PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Eduardo Cruz de Mello Franco

Predição de cancelamentos de voos dadas as condições meteorológicas do momento por modelos de machine learning

Belo Horizonte
2022

Eduardo Cruz de Mello Franco

Predição de cancelamentos de voos dadas as condições meteorológicas do momento por modelos de machine learning

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2022

SUMÁRIO

1. Introdução

1.1. Contextualização

Infelizmente, voos cancelados devido às condições climáticas são um problema comum no setor de aviação. Devido ao mau tempo, muitos voos são cancelados ou atrasados, o que pode causar grande transtorno para os passageiros. Muitas companhias aéreas fornecem informações sobre possíveis cancelamentos de voos devido ao mau tempo e estar ciente dessas informações pode ajudar a evitar surpresas desagradáveis. Além disso, é importante monitorar a situação do tempo constantemente, especialmente se o voo estiver programado para ocorrer em uma região com condições climáticas instáveis.

Por fim, é importante lembrar que os voos cancelados devido às condições climáticas não refletem a qualidade geral da companhia aérea. As companhias aéreas fazem o melhor que podem para garantir a segurança dos passageiros e minimizar o impacto dos voos cancelados. Em resumo, os voos cancelados devido às condições climáticas são uma realidade que pode causar transtorno aos passageiros, mas existem maneiras de minimizar o impacto.

1.2. O problema proposto

O objetivo deste artigo é avaliar os voos cancelados no período de 2019 e as condições meteorológicas presentes no momento do cancelamento. O registro dessas condições são avaliados por um modelo de machine learning que prediz se determinado voo será cancelado ou não dadas as condições meteorológicas.

O Problema de Classificação de Cancelamento de Voo é um problema de aprendizado de máquina supervisionado onde o objetivo é prever se um voo será cancelado ou não, com base em um conjunto de recursos de entrada. Esses recursos podem incluir informações sobre o voo em si, informações sobre as condições climáticas nos aeroportos de partida e destino e informações sobre a companhia aérea.

A variável alvo neste problema é binária, com um valor de 0 representando um voo que não foi cancelado e um valor de 1 representando um voo que foi cancelado. Para resolver este problema, um modelo de aprendizado de máquina é treinado em um conjunto de dados de voos rotulados, onde o status de cancelamento de cada voo é conhecido. O modelo usa esses dados de treinamento para aprender padrões e relações entre os recursos de entrada e a variável alvo, e esses padrões são usados para fazer previsões em novos dados de voo não vistos. O desempenho do modelo pode ser avaliado usando métricas como precisão, revocação, precisão e pontuação F1. O modelo então pode ser aperfeiçoado e melhorado ajustando a arquitetura do modelo ou usando técnicas de engenharia de recursos para gerar novos recursos mais informativos a partir dos dados existentes. Este problema é relevante na indústria de aviação, pois os cancelamentos de voo podem causar significativas interrupções e custos para os passageiros, as companhias aéreas e os aeroportos. Um modelo eficaz para prever os cancelamentos de voo pode ajudar as companhias aéreas a gerenciar proativamente suas operações e minimizar o impacto dos cancelamentos em seus clientes.

2. Coleta de Dados

O modelo tem por principal conduta estudar em que condições houveram os cancelamentos através da análise dos dados dos voos históricos registrados e disponibilizados pela ANAC no site:

https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas/historico-de-voos

O Voo Regular Ativo – VRA é uma base de dados composta por informações de voos de empresas de transporte aéreo que apresenta os cancelamentos e horários em que os voos ocorreram.

Esses dados são captados próximos aos aeroportos e trazem informações como: Temperatura, velocidade do vento, visibilidade do céu, umidade relativa e vários outros fenômenos. Esse banco de dados foi retirado da Iowa State University que possui todas as condições meteorológicas de 2019, que é atualizado a cada hora, nos 10 principais aeroportos do Brasil.

Segue o site de onde foi feita a requisição:

https://mesonet.agron.iastate.edu/request/download.phtml

As bibliotecas são coleções de códigos pré-escritos que podem ser usados para realizar tarefas comuns em um programa. Ao importar bibliotecas em um programa, você pode aproveitar o trabalho já feito por outros desenvolvedores e economizar tempo na escrita de código.

Importar as bibliotecas corretas e usá-las de maneira eficaz é uma parte importante do processo de análise de dados e pode ajudar a acelerar o processo de desenvolvimento e melhorar a qualidade dos resultados.

Importação das bibliotecas necessárias para o projeto:

```
import datetime
import string
import requests
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import imblearn


from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from imblearn.under_sampling import NearMiss
from sklearn.metrics import plot_confusion_matrix, accuracy_score, f1_score, recall_score, precision_scor
from sklearn import model_selection
from sklearn.model_selection import cross_val_score, KFold, train_test_split, GridSearchCV
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix

import scikitplot as skplt
import warnings
warnings.filterwarnings('ignore')
import io, os, sys, types, time, math, random, subprocess, tempfile
```

3. Processamento/Tratamento de Dados

A primeira situação que deve ser analisada é o banco de dados e qual é o objetivo da utilização de cada coluna. Ao final da análise o dataframe df00 não será integrado ao banco de dados final, ele apenas servirá de guia para a seleção das colunas(features).

Percebe-se que o banco de dados disponibilizado pela ANAC possui algumas variáveis que são irrelevantes para análise. Dito isso, serão removidas as colunas: 'Número voo', 'Código DI', 'Código Tipo Linha'.

De acordo com a hipótese que foi levantada não há a necessidade de utilizarmos as colunas que seriam do destino final, pois será avaliado apenas as condições meteorológicas do aeroporto de origem. Portanto, serão deletadas as seguintes colunas: 'ICAO Aeródromo Destino', 'Chegada Prevista', 'Chegada Real'. E a coluna 'Partida Real' também será removida devido ao fato de estar sendo avaliado os voos cancelados, não havendo a partida.

	ICAO Empresa Aérea	Número Voo	Código DI	Código Tipo Linha	ICAO Aeródromo Origem	ICAO Aeródromo Destino	Partida Prevista	Partida Real	Chegada Prevista	Chegada Real	Situação Voo	Código Justificativa
0	AAF	35	0	I	LFPO	SBKP	25/01/2019 06:15	25/01/2019 06:15	25/01/2019 18:15	25/01/2019 18:15	REALIZADO	NaN
1	AAF	35	0	I	LFPO	SBKP	27/01/2019 06:15	27/01/2019 06:15	27/01/2019 18:15	27/01/2019 18:15	REALIZADO	NaN
2	AAF	35	0	I	LFPO	SBKP	29/01/2019 06:15	29/01/2019 06:15	29/01/2019 18:15	29/01/2019 18:15	REALIZADO	NaN
3	AAF	36	0	I	SBKP	LFPO	25/01/2019 20:15	25/01/2019 20:15	26/01/2019 07:45	26/01/2019 07:45	REALIZADO	NaN
4	AAF	36	0	I	SBKP	LFPO	27/01/2019 20:15	27/01/2019 20:15	28/01/2019 07:45	28/01/2019 07:45	REALIZADO	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
91812	UPS	417	0	G	SBKP	SKBO	29/01/2019 01:50	29/01/2019 01:50	29/01/2019 07:22	29/01/2019 07:22	REALIZADO	NaN
91813	UPS	417	0	G	SKBO	KMIA	08/01/2019 10:11	08/01/2019 08:08	08/01/2019 13:44	08/01/2019 12:02	REALIZADO	HI
91814	UPS	417	0	G	SKBO	KMIA	15/01/2019 10:11	15/01/2019 08:21	15/01/2019 13:44	15/01/2019 11:52	REALIZADO	HI
91815	UPS	417	0	G	SKBO	KMIA	22/01/2019 10:11	22/01/2019 10:11	22/01/2019 13:44	22/01/2019 13:44	REALIZADO	NaN
91816	UPS	417	0	G	SKBO	KMIA	29/01/2019 10:11	29/01/2019 08:15	29/01/2019 13:44	29/01/2019 11:45	REALIZADO	HI

91817 rows × 12 columns

Faz a separação dos voos que foram cancelados por condições meteorológicas e verifica-se alguns gráficos. Para separar os voos cancelados devido às condições meteorológicas, você precisará primeiro determinar quais voos foram cancelados e quais não foram. Essa informação está armazenada na coluna ‘Situação Voo’. Já na coluna ‘Código Justificativa' os voos cancelados tem a explicação da motivação desse cancelamento, e neste caso estão sendo avaliadas as condições climáticas que estão representados pela sigla ‘XO’ dentro desta coluna.

Encerra-se o requerimento dos bancos de dados dos voos nos 12 meses de 2019. Logo após, ocorre a união dos 12 meses em apenas 1 Dataframe pela função concat do Pandas.

	ICAO Empresa Aérea	ICAO Aeródromo Origem	Partida Prevista	Situação Voo	Código Justificativa
4200	AZU	SBCY	31/01/2019 17:25	CANCELADO	XO
4227	AZU	SBLO	31/01/2019 19:40	CANCELADO	XO
4614	AZU	SBRJ	19/01/2019 08:05	CANCELADO	XO
4996	AZU	SBUR	26/01/2019 05:50	CANCELADO	XO
6415	AZU	SBVH	29/01/2019 15:15	CANCELADO	XO
...	...	...	...	...	...
86990	OWT	SBSP	05/12/2019 19:00	CANCELADO	XO
87293	TAM	SBSP	10/12/2019 15:00	CANCELADO	XO
87314	OWT	SSPG	05/12/2019 10:10	CANCELADO	XO
87525	OWT	SSPG	06/12/2019 10:10	CANCELADO	XO
87628	AZU	SBRJ	13/12/2019 06:30	CANCELADO	XO

2166 rows × 5 columns

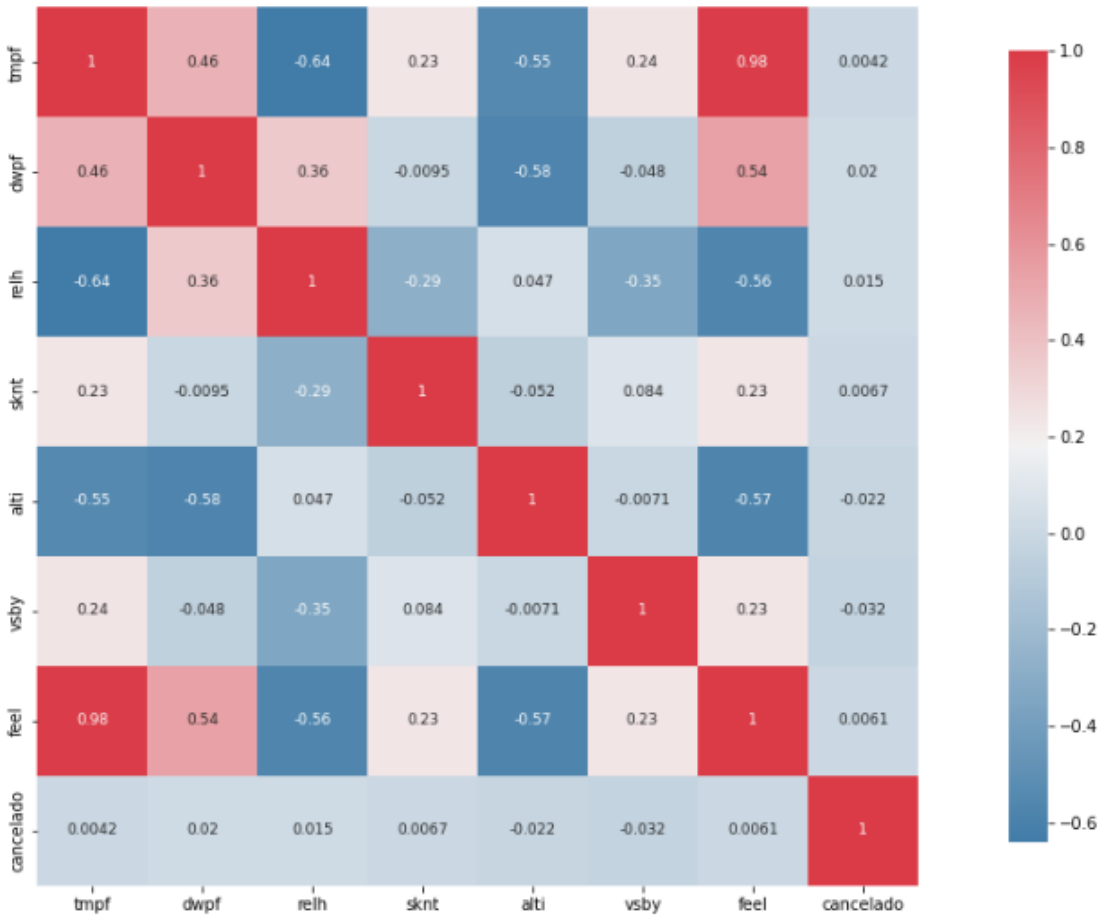
Coleta-se o banco de dados das condições meteorológicas dos 10 aeroportos com maiores quantidades de voos de 2019, banco de dados disponibilizado pela Universidade de Iowa.

	Código Justificativa	period	station	tmpf	dwpf	relh	drct	sknt	alti	vsby	...	skyc1	skyc2	skyc3	skyc4	skyl1	skyl2	skyl3	skyl4	wxc
0	NaN	SBKP2019-01-25 20:00:00	SBKP	71.60	66.20	83.09	130.00	15.00	30.00	3.11	...	FEW	FEW	NaN	NaN	4000.00	5000.00	NaN	NaN	-1
1	NaN	SBKP2019-01-25 20:00:00	SBKP	71.60	66.20	83.09	130.00	15.00	30.00	3.11	...	FEW	FEW	NaN	NaN	4000.00	5000.00	NaN	NaN	-1
2	WR	SBKP2019-01-25 20:00:00	SBKP	71.60	66.20	83.09	130.00	15.00	30.00	3.11	...	FEW	FEW	NaN	NaN	4000.00	5000.00	NaN	NaN	-1
3	RM	SBKP2019-01-25 20:00:00	SBKP	71.60	66.20	83.09	130.00	15.00	30.00	3.11	...	FEW	FEW	NaN	NaN	4000.00	5000.00	NaN	NaN	-1
4	TD	SBKP2019-01-25 20:00:00	SBKP	71.60	66.20	83.09	130.00	15.00	30.00	3.11	...	FEW	FEW	NaN	NaN	4000.00	5000.00	NaN	NaN	-1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
585723	MX	SBBR2019-01-12 02:00:00	SBBR	69.80	66.20	88.34	0.00	0.00	30.12	6.21	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
585724	NaN	SBRF2019-12-23 01:00:00	SBRF	80.60	71.60	74.11	110.00	7.00	29.94	6.21	...	SCT	NaN	NaN	NaN	2500.00	NaN	NaN	NaN	
585725	NaN	SBGL2019-12-20 02:00:00	SBGL	75.20	69.80	83.32	280.00	9.00	30.00	6.21	...	BKN	OVC	NaN	NaN	1700.00	3000.00	NaN	NaN	
585726	MX	SBRF2019-12-22 01:00:00	SBRF	80.60	69.80	69.71	120.00	9.00	29.94	6.21	...	SCT	NaN	NaN	NaN	2300.00	NaN	NaN	NaN	
585727	NaN	SBKP2019-08-12 21:00:00	SBKP	82.40	51.80	34.69	360.00	2.00	30.06	6.21	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

585728 rows × 21 columns

O banco de dados de condições meteorológicas da Universidade de Iowa é uma coleção de dados meteorológicos históricos que inclui informações sobre temperatura, umidade, pressão, vento, entre outros. Estes dados são coletados por estações meteorológicas da Universidade e são amplamente utilizados para pesquisas climáticas e análises de tendências climáticas. Além disso, também pode ser usado para modelar e prever condições climáticas futuras e sua influência na aviação, como, por exemplo, a previsão de cancelamento de voos.

Verifica-se o quadro de correlação:



A matriz de correlação na previsão de cancelamento de voo ajuda a identificar a relação entre as variáveis do modelo e a probabilidade de cancelamento de um voo. Ela mostra como as variáveis são correlacionadas entre si e se existe uma relação forte ou fraca entre elas. A partir destas informações, é possível ajustar o modelo para uma previsão mais precisa.

Na imagem abaixo, verifica-se os dados nulos e se retira algumas amostras para melhor leitura dos modelos de machine learning:

```
dft.isnull().sum(axis = 0)
```

```
Código Justificativa    394153
tmpf                      1458
dwpf                      1681
relh                      1929
sknt                       331
alti                      2169
vsby                       304
skyc1                   190340
skyc2                   393420
skyc3                   536794
skyc4                   582620
skyl1                   210395
skyl2                   393548
skyl3                   536973
skyl4                   582714
wxcodes                 516112
feel                      1929
dtype: int64
```

Retira-se algumas amostras com valores nulos para melhor avaliação futura pelo algoritmo de machine learning.

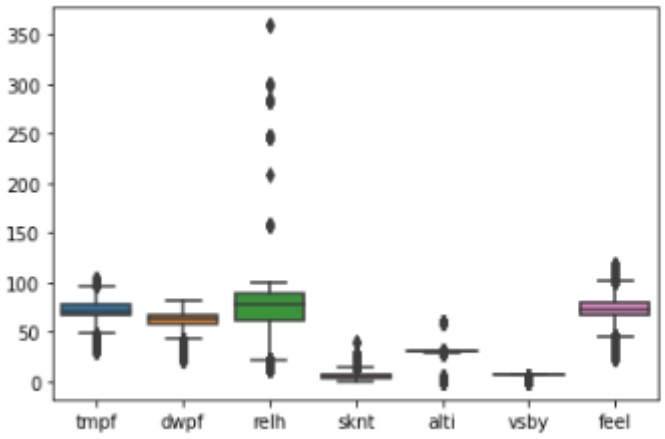
```
dft.dropna(subset=['relh'], how='all', inplace=True)
dft.dropna(subset=['alti'], how='all', inplace=True)
dft.dropna(subset=['vsby'], how='all', inplace=True)
dft.dropna(subset=['sknt'], how='all', inplace=True)
```

```
dft.isnull().sum(axis = 0)
```

```
Código Justificativa    391627
tmpf                         0
dwpf                         0
relh                         0
sknt                         0
alti                         0
vsby                         0
skyc1                   189990
skyc2                   392535
skyc3                   535384
skyc4                   578902
skyl1                   209977
skyl2                   392649
skyl3                   535468
skyl4                   578920
wxcodes                 515413
feel                         0
dtype: int64
```

Dados nulos são valores ausentes ou inválidos em um conjunto de dados. Eles podem ser causados por vários motivos, como erros de digitação, falhas de coleta de dados ou ausência de informações relevantes. A presença de dados nulos pode prejudicar a análise e a modelagem de dados, pois muitos algoritmos de aprendizado de máquina não são capazes de lidar com valores ausentes. Por essa razão, é importante lidar com dados nulos antes de prosseguir com a análise. Algumas técnicas comuns incluem remoção de linhas ou colunas com dados nulos, imputação de valores com base em outras informações ou uso de técnicas de aprendizado de máquina que são capazes de lidar com dados nulos. A escolha da técnica a ser usada dependerá da natureza dos dados e do objetivo da análise.

Na imagem abaixo, verifica-se os outliers:



Outliers são pontos de dados que estão significativamente fora da faixa da maioria dos dados. Eles podem ter um grande impacto na análise e modelagem de dados, pois podem distorcer os resultados e levar a conclusões incorretas. Na análise estatística, os outliers são frequentemente detectados usando técnicas como o box-plot, por exemplo. Uma vez detectados, os outliers podem ser manipulados de várias maneiras, incluindo removê-los, imputar valores ausentes ou transformar os dados para minimizar seu impacto. É importante considerar a natureza dos dados e o propósito da análise ao decidir sobre o método adequado para lidar com os outliers.

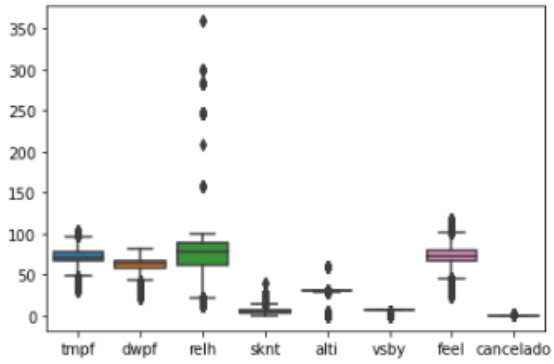
Remove-se os outliers indesejados:

```
dfto = dft.drop(dft[dft.relh > 120].index)
dfto = dfto.drop(dfto[dfto.sknt > 32].index)
dfto = dfto.drop(dfto[dfto.alti < 20].index)
dfto = dfto.drop(dfto[dfto.alti > 50].index)
dfto
```

Verifica-se os outlier de todos os voos nacionais de 2019 através do Box plot, e também se verifica os outlier dos voos cancelados de 2019 através do Box plot:

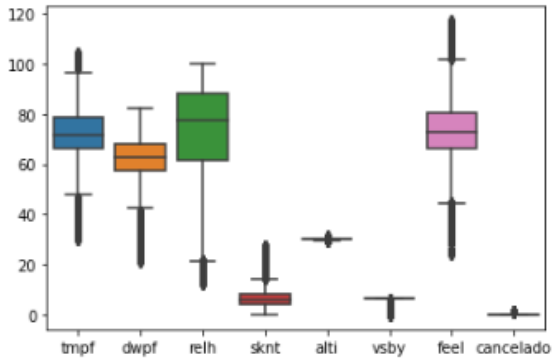
Apresenta o Box-plot de todos os voos nacionais

```
In [46]: sns.boxplot(data=dft)
Out[46]: <AxesSubplot:>
```



Apresenta o Box-plot de todos os voos nacionais sem os outliers

```
In [47]: sns.boxplot(data=dfto)
Out[47]: <AxesSubplot:>
```



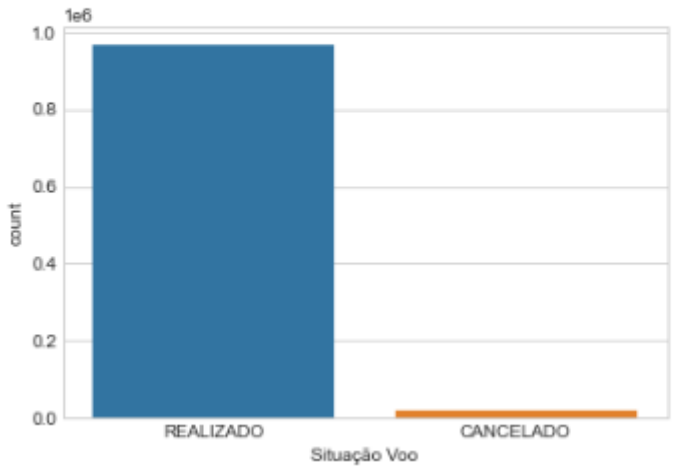
O box plot na previsão de cancelamento de voo mostra a distribuição de dados de uma variável específica, incluindo a mediana, e quartis e os valores extremos (outliers). Ele ajuda a identificar a presença de valores atípicos e a distribuição geral dos dados, o que pode ser útil na avaliação da relevância das variáveis para o modelo de previsão.

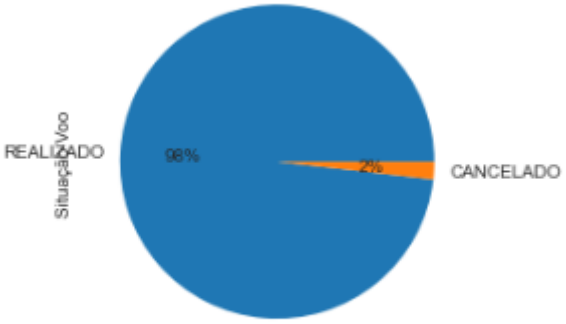
4. Análise e Exploração dos Dados

Para analisar e explorar dados de voos cancelados e realizados, é necessário coletar informações relevantes sobre cada voo, como horários, destinos, motivos de cancelamento, entre outros dados. Em seguida, é possível utilizar técnicas de análise exploratória de dados para identificar padrões, tendências e relações entre as variáveis. Algumas das técnicas de análise e exploração de dados incluem gráficos de dispersão, histogramas, box plots e tabelas de contingência. Uma vez que os dados são explorados, é possível criar modelos estatísticos para prever o número de voos cancelados ou realizados em uma determinada data ou período.

Estes modelos podem ser baseados em regressão linear, árvores de decisão, florestas aleatórias, entre outros algoritmos de aprendizado de máquina. Além disso, é possível utilizar técnicas de aprendizado de cluster para identificar grupos de voos com características similares, como rotas, horários e motivos de cancelamento. Isso pode ajudar a identificar padrões e tendências que possam ser úteis para tomar decisões estratégicas em relação à programação de voos e à gestão de riscos.

	Cancelados_Realizados	N_VOOS
0	REALIZADO	965183
1	CANCELADO	17793





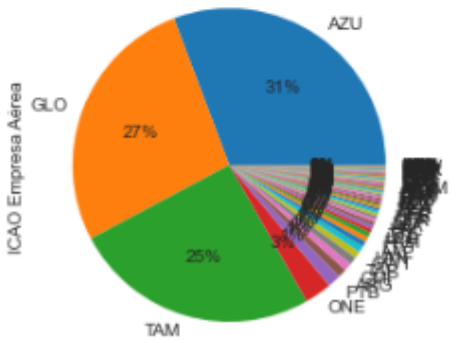
Análise das empresas aéreas por voos totais:

```
dfx['ICAO Empresa Aérea'].value_counts()
```

```
AZU    302316
GLO    265582
TAM    250403
ONE     27589
PTB     13363
       ...
SLM       214
EDW       199
CFG       164
FBZ        97
AZN        32
Name: ICAO Empresa Aérea, Length: 63, dtype: int64
```

```
dfx['ICAO Empresa Aérea'].value_counts().plot(kind='pie', autopct='%1.0f%%')
```

```
<AxesSubplot:ylabel='ICAO Empresa Aérea'>
```



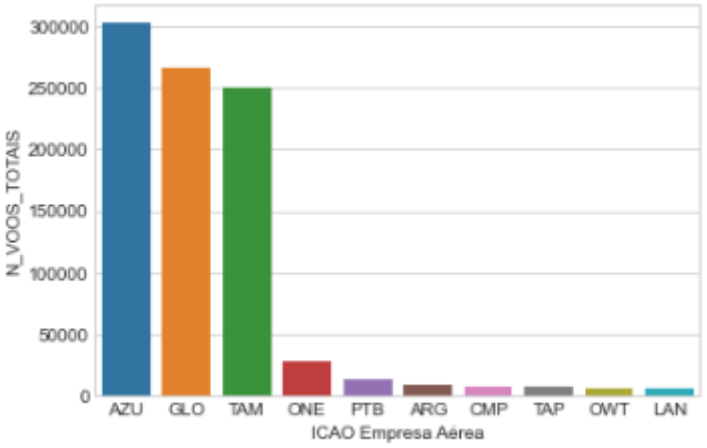
A análise das empresas aéreas por voos totais envolve avaliar o número de voos realizados por cada empresa aérea em um determinado período de tempo. Para realizar essa análise, é necessário coletar dados sobre o número de voos realizados por cada empresa aérea, bem como outras informações relevantes, como destinos, horários, entre outros.

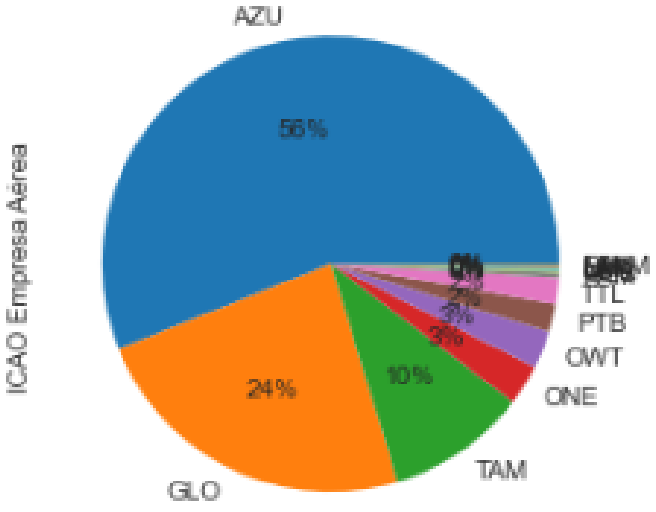
Uma vez que os dados estão disponíveis, é possível utilizar técnicas de análise de dados para visualizar e comparar o desempenho de cada empresa aérea. Algumas das técnicas incluem gráficos de barras, gráficos de setores, tabelas de dados, entre outras. É possível comparar o número total de voos realizados por cada empresa aérea, bem como o número de voos por destino, horário, entre outros aspectos.

Além disso, é possível utilizar técnicas de análise de tendências para identificar padrões e tendências no desempenho das empresas aéreas ao longo do tempo. Estas técnicas incluem regressão linear, análise de séries temporais, entre outros. A análise das empresas aéreas por voos totais é importante para avaliar a eficiência das empresas aéreas, bem como para tomar decisões estratégicas sobre investimentos, alocação de recursos e planejamento de voos.

Verifica-se a quantidade de voos das empresas aéreas:

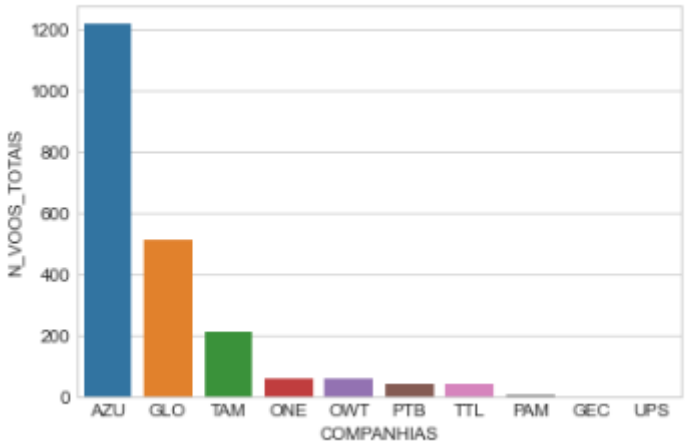
	ICAO Empresa Aérea	N_VOOS_TOTAIS
0	AZU	302316
1	GLO	265582
2	TAM	250403
3	ONE	27589
4	PTB	13363
5	ARG	8885
6	CMP	8087
7	TAP	7750
8	OWT	7007
9	LAN	6509





Verifica-se a quantidade de voos cancelados das empresas aéreas:

	COMPANHIAS	N_VOOS_TOTAIS
0	AZU	1215
1	GLO	512
2	TAM	215
3	ONE	62
4	OWT	58
5	PTB	43
6	TTL	40
7	PAM	8
8	GEC	3
9	UPS	3



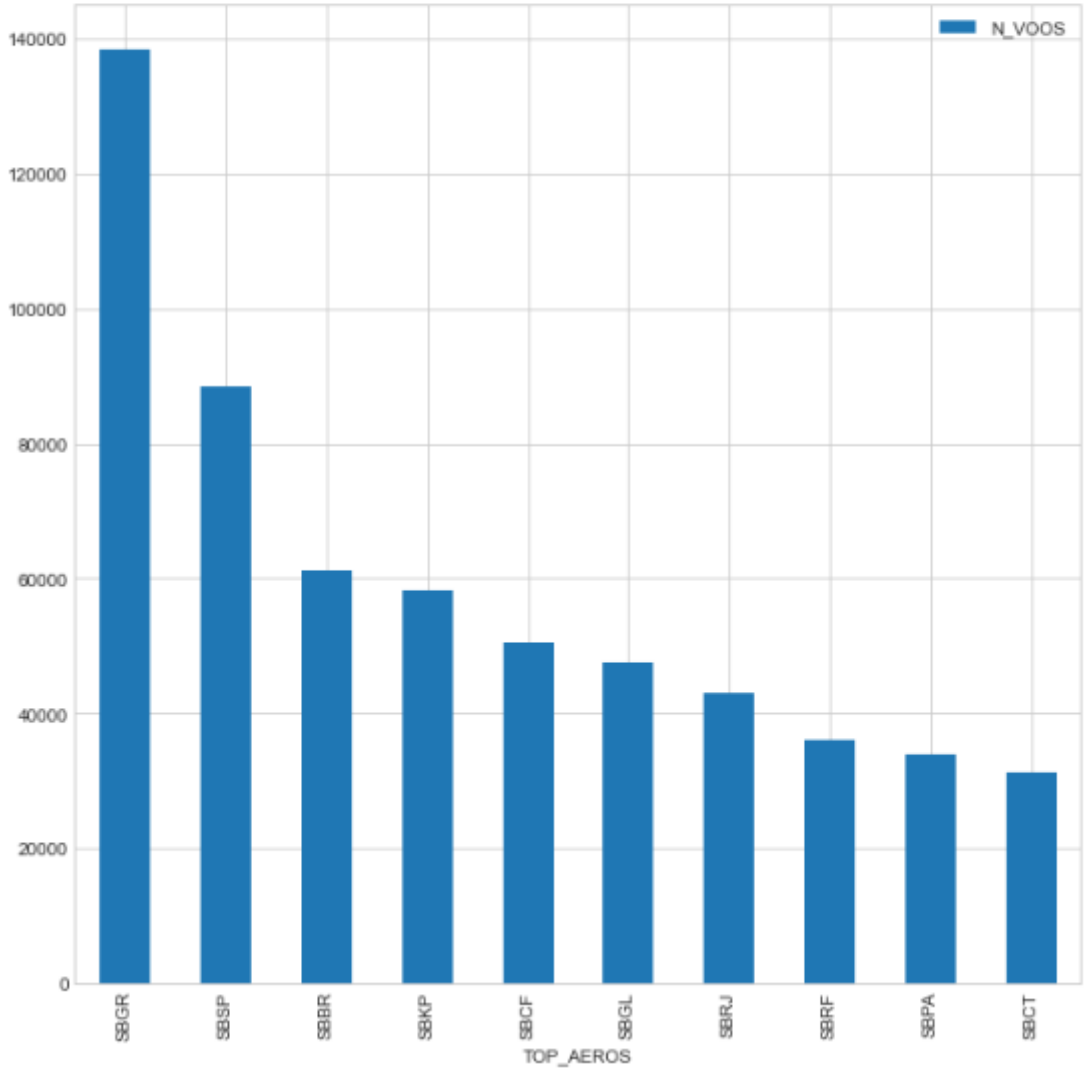
Como o banco de dados é muito extenso, decidiu-se por excluir algumas amostras, adotando a estratégia de selecionar apenas os 10 aeroportos com a maior quantidade de voos como espaço amostral.

A seleção de apenas os 10 aeroportos com a maior quantidade de voos como espaço amostral é uma técnica conhecida como amostragem por conveniência ou amostragem por julgamento. Isso significa que a amostra é selecionada com base em critérios pré-determinados, como a quantidade de voos em um aeroporto, em vez de ser selecionada aleatoriamente. Essa abordagem pode ser útil em situações em que o banco de dados é muito extenso e o objetivo é obter uma compreensão geral dos dados, sem precisar de uma análise detalhada de todos os dados. No entanto, é importante lembrar que a amostragem por conveniência pode resultar em uma amostra não representativa da população total e, portanto, os resultados obtidos a partir da amostra podem não ser generalizáveis para a população total.

Além disso, é importante verificar a adequação da amostra selecionada para o objetivo da análise. Se o objetivo é avaliar o desempenho das empresas aéreas em todo o mundo, a seleção de apenas 10 aeroportos pode não fornecer uma visão completa do desempenho das empresas aéreas. Neste caso, pode ser necessário selecionar uma amostra mais ampla ou adotar outra estratégia de amostragem.

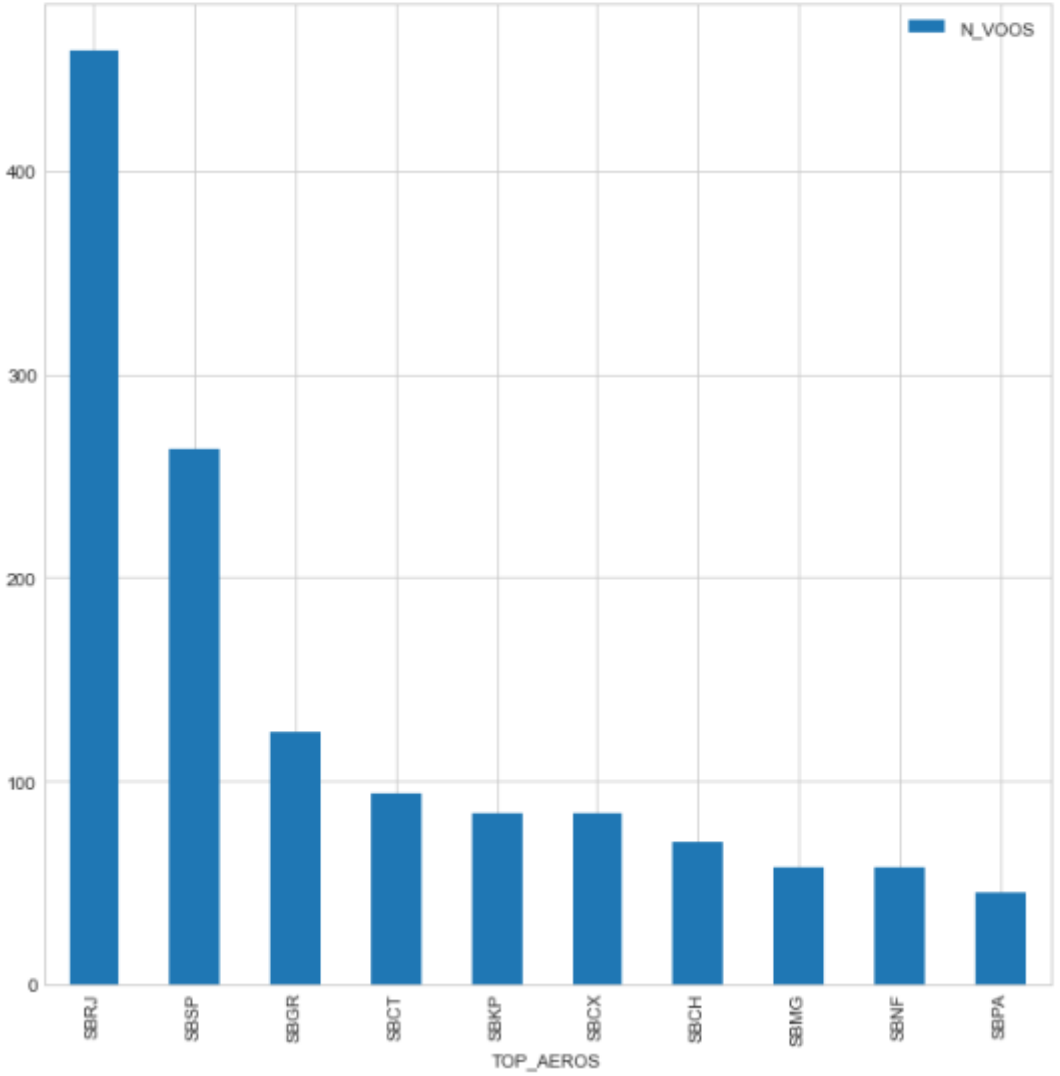
Verifica-se a quantidade de voos por aeroporto:

	TOP_AEROS	N_VOOS
0	SBGR	138254
1	SBSP	88504
2	SBBR	61260
3	SBKP	58154
4	SBCF	50509
5	SBGL	47696
6	SBRJ	43131
7	SBRF	36068
8	SBPA	33939
9	SBCT	31347



Verifica-se a quantidade de voos cancelados dos aeroportos:

	TOP_AEROS	N_VOOS
0	SBRJ	459
1	SBSP	263
2	SBGR	124
3	SBCT	94
4	SBKP	84
5	SBCX	84
6	SBCH	70
7	SBMG	58
8	SBNF	58
9	SBPA	45



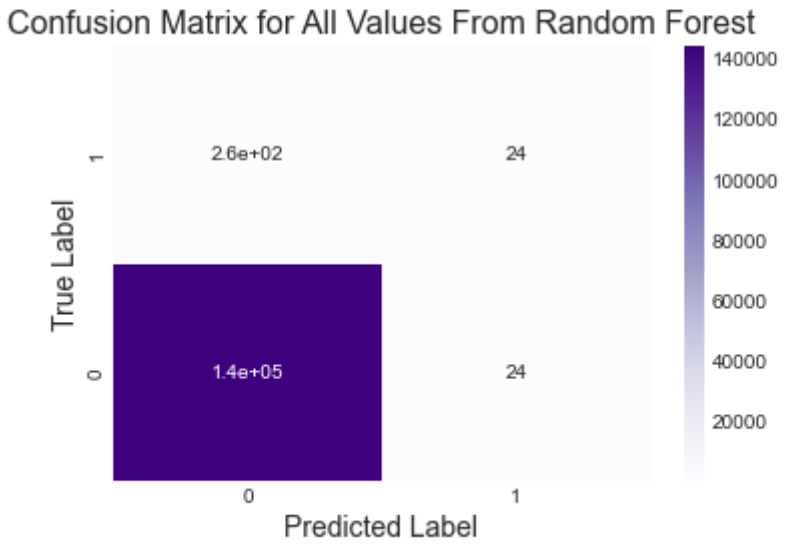
Verificar a quantidade de voos cancelados nos aeroportos selecionados é uma importante parte da análise dos dados. Isso pode fornecer uma visão geral da eficiência das empresas aéreas e dos aeroportos em questão. Para realizar essa análise, é necessário coletar dados sobre o número de voos cancelados em cada um dos 10 aeroportos selecionados. Uma vez que os dados estão disponíveis, é possível utilizar técnicas de análise de dados para visualizar e comparar a quantidade de voos cancelados em cada aeroporto. Algumas das técnicas incluem gráficos de barras, gráficos de setores, tabelas de dados, entre outras. Além disso, é possível utilizar técnicas de análise de tendências para identificar padrões e tendências no número de voos cancelados ao longo do tempo. Estas técnicas incluem regressão linear, análise de séries temporais, entre outras. A verificação da quantidade de voos cancelados pode fornecer informações valiosas sobre a eficiência das empresas aéreas e dos aeroportos, bem como sobre as causas dos cancelamentos. Isso pode ajudar a identificar problemas e a tomar medidas para melhorar a eficiência e a segurança dos voos.

5. Criação de Modelos de Machine Learning

No primeiro modelo testado abaixo, percebe-se que há uma alta porcentagem de acertos. Isso ocorre devido haver uma classe majoritária (classe dos voos REALIZADOS) e uma classe minoritária (classe dos voos CANCELADOS).

```
Training Recall Score, Random Forest:  0.1715686274509804
Test Recall Score, Random Forest:  0.08540925266903915
Training Precision Score, Random Forest:  0.7446808510638298
Test Precision Score, Random Forest:  0.5
Training Accuracy Score, Random Forest:  0.9983325656379548
Test Accuracy Score, Random Forest:  0.9980585073306893
Training F1 Score, Random Forest:  0.27888446215139445
Test F1 Score, Random Forest:  0.14589665653495443

Text(0.5, 1.0, 'Confusion Matrix for All Values From Random Forest')
```



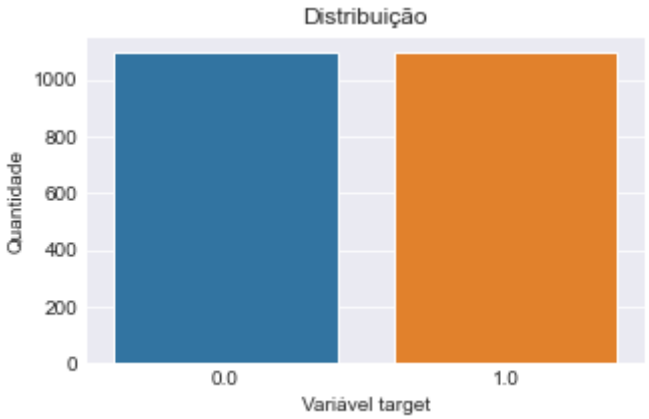
Os modelos de machine learning sempre irão procurar o meio de ter o melhor aproveitamento de performance possível. Logo, é esperado que os modelos classifiquem quase todas as amostras de uma classe majoritária, pois assim, ela terá o melhor resultado possível. Esse resultado não expressa a realidade e deve ser descartado, e procurar um outro modelo de banco de dados, pois assim será feito um balanceamento para a solução dessa situação.

O desbalanceamento de classes é um problema comum na análise de dados, especialmente quando existe uma classe dominante. No caso da previsão de voos

cancelados, isso significa que a maioria dos voos não é cancelada, o que pode tornar difícil para o modelo prever corretamente os voos cancelados.

No gráfico abaixo percebemos muitas amostras da classe 'REALIZADOS' (0.0) e pouco da classe 'CANCELADOS' (1.0)

Para a correção dessa situação é necessário fazer o balanceamento das classes. Para isso foi utilizado a função NearMiss da biblioteca Imblearn. O balanceamento das classes se dá a partir da seleção aleatória de amostras da classe majoritária equivalente a quantidade de amostras da classe minoritária, como é demonstrado no gráfico a seguir:



Equilibrar a distribuição de classes e tornar mais fácil para o modelo prever corretamente os voos cancelados. Lidar com o desbalanceamento de classes é

importante para garantir que o modelo tenha um desempenho preciso e equilibrado na previsão de voos cancelados.

O undersampling é uma técnica de balanceamento de classes que consiste em remover exemplos da classe dominante para tornar a distribuição de classes mais equilibrada. No caso da previsão de voos cancelados, isso significa remover exemplos de voos não cancelados para tornar a distribuição de voos cancelados e não cancelados mais equilibrada.

A vantagem do undersampling é que ele pode ser mais eficaz que o oversampling quando há uma grande quantidade de dados e, portanto, é possível remover exemplos sem prejudicar significativamente a quantidade de dados disponíveis para o treinamento. No entanto, é importante ter cuidado ao aplicar o undersampling, pois ele pode resultar em perda de informação importante e afetar o desempenho do modelo.

Para aplicar o undersampling, é necessário seguir os seguintes passos:

- Identificar a classe dominante: No caso da previsão de voos cancelados, a classe dominante são os voos não cancelados.
- Calcular a quantidade de exemplos na classe minoritária: Calcular a quantidade de exemplos de voos cancelados na base de dados.
- Determinar a quantidade de exemplos da classe dominante a serem removidos: A quantidade de exemplos da classe dominante a serem removidos deve ser igual à quantidade de exemplos da classe minoritária.
- Remover exemplos da classe dominante: Remover aleatoriamente exemplos da classe dominante até que a quantidade de exemplos de ambas as classes seja igual.
- Treinar o modelo: Treinar o modelo com o conjunto de dados equilibrado e avaliar o desempenho do modelo.

6. Apresentação dos Resultados

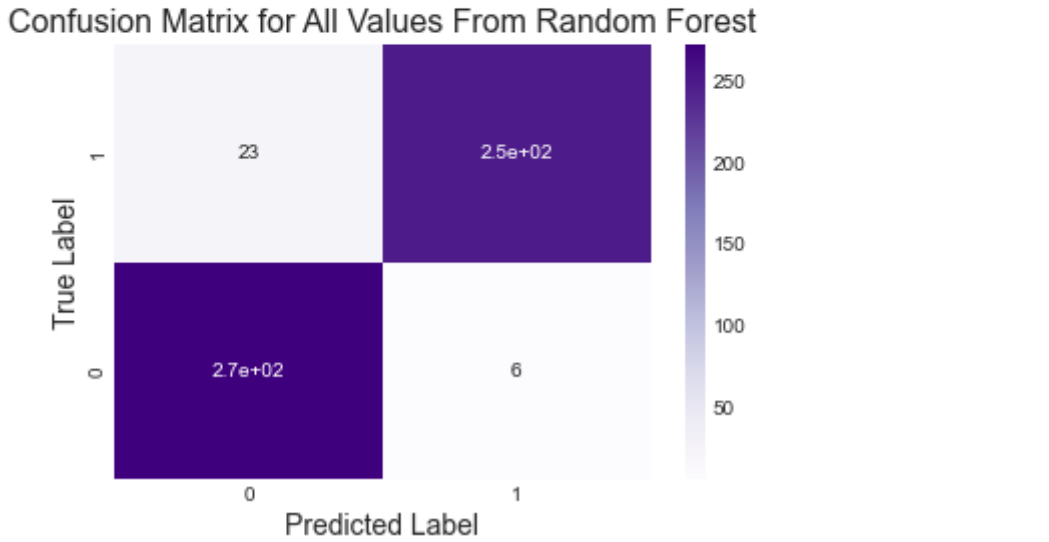
A verificação dos modelos de teste e os de treinamento são etapas importantes no processo de avaliação de modelos de aprendizado de máquina. O objetivo é verificar se o modelo está sofrendo de overfitting, ou seja, se ele está aprendendo demais os dados de treinamento e não está generalizando bem para novos dados.

Em um modelo de machine learning é sempre importante verificar as diferenças entre as métricas de desempenho nos conjuntos de treinamento e teste. E se, houver uma grande diferença entre as métricas de desempenho nos conjuntos de treinamento e teste, isso pode indicar overfitting.

Roda-se os modelos de machine learning com o banco de dados final, limpo e balanceado. Utiliza-se os modelos de Random Forest, KNN (K-Nearest Neighbors) e Logistic Regression, 3 algoritmos de machine learning bem conhecidos.

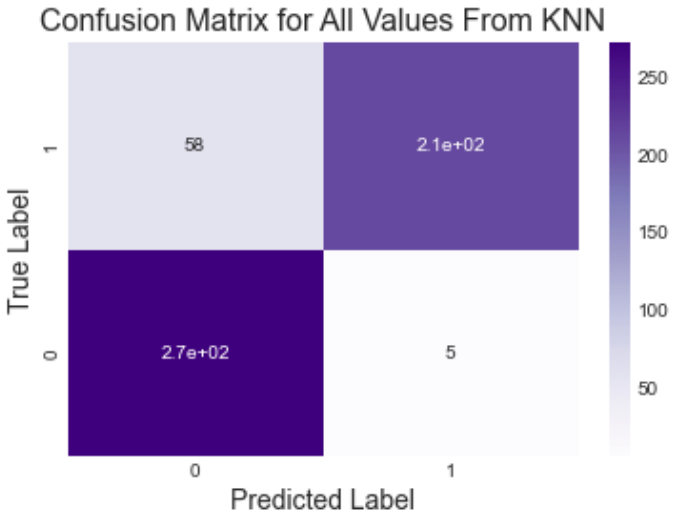
```
Training Recall Score, Random Forest:  0.8983050847457628
Test Recall Score, Random Forest:  0.915129151291513
Training Precision Score, Random Forest:  0.9867021276595744
Test Precision Score, Random Forest:  0.9763779527559056
Training Accuracy Score, Random Forest:  0.9428571428571428
Test Accuracy Score, Random Forest:  0.9471766848816029
Training F1 Score, Random Forest:  0.9404309252217997
Test F1 Score, Random Forest:  0.9447619047619048

Text(0.5, 1.0, 'Confusion Matrix for All Values From Random Forest')
```



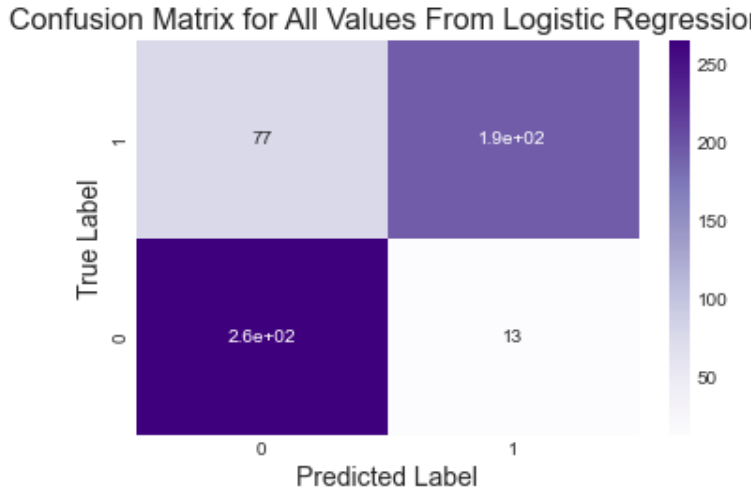
```
Training Recall Score, KNN:  0.8256658595641646
Test Recall Score, KNN:  0.7859778597785978
Training Precision Score, KNN:  0.9941690962099126
Test Precision Score, KNN:  0.9770642201834863
Training Accuracy Score, KNN:  0.9100303951367781
Test Accuracy Score, KNN:  0.8852459016393442
Training F1 Score, KNN:  0.9021164021164021
Test F1 Score, KNN:  0.8711656441717792
```

```
Text(0.5, 1.0, 'Confusion Matrix for All Values From KNN')
```




```
Training Recall Score, Logistic Regression:  0.7191283292978208
Test Recall Score, Logistic Regression:  0.7158671586715867
Training Precision Score, Logistic Regression:  0.9369085173501577
Test Precision Score, Logistic Regression:  0.9371980676328503
Training Accuracy Score, Logistic Regression:  0.8346504559270517
Test Accuracy Score, Logistic Regression:  0.8360655737704918
Training F1 Score, Logistic Regression:  0.8136986301369863
Test F1 Score, Logistic Regression:  0.8117154811715481
```

```
Text(0.5, 1.0, 'Confusion Matrix for All Values From Logistic Regression')
```



Recall, precision, accuracy e F1 score são medidas comuns de desempenho utilizadas para avaliar modelos de classificação, como modelos de previsão de voos cancelados. Aqui está uma breve descrição de cada uma dessas métricas:

- Recall (sensibilidade): É a fração de voos cancelados previstos corretamente pelo modelo em relação ao número total de voos cancelados. O recall é uma medida de quantos dos voos cancelados de fato foram previstos corretamente pelo modelo.
- Precision (precisão): É a fração de voos cancelados previstos corretamente pelo modelo em relação ao número total de voos previstos como cancelados. A precisão mede a capacidade do modelo de prevê-los corretamente.
- Accuracy (acurácia): É a fração de voos previstos corretamente pelo modelo em relação ao número total de voos. A acurácia mede o quanto o modelo é preciso na previsão dos voos, independentemente de serem cancelados ou não.
- F1 score: É a média harmônica entre precision e recall. É uma medida balanceada que leva em conta tanto a precisão quanto o recall, o que a torna útil quando ambos são importantes para a avaliação do modelo.

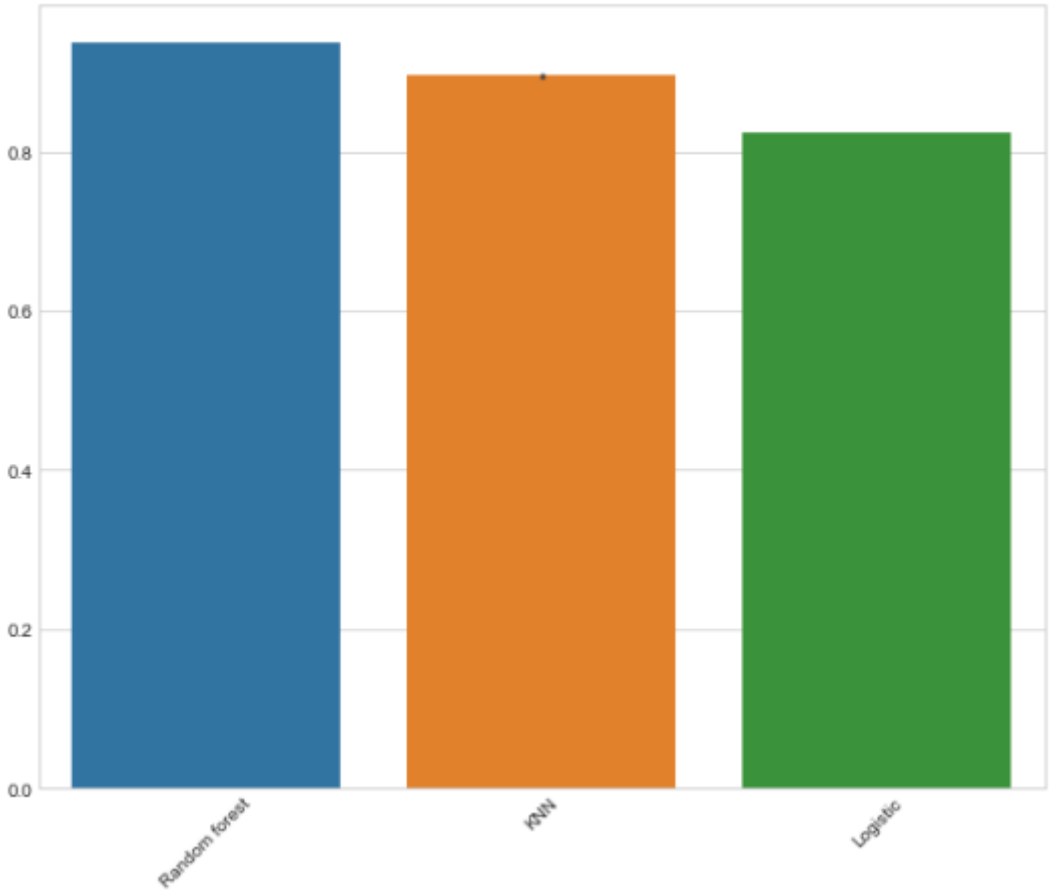
Ao avaliar um modelo de previsão de voos cancelados, é importante considerar tanto o recall quanto a precisão, pois um modelo pode ter uma boa precisão mas baixo recall ou vice-versa. O F1 score pode ser uma boa medida para avaliar o desempenho geral do modelo, já que leva em conta tanto a precisão quanto o recall.

K-Fold Cross Validation

K-Fold Cross-Validation é um método de validação de modelos de aprendizado de máquina que divide os dados em k partições e, em seguida, treina o modelo k vezes, cada vez usando uma dessas partições como conjunto de validação e as outras k-1 partições como conjunto de treinamento. A performance do modelo é então avaliada pela média dos resultados dos k experimentos. Esse método permite uma avaliação mais precisa e robusta do modelo, pois permite que todos os dados sejam usados tanto para treinamento quanto para validação.

	Random forest	KNN	Logistic
0	0.937121	0.895185	0.826812
1	0.937549	0.895602	0.826804
2	0.936644	0.890606	0.822231
3	0.938001	0.898352	0.824508
4	0.936200	0.896092	0.822254
5	0.938904	0.899234	0.821768
6	0.935745	0.893354	0.824095
7	0.937547	0.896069	0.826791
8	0.935749	0.894718	0.822688
9	0.936669	0.889265	0.824985
10	0.933898	0.895612	0.821768
11	0.934367	0.895166	0.822221
12	0.938014	0.899257	0.819518
13	0.935286	0.896098	0.823645
14	0.935743	0.894276	0.824502
15	0.933447	0.896061	0.824049
16	0.937534	0.899701	0.825438
17	0.936187	0.895629	0.819487
18	0.936198	0.892912	0.826351
19	0.934379	0.897464	0.825880
20	0.938933	0.898348	0.825448
21	0.937105	0.891986	0.822254
22	0.937594	0.892457	0.824070
23	0.938925	0.894253	0.822246
24	0.934828	0.897433	0.824969
25	0.935731	0.894724	0.821351
26	0.935293	0.893825	0.822731
27	0.937567	0.897910	0.826816
28	0.937100	0.897904	0.819072
29	0.936206	0.896544	0.828188

No gráfico abaixo avalia-se que o algoritmo Random forest foi o que apresentou melhor aproveitamento:



Média de performance de cada algoritmo de machine learning apresentado:

```
Random forest    0.936482
KNN              0.895535
Logistic         0.823765
```

É possível avaliar esse algoritmo de forma positiva, podendo apresentar um 'diagnóstico artificial' com cerca de 90% de precisão, se determinado voo será cancelado ou não dadas as condições meteorológicas presentes no momento do cancelamento.

7. Links

Link para o vídeo de 5 minutos:

https://youtu.be/oHPsZgWVc3A

Link para o vídeo de 20 minutos:

https://youtu.be/MAJeoAAN5EI

Link para o repositório:

https://github.com/eduardocruzmf/TCC-PUC-MINAS-EDUARDO

Link para os datasets:

https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas/historico-de-voos
https://mesonet.agron.iastate.edu/request/download.phtml