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 dada 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

De vez em quando ocorrem acidentes devido às condições meteorológicas. E por isso é importante estudar a segurança que esses fenômenos perturbam.

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.

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

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
```

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', 'Partida Real', 'Chegada Prevista', 'Chegada Real'

	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.

	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

Encerra-se o requerimento dos bancos de dados dos voos nos 12 meses de 2019.

Ocorre a união dos 12 meses em apenas 1 Dataframe pela função concat do Pandas.

Coleta-se o banco de dados das condições meteorológicas dos 10 aeroportos com maiores quantidades de voos de 2019, banco de dados disponilizado 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

3. Processamento/Tratamento de Dados

Verifica-se o quadro de correlação:

Verifica-se os dados nulos e se retira alguns:

```
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
```

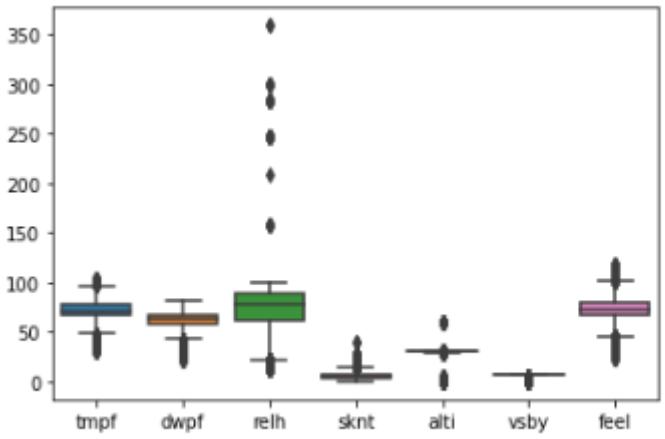
Retira-se algumas amostras com valores nulos para melhor avaliação futura pelo algoritmo de machine learning.

```
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
```

Verifica-se os outliers:



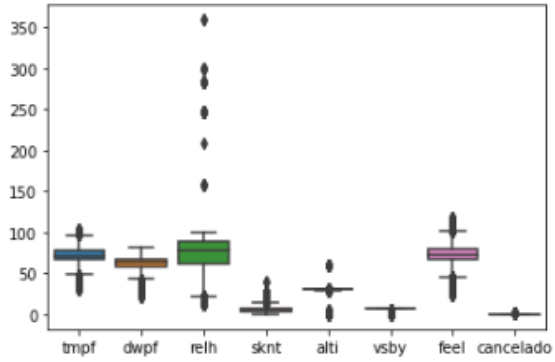
Remove os outliers indesejados:

```
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:

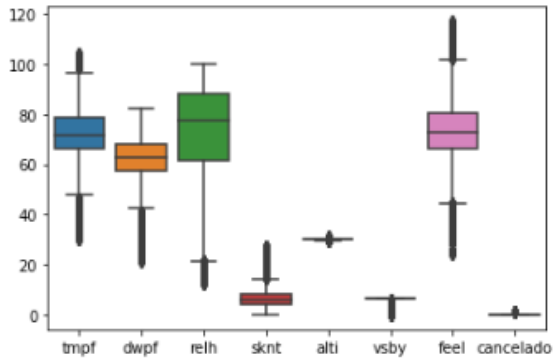
Apresenta o Box-plot de todos os voos nacionais

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



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

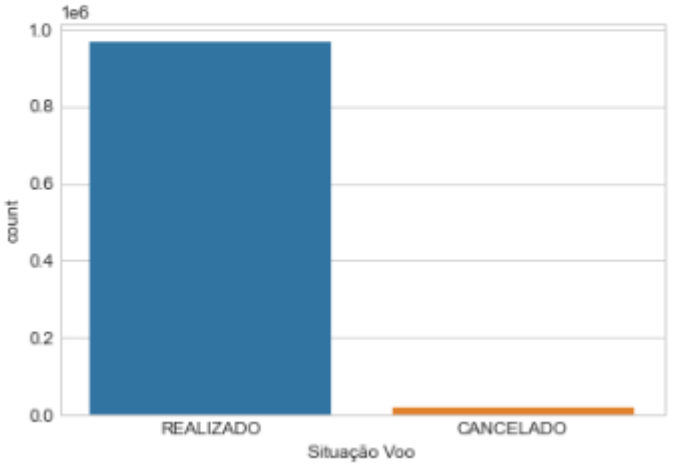
```
In [47]: sns.boxplot(data=dfto)
Out[47]: <AxesSubplot:>
```

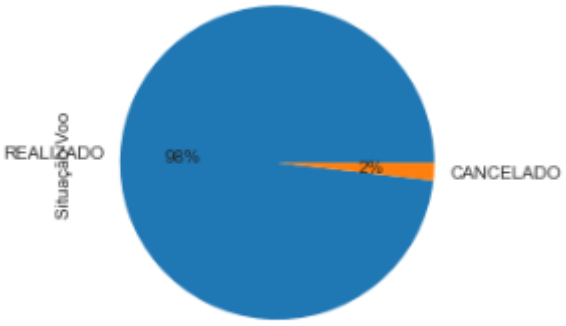


4. Análise e Exploração dos Dados

Verifica-se os voos totais que foram cancelados ou realizados.

	Cancelados_Realizados	N_VOOS
0	REALIZADO	965183
1	CANCELADO	17793





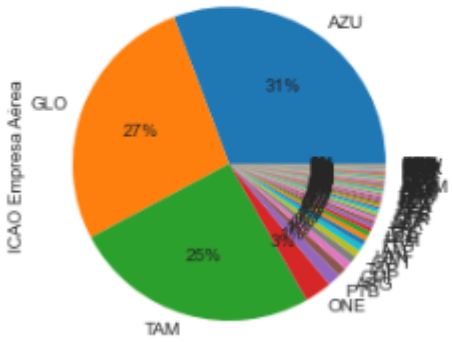
Análise das empresas aéreas por voos totais:

```
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
```

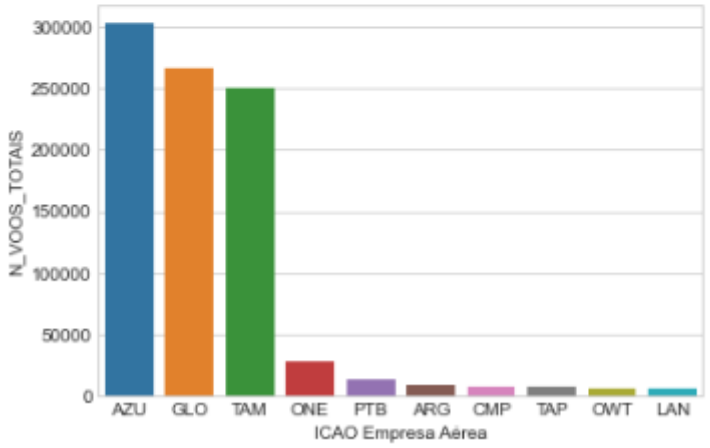
```
dfx['ICAO Empresa Aérea'].value_counts().plot(kind='pie', autopct='%1.0f%%')
```

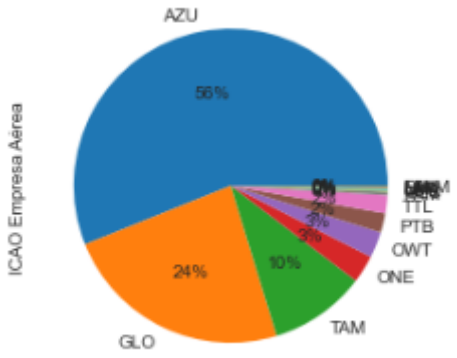
```
<AxesSubplot:ylabel='ICAO Empresa Aérea'>
```



Verifica-se a quantidade de voos das empresas aéreas:

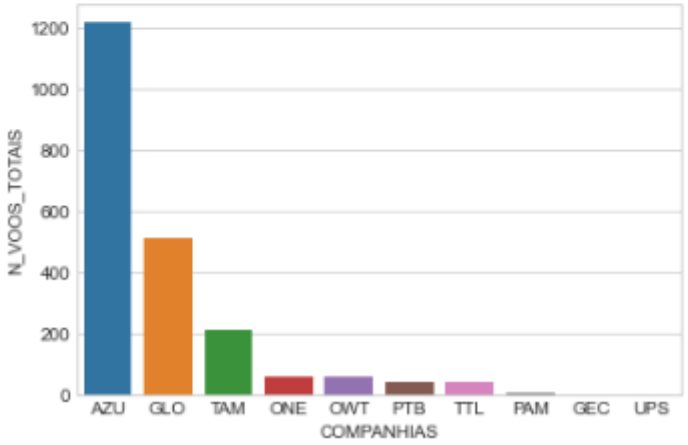
	ICAO Empresa Aérea	N_VOOS_TOTAIS
0	AZU	302316
1	GLO	265582
2	TAM	250403
3	ONE	27589
4	PTB	13363
5	ARG	8885
6	CMP	8087
7	TAP	7750
8	OWT	7007
9	LAN	6509





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

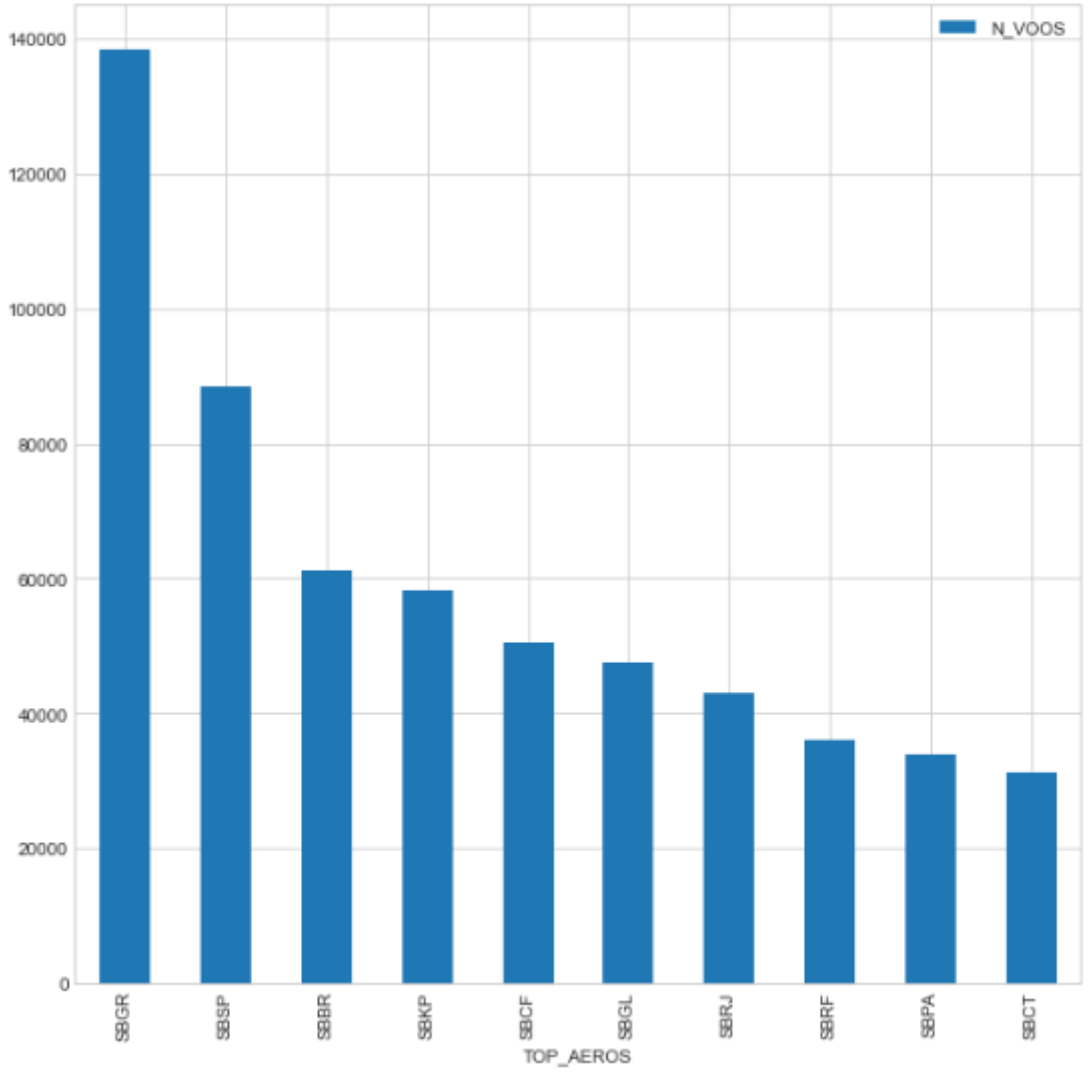
	COMPANHIAS	N_VOOS_TOTAIS
0	AZU	1215
1	GLO	512
2	TAM	215
3	ONE	62
4	OWT	58
5	PTB	43
6	TTL	40
7	PAM	8
8	GEC	3
9	UPS	3



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.

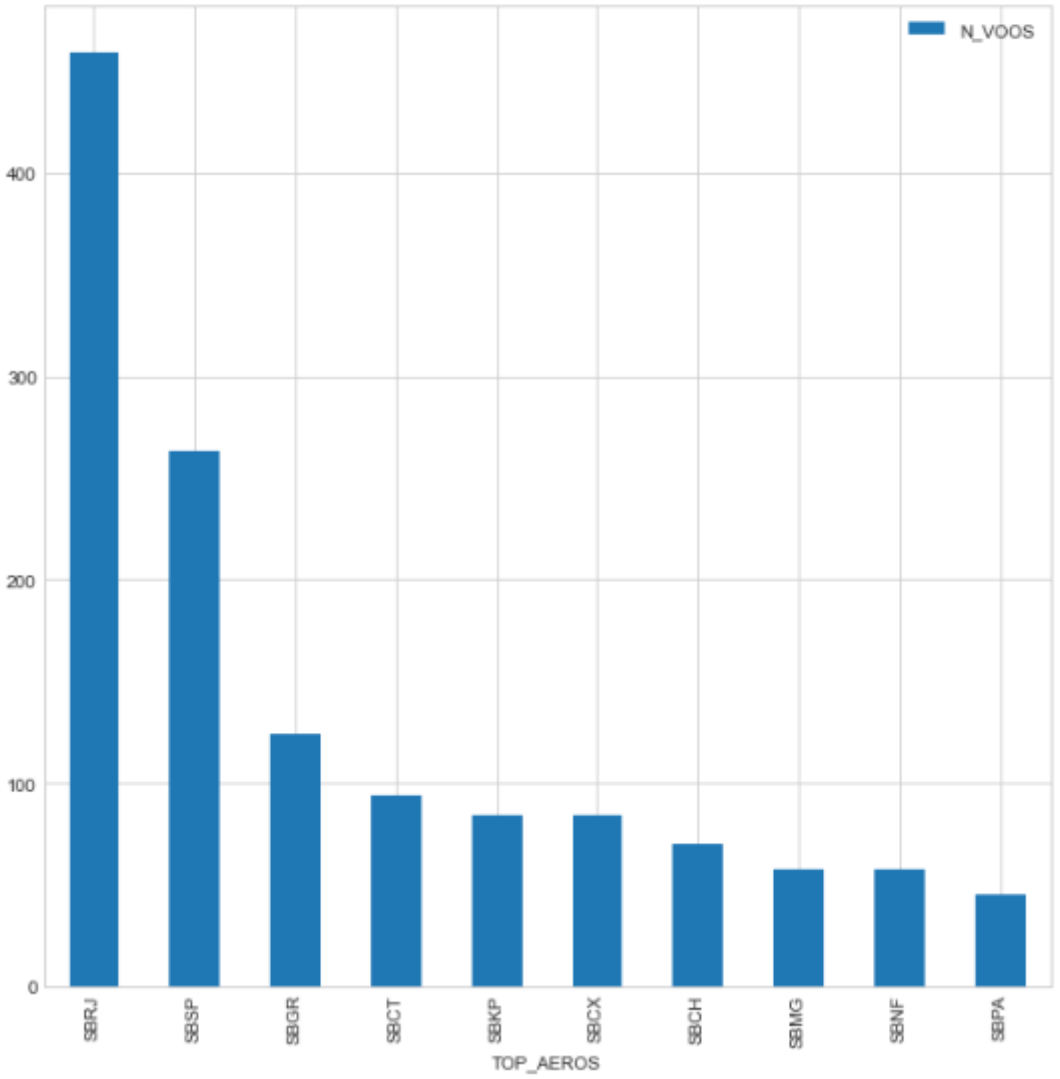
Verifica-se a quantidade de voos por aeroporto:

	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:

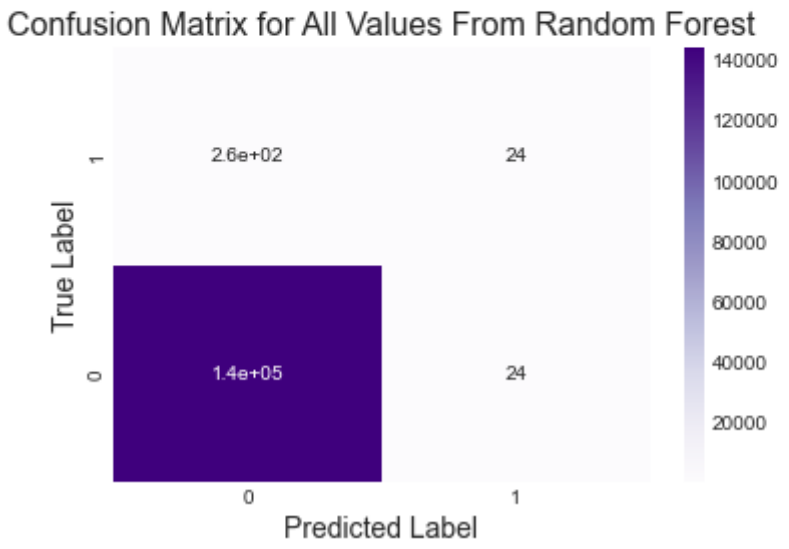
	TOP_AEROS	N_VOOS
0	SBRJ	459
1	SBSP	263
2	SBGR	124
3	SBCT	94
4	SBKP	84
5	SBCX	84
6	SBCH	70
7	SBMG	58
8	SBNF	58
9	SBPA	45



5. Criação de Modelos de Machine Learning

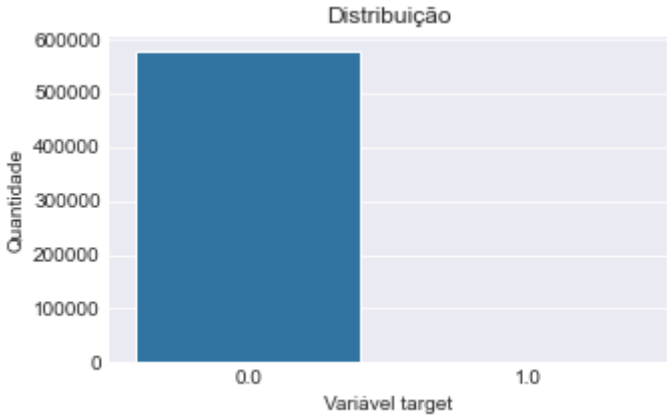
No primeiro modelo testado abaixo, percebe-se que há uma alta porcentagem de acertos. Isso ocorre devido haver uma classe majoritária (classe dos voos REALIZADOS) e uma classe minoritária (classe dos voos CANCELADOS).

```
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')
```

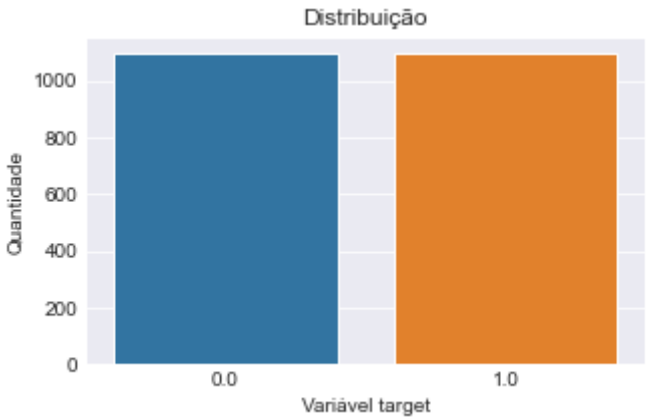


Os modelos de machine learning sempre irão procurar o meio de ter o melhor aproveitamento de performance possível. Logo, é esperado que os modelos classifiquem quase todas as amostras de uma classe majoritária, pois assim, ela terá o melhor resultado possível. Esse resultado não expressa a realidade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

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

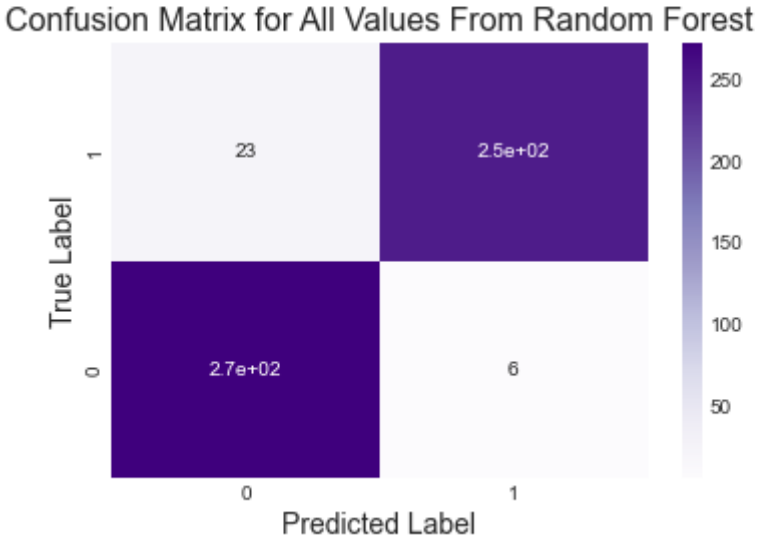


O balanceamento das classes se dá a partir da seleção aleatória de amostras da classe majoritária equivalente a quantidade de amostras da classe minoritária, como é demonstrado no gráfico a seguir:



PERCEBE-SE QUE O ALGORITMO IDENTIFICA MUITAS AMOSTRAS DA CLASSE COVID (CLASSE MAJORITÁRIA) E POUCA DA CLASSE INFLUENZA (CLASSE MINORITÁRIA)

```
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')
```



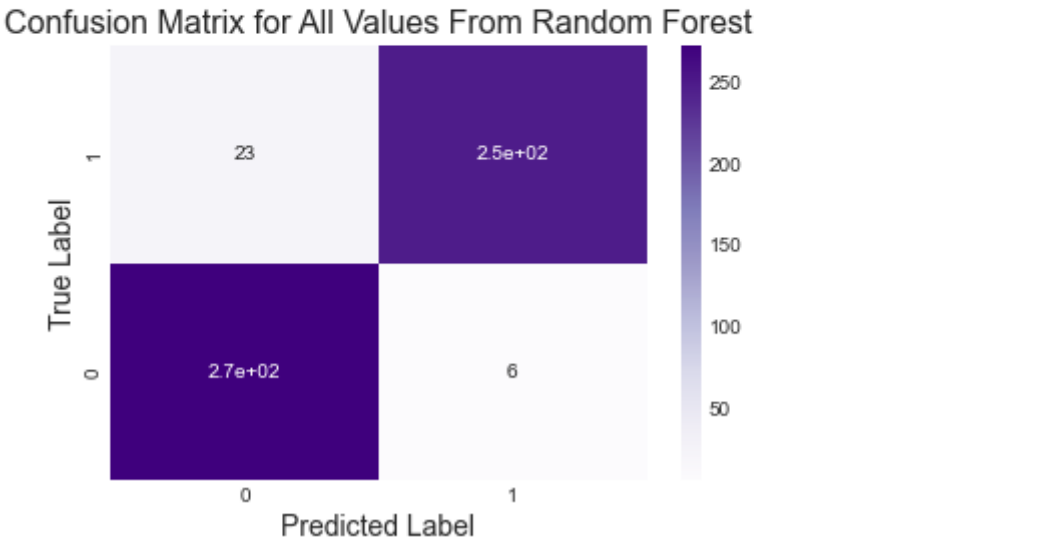
PARA CORREÇÃO DESSE FATO É NECESSÁRIO FAZER O BALANCEAMENTO DE CLASSES E PARA ISSO FOI UTILIZADO A FUNÇÃO NEARMISS PELA BIBLIOTECA IMBLEARN.

Para a correção dessa situação é necessário fazer o balanceamento das classes. Para isso foi utilizado a função NearMiss da biblioteca Imblearn.

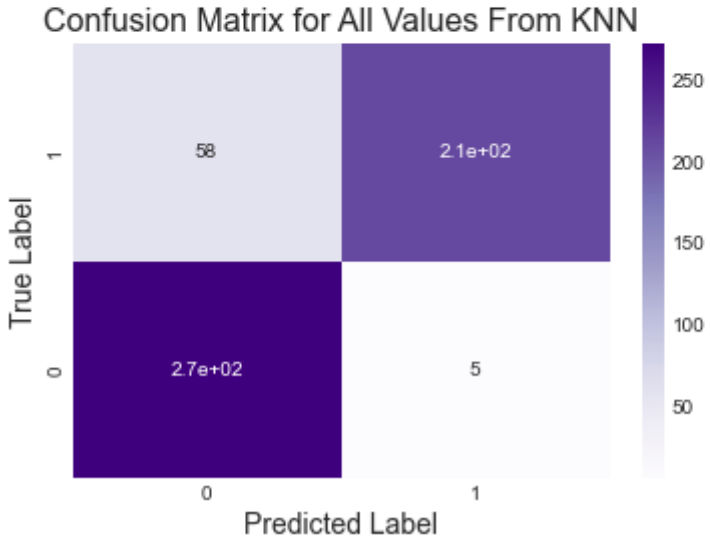
6. Apresentação dos Resultados

Roda-se os modelos de machine learning com o banco de dados final, limpo e balanceado.

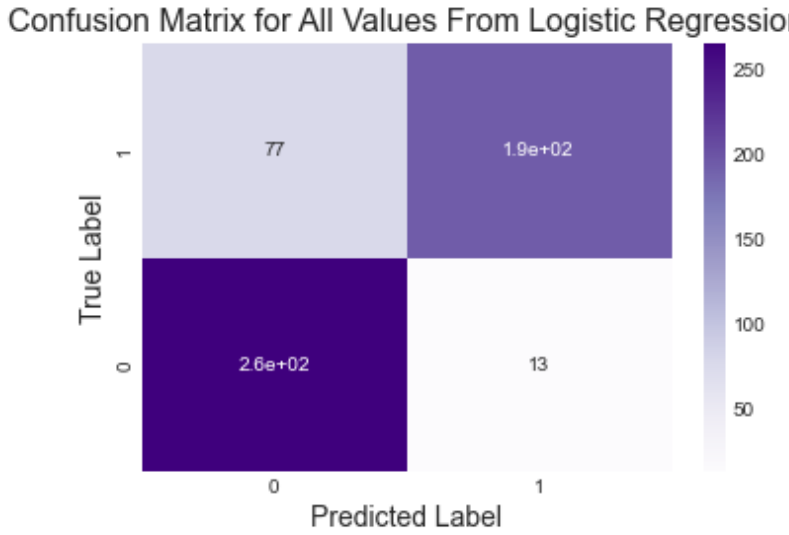
```
Training Recall Score, Random Forest:  0.8983050847457628
Test Recall Score, Random Forest:  0.915129151291513
Training Precision Score, Random Forest:  0.9867021276595744
Test Precision Score, Random Forest:  0.9763779527559056
Training Accuracy Score, Random Forest:  0.9428571428571428
Test Accuracy 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')
```



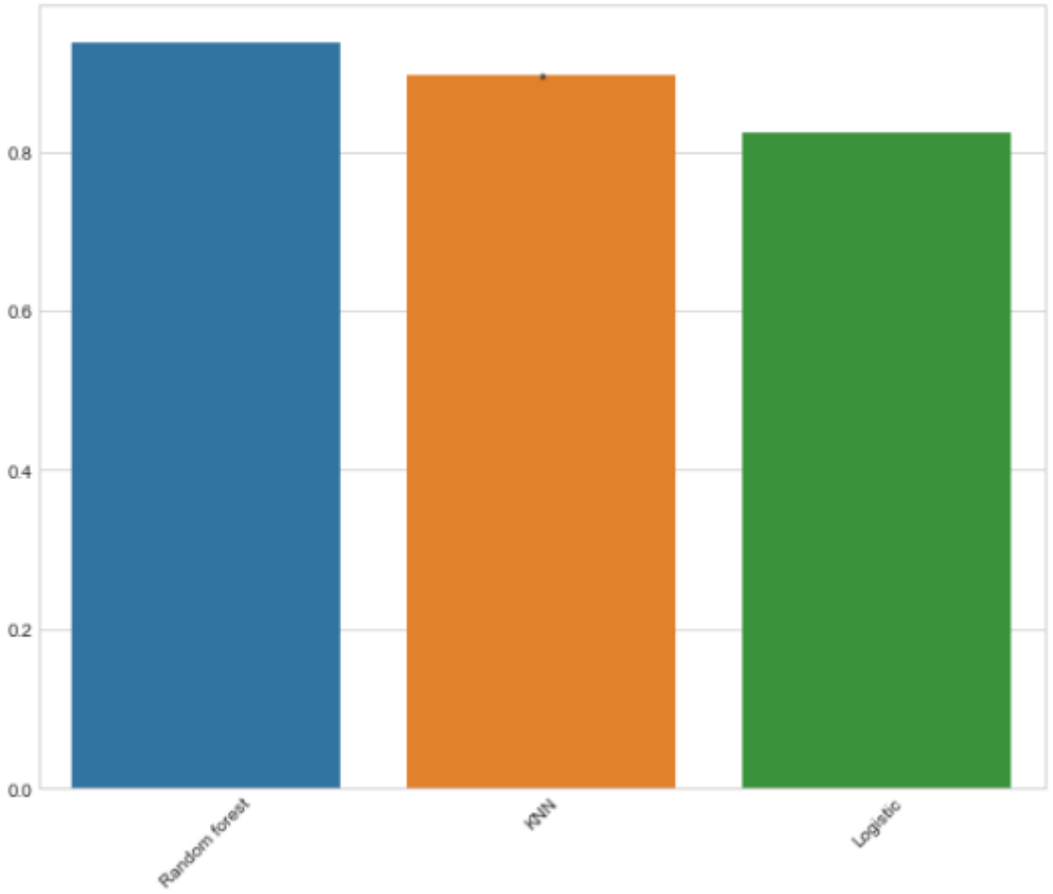
```
Training Recall Score, Logistic Regression:  0.7191283292978208
Test Recall Score, Logistic Regression:  0.7158671586715867
Training Precision Score, Logistic Regression:  0.9369085173501577
Test Precision Score, Logistic Regression:  0.9371980676328503
Training Accuracy Score, Logistic Regression:  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')
```



K-Fold Cross Validation

	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