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

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Pós-graduação Lato Sensu em Ciência de Dados e Big Data

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.

Importação das bibliotecas

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.

```
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
```

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

	ICAO Empresa Aérea	ICAO Aeródromo Origem	Partida Prevista	Situação Voo	Código Justificativa
0	AAF	LFPO	25/01/2019 06:15	REALIZADO	NaN
1	AAF	LFPO	27/01/2019 06:15	REALIZADO	NaN
2	AAF	LFPO	29/01/2019 06:15	REALIZADO	NaN
3	AAF	SBKP	25/01/2019 20:15	REALIZADO	NaN
4	AAF	SBKP	27/01/2019 20:15	REALIZADO	NaN
...	...	...	...	...	...
88212	GLO	SBSP	25/12/2019 10:45	REALIZADO	NaN
88213	GLO	SBPL	31/12/2019 02:45	REALIZADO	NaN
88214	GLO	SBNF	06/12/2019 18:55	REALIZADO	RI
88215	GLO	SBRJ	10/12/2019 17:00	REALIZADO	NaN
88216	TAM	SBSP	04/12/2019 07:55	REALIZADO	NaN

982976 rows × 5 columns

Coleta de dados

Os dados utilizados 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

	station	Date	tmpf	dwpf	relh	drct	sknt	alti	vsby	gust	skyc1	skyc2	skyc3	skyc4	skyl1	skyl2	skyl3	skyl4	wxcodes	feel
0	SBCT	2019-01-01 00:00:00	69.80	66.20	88.34	110.00	7.00	30.06	6.21	NaN	BKN	BKN	NaN	NaN	1300.00	2000.00	NaN	NaN	NaN	69.80
1	SBCF	2019-01-01 00:00:00	69.80	64.40	82.98	50.00	6.00	30.03	6.21	NaN	FEW	SCT	NaN	NaN	4000.00	8000.00	NaN	NaN	VCTS	69.80
2	SBBR	2019-01-01 00:00:00	68.00	64.40	88.26	NaN	2.00	30.06	6.21	NaN	FEW	OVC	NaN	NaN	1500.00	10000.00	NaN	NaN	-RA	68.00
3	SBGL	2019-01-01 00:00:00	78.80	69.80	73.95	310.00	5.00	29.88	6.21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	81.39
4	SBGR	2019-01-01 00:00:00	73.40	66.20	78.19	120.00	6.00	30.03	6.21	NaN	SCT	NaN	NaN	NaN	2000.00	NaN	NaN	NaN	NaN	73.40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
87101	SBCT	2019-12-30 23:00:00	68.00	68.00	100.00	90.00	7.00	29.94	6.21	NaN	FEW	NaN	NaN	NaN	900.00	NaN	NaN	NaN	NaN	68.00
87102	SBGL	2019-12-30 23:00:00	78.80	69.80	73.95	80.00	11.00	29.83	6.21	NaN	FEW	FEW	NaN	NaN	2000.00	3000.00	NaN	NaN	NaN	78.80
87103	SBKP	2019-12-30 23:00:00	75.20	62.60	64.91	90.00	10.00	29.85	6.21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	75.20
87104	SBGR	2019-12-30 23:00:00	71.60	66.20	83.09	150.00	4.00	29.94	6.21	NaN	FEW	NaN	NaN	NaN	2000.00	NaN	NaN	NaN	NaN	71.60
87105	SBPA	2019-12-30 23:00:00	84.20	71.60	65.95	130.00	7.00	29.74	4.97	NaN	NSC	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	89.60

87106 rows × 20 columns

Processamento/Tratamento de Dados

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.

```
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.

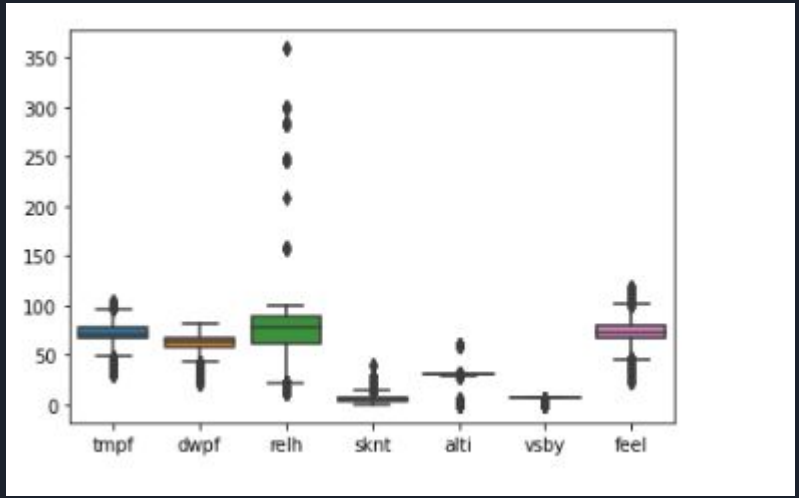
```
dft.dropna(subset=['relh'], how='all', inplace=True)
dft.dropna(subset=['alti'], how='all', inplace=True)
dft.dropna(subset=['vsby'], how='all', inplace=True)
dft.dropna(subset=['sknt'], how='all', inplace=True)
```

```
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
```

Processamento/Tratamento de Dados

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.



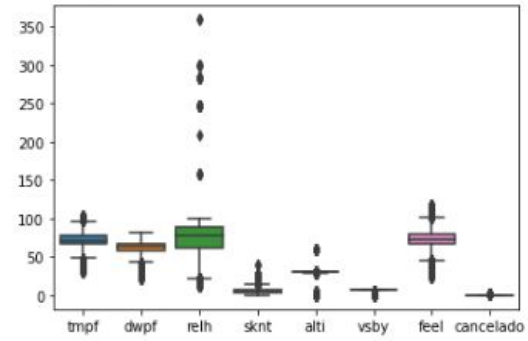

```
dfto = dft.drop(dft[dft.relh > 120].index)
dfto = dfto.drop(dfto[dfto.sknt > 32].index)
dfto = dfto.drop(dfto[dfto.alti < 20].index)
dfto = dfto.drop(dfto[dfto.alti > 50].index)
dfto
```

Processamento

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.

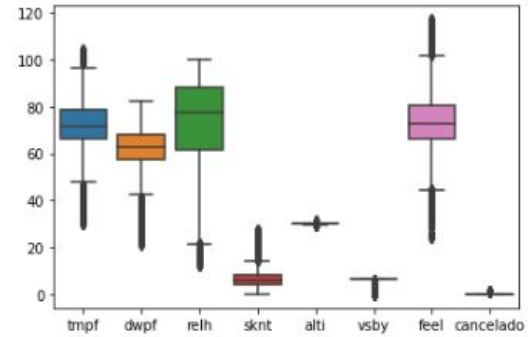
Apresenta o Box-plot de todos os voos nacionais

```
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:>
```



Processamento/Tratamento de Dados

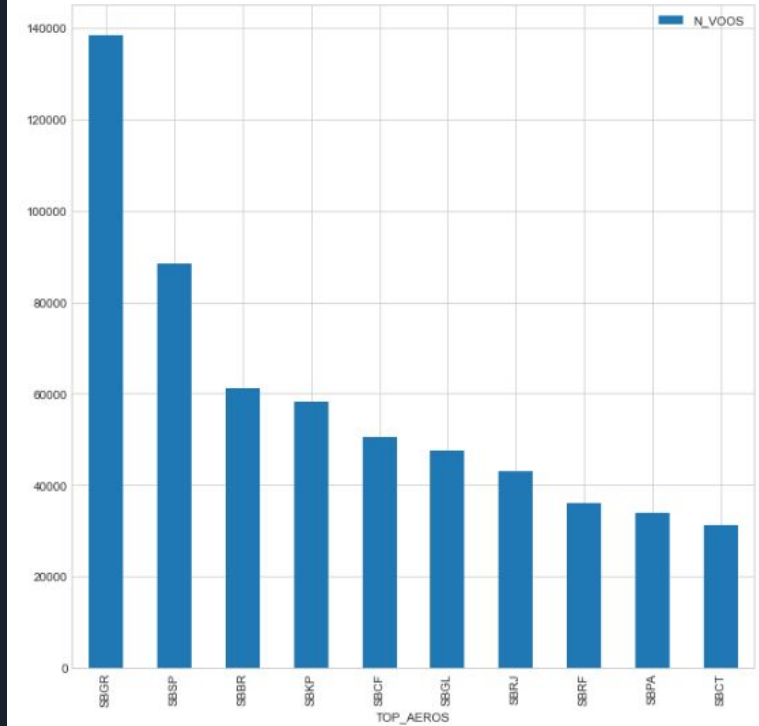
Analisa-se a disparidade de casos de voos cancelados e realizados

	Cancelados_Realizados	N_VOOS
0	REALIZADO	965183
1	CANCELADO	17793

Análise e Exploração dos D

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.

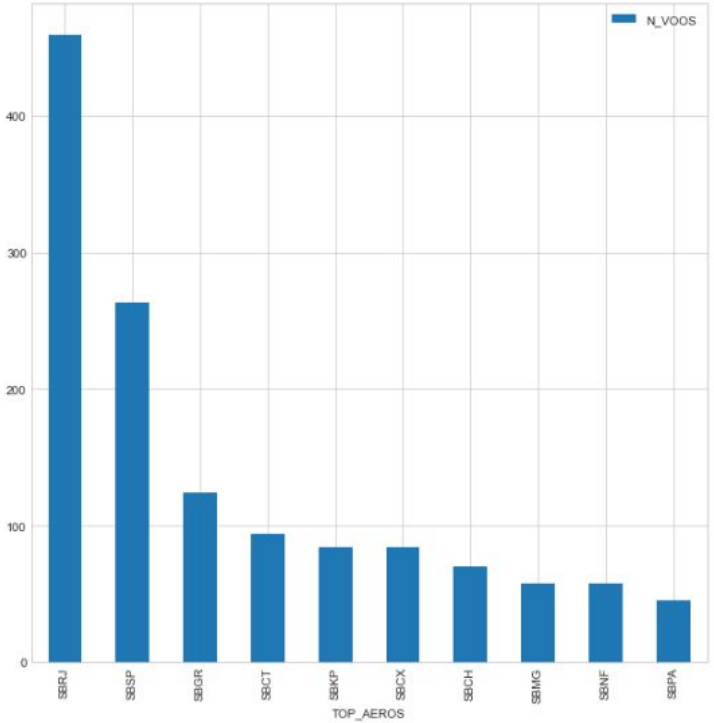
	TOP_AEROS	N_VOOS
0	SBGR	138254
1	SBSP	88504
2	SBBR	61260
3	SBKP	58154
4	SBCF	50509
5	SBGL	47696
6	SBRJ	43131
7	SBRF	36068
8	SBPA	33939
9	SBCT	31347



Análise e Exploração dos Da

Apresenta-se os voos cancelados nos 10 aeroportos de maior movimento

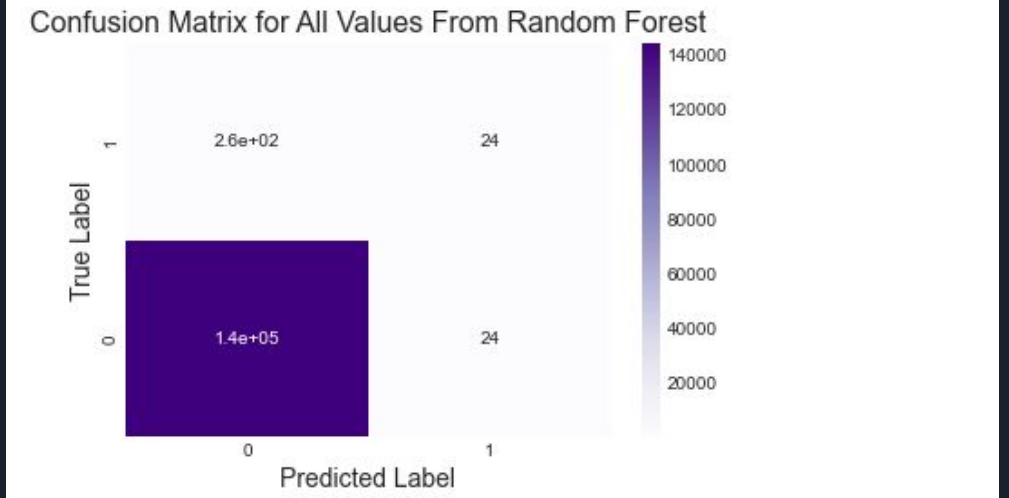
	TOP_AEROS	N_VOOS
0	SBRJ	459
1	SBSP	263
2	SBGR	124
3	SBCT	94
4	SBKP	84
5	SBCX	84
6	SBCH	70
7	SBMG	58
8	SBNF	58
9	SBPA	45



Criação de Modelos de Machine Learning

No primeiro modelo testado, 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')
```



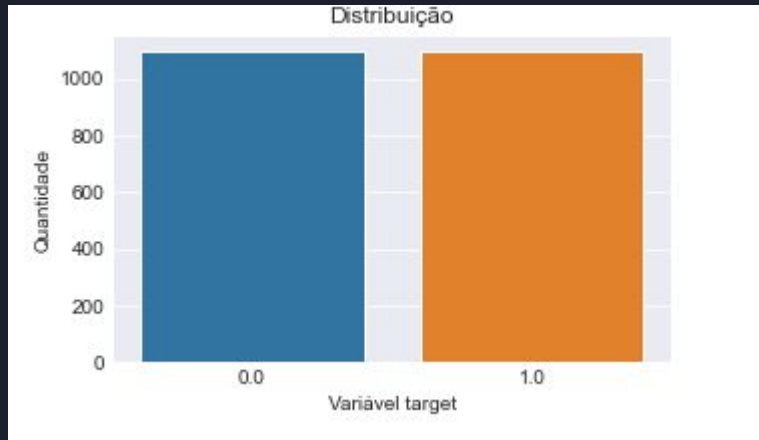
Processamento/Tratamento de Dados

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.

	Cancelados_Realizados	N_VOOS
0	REALIZADO	965183
1	CANCELADO	17793

Processamento/Tratamento de Dados

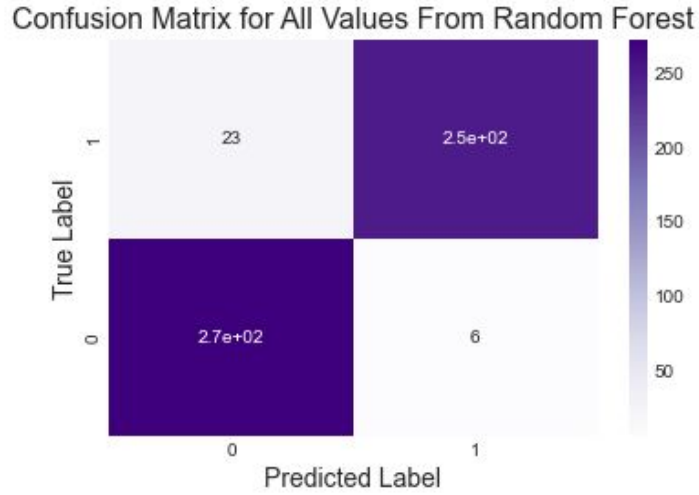
O undersampling é uma técnica de balanceamento de classes que consiste em remover exemplos da classe dominante para tornar a distribuição de classes mais equilibrada. No caso da previsão de voos cancelados, isso significa remover exemplos de voos não cancelados para tornar a distribuição de voos cancelados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



Criação de Modelos de Machine Learning

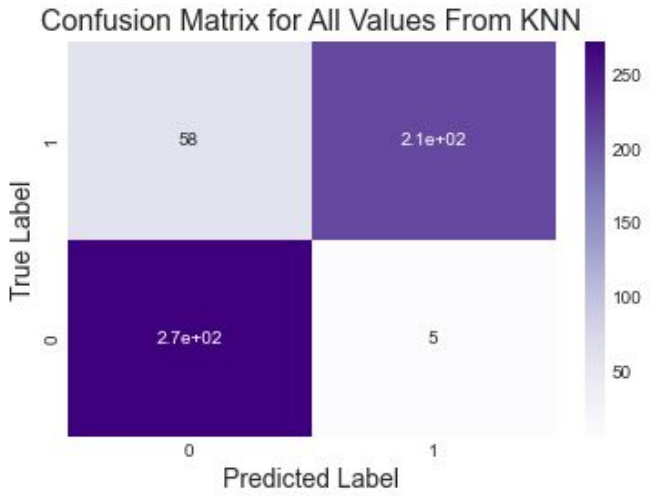
Roda-se os modelos de machine learning com o banco de dados final, limpo e balanceado. Utiliza-se os modelos de Random Forest, KNN (K-Nearest Neighbors) e Logistic Regression, 3 algoritmos de machine learning bem conhecidos

```
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')
```



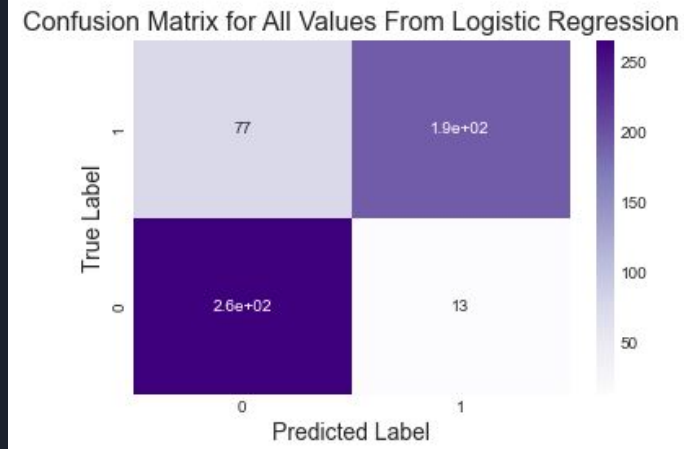
Criação de Modelos de Machine Learning

```
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')
```



Criação de Modelos de Machine Learning

```
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')
```



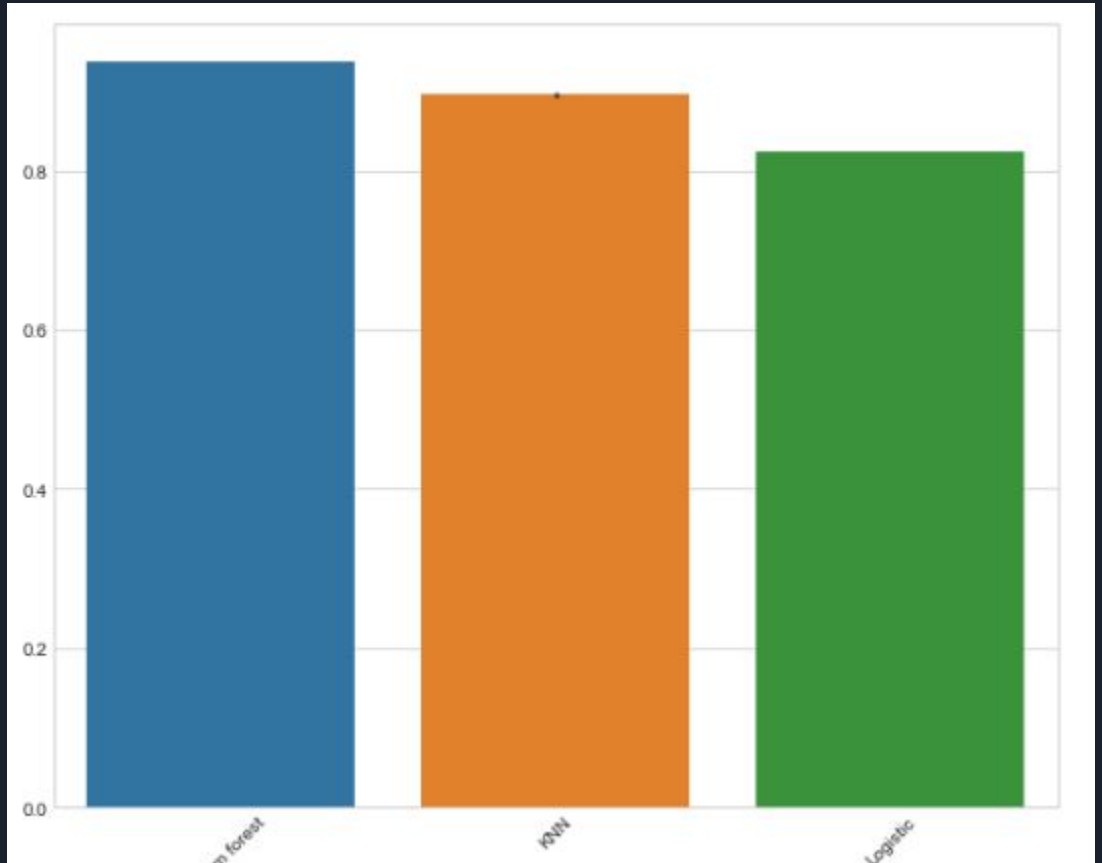
Apresentação dos Resultados

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

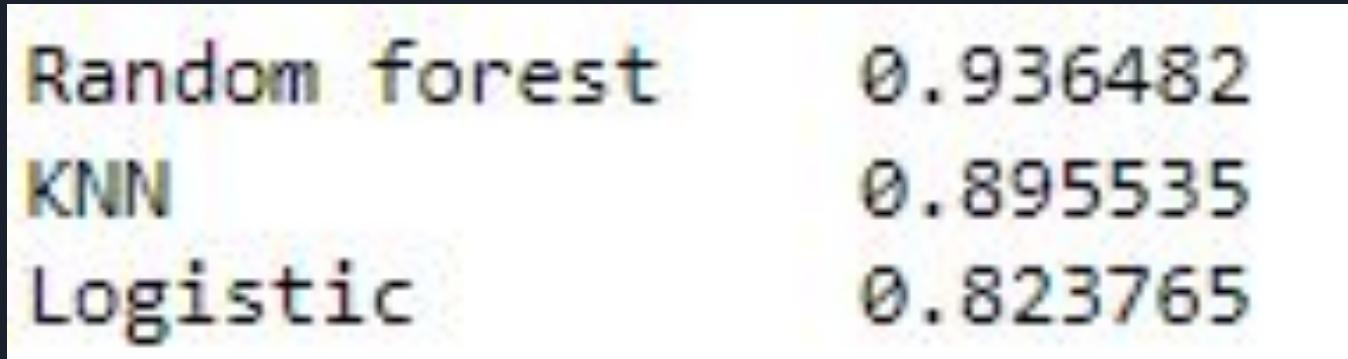
Apresentação dos Resultados

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



Conclusão

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.



Links

Link para o vídeo de 5 minutos: https://youtu.be/oHPsZgWVc3A

Link para o vídeo de 20 minutos: https://youtu.be/GOD33CiFMHQ

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