Consultoria A3Data

Teste Técnico Cientista de Dados - A3Data

Introdução

Atualmente a aviação é algo que faz parte do nosso dia a dia. Visitar amigos e parentes do outro lado do país, ou até mesmo do outro lado do planeta. Fazer compras pela internet e em poucos dias recebê-las em casa através de um voo comercial. Os voos internacionais promoveram a disseminação de um vírus, mas agora seguem com a distribuição das vacinas que o combatem.

A aviação é uma atividade que envolve ciência, tecnologia, transporte, economia, aeronáutica, militarismo, lazer e tantas outras.

Mas, por mais que hoje seja uma atividade que de certa forma é banal, (em 2019 o recorde de voos comerciais foi batido com 230.000 mil voos em 24 horas) sempre que ocorre um incidente, ou até uma tragédia maior, se gera muita dúvida e apreensão quanto a esta atividade.

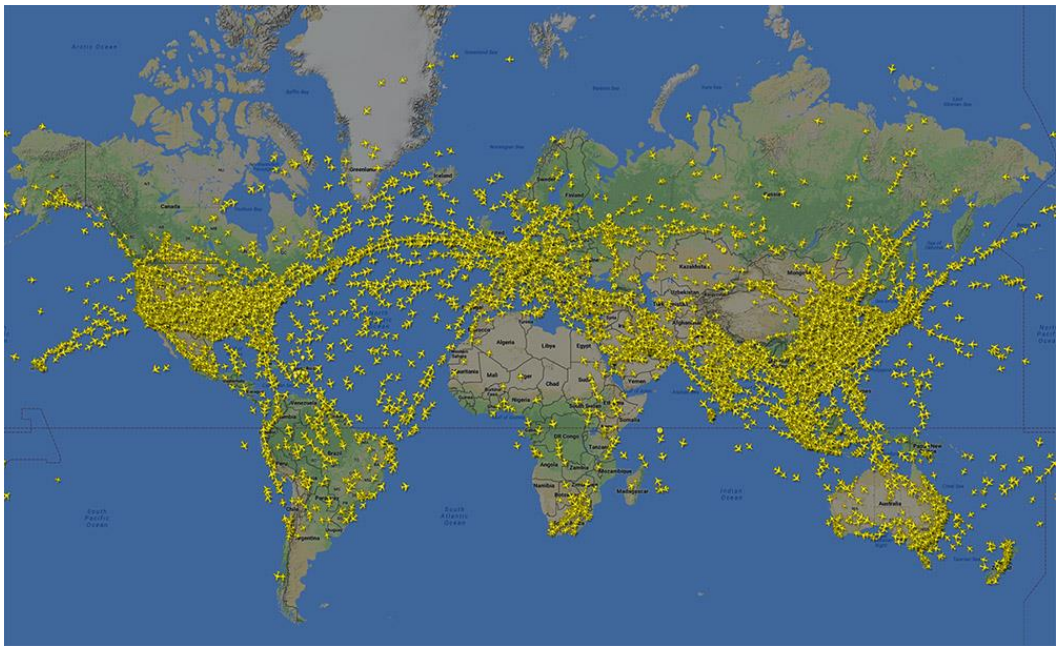


Figura 1: Tráfego aéreo no dia 25 de julho de 2019

Objetivo

Este estudo tem como objetivo tentar prever se há uma relação entre incidentes e acidentes aéreos com os tipos de veículo e seus motores.

Acidente: é qualquer evento súbito e não planejado, que cause ou possa vir a causar ferimento a pessoas ou danos a edifícios, instalações, materiais ou ao meio ambiente.

Incidente: No incidente o fato inesperado e potencialmente perigoso acontece, mas graças a alguma circunstância favorável ele não causa danos a ninguém.

Base Utilizada

Para o estudo será utilizada a base de dados de ocorrências aeronáuticas que é gerenciada pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA). Constam nesta base de dados as ocorrências aeronáuticas notificadas ao CENIPA nos últimos 10 anos e que ocorreram em solo brasileiro.

URL: https://dados.gov.br/dataset/ocorrencias-aeronauticas-da-aviacao-civil-brasileira

BASES DE DADOS UTILIZADAS:

- OCORRÊNCIA.csv - Informações sobre as ocorrências.
- AERONAVE.csv - Informações sobre as aeronaves envolvidas nas ocorrências.

Análise inicial da base

O arquivo fornecido pelo CENIPA é extenso e possui muitas informações relevantes como pode ser visto:

```
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   codigo_ocorrencia               5752 non-null   int64
 1   codigo_ocorrencia1              5752 non-null   int64
 2   codigo_ocorrencia2              5752 non-null   int64
 3   codigo_ocorrencia3              5752 non-null   int64
 4   codigo_ocorrencia4              5752 non-null   int64
 5   ocorrencia_classificacao        5752 non-null   object
 6   ocorrencia_latitude             4187 non-null   object
 7   ocorrencia_longitude            4187 non-null   object
 8   ocorrencia_cidade               5752 non-null   object
 9   ocorrencia_uf                   5752 non-null   object
 10  ocorrencia_pais                 5752 non-null   object
 11  ocorrencia_aerodromo            5752 non-null   object
 12  ocorrencia_dia                  5752 non-null   object
 13  ocorrencia_hora                 5751 non-null   object
 14  investigacao_aeronave_liberada  5411 non-null   object
 15  investigacao_status             5412 non-null   object
 16  divulgacao_relatorio_numero     4887 non-null   object
 17  divulgacao_relatorio_publicado  5752 non-null   object
 18  divulgacao_dia_publicacao       1494 non-null   object
 19  total_recomendacoes             5752 non-null   int64
 20  total_aeronaves_envolvidas      5752 non-null   int64
 21  ocorrencia_saida_pista          5752 non-null   object
dtypes: int64(7), object(15)
memory usage: 988.8+ KB
```

É possível ver que a maioria das colunas tem o tipo ‘object’ o que quer dizer que para essa base de dados temos bastante variáveis categóricas.

Tendo em mente o objetivo, pode-se reduzir bastante os atributos da base inicial para facilitar no cálculo e nas análises. Após fazer uma série de transformações, a base agora te apenas 3 variáveis:

codigo_ocorrencia	ocorrencia_classificacao	ocorrencia_saida_pista
39115	ACIDENTE	NÃO
39155	INCIDENTE	NÃO
39156	INCIDENTE GRAVE	NÃO
39158	INCIDENTE	NÃO
39176	INCIDENTE	NÃO
...	...	...
79802	INCIDENTE	NÃO
79804	INCIDENTE	NÃO
79824	ACIDENTE	NÃO
79844	INCIDENTE	NÃO
79874	ACIDENTE	NÃO

Código ocorrência: traz o código da ocorrência. Iremos utilizá-lo em breve para agregar outras informações à base.

Ocorrência classificação: classifica se a ocorrência registrada foi um incidente, um incidente grave ou se foi de fato um acidente.

Ocorrência saída pista: informa com SIM ou NÃO se a ocorrência registrada aconteceu durante a decolagem

Enriquecimento da base

Ao carregar a nova base com os tipos de aeronaves temos um novo dataframe:

codigo_ocorrencia	ocorrencia_classificacao	ocorrencia_saida_pista	aeronave_tipo_veiculo	aeronave_motor_tipo
39115	ACIDENTE	NÃO	AVIÃO	PISTÃO
39155	INCIDENTE	NÃO	AVIÃO	TURBOÉLICE
39156	INCIDENTE GRAVE	NÃO	AVIÃO	TURBOÉLICE
39158	INCIDENTE	NÃO	AVIÃO	JATO
39176	INCIDENTE	NÃO	AVIÃO	JATO

Tratando a base enriquecida

Foram feitos 3 novos tratamentos para a base:

1. Tratar os valores que estavam faltando e os rotular como "NÃO INFORMADO"
2. Criar coluna “acidente” a partir da ocorrencia_classificacao com os rótulos “SIM” ou “NÃO” para calcular a predição
3. Limpar o dataframe das colunas que não serão mais utilizadas

	ocorrencia_saida_pista	aeronave_tipo_veiculo	aeronave_motor_tipo	acidente
0	NÃO	AVIÃO	PISTÃO	SIM
1	NÃO	AVIÃO	TURBOÉLICE	NÃO
2	NÃO	AVIÃO	TURBOÉLICE	NÃO
3	NÃO	AVIÃO	JATO	NÃO
4	NÃO	AVIÃO	JATO	NÃO
...	...	...	...	...
5747	NÃO	AVIÃO	JATO	NÃO
5748	NÃO	NÃO INFORMADO	JATO	NÃO
5749	NÃO	AVIÃO	PISTÃO	SIM
5750	NÃO	HELICÓPTERO	TURBOEIXO	NÃO
5751	NÃO	NÃO INFORMADO	NÃO INFORMADO	SIM

Preparação

Foi feita a separação da variável independente “acidente” das demais dependentes.

Para os modelos funcionarem foi preciso transformar os dados categóricos em uma matriz de zeros e uns. O resultado foi uma matriz de 17 colunas.

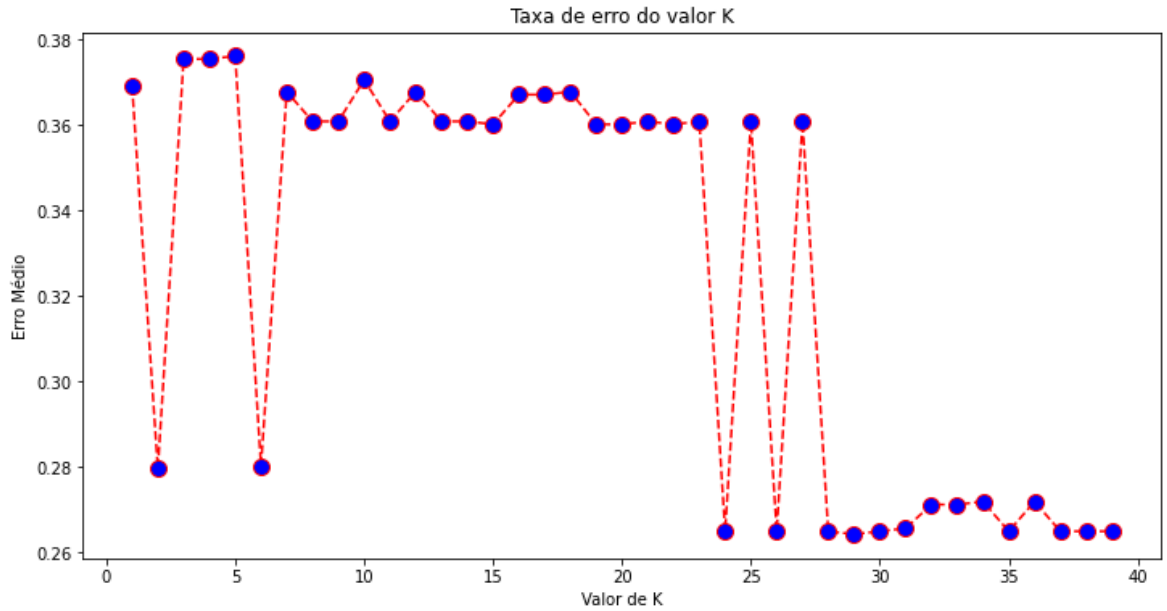
Depois foi necessário separar os dados. Quais seriam os dados de treino e quais os de teste.

Modelo KNN

Primeiro modelo utilizado para a análise, utilizando os parâmetros padrão teve como resultado uma matriz de confusão:

Predito	NÃO	SIM	All
Real			
NÃO	560	428	988
SIM	113	337	450
All	673	765	1438

Para garantir um melhor resultado, foi feita uma pequena análise alterando os valores do parâmetro K a fim de achar quais parâmetros demonstravam um menor erro médio:



Alterando o K padrão (5) para o novo encontrado (29) obteve-se

Predito	NÃO	SIM	All
Real			
NÃO	911	77	988
SIM	303	147	450
All	1214	224	1438

Pela análise das matrizes de confusão foi possível ver uma leve melhoria nos acertos, mas ainda assim, com bastante erro na predição.

Modelo Random Forest

Após configurar o modelo foram encontrados os seguintes valores:

Predito	NÃO	SIM	All
Real			
NÃO	916	72	988
SIM	306	144	450
All	1222	216	1438

	precision	recall	f1-score	support
NÃO	0.75	0.93	0.83	988
SIM	0.67	0.32	0.43	450
accuracy			0.74	1438
macro avg	0.71	0.62	0.63	1438
weighted avg	0.72	0.74	0.70	1438

É possível ver uma melhora na predição, mas ainda podem ser feitos mais estudos para melhorar o modelo.

Analisando a importância de cada variável:

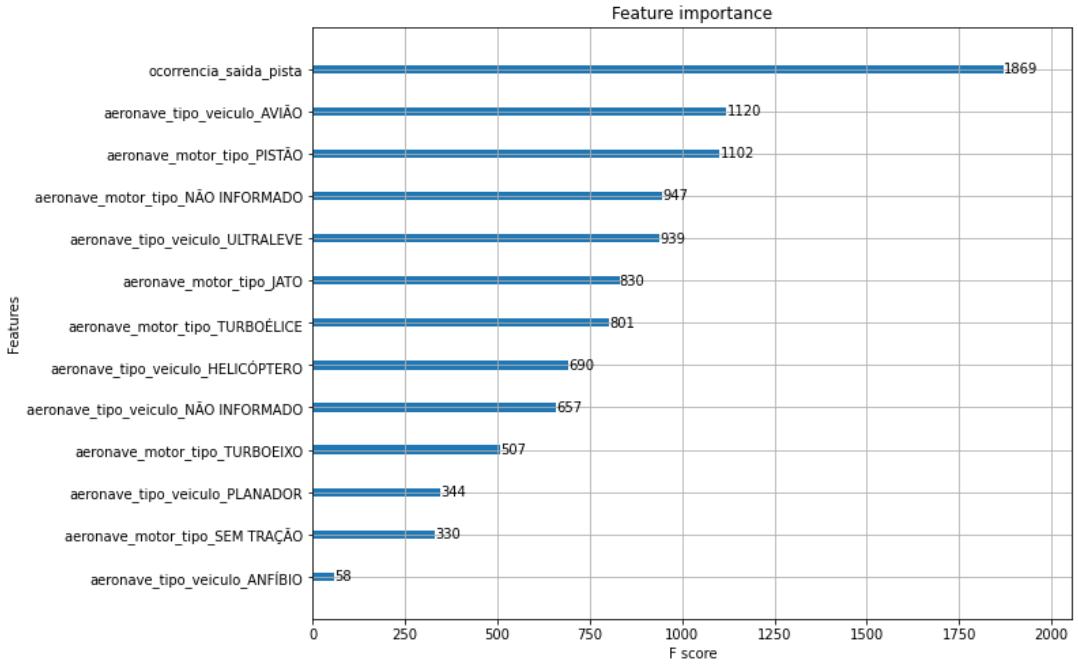
	importância
aeronave_motor_tipo_PISTÃO	0.378
aeronave_motor_tipo_JATO	0.216
aeronave_tipo_veiculo_ULTRALEVE	0.082
aeronave_tipo_veiculo_AVIÃO	0.081
ocorrencia_saida_pista	0.071
aeronave_motor_tipo_TURBOÉLICE	0.052
aeronave_motor_tipo_TURBOEIXO	0.039
aeronave_tipo_veiculo_HELICÓPTERO	0.030
aeronave_motor_tipo_NÃO INFORMADO	0.019
aeronave_tipo_veiculo_NÃO INFORMADO	0.009
aeronave_tipo_veiculo_ANFÍBIO	0.006
aeronave_tipo_veiculo_PLANADOR	0.005
aeronave_motor_tipo_SEM TRAÇÃO	0.005
aeronave_tipo_veiculo_TRIKE	0.004
aeronave_tipo_veiculo_BALÃO	0.001
aeronave_tipo_veiculo_HIDROAVIÃO	0.000
aeronave_tipo_veiculo_DIRIGÍVEL	0.000

Modelo XGBoost

O modelo XGBoost é mais robusto e consegue entregar melhores resultados pois não sofre tanta influência caso as amostras estejam desbalanceadas

Predito	NÃO	SIM	All
Real			
NÃO	912	76	988
SIM	303	147	450
All	1215	223	1438

Ao gerar um gráfico que mostra a importância das variáveis para o resultado acidente = sim ou não



Método Ensemble

Após feito a análise em cima de 3 modelos diferentes, foi utilizado o método de ensemble que irá combinar os resultados dos modelos Random Forest e XGBoost.

Fora obtidos os seguintes resultados:

Predito	NÃO	SIM	All
Real			
NÃO	912	76	988
SIM	302	148	450
All	1214	224	1438

	precision	recall	f1-score	support
NÃO	0.75	0.92	0.83	988
SIM	0.66	0.33	0.44	450
accuracy			0.74	1438
macro avg	0.71	0.63	0.63	1438
weighted avg	0.72	0.74	0.71	1438

Visualização e considerações finais:

Após modelar a predição da melhor forma possível, foi feito o cálculo de predição para se uma ocorrência pode vir a ser um acidente, e qual a sua relação com os tipos de veículo, motor e se o mesmo ocorreu ao decolar:

	Tipo de Veículo	Tipo de Motor	Acidente na saída de pista	Probabilidade de ser Acidente
0	TRIKE	NÃO INFORMADO	NÃO	0.985
1	PLANADOR	NÃO INFORMADO	NÃO	0.902
2	PLANADOR	SEM TRAÇÃO	SIM	0.901
3	AVIÃO	NÃO INFORMADO	SIM	0.895
4	BALÃO	SEM TRAÇÃO	NÃO	0.893
5	HELICÓPTERO	PISTÃO	NÃO	0.747
6	ANFÍBIO	TURBOÉLICE	NÃO	0.717
7	ULTRALEVE	PISTÃO	NÃO	0.716
8	ULTRALEVE	NÃO INFORMADO	NÃO	0.626
9	PLANADOR	SEM TRAÇÃO	NÃO	0.597
10	AVIÃO	PISTÃO	SIM	0.559
11	ULTRALEVE	PISTÃO	SIM	0.507
12	TRIKE	PISTÃO	NÃO	0.507
13	DIRIGÍVEL	PISTÃO	NÃO	0.460
14	HIDROAVIÃO	PISTÃO	NÃO	0.460
15	AVIÃO	NÃO INFORMADO	NÃO	0.439
16	AVIÃO	PISTÃO	NÃO	0.409
17	AVIÃO	TURBOÉLICE	SIM	0.352
18	AVIÃO	JATO	SIM	0.323
19	HELICÓPTERO	TURBOEIXO	NÃO	0.227
20	HELICÓPTERO	TURBOÉLICE	NÃO	0.224
21	ANFÍBIO	PISTÃO	NÃO	0.202
22	AVIÃO	TURBOÉLICE	NÃO	0.144
23	ANFÍBIO	PISTÃO	SIM	0.100
24	AVIÃO	JATO	NÃO	0.015
25	AVIÃO	TURBOEIXO	NÃO	0.003

Como a base toda se trata de ocorrências, a análise em questão não é se um veículo tem uma determinada chance de se acidentar. Mas sim, caso haja uma ocorrência, qual a chance de ser um acidente.