Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística
Disciplina de Itrodução à Ciência de Dados

CLASSIFICAÇÃO DE OCORRÊNCIAS AERONÁUTICAS: UMA ABORDAGEM COM ÁRVORES DE DECISÃO E FLORESTAS ALEATÓRIAS

Ana V. R. Araújo e Hélio V. R. Neto

1 Introdução

A análise de dados é crucial para entender fenômenos complexos e tomar decisões informadas. Na aviação, a segurança operacional é essencial, e a investigação de ocorrências aeronáuticas ajuda a identificar padrões e fatores de risco. Este estudo utiliza técnicas de aprendizagem de máquina para analisar dados detalhados sobre ocorrências aeronáuticas, visando identificar características associadas à sua classificação. O conjunto de dados inclui informações técnicas e operacionais, como tipo de motor, fase do voo e segmento de aviação, com foco na variável *ocorrencia_classificacao*, que categoriza a gravidade das ocorrências. A aplicação de modelos preditivos busca não apenas entender os fatores de risco, mas também auxiliar na identificação precoce de problemas e na implementação de medidas preventivas, contribuindo para a melhoria da segurança na aviação.

2 Metodologia

Para este estudo, escolhemos dois métodos: Árvore de Decisão e Floresta Aleatória, devido à sua capacidade de lidar com dados estruturados, interpretabilidade e eficácia em problemas de classificação. A metodologia adotada foi dividida em etapas, conforme descrito a seguir.

2.1 Pré-processamento dos Dados

O conjunto de dados utilizado contém informações detalhadas sobre ocorrências aeronáuticas. A variável resposta é *ocorrencia_classificacao*, que categoriza as ocorrências de acordo com sua gravidade. Para garantir a qualidade da análise, foram realizadas algumas etapas de pré-processamento. Estas etapas foram:

- Tratamento de Valores Ausentes: Valores faltantes foram identificados e excluidos para manter a integridades dos dados.
- Tratamento de duplicatas: Observações repetidas foram identificadas e excluidas.
- padronização dos dados: Valores anômalos foram identificados e excluídos.
- Divisão dos Dados: O conjunto de dados foi dividido em treino (80%) e teste (20%) para garantir a validação adequada dos modelos.

2.2 Aplicação da Árvore de Decisão

A Árvore de Decisão foi escolhida por sua simplicidade e facilidade de interpretação. O modelo foi treinado utilizando a biblioteca Pycaret. A princípio, uma árvore foi treinada e avaliada, depois o modelo foi submetido a um ajuste fino. O desempenho dos modelos foram avaliado no conjunto de teste utilizando métricas como Acurácia, Precisão, Recall, F1-score e Matriz de Confusão. A importância de cada variável foram analisadas para compreender as variáveis que mais influenciam na classificação.

2.3 Aplicação da Floresta Aleatória (Random Forest)

A Floresta Aleatória foi utilizada para melhorar a precisão e robustez do modelo, combinando múltiplas árvores de decisão. O modelo foi treinado com o conjunto de treino, utilizando um número definido de árvores n_estimators e outros parâmetros como profundidade máxima max_depth. O desempenho foi avaliado no conjunto de teste, utilizando as mesmas métricas aplicadas à Árvore de Decisão. A importância de cada variável foi calculada para identificar os principais fatores que influenciam a classificação das ocorrências.

3 Desenvolvimento

3.1 Árvore de Decisão

O modelo de Árvore de Decisão foi treinado com profundidade máxima de 3 (max_depth=3) e mínimo de 2 amostras para dividir um nó (min_samples_split=2), sem validação cruzada (cross_validation=False). O conjunto de dados foi dividido em 80% para treino e 20% para teste. No conjunto de teste, o modelo alcançou acurácia de 83,19%, AUC de 0,8935, recall de 83,19%, precisão de 79,46%, F1-Score de 79,79%, Kappa de 0,6900 e MCC de 0,7041, indicando boa capacidade de distinguir entre as classes (incidentes, acidentes e incidentes graves), com equilíbrio entre precisão e recall. A curva de aprendizado mostra convergência entre as pontuações de treinamento (0,86) e validação cruzada (0,85), sugerindo ausência de overfitting e que aumentar o tamanho do conjunto de treinamento não traria ganhos significativos. A limitação da profundidade da árvore manteve o modelo interpretável e eficiente. A variável mais importantes para o modelo é a aeronave_nivel_dano que tem importancia acima de 0,6.

A matriz de confusão e o relatório de classificação confirmam o desempenho sólido, com acurácia de 83,19% e Kappa de 0,69. Para a classe ACIDENTE, o modelo obteve alta precisão (0,88) e recall (0,88). Para INCIDENTE, o recall foi excelente (0,97), mas a precisão foi menor (0,83), indicando erros na classificação. A classe INCIDENTE GRAVE teve desempenho inferior, com precisão de 0,43 e recall de 0,10, mostrando dificuldade em classificar corretamente esses casos. O F1-Score médio ponderado foi de 0,83, destacando bom desempenho nas classes majoritárias, mas a necessidade de melhorias na classificação de INCIDENTE GRAVE.

Após o *model tuning*, o modelo foi configurado com criterion = 'entropy', min_samples_split = 10, min_samples_leaf = 3 e max_features = 1. A curva de aprendizado mostrou convergência entre treinamento (0,86) e validação cruzada (0,85), confirmando ausência de overfitting. O ajuste melhorou ligeiramente o desempenho, com acurácia de 84,28%, AUC de 0,8548, recall de 84,28%, precisão de 73,83%, F1-Score de 78,69%, Kappa de 0,7032 e MCC de 0,7250, indicando que o modelo generaliza bem

para diferentes subconjuntos dos dados. A variável mais importantes para o modelo é a aeronave_nivel_dano que tem importancia acima de 0,5.

A matriz de confusão e o relatório de classificação após o ajuste mostram acurácia de 83,41% e Kappa de 0,69. Para ACIDENTE, o modelo manteve alta precisão (0,85) e recall (0,92). Para INCIDENTE, o recall foi excelente (0,97), mas a precisão menor (0,83) sugere erros de classificação. Para INCIDENTE GRAVE, o modelo não fez previsões corretas (precisão e recall de 0,00). O F1-Score médio ponderado foi de 0,78, reforçando o bom desempenho nas classes majoritárias, mas a necessidade de melhorias na classificação de INCIDENTE GRAVE. A matriz de confusão confirma que a maioria dos erros ocorreu na confusão entre INCIDENTE e INCIDENTE GRAVE.

O modelo de Árvore de Decisão antes do *model tuning* apresentou desempenho sólido em métricas gerais (acurácia de 83,19%, AUC de 0,8935), mas teve dificuldades com a classe INCIDENTE GRAVE, com precisão de 0,43 e recall de 0,10. Após o ajuste, o modelo melhorou ligeiramente (acurácia de 84,28%, AUC de 0,8548), mas falhou completamente na classificação da classe INCIDENTE GRAVE, com precisão e recall de 0,00. Embora o modelo ajustado tenha mostrado uma ligeira melhora em algumas métricas, o erro total em uma classe é um problema crítico, especialmente em um contexto prático. Portanto, apesar das melhorias, o modelo ainda não é confiável para classificação completa das ocorrências aeronáuticas, necessitando de ajustes adicionais. Desta forma, o modelo da árvore de decisão antes do ajuste parece ser mais adequado para a classificação neste contexto.

3.1 Floresta aleatória

Os resultados indicam um desempenho sólido e equilibrado. O modelo alcançou uma acurácia e Recall de 83,82%, demonstrando capacidade para classificar corretamente a maioria das ocorrências e identificar bem os casos positivos. A Precisão de 79,64% e o F1-Score de 80,33% mostram um bom equilíbrio entre evitar falsos positivos e capturar ocorrências corretas. Além disso, o Kappa de 70,06% e o MCC de 71,32% confirmam uma concordância substancial entre as previsões e os valores reais.

As variáveis mais importantes para o modelo foram: nível do dano, tipo da ocorrência,

modelo da aeronave, peso máximo de decolagem e quantidade de motores. Essas variáveis tiveram maior influência na classificação das ocorrências aeronáuticas, indicando que estão diretamente relacionadas à gravidade e ao tipo de evento registrado.

A curva de aprendizado mostra que a curva do conjunto de treino está acima da curva do conjunto de testes, o que é um comportamento esperado, isso indica que o modelo tem um desempenho melhor nos dados de treinamento em comparação com os dados de teste.

O modelo ajustado apresentou melhorias na Acurácia, Recall, Kappa e MCC, indicando uma capacidade superior de classificação e generalização. No entanto, houve uma queda na Precisão e no F1-Score. As variaveis mais importantes do modelo após o ajuste permaneceram as mesmas do modelo normal.

No início, a curva do conjunto de treino começou abaixo da curva do conjunto de teste, mas, a partir da metade do processo, a curva de treino ultrapassou a curva de teste. Isso pode indicar que conforme o modelo foi exposto a mais dados, o modelo começou a aprender os padrões específicos do conjunto de treino.

O modelo performa bem para as classes ACIDENTE e INCIDENTE, com altos valores de Recall e F1-Score, mas falha completamente na classificação de INCIDENTE GRAVE. No geral, o modelo é eficaz para a maioria das ocorrências, mas precisa de ajustes para lidar com casos graves.

4 Conclusão

Tanto a Árvore de Decisão quanto a Floresta Aleatória demonstraram desempenho sólido na classificação de ocorrências aeronáuticas, com métricas como acurácia, recall e F1-Score indicando boa capacidade de generalização. Após o ajuste, ambos os modelos apresentaram melhorias em algumas métricas, como recall e capacidade de generalização. No entanto, um problema crítico persiste: tanto a Árvore de Decisão quanto a Floresta Aleatória falharam completamente na classificação da classe INCIDENTE GRAVE. Isso indica uma limitação grave, especialmente em um contexto prático, onde a identificação correta de ocorrências graves é essencial. As variáveis mais importantes, como nível do dano e tipo de

operação, foram consistentes em todos os modelos, destacando sua relevância para a classificação. No geral, a Floresta Aleatória ajustada se mostrou superior em termos de consistência e capacidade de generalização, mas ainda requer ajustes significativos, como técnicas de balanceamento de classes ou coleta de mais dados, para melhorar a classificação de INCIDENTE GRAVE e garantir um desempenho confiável em todas as classes.

5 Participação dos membros do grupo

Hélio V. R. Neto foi responsável pela redação da introdução, parte da metodologia e pelo treinamento e ajuste dos modelos de Floresta Aleatória. Além disso, contribuiu com a análise dos resultados e a interpretação das métricas desses modelos.

Ana V. R. Araújo realizou o processamento dos dados, garantindo a preparação e limpeza do conjunto de dados para análise. Também escreveu uma menor parte da metodologia e foi responsável pelo treinamento e ajuste dos modelos de Árvore de Decisão, além de contribuir com a análise dos resultados e a interpretação das métricas desses modelos.

Referências

Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). Classification and Regression Trees (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Morettin, P.A., & Singer, J.M. (2022). Estatística e ciência de dados.

Izbicki, R., & Santos, T.M. (2020). Aprendizado de máquina: uma abordagem estatística.