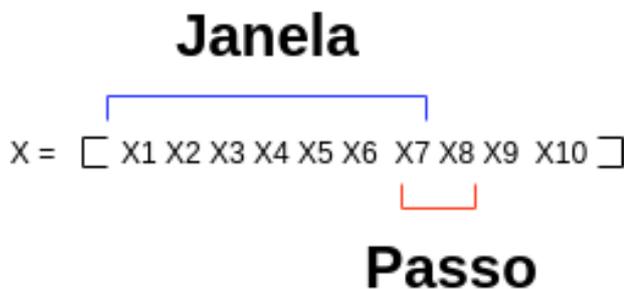


Figura 2. Exemplo da janela e passo

A janela representa um intervalo de dias que será considerado para agrupar e contabilizar os acidentes de um período. Considerando uma data inicial X_1 , e se for definida uma janela de 7 dias, serão agrupados os dados relativos aos dias X_1 a X_7 . Além da janela também existe o conceito de passo, esse representa a quantidade de dias que a janela será deslocada para frente. Complementando o exemplo anterior, considerando X_1 e uma janela de 7 dias, a primeira amostra será formada do agrupamento dos acidentes de X_1 a X_7 . Ao ser definido um passo de 1 dia, a janela será deslocada e a segunda amostra será composta pelo agrupamento de acidentes de X_8 até X_{14} . Na Figura 2 são representados esses dois conceitos.

Após agrupar os dados da base por período de tempo (Ex: 3, 5 ou 7 dias), foram retiradas do conjunto original apenas duas características para compor o modelo de dados, são elas: o tipo de acidente e a sua classificação *holding* que o enquadrava como grave ou não.

Durante o processamento da janela deslizante, foram agrupados os acidentes por tipo. Levando em conta uma janela de 7 dias, seria contabilizado e agrupados quantos acidentes de cada tipo ocorreram de cada nesse período. Assim, gerou-se uma base pré-processada com 26 colunas, sendo uma para cada tipo de acidente. Similarmente, as classificações *holding* foram contabilizadas para servir como rótulo de saída/predição para os algoritmos de ML. Dessa forma, gerou-se uma matriz de características com 27 colunas como entrada para os modelos de classificação.

Tabela 1. Cenário 01: Janela 3 dias - Passo 1 - Exemplo com 2 características e o rótulo.

trânsito/carro	movimentação de carga	Acidentes	Computáveis
4.0	1.0	0.0	
2.0	0.0	0.0	
2.0	0.0	0.0	
4.0	0.0	1.0	
4.0	1.0	0.0	
1.0	0.0	0.0	
1.0	0.0	0.0	

Na tabela 1, é mostrado um exemplo do modelo de dados final, em que foi utilizado uma janela de três dias com

passo de um dia. Para simplificar a tabela, foram representadas apenas duas das 26 possíveis colunas dos tipos de acidente e o rótulo, indicando se no próximo período contabilizado pela janela poderá ocorrer um acidente grave ou não.

Após a construção dos modelos em diferentes janelas e passos, notou-se que na maioria dos casos ocorria um desbalanceamento dos dados, Tabela 2, prejudicando a generalização do modelo, uma vez que existiam mais exemplos da classe grave, resultando erros no classificador, uma vez que tende a acertar mais para um rótulo do que para outro, generalizando uma classe e errando todos os exemplos da outra.

A fim de solucionar este problema, foi adicionado uma nova etapa de *oversampling*, em que o objetivo principal foi aumentar o número de dados da classe com menos exemplos. Apenas foram gerados novos dados no treinamento do classificador. Foi utilizado o *Synthetic Minority Oversampling Technique* (SMOTE).

Cenários aplicados Com a geração do modelo de dados a ser utilizado, definiu-se algumas janelas e passos padrões para testes de iniciais de predição. Na Tabela 2 são especificados os cenários executados sem *oversampling*.

Tabela 2. Cenários utilizados.

Cenário	Janela (w)/Passo (s)	N	$y_i = 1(\%)$	$y_i = 0(\%)$
01	3/1	400	39,50	60,50
02	3/2	318	46,22	53,77
03	5/1	235	63,82	36,17
04	5/2	219	60,73	39,26
05	7/1	178	72,47	27,52
06	7/2	168	70,23	29,76

Configuração dos testes A configuração das simulações realizadas em cada cenário foi, hold-out de 80/20, que consiste em selecionar de maneira aleatória 80% do dados para o conjunto de treinamento e 20% para o de teste. Além disso, foi aplicada uma etapa de validação, na qual foi feita com uma parte dos dados de treino. O objetivo da etapa de validação é obter os melhores hiperparâmetros, para isso foi utilizado o algoritmo *K-Fold Cross Validation*, em que o $K=10$. Por fim, todos os classificadores foram rodados 10 vezes e retirou-se os valores médios e o desvio padrão de cada métrica.

4.3 Avaliação

Para a avaliação deste tipo de problema existem diversas métricas, as aplicadas a esta pesquisa foram; acurácia (ACC) que mede a taxa de acerto, área sob a curva ROC (AUC) que determina o desempenho do modelo em relação ao problema, a precisão (P) que verifica a proporção de exemplos positivos que foram classificados corretamente dentro do número total de instâncias preditas positivas, a sensibilidade ou recall (R) que indica o número de classificações corretas da classe positiva na classe positiva, F1-score (F1) que faz uma média harmônica da precisão e da sensibilidade e o coeficiente de Matthews (MCC) que mede a qualidade de classificações binárias, podendo ser aplicada mesmo se as classes possuírem tamanhos diferentes (Tharwat, 2020).

5. RESULTADOS

5.1 Interface do usuário

O protótipo SISPREV possui cinco funcionalidades, sendo elas: Login, Cadastro do Usuário, Cadastro de *Dataset*, Incremento de *Dataset* e predição. A tela de Login, Figura 4, contém dois campos de texto, dois botões e um link para a tela de Cadastro, essa oferecerá como opções de autenticação um login padrão de email e senha com autenticação JWT, e, um login *oAuth* pelo provider Google.

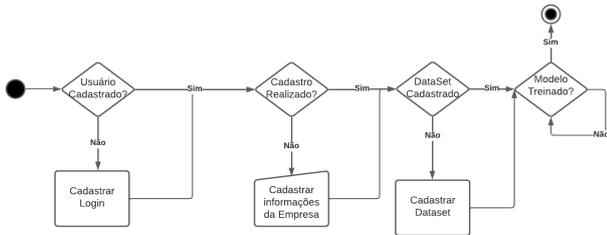


Figura 3. Fluxo de atividades do cadastro.

Figura 4. Tela de login.

A tela de Cadastro do Usuário possui um formulário, feito por campos de texto e um botão de confirmação. O fluxo representado na Figura 3, descreve como é feito o registro completo. O usuário terá a opção de cadastrar-se pelo *oAuth* do Google ou direto pelo sistema, caso escolha a primeira opção, terá que compor o seu cadastro com algumas informações adicionais, já se escolher o modo direto, isso não será necessário.

Logo após de completar o cadastro, o usuário poderá enviar o seu *DataSet* com dados de acidentes. Após o *upload* da base de dados, o sistema irá realizar o treinamento do modelo e o informará quando o processo for finalizado. A Figura 5, representa as informações sobre o status do *dataset* e seu treinamento, são elas: nome de arquivo, data de envio, resultado e observação das bases que estão atualmente treinando.

Com resposta recebida do *backend cloud* que o treinamento do modelo foi finalizado, o usuário tem acesso à tela de predição, onde é possível enviar dados para obter a resposta de possibilidade de ocorrência ou não de acidente grave no período futuro, conforme ilustrado na Figura 6.

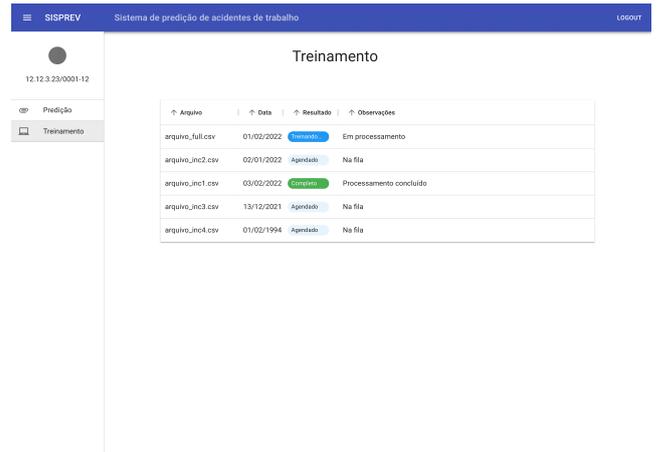


Figura 5. Tela de treinamento.

Adicionalmente, o usuário pode indicar se os dados da predição são um incremento para serem adicionados a base já existente. As informações exibidas pela tabela de predições são: ID do acidente, data de envio do arquivo, janela, período dos dados e o resultado.

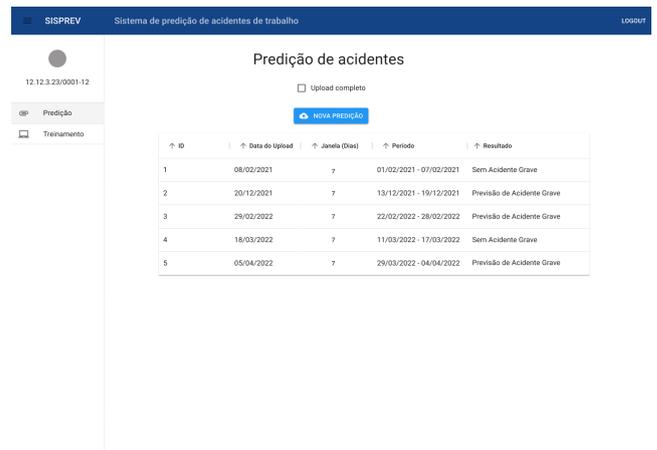


Figura 6. Tela de predição.

Além da opção do registro completo, também é possível utilizar o modo incremental. Nesse a empresa conseguirá enviar pequenas porções de arquivos e obter informações que indiquem se pode ou não ocorrer um acidente grave. A Figura 7 descreve como o sistema se comporta no modo incremental, inicialmente é verificado se o usuário deseja informar o incremento, se sim os dados serão salvos, enviados para predição e por fim exibidos na tela de predição. O registro de incrementos é possível a partir da tela de predição, esta contém um botão superior para o envio de novos arquivos e uma checkbox que determina o tipo de arquivo enviado.

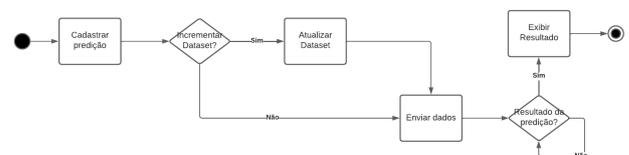


Figura 7. Fluxo de atividades do modo incremental.

5.2 Arquitetura

A aplicação é projetada em torno de serviços disponibilizados pela **Amazon Web Services**, conhecida mundialmente como **AWS**, um dos serviços cloud mais utilizados do mundo pelo preço competitivo e qualidade.

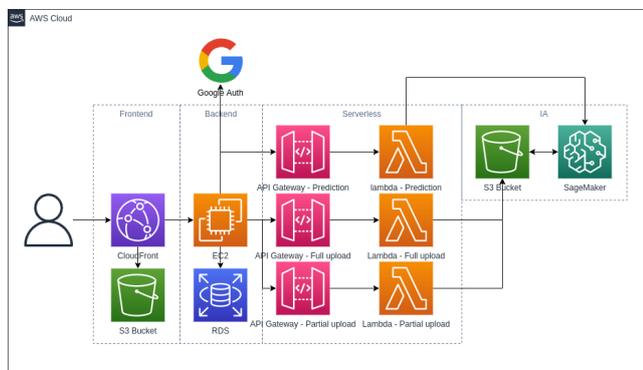


Figura 8. Arquitetura do sistema.

Dá-se como fluxo do sistema: o usuário interage com o **AWS CloudFront**, um serviço de **CDN – Content Delivery Network** – destinado a fazer caching em pontos estratégicos chamados de *edge locations*, que são localizações mais próximas do usuário, diminuindo a latência tornando assim o acesso ao aplicativo mais rápido. O **CloudFront** recebe o conteúdo de origem de um **bucket AWS S3 – AWS Simple Storage Service** – o serviço da AWS para armazenamento de conteúdo estático (a localização dentro do S3 é chamado de *bucket*), que em primeira execução encaminha o conteúdo do S3 para o usuário ao mesmo tempo que faz o *caching* para próximas consultas, mantendo o conteúdo na *edge location* por um tempo configurável, chamado de **TTL (time to live)**. **CloudFront** juntamente com o primeiro **bucket S3** formam o que chamamos de *frontend*: conteúdo na qual o usuário final tem contato direto. Em seguida temos o *backend*, que é de onde a aplicação executada no navegador faz requisições para possibilidade de um conteúdo dinâmico dentro da aplicação, possibilitando autenticação, persistência e consulta de dados. Dada a autenticação de usuário, o mesmo pode enviar arquivos para serem treinados, e esses arquivos são enviados através do serviço **AWS EC2 – AWS Elastic Compute Cloud** - na qual ficará a lógica de comunicação entre o usuário e o sistema, sendo o responsável pela persistência de dados através do serviço **AWS RDS – AWS Relational Database Service** - para banco de dados para utilização de autenticação e envio de arquivos que serão salvos em um outro *bucket* no AWS S3 através do serviço **AWS Lambda** (um tipo de serviço *serverless* onde não gerenciamos o servidor, somente a função), e estes para serem chamados através de um link na internet utilizam um outro serviço chamado de **AWS API Gateway**, um ponto de entrada para **APIs – Application Programming Interfaces** - que fazem a ponte da internet com as funções dentro do serviço **AWS Lambda**. O sistema dispõe de duas funções que fazem o serviço de enviar os arquivos para um *bucket* do S3, sendo uma das funções destinadas ao envio de um arquivo completo para predição, enquanto a outra função é responsável pelo envio de um arquivo incremental, na qual será feito a predição e inclusão dos novos dados no treinamento. Por último, utilizamos o

serviço **AWS SageMaker** responsável pelo treinamento de um algoritmo utilizando os arquivos que foram enviados pelas funções para um *bucket* S3. Após o processamento dos dados, o resultado é salvo no mesmo *bucket* na qual os arquivos de treinamento estão guardados, possibilitando o treinamento posteriormente

5.3 Desempenho do Modelo de Predição

Nesta seção, serão mostrados os resultados obtidos em cada cenário com aplicação do algoritmo de *oversampling* no treino. Devido ao escopo do trabalho e limitações de tempo, a proposta não foi validada em ambiente real e foi testada com os seguintes algoritmos: *Classification and regression trees (CART)*, *Random Forest (RF)*, *Extra Trees (ET)* e *Gradient Boosting Classifier (GBC)*. Os classificadores aplicados foram escolhidos mediante a um estudo científico, na qual foram testados vários modelos de aprendizado de máquina, esses por sua vez obtiveram o melhor desempenho. Nas tabelas 3 a 8 estão as respectivas métricas obtidas com seu desvio padrão a direita.

Tabela 3. Cenário 01: Janela 3 dias - Passo 1.

	CART	RF	ET	GBC
AUC	58.60 ± 0.049	67.27 ± 0.018	65.15 ± 0.024	61.46 ± 0.026
ACC	60.0 ± 0.061	68.6 ± 0.017	67.59 ± 0.023	65.0 ± 0.030
F1	47.23 ± 0.068	61.23 ± 0.032	56.63 ± 0.041	50.0 ± 0.030
MCC	15.79 ± 0.10	26.22 ± 0.035	19.0 ± 0.049	23.36 ± 0.056
P	49.91 ± 0.069	63.39 ± 0.028	60.74 ± 0.035	58.33 ± 0.042
R	46.54 ± 0.12	59.38 ± 0.055	53.15 ± 0.061	43.75 ± 0.041

Tabela 4. Cenário 02: Janela 3 dias - Passo 2.

	CART	RF	ET	GBC
AUC	57.33 ± 0.030	63.23 ± 0.060	65.10 ± 0.037	63.40 ± 0.050
ACC	57.94 ± 0.022	64.06 ± 0.056	65.62 ± 0.035	63.81 ± 0.053
F1	50.52 ± 0.098	56.60 ± 0.083	60.71 ± 0.048	59.65 ± 0.042
MCC	15.39 ± 0.057	19.81 ± 0.11	18.42 ± 0.074	15.06 ± 0.10
P	54.73 ± 0.027	65.22 ± 0.054	65.38 ± 0.032	62.96 ± 0.064
R	50.0 ± 0.17	50.0 ± 0.11	56.67 ± 0.07	56.67 ± 0.04

No primeiro cenário avaliado, conforme apresentado na Tabela 3, as duas maiores taxas de acerto foram 68.6% e 67.59% com os classificadores RF e ET, já a área sob a curva ROC foi nos dois melhores casos 67.27% e 65.15% para os mesmos classificadores.

O próximo cenário a janela ficou fixa e o passo foi aumentado, com esse notou-se um decaimento nas taxas, a melhor acurácia foi 65.62%, diminuição de 2.98% quando comparado com as taxas do primeiro cenário. Além disso, assim como a acurácia, a área sob a curva ROC diminuiu 2%. Para finalizar a avaliação da janela de 3 dias foi utilizado uma métrica que indica a qualidade das predições, a MCC, na qual as maiores foram 26% e 19%, respectivamente, para os cenários 1 e 2. A Tabela 4 apresenta os resultados obtidos.

Tabela 5. Cenário 03: Janela 5 dias - Passo 1.

	CART	RF	ET	GBC
AUC	62.33 ± 0.036	57.88 ± 0.052	58.39 ± 0.030	53.90 ± 0.067
ACC	62.97 ± 0.041	65.10 ± 0.043	66.30 ± 0.031	58.72 ± 0.056
F1	68.26 ± 0.073	75.34 ± 0.035	74.98 ± 0.033	68.77 ± 0.041
MCC	24.39 ± 0.073	21.01 ± 0.090	19.88 ± 0.067	17.28 ± 0.11
P	75.32 ± 0.058	68.66 ± 0.031	69.01 ± 0.023	66.73 ± 0.048
R	64.66 ± 0.12	84.0 ± 0.074	82.66 ± 0.082	71.33 ± 0.061

Tabela 6. Cenário 04: Janela 5 dias - Passo 2.

	CART	RF	ET	GBC
AUC	58.63 ± 0.002	62.42 ± 0.065	65.21 ± 0.080	67.39 ± 0.052
ACC	59.32 ± 0.039	65.91 ± 0.064	70.45 ± 0.064	70.45 ± 0.064
F1	62.28 ± 0.056	73.68 ± 0.055	77.19 ± 0.037	75.47 ± 0.028
MCC	—	11.26 ± 0.12	18.06 ± 0.18	24.94 ± 0.11
P	70.38 ± 0.036	70.0 ± 0.045	73.33 ± 0.051	76.92 ± 0.045
R	56.29 ± 0.079	77.78 ± 0.072	81.48 ± 0.036	74.07 ± 0.050

Em outro cenário executado, foi aplicada uma janela de cinco dias. Na Tabela 5, estão os resultados com essa configuração. A melhor taxa de acerto foi 66.30%. Ao compará-la com a obtida na primeira abordagem, notou-se uma diminuição de 2.3%. Em contrapartida, ocorreu um aumento nas medidas F1, P e R, indicando uma melhora nas classificações dos acidentes graves. No quarto cenário, representado na Tabela 6, de forma similar ao cenário 2 a janela ficou fixa e o passo foi aumentado, nesse a melhor acurácia foi 70.45%, aumento de 4.83% em relação ao cenário dois. As outras métricas F1, P (*precision*) e R (*recall*), do cenário 4 foram satisfatórias, na qual a maioria ficou acima dos 70%. Assim como nos cenários 1 e 2, foi analisada a taxa MCC, para verificar a qualidade das predições com a janela de cinco dias, os valores dessa métrica foram similares ao caso com janela de três dias.

Tabela 7. Cenário 05: Janela 7 dias - Passo 1.

	CART	RF	ET	GBC
AUC	57.0 ± 0.044	53.07 ± 0.066	58.07 ± 0.044	57.15 ± 0.078
ACC	62.77 ± 0.045	67.77 ± 0.054	70.55 ± 0.037	68.3 ± 0.055
F1	72.78 ± 0.050	79.39 ± 0.036	80.85 ± 0.025	78.90 ± 0.047
MCC	10.38 ± 0.14	—	15.85 ± 0.12	15.78 ± 0.16
P	76.69 ± 0.026	73.70 ± 0.032	76.18 ± 0.022	75.86 ± 0.040
R	70.0 ± 0.092	86.15 ± 0.052	86.13 ± 0.030	82.3 ± 0.062

Tabela 8. Cenário 06: Janela 7 dias - Passo 2.

	CART	RF	ET	GBC
AUC	50.08 ± 0.12	48.0 ± 0.051	55.6 ± 0.062	51.25 ± 0.060
ACC	55.88 ± 0.11	61.17 ± 0.070	67.06 ± 0.028	59.99 ± 0.060
F1	66.97 ± 0.095	74.12 ± 0.058	78.03 ± 0.022	71.67 ± 0.052
MCC	—	—	10.36 ± 0.12	11.06 ± 0.11
P	70.18 ± 0.077	69.36 ± 0.025	73.97 ± 0.040	71.34 ± 0.034
R	64.16 ± 0.11	79.97 ± 0.10	83.3 ± 0.074	72.5 ± 0.089

No última cenário executado, foi aplicada uma janela de sete dias. Nas Tabelas 7 e 8 estão os resultados, com passo de um e dois dias respectivamente. Nota-se que houve um aumento na acurácia, porém há um decréscimo na taxa AUC, essa diminuição da taxa AUC pode ser explicada pelo desbalanceamento ocasionado com o aumento da janela. Por fim, ao avaliar todos os cenários em conjunto, nota-se que os classificadores treinados a partir modelo genérico apresentaram resultados satisfatórios.

6. CONCLUSÃO

Neste trabalho, foi apresentado uma abordagem para identificar padrões que possibilitam a predição de acidentes graves a partir do uso de algoritmos de *Machine learning*. Adicionalmente, foi proposto um protótipo de um sistema web inteligente que emprega o modelo de predição criado. Com uma interface simples, o protótipo possibilita o envio de bases de acidentes e conseguir uma predição de acidentes graves em um futuro próximo. Vale ressaltar os desafios que tiveram que ser vencidos para a construção

desse artigo, por exemplo; uma quantidade pequena de dados sobre o assunto.

Dessa forma, pretende-se servir de base para contribuir com futuros trabalhos para a diminuição de acidentes graves, evitando mortes e prejuízos para empresas brasileiras ou mundiais. Para pesquisas futuras, pretende-se testar a solução em ambientes reais, assim como aplicar novos algoritmos de *Machine learning* ao modelo de dados proposto.

REFERÊNCIAS

- Abad, A., Gerassis, S., Saavedra, Á., Giráldez, E., García, J.F., and Taboada, J. (2018). A bayesian assessment of occupational health surveillance in workers exposed to silica in the energy and construction industry. *Environmental Science and Pollution Research*, 26(29), 29560–29569. doi:10.1007/s11356-018-2962-6. URL <https://doi.org/10.1007/s11356-018-2962-6>.
- Abu-Mostafa, Y.S., Magdon-Ismail, M., and Lin, H.T. (2012). *Learning from data: a short course*. AML-book.com, S.l. OCLC: 808441289.
- Alawad, H., Kaewunruen, S., and An, M. (2020). Learning from accidents: Machine learning for safety at railway stations. *IEEE Access*, 8, 633–648. doi:10.1109/ACCESS.2019.2962072.
- Alzubi, J., Nayyar, A., and Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series*, 1142, 012012. doi:10.1088/1742-6596/1142/1/012012. URL <https://doi.org/10.1088/1742-6596/1142/1/012012>.
- Basilio, P. (2021). Brasil é 2º país do G20 em mortalidade por acidentes no trabalho. <https://g1.globo.com/economia/noticia/2021/05/01/brasil-e-2o-pais-do-g20-em-mortalidade-por-acidentes-no-trabalho.ghtml>. [Online; acessado em 20/04/2022].
- Brasil (1988). Inciso xxviii do artigo 7 da constituição federal de 1988. URL <https://www.jusbrasil.com.br/topicos/10726047/inciso-xxviii-do-artigo-7-da-constituicao-federal-de-1988>.
- Brasil (2018). Boletim epidemiológico: Acidentes de transporte relacionados ao trabalho no Brasil, 2007-2016. URL <http://renastonline.ensp.fiocruz.br/recursos/boletim-epidemiologico-acidentes-transpor-te-relacionados-trabalho-brasil-2007-2016>.
- Cardoso, C.T.S. and Batista, K.R. (2019). A responsabilidade do empregador diante do acidente de trabalho.
- Cordeiro, R., Luz, V.G., Hennington, É.A., Martins, A.C.A., and Tófoli, L.F. (2017). A violência urbana é a maior causa de acidente de trabalho fatal no brasil. *Revista de Saúde Pública*, 51, 123. doi:10.11606/s1518-8787.2017051000296. URL <https://doi.org/10.11606/s1518-8787.2017051000296>.
- Gerassis, S., Martín, J.E., García, J.T., Saavedra, A., and Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. *Journal of Construction Engineering and Management*, 143(2), 04016093. doi:10.1061/(asce)co.1943-7862.0001225. URL [https://doi.org/10.1061/\(asce\)co.1943-7862.0001225](https://doi.org/10.1061/(asce)co.1943-7862.0001225).

- Haq, A.U., Li, J.P., Memon, M.H., Nazir, S., and Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018, 1–21. doi:10.1155/2018/3860146. URL <https://doi.org/10.1155/2018/3860146>.
- Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., Deng, L., Zheng, C., Zhou, J., Shi, H., and Feng, J. (2020). Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nature Communications*, 11(1). doi:10.1038/s41467-020-18685-1. URL <https://doi.org/10.1038/s41467-020-18685-1>.
- Mosquera, R., Parra, L., Ledesma, A.J., and Bonilla, H.A.F. (2021). Predicci de la accidentalidad laboral en la industria de pulpa y papel usando algoritmos de clasificaci . *Informaci tecnol *, 32, 133 – 142. URL http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642021000100133&nrm=iso.
- Norvig, P. and Russell, S. (2013). *Intelig ncia Artificial*. ELSEVIER EDITORA. URL <https://books.google.com.br/books?id=KhUQvgAACAAJ>.
- Ramos, E. (2020). N mero de acidentes de trabalho no brasil e no rs segue alto. <https://www.trt4.jus.br/portais/trt4/modulos/noticias/305976>.
- Scott, E., Hirabayashi, L., Levenstein, A., Krupa, N., and Jenkins, P. (2021). The development of a machine learning algorithm to identify occupational injuries in agriculture using pre-hospital care reports. *Health Information Science and Systems*, 9(1). doi:10.1007/s13755-021-00161-9. URL <https://doi.org/10.1007/s13755-021-00161-9>.
- Takala, J., H m l inen, P., Saarela, K.L., Yun, L.Y., Marnickam, K., Jin, T.W., Heng, P., Tjong, C., Kheng, L.G., Lim, S., and Lin, G.S. (2014). Global Estimates of the Burden of Injury and Illness at Work in 2012. *Journal of Occupational and Environmental Hygiene*, 11(5), 326–337. doi:10.1080/15459624.2013.863131. URL <http://www.tandfonline.com/doi/abs/10.1080/15459624.2013.863131>.
- Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, ahead-of-print(ahead-of-print). doi:10.1016/j.aci.2018.08.003. URL <https://doi.org/10.1016/j.aci.2018.08.003>.
- Wang, C., Kim, Y., Kim, D.G., Lee, S.H., and Min, S.D. (2020). Smart helmet and insole sensors for near fall incidence recognition during descent of stairs. *Applied Sciences*, 10(7). doi:10.3390/app10072262. URL <https://www.mdpi.com/2076-3417/10/7/2262>.