Teste Técnico Cientista de Dados - A3Data

Caio Gabriel Barreto Balieiro

18 de Abril de 2022

Belo Horizonte - MG

Conteúdo

Introdução

Conjunto de dados

- Este teste consiste na exploração da base dados "Ocorrências Aeronáuticas na Aviação Civil Brasileira"dos dados abertos do governo (`https://dados.gov.br/dataset/ocorrencias-aeronauticas-da-av`
- Neste sentido, foi decido trabalhar com perguntas de investigação das ocorrências e duas aplicações de modelos via modelo de previsão de séries temporais e modelos de classificação de Machine Learning.
- Além disso, os modelos citados acima foram colocados em produção utilizando ferramenta de ShinyDeashboard para que as respostas sejam expostas de maneira mais dinâmicas.

Perguntas do projeto

- A priori os questionamentos levantados foram:
  - 1 - Quais são as classificações de ocorrência no CENIPA?
  - 2 - Quais são os Estados que possuem ocorrências? E quais são os top 10 dos Estados com maiores ocorrências de acidentes?
  - 3 - Quantas ocorrências envolvendo acidentes os relatórios foram divulgados?
  - 4 - Qual o número de aeronaves envolvidas em ocorrências no CENIPA? E quantas delas são classificadas como acidente?
  - 5 - Quais são as categorias dos operadores das aeronaves que apresentaram ocorrência?
  - 6 - Quais são os tipos de veículos que tiveram ocorrência?
  - 7 - Em quais tipos de veículo ocorreram mais acidentes?
  - 8 - Quais são os tipos de motor de aeronaves em que ocorreram mais acidentes?
  - 9 - Qual é o top 5 de tipos de ocorrências registrados no CENIPA? E quais são os top 5 de acidentes?

Conjunto de dados do CENIPA

- Não o bastante, foi apresentado duas modelagens envolvendo modelos de séries temporais e modelo de classificação de Machine Learning.
  - Para o modelo de séries temporais foi utilizado os modelos Arima para a prever o número de ocorrências classificadas por: Acidente, Incidente e Incidente Grave.
  - Para o modelo de classificação foi apresentado um modelo de aprendizagem de máquina capaz de predizer se uma dada ocorrência será Acidente, Incidente, e Incidente grave.
- A escolha desses modelos se dá ao fato de que, prever com antecedência o número de ocorrências, pode ajudar o CENIPA a tomar decisões antecipadamente.
- Além disso, classificar uma ocorrência por Acidente, Incidente ou Incidente grave, leva-se um conjunto de fatores humanos que tornam essa escolha razoável. No entanto, deixar esse processo automatizado pode economizar bastante tempo na hora de se trabalhar com tomadas de decisões.

Conjunto de dados

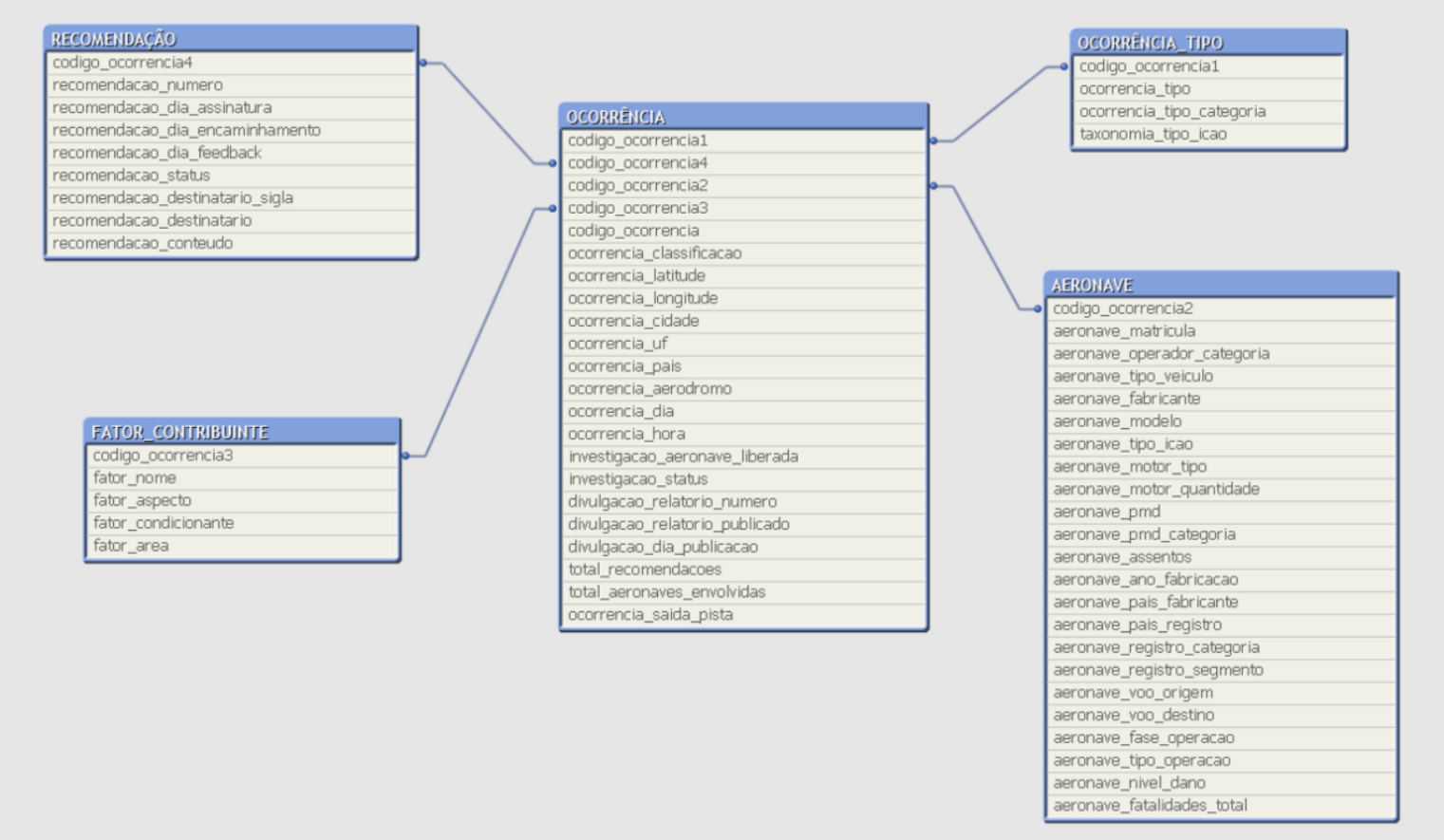

Figura 1: Conjuntos de dados do CENIPA.

- Para este projeto foi utilizado os conjuntos de dados descritos acima.

Conjunto de dados

- Para o projeto foi trabalho especialmente com as tabelas referente a ocorrência (geral), aeronave, ocorrência tipo, as demais tabelas apresentavam diversos textos, por essa razão foi decidido trabalhar apenas com essas três tabelas.
- Para responder as 9 perguntas relacionadas a investigação das causas de ocorrência, foi trabalhado com linguagem python, cujo o código desta análise é denominado Cod.Desafio.ipynb.
- A escolha das perguntas é justificada pelo interesse de avaliar e estudar um pouco mais sobre as ocorrências que são registradas no CENIPA. Mas, é importante ressaltar que essas perguntas poderiam ser estendidas para diversos contextos.

Respostas das Perguntas de investigação

Respostas das Perguntas de investigação

- 1 - Quais são as classificações de ocorrência no CENIPA?

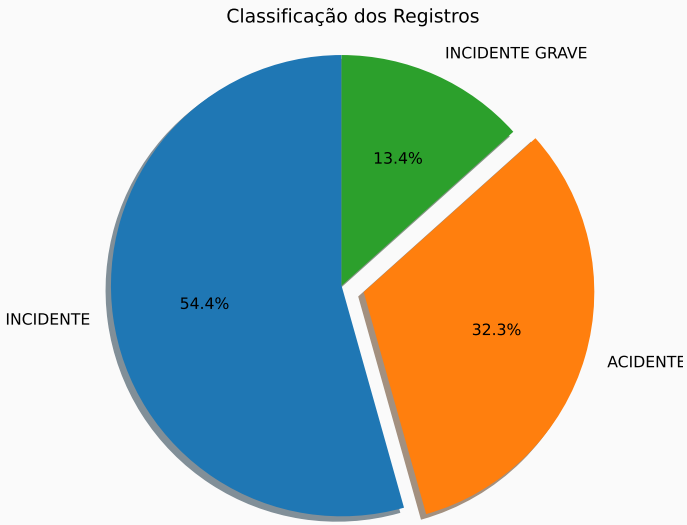

Figura 2: Gráfico de setores ocorrência por classificação.

- Note que a classificação de ocorrências é dada por Incidente, Acidente e Incidente grave, sendo que corresponde a incidentes, 32% a acidentes e apenas 13,4% para incidente grave.

Respostas das Perguntas de investigação

- 2 - Quais são os Estados que possuem ocorrências? E quais são os top 10 dos Estados com maiores ocorrências de acidentes?

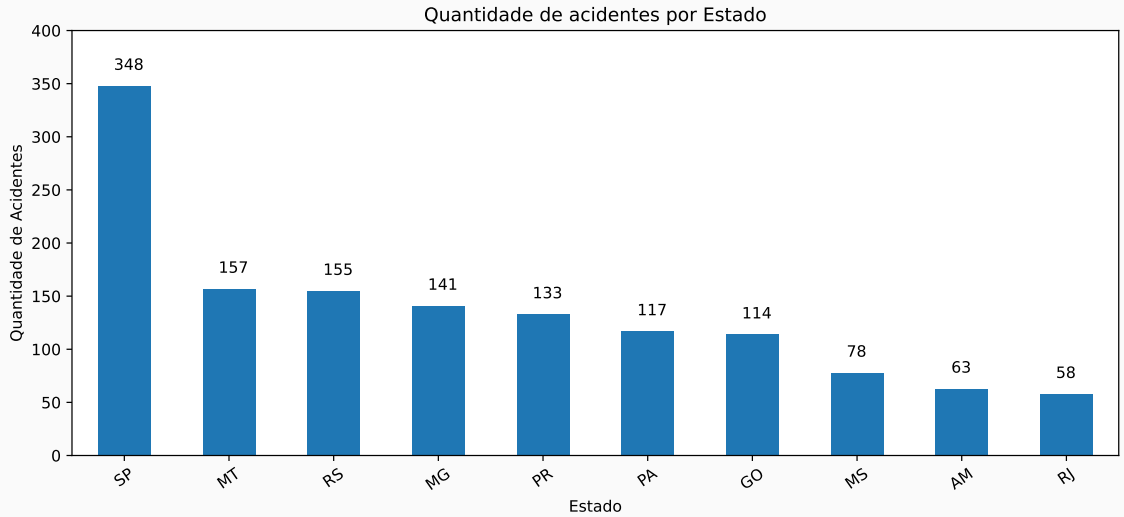

Figura 3: Gráfico de barras top 10 dos acidentes por Estados.

- Todos os Estados do Brasil apresentam ocorrências. Com base no gráfico acima, tem-se o Top 10 dos estados com maiores registros, dentre eles são: SP com 348 acidentes MT em segundo com 157 acidentes, e o último é a RJ com 58 registros de acidentes.

Respostas das Perguntas de investigação

- 3 - Quantas ocorrências envolvendo acidentes os relatórios foram divulgados?

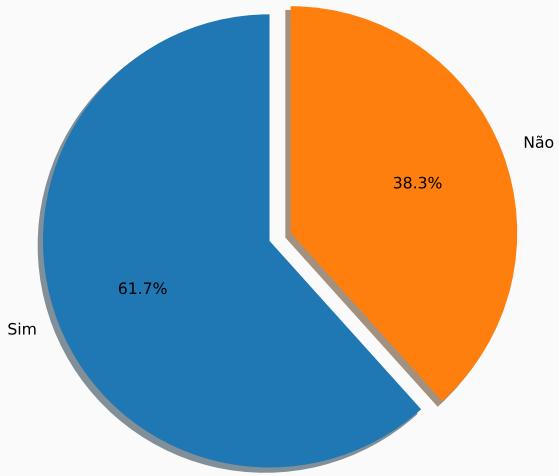

Figura 4: Gráfico de setores sobre divulgação dos relatórios de acidentes.

- No gráfico acima, tem-se que 61,7% dos relatórios técnicos foram apresentados e 38% não divulgaram.

Respostas das Perguntas de investigação

- 4 - Qual o número de aeronaves envolvidas em ocorrências no CENIPA? E quantas delas são classificadas como acidente?

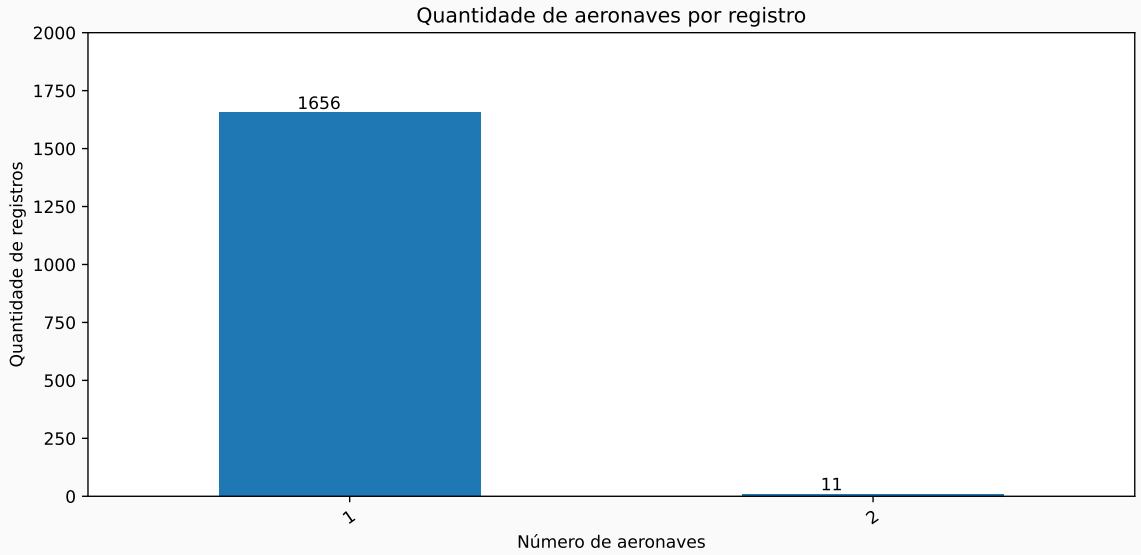

Figura 5: Gráfico de barras de aeronave por acidente.

- O número de aeronaves envolvidas em acidente é no máximo 3. Com base no gráfico, nota-se que o número de ocorrências classificadas como acidente tiveram mais casos com apenas uma aeronave 1656.

Respostas das Perguntas de investigação

- 5 - Quais são as categorias dos operadores das aeronaves que apresentaram ocorrência?

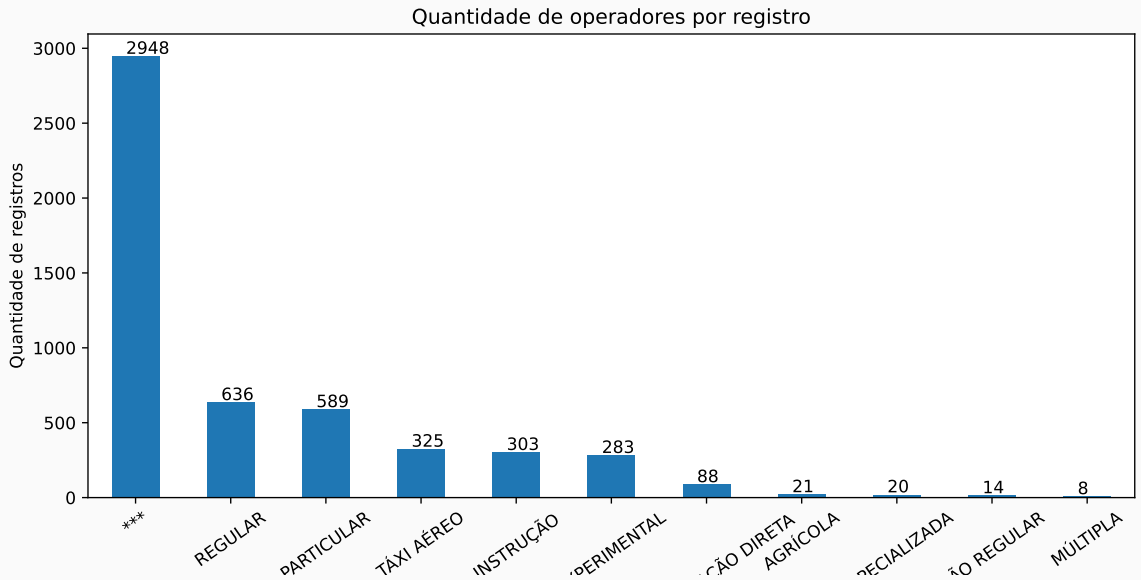

Figura 6: Gráfico de barras de categoria de operadores.

- Observa-se na Figura 5 que a categoria maior dos operadores é ocupada por uma não foi divulgada pelo conjunto de dados, em segundo lugar temos a regular e por último temos a múltipla.

Respostas das Perguntas de investigação

- 6 - Quais são os tipos de veículos que tiveram ocorrência?

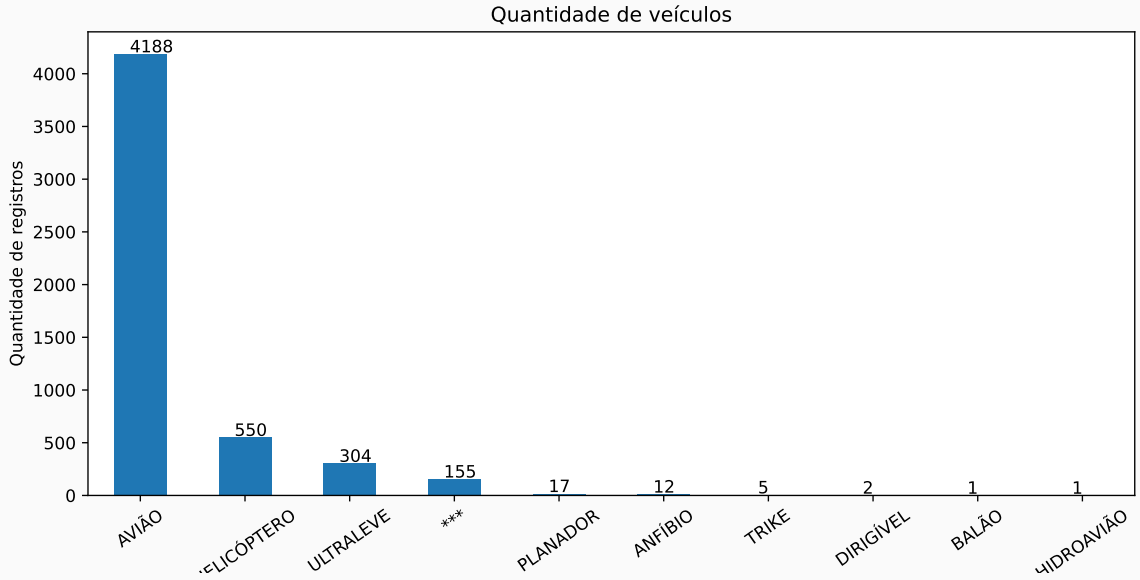

Figura 7: Gráfico de barras dos tipos de veículo com mais ocorrências.

- No que se refere a quantidade de veículos, tem-se que 80% dos veículos são classificados como avião, 10% como helicóptero, e os demais apresentam uma porcentagem abaixo de 6%.

Respostas das Perguntas de investigação

- 7 - Em quais tipos de veículo ocorreram mais acidentes?

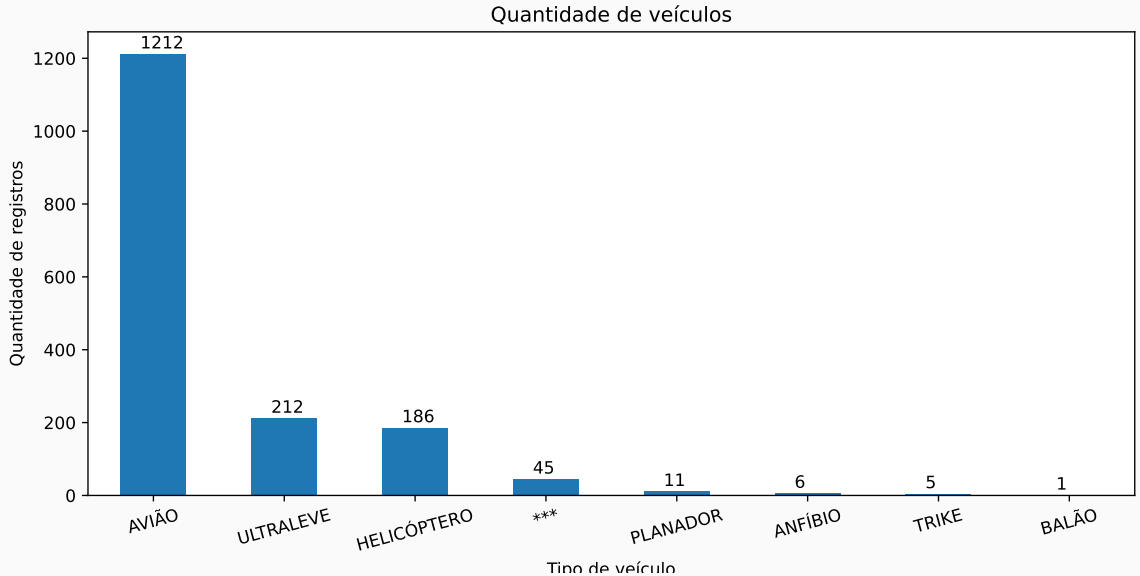

Figura 8: Gráfico de barras dos tipos de veículo que mais sofrem acidente.

- Observa-se na Figura acima que 72% dos acidentes são com avião, 12% são com Ultraleve e 11% com Helicóptero, e os demais tipos ficam abaixo de 3%.

Respostas das Perguntas de investigação

- 8 - Quais são os tipos de motor de aeronaves em que ocorreram mais acidentes?

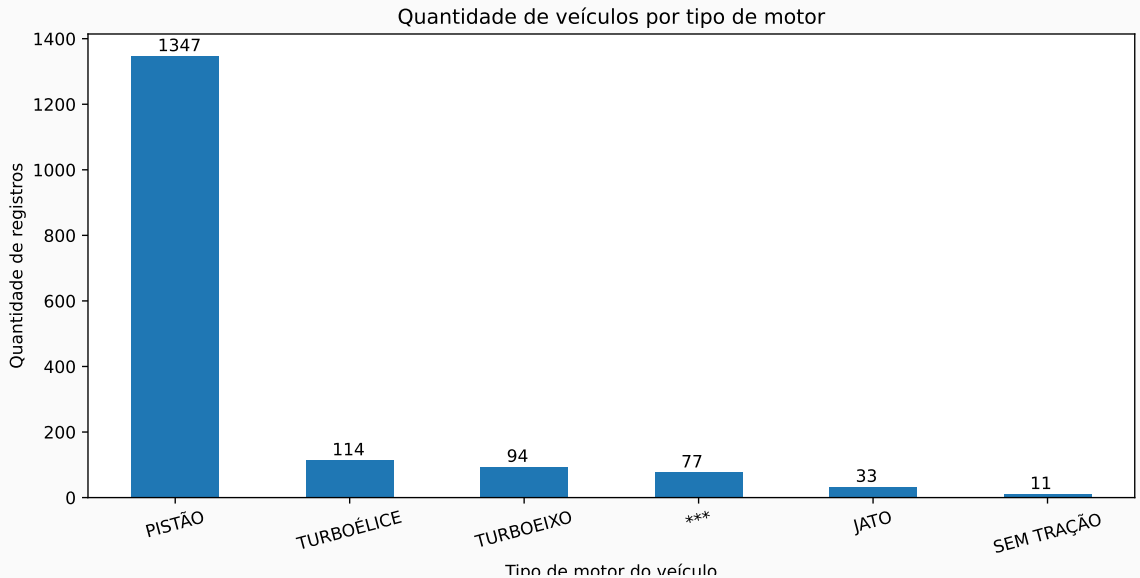

Figura 9: Gráfico de barras tipos de motor de aeronaves que mais sofrem acidente.

- Na Figura acima, tem-se que 1347 dos acidentes com veículos ocorrem com as que possuem motor pistão, enquanto que os demais motores representam menos de 150 ocorrências de acidente.

Respostas das Perguntas de investigação

- 9 - Qual é o top 5 de tipos de ocorrências registrados no CENIPA? E quais são os top 5 de acidentes?

Tabela 1: Tabela de frequência do top 5 tipo de ocorrência registrado no CENIPA.

Tipo de Ocorrência	Porcentagem(%)
FALHA DO MOTOR EM VOO	26,19
FALHA DE SISTEMA	23,33
ESTOURO DE PNEU	22,63
PERDA DE CONTROLE NO SOLO	15,11
PERDA DE CONTROLE EM VOO	12,74

- Na Tabela acima, tem-se que a maior causa de ocorrência é dada por falha de motor em voo, e a que menos ocorre é perda de controle de voo.

Respostas das Perguntas de investigação

- 9 - Qual é o top 5 de tipos de ocorrências registrados no CENIPA? E quais são os top 5 de acidentes?

Tabela 2: Tabela de frequência do top 5 tipo de acidentes registrados no CENIPA.

Tipo de Ocorrência	Porcentagem (%)
FALHA DO MOTOR EM VOO	31,84
PERDA DE CONTROLE EM VOO	29,06
PERDA DE CONTROLE NO SOLO	17,92
COLISÃO COM OBSTÁCULO	11,04
INDETERMINADO	10,14

- Na Tabela acima, tem-se que a maior causa de acidente é dada por falha de motor em voo, e a que menos ocorre é por evento indeterminado.

Modelos de Aprendizagem de Máquinas para classificação de ocorrência

Modelos de Aprendizagem de Máquinas

- Os modelos que foram utilizados para a classificação das ocorrências do CENIPA foram: Modelo Linear Generalizado (MLG) Multinomial, *KNN*, Árvore de decisão e *Random forest*.
- O modelo MLG Multinomial foi incluído pois os dados apresentados não tinham apenas duas classes.
- Como base na Figura 2, os dados de classificação de ocorrência são desbalanceados. Para resolver este problema foi utilizado a técnica de undersampling, isto é, o conjunto de dados foi reduzido (aleatoriamente) para o tamanho da classe menor (Incidente Grave), então o tamanho dos dados é de n = 2136 ocorrências.

Modelos de Aprendizagem de Máquinas

- As variáveis escolhias foram: Total aeronaves envolvidas, ocorrência de saída pista?, aeronave operador categoria, aeronave tipo de veiculo, tipo de motor, quantidade de motor, categoria de registro, registro de segmento, tipo de operação, nível de dano, e fatalidade total.
- A escolha dessas variáveis foi baseada nas informações disponíveis nos conjunto de dados, e são de escolhas do autor, devida as informações estarem completas e fazerem sentido para a classificação final.
- Após a definição dos modelos a serem utilizados, foi separado a amostra dos dados de ocorrência em amostra treino (70%) e amostra teste (30%).
- Com esta partição dos dados os modelos citados acima foram treinados (realizado o tunning dos parâmetros do modelo) e testados e os resultados das métricas de Acurácia e AUC (Área abaixo da Curva Roc) são dadas na Tabela 8.

Modelos de Aprendizagem de Máquinas

Tabela 3: Modelos de Aprendizagem de máquinas utilizados para predição da classificação de ocorrência do CENIPA.

Modelos	Acurácia	AUC
Multinomial	0,8069	0,9332
KNN	0.7290	0,8958
Arvore de Decisão	0,8022	0,9154
Random Forest	**0,8115**	**0,9224**

Tabela 4: Matriz de confusão do modelo *Random Forest*.

-	ACIDENTE	INCIDENTE	INCIDENTE GRAVE
Acidente	187	2	34
INCIDENTE	2	181	28
INCIDENTE GRAVE	25	31	152

- Observa-se na Tabela 3 e 4 que o modelo *Random Forest* é o melhor dentre os modelos para realizar a predição de classificação de ocorrências do CENIPA.
- Dessa forma, o modelo *Random Forest* será colocado em produção, para tal, será foi realizado um Deploy do modelo, utilizando a ferramenta Shinydashboard da linguagem R com o objetivo de tornar a análise mais facilitadora na prática. O arquivo encontra-se em 'CAM.R' e uma base de dados para testar o APP encontra-se em 'tab test.csv'.

Modelos de Previsão de Série Temporal

Modelos de Previsão de Série Temporal

- Para fazer previsão do número de ocorrência por Acidente, Incidente, ou Incidente Grave (univariado) foi considerado modelos de previsão de séries temporais, inicialmente foi considerado o modelo Arima.
- Modelos Arima são conhecidos na literatura por acomodar os principais componentes de séries, sendo eles: Tendência e Sazonalidade, por esta razão esses modelos foram considerados.
- Dessa forma, foi colocado em produção o modelo de séries temporais para a previsão de acidentes, incidentes e incidentes graves, o deploy está disponível em `https://gx1jfd-caio-balieiro.shinyapps.io/A3data/`.
- No que se refere a ajuste de modelo os três modelos se ajustaram bem, geraram o ruído branco, por essa forma, essa Deploy ajuda de maneira direta com previsões, desta forma, o CENIPA pode tomar ações de antemão para reduzir o número de ocorrências.

Resultados e discussões

Resultados e discussões

- O conjunto de dados apresentado neste projeto contempla uma gama de variáveis que permitem uma série de estudos paralelos. No entanto, nessa apresentação foi apresentados resultados envolvendo ocorrências do CENIPA, principalmente analisando os acidentes.
- No que se refere a modelagem, foi proposto uma análise descritiva para responder questões de investigação do conjunto de dados do CENIPA, em seguida foi proposto duas modelagens de séries temporais e de aprendizagem de máquina, cujo o objetivo é ajudar a automatizar os processos de tomada de decisão.
- Excepcionalmente quando trata-se de acidentes aéreos é importante que as informações de relatório sejam passadas de maneira clara, transparente e rápida. Com as aplicações desenvolvidas neste projeto, espera-se que os APP desenvolvidos ajudem a melhorar capacidade de classificação (automatizar) de ocorrências e a prever com precisão sobre quantas ocorrências irão ocorrer nos próximos meses.