Classificador de notificações de acidentes ambientais no Sistema Nacional de Emergências Ambientais do IBAMA

Filipe de M. Santos, Lucas M. Aguiar, Gilberto F. de Sousa Filho, Bruno J. Sousa Pessoa

[1]Centro de Informática - Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

{lucas.aguiar,filipemedeiros}@academico.ufpb.br, {gilberto,bruno}@ci.ufpb.br

Abstract. *Siema is the IBAMA system responsible for registering reports of environmental accidents from citizens throughout Brazil. According to IBAMA's own database, 28% of the reports are wrong, generating a financial and human cost for the institution during the verification of these supposed occurrences. This work proposes the use of machine learning techniques to classify future reports of accidents as valid or not, aiming at the best use ofpublic resources in the analysis of notifications received by the agency. Three machine learning models were employed and classification metrics were presented regarding Siema records, obtaining a classifier that is able to correctly identify 91% of invalid accident reports, in addition to an overall accuracy of 89%.*

Resumo. *O Siema é o sistema do IBAMA responsável pelo cadastro dos relatos de acidentes ambientais de cidadãos de todo o Brasil. Segundo a base de dados do próprio IBAMA, 28% dos relatos são equivocados, gerando um custo financeiro e humano para a instituição durante a verificação destas supostas ocorrências. Este trabalho propõe a utilização de técnicas de aprendizado de máquina para classificar futuros relatos de acidentes como válidos ou não, objetivando a melhor utilização dos recursos públicos na análise das notificações recebidas pelo órgão. Foram utilizados três modelos de aprendizagem de máquina e apresentadas métricas de classificação a respeito dos registros do Siema, obtendo-se um classificador que é capaz de identificar corretamente 91% dos relatos de acidentes inválidos, além de uma acurácia geral de 89%.*

1. Introdução

O Sistema Nacional de Emergências Ambientais (Siema) foi instituído pela Instrução Normativa nº 15, de 6 de outubro de 2014, sendo responsável por registrar comunicações de acidentes ambientais. O Siema constitui ambiente automatizado para o registro de acidentes ambientais por cidadãos através de preenchimento online de formulário próprio ou, excepcionalmente, por *email*, sendo a consolidação destas informações apresentadas através de visualizações de mapas interativos e de estatísticas descritivas dos acidentes relatados.

A equipe do Siema disponibilizou no Portal Brasileiro de Dados Abertos uma base de dados contendo os registros de comunicação sobre acidentes ambientais em todo o Brasil, a qual é composta por informações do acidente, do meio em que se efetivou a

comunicação e dados não identificáveis do denunciante. Em análise simples, é possível constatar que 28% das comunicações foram equivocadas, isto é, por não conhecerem o protocolo do IBAMA que define o que é ou não um crime ambiental, são relatadas ocorrências que não cumprem os requisitos necessários para serem consideradas como crimes. Como o órgão não sabe, a princípio, que trata-se de uma ocorrência equivocada, recursos materiais e humanos são alocados para a análise do suposto crime ambiental, que, embora mereça alguma atenção, por não ser prioritário, pode ser tratado em um momento posterior.

Para fazer uma melhor triagem das notificações recebidas, dando prioridade a ocorrências que estão dentro do rol de situações consideradas como crimes ambientais pelo IBAMA e assim alocando melhor os recursos do órgão, propomos a utilização de três modelos de Aprendizagem de Máquina (AM) capazes de prever se determinada ocorrência consiste em um relato válido ou não. Além disso, é proposta uma arquitetura de sistema que integra os modelos de AM com o Siema.

A área de Aprendizagem de Máquina foco deste trabalho se utiliza de algoritmos que constroem funções matemáticas cujo objetivo é identificar padrões presentes nas bases de dados de um determinado contexto, a fim de generalizar o padrão aprendido para novos dados. Neste trabalho, utilizamos dois modelos de aprendizagem com grande poder de classificação, como as Redes Neurais Artificial e as Máquinas de Vetores de Suporte, e um terceiro, baseado em Árvore de Decisão, capaz de interpretar os resultados obtidos na classificação.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta trabalhos da literatura que abordam a validação na comunicação dos usuários, a Seção 3 apresenta informações da base de dados disponibilizados pelo Siema e o mapeamento deste como entrada para os algoritmos de AM; na Seção 4 descrevemos uma arquitetura de integração do classificador proposto ao Siema e como os algoritmos de AM foram usados para construir o classificador de relatos de acidentes; a Seção 5 apresenta resultados computacionais encontrados com os classificadores propostos; e por fim a seção 6 apresenta as conclusões e trabalhos futuros.

2. Trabalhos relacionados

Validar as comunicações de clientes sobre ocorrências de eventos é um grande desafio para as companhias de seguro. Em [Severino and Peng 2021], é proposto um classificador de fraudes nas reivindicações de seguros de propriedades através de técnicas de AM que tem como entrada uma base de dados real da principal companhia de seguros brasileira. São analisados 9 modelos de AM, sendo os melhores resultados médios obtidos por: *random forest*, *gradiente boosting* e *deep learning*. Uma análise sobre as principais características da base de dados que melhor identificam reivindicações fraudulentas foi realizada, permitindo assim explicar o motivo das classificações obtidas e ajudando profissionais a construir suas regras de decisão na análise futura de apólices de seguro.

No ramo de seguro de veículos, [Frempong et al. 2017] propõem o desenvolvimento de um modelo preditivo para prever a probabilidade de sinistros com base em fatores de risco como idade do segurado e do veículo. Foi utilizada a técnica de árvores de decisão e como consequência foram obtidos resultados interpretáveis que demonstram, por exemplo, que os segurados corporativos com veículos com idade até 8 anos têm

maior probabilidade de sinistro. Já em [YÜCEL 2022], é proposto o uso de modelos de AM para a classificação de sinistros de veículos na tomada de decisão de seu pagamento. Para tanto, foi utilizada uma base de dados disponível no *kaggle.com* para a identificação de fraudes na descrição de danos físicos relacionados a acidentes. Inicialmente, aplicou-se a técnica de Processamento de Linguagem Natural para computar a frequência de uso de cada termo na descrição dos acidentes. Tais dados então foram testados em 5 modelos de AM, dos quais a Rede Neural Artificial foi técnica que obteve melhor acurácia de classificação.

A validação/classificação automática de propostas para alocação de recursos é algo que o setor público também tem enfrentado em várias situações. [Chand and Zhang 2022] trataram um problema enfrentado pelo governo australiano por meio do programa NDIS (do inglês, *Australian National Disability Insurance Scheme*), que aloca recursos financeiros para ajudar pessoas com deficiências. Foi constatado que apenas 1% dos participantes do programa gastava todos os recursos alocados pelo governo, reduzindo assim a eficiência no gasto público. Com base nos dados presentes nas declarações dos solicitantes, os autores empregaram modelos de AM para realizar uma previsão de gastos que minimizasse o *gap* entre recursos alocados e gastos, obtendo resultados promissores utilizando árvores de decisão.

Por sua vez, [Haritsah Luthfi and Hartoyo 2023] utilizaram os modelos *Support Vector Machine* (SVM) e Regressão Logística, combinados com o método explicativo de inteligência artificial, LIME, para realizar um estudo sobre a detecção de notícias falsas relacionadas à Covid-19. O método LIME permitiu explicar as razões pelas quais uma notícia foi classificada como verdadeira ou falsa, ao demonstrar que notícias falsas sobre Covid-19 tendem a ter palavras relacionadas a religião e política, enquanto notícias verdadeiras têm mais palavras relacionadas ao governo e à medicina.

À luz dos trabalhos mencionados e também considerando a finalidade de classificar comunicações realizadas por usuários de serviços, este trabalho tem como sua principal contribuição a utilização de modelos de Aprendizagem de Máquina para alocação mais eficiente de recursos humanos e financeiros no tratamento de crimes ambientes por parte do IBAMA.

3. *Dataset* SIEMA

A base de dados do Siema[1] disponibilizada pelo IBAMA está sob licença de uso Open Data Commons Open Database License (ODbL), e a versão deste trabalho foi extraída do *site* no dia 02 de fevereiro de 2023. A base possui 13.127 registros de acidentes ambientais relatados por cidadãos de todo o Brasil. Removemos os registros que estão sem conteúdo em alguma das seguintes colunas: periodo_ocorrencia, uf e origem. Ao final permaneceram 11.616 registros, sendo 8.370 destes relatos considerados válidos pela equipe do Siema e 3.246 relatos de acidentes rotulados como inválidos.

A base se encontra em formato tabular com 68 colunas com diversos tipos de informação sobre a comunicação registrada. Inicialmente, destacamos a coluna *validado* cujos valores são rótulos binários indicando que o relato foi validado pela equipe

[1]Link para acesso à base de dados do SIEMA: https://dados.gov.br/dados/conjuntos-dados/siema-comunicado-de-acidente-ambiental

do IBAMA, ou não foi validado, tratando-se de uma comunicação a priori menos relevante, assumindo valores 1 ou 0, respectivamente.

Como o objetivo deste trabalho é classificar novas comunicações de acidentes como válidos ou inválidos, descartamos as características de baixo valor preditivo, assim classificadas pelo seu valor de correlação com o rótulo da coluna *validado* que obtenham valor menor que 0,15. Seguindo esta regra as seguintes colunas da base foram descartados: FID, uuid, id_ocorrencia, id_município, id_uf, id_responsavel, des_complemento_tipo_localizaca, des_complemento_tipo_evento, cpf_contato, des_obs, des_danos, des_complemento_instituicao_atu, des_complemento_tipo_dano_ident, dt_ocorrencia, hr_ocorrencia, dt_registro, nro_ocorrencia, des_informacoes_adicionais, plano_emergencia, des_outras_providencias, dt_carga, dt_primeira_obs, hr_primeira_obs, acao_inicial_tomada, des_causa_provavel, situacao_atual_descarga, endereco_ocorrencia, nome_instituicao_atuando, geom, telefone_ instituicao_atuando, tipo_substancia, volume_estimado, produtos_onu, produto_nao_se_aplica, produto_perigoso, tipo_evento, produto_nao_especificado, id_bacia_sedimentar, legado, des_instituicao_empresa, nome_comunicante, telefone_contato, email_comunicante, des_funcao_comunicante, ip_contato, tipo_comunicado, periodo_primeira_obs, municipio, feicao_proxima, bacia_sedimentar, dt_ocorrencia_feriado, dia_semana, dia_semana_primeira_obs, dia_semana_registro.

As técnicas de AM utilizadas, durante a fase de treinamento, esperam receber informações numéricas, entretanto, muitas das informações da base original são compostas por textos descritivos, como, por exemplo, as colunas des_ocorrencia, origem, instituicoes_atuando_local, tipos_fontes_informacoes e tipos_danos_identificados.

Para viabilizar a utilização das colunas acima foram utilizadas as seguintes estratégias:

1. Os dados da coluna des_ocorrencia foram transformados em dados binários que representam a informação se o relato de acidente possui ou não uma descrição da ocorrência;
2. Para tratar os dados da coluna origem partimos do pressuposto de que a informação do local de origem do acidente segue uma organização hierárquica em ordem decrescente de relevância e filtramos o primeiro registro presente em cada entrada;
3. As colunas instituicoes_atuando_local, tipos_fontes_informacoes e tipos_danos_identificados foram pré-processadas de maneira uniforme, seus dados foram mapeados para se obter a quantidade de informações registradas e, com isso, foram transformadas em quant_instituicoes_atuando_local, quant_fontes_informacoes e quant_tipos_danos_identificados.

Sendo assim, mapeamos os textos descritivos e dados categóricos nominais em variáveis numéricas que estão detalhadas no Quadro 1. Essas 12 variáveis compõem as amostras a serem utilizadas pela técnica na fase de treinamento, e também para predizer novos relatos de acidentes como reais ou trotes.

4. Classificador de comunicações do Siema

O componente de classificação de comunicações do Siema proposto neste trabalho utiliza as técnicas de aprendizado de máquina sobre a base de dados descrita na Seção 3 e tem como objetivo classificar os relatos dos usuários do Siema como válidos ou inválidos.

Variável	Descrição	Valores
plano_emergencia_acionado	Indica se algum plano de emergência foi acionado, podendo ele ser oriundo de instituições públicas ou privadas.	{sim : 1} {não : 0}
iniciadas_outras_providencias	Indica se alguma instituição especializada, seja pública ou privada, iniciou alguma ação para conter ou minimizar os danos, como: cordões de contenção, mantas de absorção, almofadas, entre outros.	{sim : 1} {não : 0}
informacao_geografica	Indica se foi informado o local do acidente.	{sim : 1} {não : 0}
informacao_responsavel	Indica se foi fornecida as informações sobre o responsável pelo acidente.	{sim : 1} {não : 0}
ocorrencia_oleo	Indica se o acidente se trata de um incidente de poluição por óleo em águas sob jurisdição nacional.	{sim : 1} {não : 0}
des_ocorrencia	Indica se foi fornecida pelo notificante uma descrição do acidente.	{sim : 1} {não : 0}
periodo_ocorrencia	Turno em que ocorreu o acidente ambiental.	{matutino : 0} {noturno : 1} {sembrol : 2} {vespertino : 3}
uf	Unidade federativa em que ocorreu o acidente.	0 até 26
origem	Define o tipo de local onde se iniciou o acidente. São categorizadas em: Rodovias, Indústrias, Armazenamento/depósito, Plataforma entre outros.	0 até 11
quant_institiuicoes_atuando_local	Quantidade de instituições cujo notificante sabe que estão atuando no local do acidente.	número inteiro
quant_tipos_fonte_informação	Quantidade de fontes de informação pela qual o acidente foi comunicado.	número inteiro
quant_tipos_danos_identificados	Quantidade de tipos de danos identificados pelo notificante no local do acidente.	número inteiro

Quadro 1. Descrição das variáveis utilizadas do *dataset* do SIEMA.

A seguir, descrevemos uma proposta de integração deste componente com o Siema e detalhes da implementação de seu motor de aprendizado.

4.1. Integração ao Siema

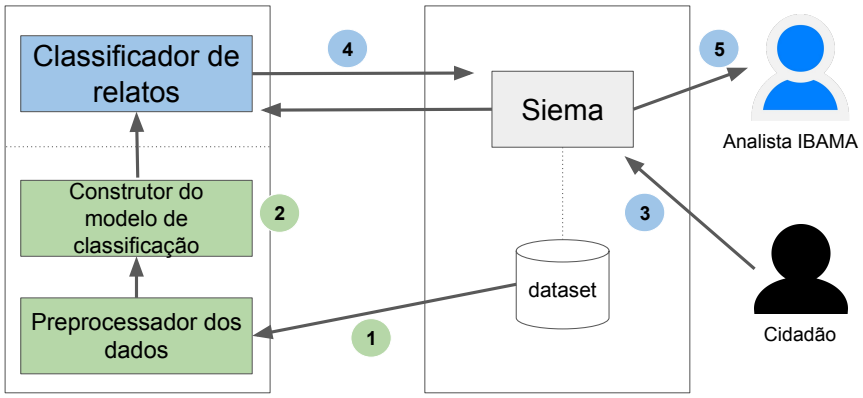
São propostas duas fases distintas de execução para o componente prosto: treinamento/aprendizado e classificação/predição. A Figura 1 apresenta uma arquitetura de integração ao Siema, ilustrando a comunicação entre os componentes que atuam em cada fase. As etapas de cada fase são rotuladas por números.

A fase de aprendizado é dividida em duas etapas. Na Etapa 1 a base de dados (*dataset*) é requisitada pelo componente *Preprocessador* dos dados ao Siema, eliminando dados inválidos e transformando cada amostra no formato descrito no Quadro 1. Na Etapa 2, o componente *Construtor do modelo de classificação* recebe os dados manipulados pelo *Preprocessador*, aplica o algoritmo de aprendizado para se ajustar aos dados de entrada e fornece como saída uma função hipótese que identifica o padrão e é utilizada na fase de

classificação. A fase de aprendizado deve ser realizada uma única vez ou toda vez que a base de dados estiver atualizada com uma quantidade grande de novos dados.

A fase de classificação será executada toda vez que um usuário cadastrar um relato de ocorrência de um novo acidente ambiental e consiste nas três etapas 3, 4 e 5. Na Etapa 3, o usuário informa a ocorrência de um novo acidente pelo formulário *online* presente no sítio eletrônico do sistema ou via *email* e este relato é cadastrado no Siema. Na etapa 4, o Siema informa os dados do relato cadastrado ao componente *Classificador de relatos* que utiliza a função hipótese criada para classificar esse relato como válido ou não. Tal classificação é fornecida como resposta ao Siema, que repassa a informação aos analistas do IBAMA através de notificações ou relatórios estatísticos (Etapa 5).

Figura 1. Arquitetura de integração do classificador de comunicações ao Siema.



4.2. Modelos de Aprendizagem de Máquina

Empregamos neste trabalho três modelos de aprendizado de máquina para a construção de classificadores dos relatos de acidentes: Redes Neurais Artificiais, Máquina de Vetores de Suporte e Árvore de Classificação. As métricas de classificação de tais modelos são comparadas na Seção 5.1.

Todo modelo de AM possui parâmetros de configuração que devem ser ajustados aos dados de treinamento com o objetivo de evitar aprender padrões errados ou pouco frequentes no universo de valores, os chamados ruídos de dados. Aprender os ruídos leva ao problema conhecido como *overfitting*, quando o classificador se ajusta demais aos dados de treino, tendo sua qualidade de classificação reduzida para dados desconhecidos (teste). Para tratar o fenômeno de *overfitting*, a base destinada ao treinamento é dividida em base de treino e base de validação e, por meio da estratégia de validação cruzada, são escolhidos os valores dos hiperparâmetros de cada modelo que melhor generalizam a aprendizagem obtida na fase treinamento. A seguir faremos uma descrição resumida de cada modelo utilizado e listamos os valores dos hiperparâmetros selecionados.

Support Vector Machine (SVM)

A Máquina de Vetores de Suporte (do inglês *Support Vector Machine*) é um algoritmo que, no contexto de classificação, busca obter uma função fronteira que proporcione a melhor separação dos dados de diferentes classes, isto é, os dados são classificados conforme o lado da fronteira em que se encontram. O SVM possui dois parâmetros a serem

ajustados: o γ que é responsável por ponderar a contribuição da proximidade dos dados na composição da função fronteira que será utilizada na classificação de uma nova amostra, e o parâmetro C responsável pela flexibilidade da função fronteira com relação a dados mal posicionados (ruídos). Para a validação destes parâmetros foi utilizado a técnica *cross-validation* com 5 *slots* ([Cichosz 2011]) e valores para $\gamma \in \{0,01; 0,1; 1; 5\}$ e $C \in \{1; 2; 5; 10\}$.

Rede Neural Artificial (RNA)

É uma técnica inspirada no cérebro humano que consiste em vários nós de processamento interconectados, distribuídos em camadas e com as conexões entre nós associadas a pesos. Para todo nó da rede, os valores das conexões de entrada deste são agrupados, e caso o valor da operação resultante desse agrupamento seja inferior a um limiar, esse nó não passa nenhuma informação à camada seguinte, caso contrário, é transmitido o resultado da operação.

Durante o treinamento, busca-se ajustar os valores dos limiares e dos pesos das conexões para classificar os dados de entrada de forma consistente. Em todo caso, os dados percorrem todas as camadas da rede, sendo submetidos a várias transformações não-lineares, até chegar à camada de saída, que classifica o exemplo examinado pela rede.

Com base na teoria da dimensão VC das redes neurais e na quantidade mínima de exemplos associada à tal métrica para atingir uma boa generalização ([Abu-Mostafa et al. 2012]), foram testadas duas arquiteturas, uma com uma única camada escondida de 58 neurônios e outra com duas camadas escondidas de 116 neurônios.

Para a segunda arquitetura, a fim de evitar *overfitting*, foram utilizadas técnicas de regularização como *Dropout*, *Early Stopping* e *Weight Decay*. Comparando-se resultados preliminares das duas arquiteturas, a primeira obteve os melhores resultados no que diz respeito à generalização.

Árvore de Classificação (AC)

É um algoritmo que possui como grande vantagem o fato de suas classificações serem facilmente interpretáveis. Na árvore, cada nó interno, ou nó de ramificação, representa um teste a ser feito com uma das variáveis de entrada, sendo definido um limiar l para a variável v escolhida e ramificando o nó em duas subárvores, a subárvore da esquerda que recebe as amostras cujo $v < l$ e a subárvore da direita com as amostras contendo $v \geq l$. Por sua vez, cada nó folha representa a classificação final da amostra que alcançá-la.

Para o algoritmo AC o principal parâmetro de definição da complexidade de uma árvore é seu número de ramificações. Quanto mais ramificações a árvore tem, mais o algoritmo tem a capacidade de aprender detalhes sobre os dados, podendo levar ao fenômeno do *overfitting*. Sendo assim, utilizamos o algoritmo de regularização *Minimal Cost-Complexity Pruning* [Breiman et al. 1984], que realiza uma busca exaustiva pelo valor do parâmetro α da árvore T que minimiza a função de classificação

$$R_\alpha(T) = R(T) + \alpha \cdot |T|,$$

procurando ajustar o T aos dados, equilibrando os erros de classificação $R(T)$ medidos

na base de validação e o tamanho $\alpha \cdot |T|$ da árvore.

5. Resultados computacionais

Os algoritmos de aprendizagem de máquina, listados na Seção 4.2, foram implementados na linguagem de programação *Python*, versão 3.9.6, com auxílio da biblioteca *scikit-learn* [du Boisberranger et al. 2022]. Todos os experimentos computacionais deste trabalho foram executados em um computador com processador Intel Core i5, de 4 núcleos de 2,3 GHz, a máquina dispõe de 8 GB de memória RAM, e possui como sistema operacional o MacOS Ventura.

5.1. Métricas de classificação

Para medir corretamente os resultados obtidos pelos modelos de AM, dividimos a base original em duas partes: base de treino com 70% dos dados e base de teste com 30% dos dados. A base de treino foi utilizada pelos algoritmos como descrito na Seção 4.2, já a base de teste foi utilizada para medir a qualidade dos classificadores treinados e seus resultados estão detalhados na Tabela 1, cujas colunas são descritas a seguir:

- $modelo$: define o modelo de AM cujas métricas estão sendo apresentadas;
- E_{out}: porcentagem de classificações erradas do algoritmo na base de teste;
- E_{in}: porcentagem de classificações erradas do algoritmo na base de treino;
- ΔE : diferença entre E_{out} e E_{in};
- $classe$: classe $c \in \{0, 1\}$ da comunicação avaliada. O valor 0 indica classificação de comunicações como inválidas (trote), e valor 1 indica classificação de comunicações válidas;
- *precisão*: razão entre a quantidade de amostras corretamente classificadas pelo algoritmo como c e a quantidade total de amostras classificadas como c;
- $recall$: razão entre a quantidade de amostras classificadas corretamente pelo algoritmo como c e o número total de objetos da classe c.

Existem duas etapas principais durante o processo de aprendizagem de máquina: a primeira é identificar o padrão no conjunto de dados de treinamento, sendo esta medida pelo E_{in} de cada modelo. Quem tiver o menor E_{in}, obteve o melhor aprendizado nessa fase. E a segunda é a generalização do aprendizado que tem por objetivo verificar se acurácia obtida na classificação dos dados de treinamento também ocorre nos dados de teste.

É possível perceber que os algoritmos RNA e AC obtiveram os menores valores para E_{out}, com $10,67\%$ e $10,7\%$ respectivamente, alcançando acurácias de 89% em suas classificações de relatos de acidente ambiental. O algoritmo RNA obteve a melhor generalização de aprendizado com o $\Delta E = 0,0038$.

A precisão de classe '1', que indica a taxa de acerto ao identificar a comunicação como válida, foi superior a 95% para todos os modelos, ou seja, com a adoção do classificador proposto, no máximo 5% das comunicações serão tratadas como relevantes sem necessidade.

5.2. Classificação interpretável

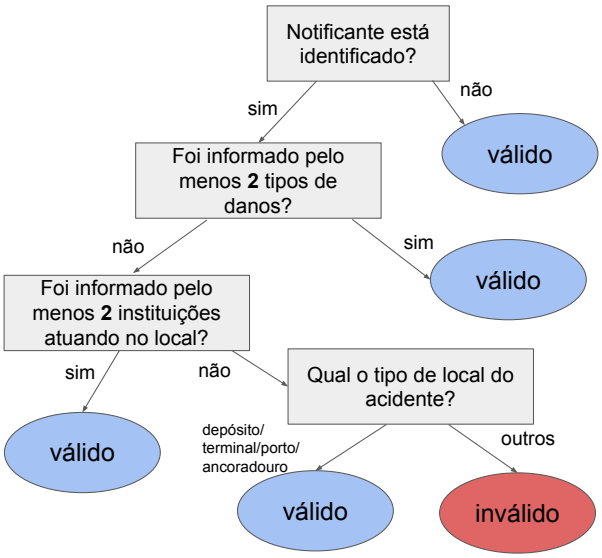
Há um *trade-off* muito comum quando se trabalha com modelos de aprendizagem de máquina. Por um lado, pode-se priorizar a qualidade do aprendizado, por outro pode-se

Tabela 1. Resultado das classificações obtidas pelos algoritmos RNA, SVM e Árvore de Classificação sobre a base de dados do Siema.

modelo	E_{out}**(%)**	E_{in}**(%)**	ΔE	**classe**	**precisão (%)**	***recall* (%)**
RNA	10,67	10,29	**0,0038**	0	77	88
				1	**95**	90
SVM	10,82	10,06	0,0072	0	75	91
				1	**96**	88
AC	10,70	10,17	0,0053	0	76	91
				1	**96**	89

focar na interpretabilidade dos resultados obtidos. Como já foi observado, o modelo RNA obteve a melhor generalização de aprendizado, entretanto, ele é considerado um modelo ”caixa preta”, pois não é capaz de descrever os motivos, por exemplo, que fizeram uma comunicação ser considerada como válida ou inválida. Por outro lado, modelos como a Árvore de Classificação, apesar de não ter a capacidade de aprendizado de outras técnicas, constrói uma estrutura em árvore capaz de descrever, com uma linguagem humana, os motivos que a conduziram para uma classificação. A Figura 2 apresenta a estrutura da árvore de classificação inferida pelo modelo AC sobre os dados do Siema.

Figura 2. Árvore de classificação inferida do *dataset* do Siema/IBAMA.



Para ilustrar a interpretação da classificação de notificações, propomos o seguinte conjunto de valores $m = \{$*não, não, sim, sim, não, sim, vespertino, PB, rodovia,*$0, 2, 1\}$, atribuídos conforme a ordem das variáveis descrita no Quadro 1. Classificando m a partir da árvore da Figura 2, temos que a notificação em questão é inválida, dado que m possui a identificação do notificante ($informacao\_responsavel = sim$); possui menos de 2 tipos de danos no acidente ($quant\_tipos\_danos\_identificados = 1$); menos de duas instituições foram identificadas atuando no local ($quant\_institiuicoes\_atuando\_local = 0$); e o tipo local do acidente é uma rodovia ($origem =$”*rodovia*”). Esta invalidade faz

sentido pois uma notificação de acidente que ocorre numa rodovia, que é um local público, não ter a presença de nenhuma instituição pública como a PRF ou órgão estadual equivalente não parece precisa.

6. Considerações finais e trabalhos futuros

Propomos neste trabalho um classificador de comunicações/notificações de acidentes ambientais relatados ao Siema do IBAMA, utilizando técnicas de aprendizado de máquina alimentadas pela base de dados disponível no Portal de Dados do Governo Federal (*dados.gov.br*). Foram removidas colunas de baixo valor preditivo e transformados dados textuais em valores categóricos ou numéricos para alimentar os três modelos propostos: SVM, RNA e AC (Árvore de Classificação). O RNA obteve a melhor generalização ($\Delta E = 0,0038$) e a melhor acurácia de classificação (89%) durante o teste de validação cruzada. Já o modelo AC obteve métricas de classificação muito próximas do RNA, porém com uma melhor capacidade de interpretabilidade dos motivos da classificação nas comunicações, sendo sua adoção mais indicada por trazer aos analistas do IBAMA mais informações para sua tomada de decisão no tratamento das comunicações.

Como trabalho futuro, propomos a adoção de técnicas de Processamento de Linguagem Natural para melhor analisar os dados de entrada textual das comunicações, procurando assim associar o uso de termos específicos em mensagens válidas ou inválidas, a fim de melhorar sua acurácia de classificação.

Referências

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data*. AMLBook.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Chand, S. and Zhang, Y. (2022). Learning from machines to close the gap between funding and expenditure in the australian national disability insurance scheme. *International Journal of Information Management Data Insights*, 2(1):100077.

Cichosz, P. (2011). Assessing the quality of classification models: Performance measures and evaluation procedures. *Open Engineering*, 1(2):132–158.

du Boisberranger, J., Van den Bossche, J., Estève, L., and J. Fan, T. (2022). scikit-learn: machine learning in python.

Frempong, N. K., Nicholas, N., and Boateng, M. (2017). Decision tree as a predictive modeling tool for auto insurance claims. *International Journal of Statistics and Applications*, 2017:117–120.

Haritsah Luthfi, N. and Hartoyo, A. (2023). Ai explanation related covid hoax detection using support vector machine and logistics regression methods. *Jurnal Media Informatika Budidarma*, 7:170–177.

Severino, M. K. and Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, 5:100074.

YÜCEL, A. (2022). A novel data processing approach to detect fraudulent insurance claims for physical damage to cars. *Journal of New Results in Science*, 11:120–131.