

Comparativo de Algoritmos de *Machine Learning* para Classificação de Acidentes e Incidentes Aéreos: Estudo de Caso Brasileiro (2014-2024)

Maria Helena¹, George Souza¹, Ayllen Aryelle¹

¹Faculdade Nova Roma

Av. Adjar da Silva Casé, 800 - Indianópolis, Caruaru - PE, 55024-740 - Brasil

Abstract. *Air transport plays a crucial role in the economy, boosting several sectors. Despite being generally considered safe, there are still occurrences of aviation accidents and incidents, resulting in both personal and material damage. This study seeks to contribute to the safety and continuous improvement of the airline sector, identifying patterns and risks associated with these events. For this, different Machine Learning algorithms were compared, such as Random Forest, XGBoost and LightGBM, with the aim of classifying occurrences between incidents and air accidents. The data used was collected in Brazil over the period between 2014 and 2024.*

Resumo. *O transporte aéreo desempenha um papel crucial na economia, impulsionando diversos setores. Apesar de ser geralmente considerado seguro, ainda há ocorrências de acidentes e incidentes aéreos, resultando em danos tanto pessoais quanto materiais. Este estudo busca contribuir para a segurança e aprimoramento contínuo do setor aéreo, identificando padrões e riscos associados a esses eventos. Para isso, foram comparados diferentes algoritmos de Machine Learning, como Random Forest, XGBoost e LightGBM, com o objetivo de classificar as ocorrências entre incidentes e acidentes aéreos. Os dados utilizados foram coletados no Brasil ao longo do período entre 2014 e 2024.*

1. Introdução

O transporte aéreo desempenha um papel fundamental na economia de um país, impulsionando a movimentação de pessoas, mercadorias e o crescimento do turismo. Além de facilitar o deslocamento e conectar regiões distantes, esse setor também contribui para a atração de investimentos e a criação de empregos. Em conjunto com as demais atividades turísticas, a aviação e sua cadeia de abastecimento representam uma parcela relevante do Produto Interno Bruto (PIB) nacional, alcançando cerca de 2,9% em 2021[1].

Embora o transporte aéreo seja amplamente reconhecido como uma das formas mais seguras de viagem, os acidentes e incidentes aéreos continuam a ocorrer em todo o mundo, apesar do uso de tecnologias avançadas na fabricação de aeronaves[2].

Um acidente, derivado do termo latino "accidens" (acazos), é qualquer evento que perturba o curso normal de uma ação ou evento, causando lesões ou danos pessoais, materiais, ou ambos. Já o incidente é um evento que tem o potencial de causar danos a alguém ou a algo, mas que não resulte em nenhum dano visível ou mensurável. É uma situação que poderia ter levado a um acidente ou que deu origem a um acidente. Ambos

podem ser causados por uma variedade de fatores, como erros humanos, falhas mecânicas, entre outros[3].

Na década de 90, a taxa média de acidentes era de uma ocorrência a cada 1,54 milhão de voos. Já entre 2010 e 2015, essa taxa diminuiu consideravelmente, apresentando apenas um acidente a cada 2,75 milhões de voos.[4]. Apesar da redução no número de ocorrências, o crescimento da aviação comercial tem suscitado uma crescente preocupação entre pesquisadores de todo o mundo, que buscam ampliar a compreensão dos fatores que podem desencadear um acidente aéreo[5]. A indústria da aviação reconhece a identificação de problemas e o gerenciamento de risco de segurança como processos fundamentais para a gestão de segurança[6]. No entanto, mesmo com os avanços técnico-científicos e na gestão de segurança, os acidentes aéreos continuam a resultar em inúmeras fatalidades globalmente, destacando a necessidade de novas abordagens para prevenção desses acidentes[7].

Dentre as técnicas de modelagem e abordagens aplicadas em nível mundial, a utilização de algoritmos de *Machine Learning* vem tendo destaque. Estes modelos possuem diversas aplicações no contexto de predição, podendo prever desde variação cambial até demanda energética[8]. Mundialmente no ramo da aviação, inúmeros estudos avaliando estes algoritmos vêm sendo aplicados para desenvolver modelos de classificação de acidentes aéreos[9], porém poucos estudos com este propósito foram desenvolvidos no Brasil.

O Brasil reportou 211 acidentes aéreos entre 2006 e 2015, com uma média de 133 mortes por ano no segmento da aviação[10]. Esse volume de acidentes evidencia a relevância de estudos que possam auxiliar gestores e tomadores de decisão para minimizar a ocorrência destas fatalidades. Entretanto, no Brasil algoritmos de *Machine Learning* vem sendo pouco empregados no cerne de avaliação de fatores que possam desencadear acidentes aéreos, com estas abordagens sendo utilizadas para identificar erros de natureza humana após a ocorrência dos acidentes[11] ou para avaliar a natureza de processos judiciais para companhias aéreas[12].

Este estudo pretende avaliar diferentes algoritmos de *Machine Learning* para a classificação de acidentes aéreos no Brasil, utilizando dados fornecidos pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA) entre 2014 e 2024.

2. Fundamentação Teórica

2.1. Acidentes Aéreos e Machine Learning

Acidentes e incidentes aéreos são historicamente alvos de análises causais com objetivo de avaliar quais fatores foram diretamente responsáveis em causar o evento (fatores de causa primária) ou quais fatores contribuíram com o evento (fatores de causa secundária)[13]. Este tipo de análise possui a vantagem de poder identificar os fatores mais comuns em múltiplos casos, fazendo com que sistemas de segurança possam concentrar suas atenções nos fatores que possam oferecer maior risco para as operações aéreas[14].

Análises causais de acidentes aéreos contribuíram para identificar a natureza de diversos desastres da aviação, identificando problemas como perda de controle[15] e condições climáticas[16]. Entretanto, a maior parte destes estudos realizam análises univariadas para se ter um entendimento específico acerca de como um fator pode impactar

a aviação, porém a operação aérea se caracteriza por possuir uma elevada complexidade, estando sujeita a inúmeros fatores simultaneamente[17].

Dentre as técnicas de análise multivariada, a avaliação de algoritmos de *Machine Learning* vem sendo uma das mais estudadas contemporaneamente. Esta atenção se deve principalmente à ampla gama de aplicabilidade destes algoritmos, trazendo alta capacidade preditiva em contextos que vão desde análises de origem de produtos agrícolas[18] e pesquisas de paralisia cerebral[19] a preços de commodities pesqueiras[20]. A utilização destes algoritmos já vem sendo estudada no contexto da aviação para identificação de anomalias em aviões comerciais[21], riscos de turbulência[22] e prognósticos de falha[23].

2.2. Modelos e Métricas de Avaliação

Os algoritmos escolhidos esta pesquisa foram *Random Forest*, *XGBoost* e *LightGBM*, que são conhecidos por sua capacidade de trabalhar com dados complexos, aplicados aqui para analisar padrões e tendências nos registros de acidentes e incidentes aéreos.

O algoritmo *Random Forest* é um classificador que utiliza o método de árvores de decisão, para realizar a mineração de dados. Ao contrário das abordagens convencionais de árvores de decisão, o *Random Forest* busca criar várias árvores de decisão, cada uma utilizando um subconjunto aleatório de atributos selecionados a partir do conjunto original. Essas árvores são construídas usando uma técnica de amostragem chamada *Bootstrap*, que seleciona os dados com reposição. Essa abordagem diversificada permite uma análise mais robusta dos dados, facilitando a classificação[24].

O *XGBoost* é um algoritmo de aprendizado de máquina escalável e preciso, implementando o processo de *Boosting* para gerar modelos precisos de árvore de decisão. Ele corrige as deficiências de modelos anteriores a cada nova iteração. Sua escalabilidade permite sua execução em uma ampla variedade de cenários, desde uma única máquina até configurações distribuídas para bilhões de exemplos, graças a otimizações algorítmicas e sistemas importantes, como computação paralela e distribuída. Isso possibilita uma exploração mais rápida do modelo e o processamento de grandes volumes de dados em computadores[25].

O *LightGBM* é um algoritmo eficiente usado para classificação que melhora a precisão e a eficiência do modelo de árvores de decisão. É usado uma técnica chamada *GOSS (Gradient-based One-Side Sampling)* para determinar pontos de divisão e o *EFB (Exclusive Feature Bundling)* para acelerar o treinamento, agrupando recursos exclusivos. Isso torna o processo mais rápido e eficaz, mesmo com conjuntos de dados grandes e complexos[26].

Na avaliação do desempenho dos modelos de *Machine Learning*, utilizamos métricas como acurácia, precisão, *recall* e *F1-score*.

A acurácia mede a proporção de previsões corretas feitas pelo modelo em relação ao total. A precisão indica a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo[27]. O *recall* mede a proporção de previsões positivas corretas em relação ao total de esferas positivas na base de dados. O *F1-score* é uma média harmônica entre precisão e *recall*, fornecendo uma medida única que leva em consideração tanto os falsos positivos quanto os falsos negativos[28].

Essas métricas são cruciais para avaliar a eficácia dos modelos de *Machine Learning* na previsão de acidentes aéreos, fornecendo insights valiosos sobre sua confiabilidade na identificação de eventos críticos.

2.3. Trabalhos Relacionados

Estudos recentes utilizaram o aprendizado de máquina para prever e analisar os riscos relacionados aos acidentes aéreos. Essas pesquisas não só fornecem informações cruciais para a segurança na aviação, mas também ressaltam os desafios enfrentados ao lidar com a complexidade dos dados e aos eventos imprevisíveis associados aos acidentes aéreos.

Um estudo recente [21] abordou a eficácia dos modelos de *Machine Learning* na prevenção de acidentes aéreos, ressaltando sua importância para a segurança na aviação. A pesquisa incluiu diversas etapas, desde a seleção e rotulagem dos dados até o desenvolvimento e teste dos modelos preditivos. O modelo mais eficiente, a análise discriminante quadrática, alcançou uma impressionante precisão de 93%, evidenciando sua adequação. Além disso, a ampla área sob a curva para detecção de anomalias e normalidades confirma a robustez do modelo.

Shinya Mizuno, Haruka Ohba e Koji Ito [22] usaram dados abertos para prever as turbulências, visando a segurança dos voos. Utilizou-se clusters de risco como supervisão para prever eventos de baixa frequência. A análise de fatores influenciadores ajudou na modelagem dos clusters. Os resultados foram confirmados como precisos para uso pelos pilotos. Além disso, destacou-se a viabilidade de prever turbulências com dados abertos e meteorológicos comumente utilizados.

Alan Burnett e Dong Si[29] exploraram o uso de aprendizado de máquina para prever condições que aumentam a probabilidade de acidentes graves na aviação. Embora os resultados mostrem potencial na melhoria dos modelos de previsão, também destacam desafios ao lidar com dados irregulares e a complexidade inerente à previsão de acidentes aéreos e gravidade das lesões.

Xiaoge Zhang e Sankaran Mahadevan[30] desenvolveram um modelo híbrido usando *SVM (Support Vector Machine)* e *RNN (Recurrent Neural Network)* para avaliar o risco de eventos perigosos no transporte aéreo, categorizando eventos em cinco grupos de risco, onde propõem explorar informações sobre ações dos operadores e técnicas avançadas de mineração de dados para identificar padrões e interações entre eventos, além de uma análise de causa raiz para identificar fatores cruciais.

3. Metodologia

A pesquisa é um processo sistemático e controlado que permite encontrar novos fatos e conexões em qualquer área. É uma atividade que envolve coletar informações para resolver problemas específicos. O método científico é essencial para definir as ciências, usando técnicas que garantem que o conhecimento obtido seja válido e verdadeiro. Isso inclui formas de planejar a pesquisa, encontrar erros e ajudar nas decisões do pesquisador, garantindo que o trabalho seja preciso e confiável[31].

Neste estudo, será utilizada a metodologia *CRISP-DM (Cross-Industry Standard Process for Data Mining)*, uma das abordagens mais comuns e amplamente aceita pela indústria para criar modelos com base em dados, sendo composta por seis etapas bem estruturadas e definidas[32].

Na primeira etapa é realizada a Compreensão do Negócio, onde é identificado metas e necessidades para elaborar um plano inicial de *Data Mining*. Na segunda etapa ocorre a Compreensão dos Dados, que prioriza a identificação das informações pertinentes da base de dados e a avaliação de sua utilidade. Em seguida é feita a Preparação dos Dados, que engloba atividades como extração, limpeza e transformação para obter um conjunto final de dados para modelagem. Na quarta etapa ocorre a Modelagem, onde são selecionadas e aplicadas as técnicas adequadas para atingir os objetivos propostos. Posteriormente, é realizada a Avaliação, que verifica se os modelos desenvolvidos atendem aos objetivos de negócio, assim, decidindo se o processo deve prosseguir ou ser ajustado. Por fim, na sexta etapa, ocorre o Deploy, onde é aplicado os resultados das técnicas para beneficiar o negócio.[33].

A figura 1 abaixo representa a sequência de cada uma das fases do modelo de processo do CRISP-DM.

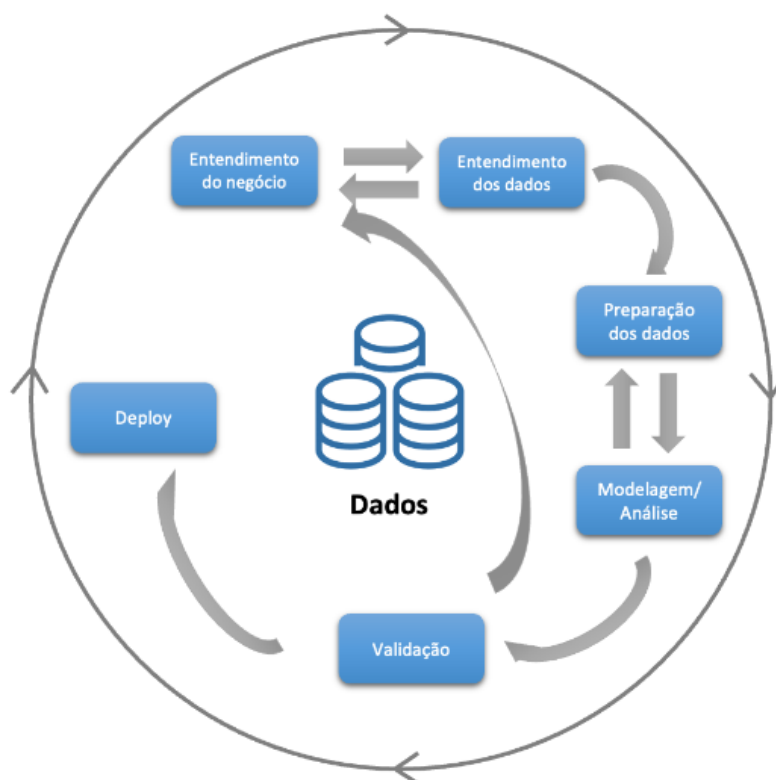


Figure 1. Processo Crisp-DM. Fonte: [34]

O banco de dados utilizado foi obtido através de download no site do Painel Sipaer, que é uma ferramenta de visualização de dados desenvolvida pelo CENIPA (Centro de Investigação e Prevenção de Acidentes Aeronáuticos) que permite pesquisar as ocorrências aeronáuticas ocorridas no Brasil no período de 02/01/2014 à 09/05/2024. O banco de dados contém 7.160 linhas e 10 colunas com dados referentes a ocorrências aeronáuticas no Brasil.

Durante a etapa de preparação dos dados, foram selecionadas apenas as colunas que continham informações anteriores às ocorrências. Portanto, as colunas "Tipo" e "Sta-

tus” foram excluídas. Além disso, a coluna ”Matrícula” foi removida, pois está diretamente associada a características específicas de cada aeronave, o que poderia introduzir vieses nos resultados e comprometer a identificação precisa de padrões ou tendências gerais.

Ao realizarmos a leitura do conteúdo restante da base de dados, o sobressalto de ”incidentes” na coluna ”alvo” que será ”Classificação” em relação aos demais resultados possíveis, poderia levar o modelo a um sobreajuste, conhecido em estatística como ”*overfitting*”. Os dados tendem a se encaixar no valor em maior evidência, comprometendo a generalização (entrada) de novos dados.

Para nivelar o conteúdo, foi aplicada a técnica de ”*undersample*”. Que consiste em remover dados da classe majoritária a fim de equilibrar o conteúdo com as demais classes, sem gerar prejuízo na qualidade dos resultados e mitiga o risco de ”*overfitting*”.

Em seguida, foi realizada uma análise detalhada dos dados restantes para garantir a integridade e a qualidade do conjunto de dados. Foi verificado que nenhuma das colunas apresentava valores nulos, assegurando a integridade das informações disponíveis. Adicionalmente, foi confirmado que não havia linhas duplicadas na base de dados, eliminando a possibilidade de redundâncias que poderiam distorcer os resultados das análises subsequentes. Essa verificação rigorosa é fundamental para assegurar a precisão e a confiabilidade das modelagens a serem realizadas.

Para preparar as variáveis categóricas para a modelagem, foi aplicada a técnica de *One-Hot Encoding*. Esse método transforma variáveis categóricas em uma série de variáveis binárias, permitindo que algoritmos de aprendizado de máquina processem essas informações de forma eficaz[35]. As variáveis categóricas selecionadas para essa transformação incluíram as colunas ”Localidade”, ”UF”, ”Aeródromo” e ”Operação”. A técnica utilizada assegura que categorias desconhecidas que possam surgir durante a aplicação do modelo sejam ignoradas, prevenindo possíveis erros.

Além disso, a coluna ”Classificação” foi convertida em um formato numérico usando a técnica de *Label Encoding*. Este método atribui um valor inteiro único a cada categoria da variável alvo, facilitando a sua utilização em algoritmos de aprendizado de máquina que requerem entradas numéricas[36].

A classificação das ocorrências possui três níveis:

- **Incidente**
- **Incidente Grave**
- **Acidente**

Essas classificações foram utilizadas como variáveis de resposta nas modelagens subsequentes. Em outras palavras, essas são as variáveis que se busca prever ou explicar com o modelo. As variáveis explicativas, ou independentes, incluíram todas as demais do conjunto de dados: ”Número da Ocorrência”, ”Data”, ”Matrícula”, ”Localidade”, ”UF”, ”Aeródromo” e ”Operação”.

Foram rodados os modelos com e sem ”*Grid Search*”, que é um método de otimização de hiperparâmetros. O Grid Search consiste em através de combinações, buscar os parâmetros mais promissores de resultados assertivos[37], mas como veremos a seguir, nem sempre o resultado será melhor que sem sua aplicação.

4. Resultados

Em cenários de investigação de acidentes aéreos, a análise precisa dos dados disponíveis é fundamental para identificar as causas e prevenir futuros incidentes. Técnicas de *Machine Learning* emergem como ferramentas poderosas para apoiar essa análise, permitindo a identificação de fatores de risco e a previsão de acidentes com base em dados operacionais.

Os modelos foram executados em notebook jupyter, em linguagem python. Disponibilizados para download no repositório github, onde estão dispostos: 2 modelos em formato .ipynb e a base de dados CSV.

Clique aqui para acessar o repositório.

As análises consistiram da aplicação de técnicas e modelos buscando os melhores resultados. Seguindo o modelo científico e descartando abordagens que não tiveram sucesso em seus resultados, como veremos a seguir.

Primeiro comparamos os ambientes com e sem a aplicação da otimização de hiperparâmetros. O método "*Grid Search*" foi aplicado em um dos modelos e comparado ao outro modelo. Como resultado, houve queda de desempenho comparando com o modelo sem sua aplicação. Conforme o comparativo vide tabela 1 abaixo:

Table 1. Comparação Detalhada: Grid Search vs. Sem Grid Search (ambos com Undersampling)

2*Modelos	Sem Grid Search			Com Grid Search		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
Random Forest	89.91%	94.05%	91.23%	64.79%	72.31%	61.42%
XGBoost	76.06%	81.24%	72.19%	63.85%	70.59%	60.23%
LightGBM	71.60%	77.34%	68.52%	62.68%	69.21%	59.07%

Baseado nessa métrica, demos prosseguimento a análise sem o "*Grid Search*" e obtivemos os seguintes resultados:

A tabela 2 abaixo compara a acurácia dos três modelos de *Machine Learning* usados neste estudo no cenário de investigação de acidentes aéreos, sendo eles: *Random Forest*, *XGBoost* e *LightGBM*.

Table 2. Comparação entre os modelos

Modelo	Acurácia do Modelo
Random Forest	89.91%
XGBoost	76.06%
LightGBM	71.60%

Os resultados mostram que o modelo *Random Forest* obteve a maior acurácia, alcançando 89.91%. Isso indica que, dentre os modelos comparados, o *Random Forest* apresentou a maior capacidade de fazer previsões corretas em relação aos dados de aci-

dentos aéreos. Em seguida, o modelo *XGBoost* obteve uma acurácia de 76.06%, seguido pelo modelo *LightGBM* com uma acurácia de 71.60%.

O modelo *LightGBM*, embora tenha alcançado uma acurácia aceitável de 72%, apresentou métricas de precisão, *recall* e *F1-score* variáveis entre as diferentes classes, indicando uma performance inconsistente vide tabela 3.

Table 3. LightGBM Classifier

	Precision	Recall	F1-score	Support
Acidente	0.68	0.73	0.70	146
Incidente	0.87	0.81	0.84	139
Incidente Grave	0.61	0.62	0.61	141
Acuraccy			0.72	426
Macro avg	0.72	0.72	0.72	426
Weighted avg	0.72	0.72	0.72	426

Por outro lado, o modelo *XGBoost* mostrou resultados mais consistentes, com métricas ligeiramente superiores em comparação ao *LightGBM*, resultando em uma acurácia de 76%, vide tabela 4.

Table 4. XGBoost Classifier

	Precision	Recall	F1-score	Support
Acidente	0.73	0.80	0.76	146
Incidente	0.89	0.83	0.86	139
Incidente Grave	0.67	0.65	0.66	141
Acuraccy			0.76	426
Macro avg	0.76	0.76	0.76	426
Weighted avg	0.76	0.76	0.76	426

No entanto, o destaque foi para o modelo *Random Forest*, que apresentou o melhor desempenho entre os três. Com uma acurácia de 90%, o *Random Forest* demonstrou uma alta precisão na previsão de acidentes aéreos. Suas métricas de precisão, *recall* e *F1-score* também foram consistentemente altas, indicando uma performance robusta e confiável, vide tabela 5.

Table 5. Random Forest

	Precision	Recall	F1-score	Support
Acidente	0.94	0.90	0.92	146
Incidente	0.91	0.94	0.92	139
Incidente Grave	0.85	0.87	0.86	141
Acuraccy			0.90	426
Macro avg	0.90	0.90	0.90	426
Weighted avg	0.90	0.90	0.90	426

Esses resultados destacam a importância de escolher o modelo aprendizado de máquina mais adequado para a análise de dados de acidentes aéreos, considerando não apenas a acurácia geral, mas também a consistência e precisão das previsões em diferentes classes de incidentes.

A matriz de confusão é aplicada com a função de avaliar os resultados do modelo quando ao índice de acertos e erros de previsão. Ela apresenta graficamente em suas linhas as classes reais dos dados e nas colunas as classes previstas pelo modelo. Baseado no cruzamento de dados fornecidos pelo modelo *Random Forest Classifier*, como podemos aferir na (Figura 2) os seguintes resultados:

- **146 Acidentes reais classificados como acidentes:** Esse número indica a quantidade de eventos reais classificados como acidentes.
- **139 Incidentes reais classificados como incidentes:** Este valor representa a quantidade de incidentes reais que foram corretamente classificados como incidentes.
- **141 Incidentes graves reais classificados como incidentes graves:** Este número representa a quantidade de incidentes graves reais que foram corretamente classificados como incidentes graves.
- **13 Acidentes reais classificados como incidentes:** Esse valor indica o número de acidentes reais que foram incorretamente classificados como incidentes.
- **0 Acidentes reais classificados como incidentes graves:** Este número representa a quantidade de acidentes reais que foram incorretamente classificados como incidentes graves.
- **1 Incidente real classificado como acidente:** Esse valor indica o número de incidentes reais que foram incorretamente classificados como acidentes.
- **9 Incidentes graves reais classificados como acidentes:** Este número representa a quantidade de incidentes graves reais que foram incorretamente classificados como acidentes.

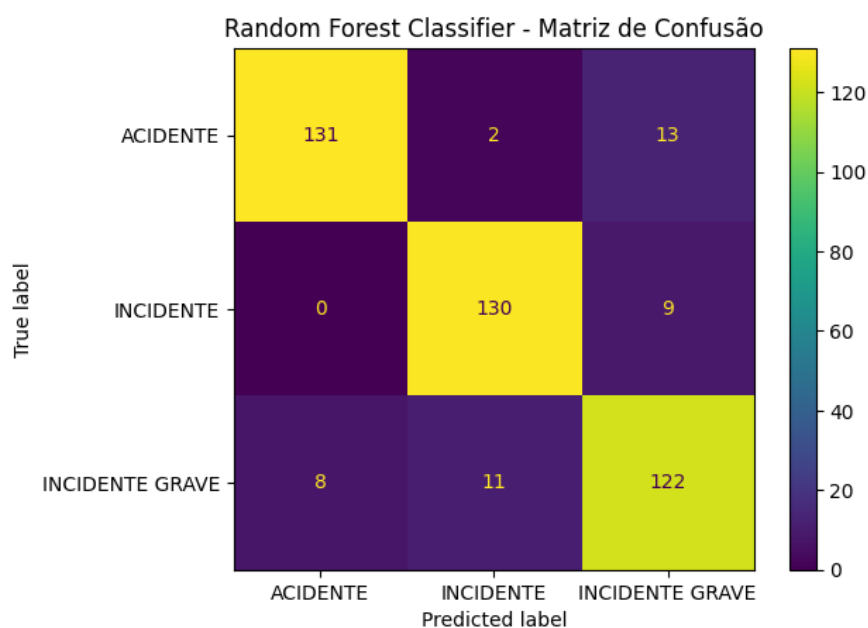


Figure 2. Matriz de Confusão do Random Forest

Os resultados destacam a viabilidade da aplicação de algoritmos de *Machine Learning* para a classificação de ocorrências aeroespaciais no contexto brasileiro. O *Random Forest Classifier* foi o modelo mais acertivo em relação a base de dados trabalhada em seu treinamento. Tornando o modelo viável como mais uma ferramenta de análise disponível aos órgãos governamentais que cuidam da segurança na aviação.

5. Conclusão

A análise comparativa dos modelos de classificação - *LightGBM*, *XGBoost* e *Random Forest* - no contexto da investigação de acidentes aéreos revelou *insights* valiosos para aprimorar a segurança na aviação. Ao examinar os resultados detalhadamente, é possível afirmar com confiança que a escolha do modelo de *Machine Learning* adequado desempenha um papel crucial na identificação de fatores de risco, previsão de acidentes e, em última instância, na prevenção de futuros incidentes.

Inicialmente, a base de dados dispunha de um predomínio dos casos de "Incidentes" em relação as demais possibilidades de classificação, em razão do alto nível de segurança existente na aviação, "Acidentes" e "Incidentes Graves" tornam-se menores. Sendo necessário o emprego de uma técnica de nivelamento dos dados para não gerar enviezamento dos resultados do treino dos modelos.

No comparativo entre modelos, observamos que o modelo *LightGBM*, embora tenha alcançado uma acurácia razoável de 72%, apresentou um desempenho variável em suas métricas de precisão, *recall* e *F1-score*, sugerindo uma certa instabilidade na sua capacidade de previsão. Embora tenha fornecido *insights* úteis, sua eficácia pode ser questionada devido à inconsistência nos resultados.

Por outro lado, o modelo *XGBoost* mostrou uma performance mais sólida e consistente, com métricas superiores em comparação ao *LightGBM*. Com uma acurácia de 76%, o *XGBoost* demonstrou ser uma opção viável para a análise de dados de acidentes aéreos. No entanto, ainda havia espaço para melhorias, especialmente considerando a necessidade de maior precisão e estabilidade nas previsões.

O destaque, sem dúvida, foi o modelo *Random Forest*. Com uma acurácia impressionante de 90% e métricas de precisão, *recall* e *F1-score* consistentemente altas, o *Random Forest* se destacou como o modelo mais eficaz na previsão de acidentes aéreos. Sua capacidade de fornecer previsões precisas e confiáveis em diferentes cenários e classes de incidentes o torna uma ferramenta indispensável para a segurança na aviação.

Portanto, com base em nossas análises, concluímos que a escolha do modelo *Random Forest* é altamente recomendada para a análise de dados de acidentes aéreos, devido à sua precisão, estabilidade e eficácia comprovadas. No entanto, é importante ressaltar que o progresso contínuo na pesquisa e desenvolvimento de técnicas de *Machine Learning* é essencial para aprimorar ainda mais a segurança na aviação e prevenir futuros incidentes.

5.1. Trabalhos Futuros

Os resultados desta pesquisa proporcionaram uma visão valiosa sobre a eficácia de diferentes modelos de *Machine Learning* na análise de dados de acidentes aéreos. No entanto, há diversas áreas para exploração em futuros estudos, visando aprimorar ainda mais nossa compreensão e prevenção de incidentes na aviação. Uma direção promissora é a

investigação de outras técnicas de *Machine Learning* além das examinadas neste estudo, como redes neurais profundas, *SVM* e *Naive Bayes*, a fim de avaliar seu desempenho em conjunto ou em comparação com os modelos analisados aqui.

Além disso, a incorporação de dados adicionais, como registros meteorológicos e de manutenção de aeronaves, pode enriquecer nossa análise e fornecer uma compreensão mais holística dos fatores que contribuem para os acidentes aéreos.

5.2. Considerações

Por fim, o desenvolvimento de sistemas de alerta precoce, baseados nos *insights* obtidos, pode ser uma estratégia eficaz para identificar padrões de comportamento de risco e implementar medidas preventivas de forma proativa, contribuindo assim para a segurança e confiabilidade da aviação. Essas são apenas algumas das possibilidades para trabalhos futuros que podem ampliar nosso conhecimento e melhorar a segurança no setor da aviação.

References

- [1] G. D. Guimarães *et al.*, “Os impactos econômicos do transporte aéreo regular para o setor do turismo nacional no período de 2019 a 2021,” 2023.
- [2] R. M. A. Valdés and F. G. Comendador, “Learning from accidents: Updates of the european regulation on the investigation and prevention of accidents and incidents in civil aviation,” *Transport policy*, vol. 18, no. 6, pp. 786–799, 2011.
- [3] M. A. F. da Costa and M. d. F. B. da Costa, “Acidente e incidente: Desmistificando o descompasso conceitual existente,” *Revista Ibero-Americana de Humanidades, Ciências e Educação*, vol. 8, no. 5, pp. 31–51, 2022.
- [4] S. K. Ekman and M. Debacker, “Survivability of occupants in commercial passenger aircraft accidents,” *Safety science*, vol. 104, pp. 91–98, 2018.
- [5] Y. Li *et al.*, “Analysis and forecast of global civil aviation accidents for the period 1942-2016,” *Mathematical Problems in Engineering*, vol. 2019, 2019.
- [6] O. Čokorilo, M. De Luca, and G. Dell’Acqua, “Aircraft safety analysis using clustering algorithms,” *Journal of Risk Research*, vol. 17, no. 10, pp. 1325–1340, 2014.
- [7] A. V. Shvetsov, “Analysis of accidents resulting from the interaction of air and ground vehicles at airports,” *Transportation research procedia*, vol. 59, pp. 21–28, 2021.
- [8] F. Kamalov, I. Gurrib, and K. Rajab, “Financial forecasting with machine learning: price vs return,” *Kamalov, F., Gurrib, I. & Rajab, K.(2021). Financial Forecasting with Machine Learning: Price Vs Return. Journal of Computer Science*, vol. 17, no. 3, pp. 251–264, 2021.
- [9] D. Truong and W. Choi, “Using machine learning algorithms to predict the risk of small unmanned aircraft system violations in the national airspace system,” *Journal of Air Transport Management*, vol. 86, p. 101822, 2020.
- [10] A. P. Gonçalves Filho, C. A. Souza, E. L. B. Siqueira, M. A. Souza, and T. P. Vasconcelos, “An analysis of helicopter accident reports in brazil from a human factors perspective,” *Reliability Engineering & System Safety*, vol. 183, pp. 39–46, 2019.

- [11] C. Morais, K. L. Yung, K. Johnson, R. Moura, M. Beer, and E. Patelli, "Identification of human errors and influencing factors: A machine learning approach," *Safety science*, vol. 146, p. 105528, 2022.
- [12] G. d. O. Torres, M. X. Guterres, and V. R. R. Celestino, "Legal actions in brazilian air transport: A machine learning and multinomial logistic regression analysis," *Frontiers in Future Transportation*, vol. 4, p. 1070533, 2023.
- [13] S. Sampath, "Safety and service difficulty reporting," *Aging Engines, Avionics, Subsystems and Helicopters*, vol. 7, p. 1, 2000.
- [14] P. Szczepaniak, G. Jastrzebski, K. Sibilski, and A. Bartosiewicz, "The study of aircraft accidents causes by computer simulations," *Aerospace*, vol. 7, no. 4, p. 41, 2020.
- [15] C. Belcastro and J. Foster, "Aircraft loss-of-control accident analysis," in *AIAA Guidance, Navigation, and Control Conference*, 2010, p. 8004.
- [16] R. K. Jenamani and A. Kumar, "Bad weather and aircraft accidents—global vis-à-vis indian scenario," *Current Science*, pp. 316–325, 2013.
- [17] S. Dekker, P. Cilliers, and J.-H. Hofmeyr, "The complexity of failure: Implications of complexity theory for safety investigations," *Safety science*, vol. 49, no. 6, pp. 939–945, 2011.
- [18] C. Maione, F. Barbosa Jr, and R. M. Barbosa, "Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review," *Computers and Electronics in Agriculture*, vol. 157, pp. 436–446, 2019.
- [19] J. Zhang, "Multivariate analysis and machine learning in cerebral palsy research," *Frontiers in neurology*, vol. 8, p. 292173, 2017.
- [20] V. F. C. d. França, L. D. d. Oliveira, and H. A. d. Andrade, "Price forecasting of a small pelagic species in a south american supply center: A machine learning approach," *Marine Resource Economics*, vol. 39, no. 2, pp. 145–162, 2024.
- [21] R. Passarella, S. Nurmaini, M. N. Rachmatullah, H. Veny, and F. N. N. Hafidzoh, "Development of a machine learning model for predicting abnormalities of commercial airplanes," *Data Science and Management*, 2024.
- [22] S. Mizuno, H. Ohba, and K. Ito, "Machine learning-based turbulence-risk prediction method for the safe operation of aircrafts," *Journal of Big Data*, vol. 9, no. 1, p. 29, 2022.
- [23] M. D. Dangut, Z. Skaf, and I. K. Jennions, "An integrated machine learning model for aircraft components rare failure prognostics with log-based dataset," *ISA transactions*, vol. 113, pp. 127–139, 2021.
- [24] T. M. Oshiro, "Uma abordagem para a construção de uma única árvore a partir de uma random forest para classificação de bases de expressão gênica," Ph.D. dissertation, Universidade de São Paulo, 2013.
- [25] M. R. Silveira, "Detecção de domínios maliciosos por meio de dns passivo utilizando xgboost," 2021.

- [26] C. Chen, Q. Zhang, Q. Ma, and B. Yu, "Lightgbm-ppi: Predicting protein-protein interactions through lightgbm with multi-information fusion," *Chemometrics and Intelligent Laboratory Systems*, vol. 191, pp. 54–64, 2019.
- [27] J. F. G. Monico, A. P. Dal Poz, M. Galo, M. C. Dos Santos, and L. C. De Oliveira, "Acurácia e precisão: revendo os conceitos de forma acurada," *Boletim de Ciências Geodésicas*, vol. 15, no. 3, pp. 469–483, 2009.
- [28] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020, pp. 79–91.
- [29] R. A. Burnett and D. Si, "Prediction of injuries and fatalities in aviation accidents through machine learning," in *Proceedings of the International Conference on Compute and Data Analysis*, 2017, pp. 60–68.
- [30] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction," *Decision Support Systems*, vol. 116, pp. 48–63, 2019.
- [31] S. C. d. SILVA *et al.*, "Elaboração de um modelo de tomada de decisões contábeis em uma instituição federal de ensino superior através da perspectiva do crisp-dm," 2023.
- [32] U. Shafique and H. Qaiser, "A comparative study of data mining process models (kdd, crisp-dm and semma)," *International Journal of Innovation and Scientific Research*, vol. 12, no. 1, pp. 217–222, 2014.
- [33] E. M. SILVA, T. G. VIOTT, C. C. ALVES, and B. A. PEREIRA, "Metodologia para desenvolvimento de sistemas de suporte a decisão crisp-dm utilizando a notação uml uma abordagem aplicada à gerência de atendimento hospitalar."
- [34] J. L. C. Ramos, R. L. Rodrigues, J. C. S. Silva, and P. L. S. de Oliveira, "Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais," in *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. SBC, 2020, pp. 1092–1101.
- [35] S. Okada, M. Ohzeki, and S. Taguchi, "Efficient partition of integer optimization problems with one-hot encoding," *Scientific reports*, vol. 9, no. 1, p. 13036, 2019.
- [36] R. Guedrez, O. Dugeon, S. Lahoud, and G. Texier, "Label encoding algorithm for mpl segment routing," in *2016 IEEE 15th International Symposium on Network Computing and Applications (NCA)*. IEEE, 2016, pp. 113–117.
- [37] P. Lerman, "Fitting segmented regression models by grid search," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 29, no. 1, pp. 77–84, 1980.