# Predição de Anormalidades em Voos Utilizando a Previsão do Tempo dos Aeroportos do Brasil

Gustavo Ramos Lima

2019

## Resumo

Projeto de conclusão do curso *Machine Learning Engineer Nanodegree* que tem como objetivo desenvolver um modelo de classificador capaz de prever a ocorrência de cancelamentos ou atrasos nos voos a partir da previsão do tempo (dos aeroportos) e o horário dos voos. Para treinar o classificador foram utilizados os registros de voos e previsões do tempo de 10 aeroportos de grande importância no Brasil.

**Palavras-chaves**: aprendizagem supervisionada. classificação. aeroportos. previsão do tempo.

## Introdução

### Histórico do Assunto

Os cancelamentos e atrasos de voo ocorrem todos os dias nos aeroportos de todo o mundo. Eles são provocados devido a diversos fatores como problemas técnicos nas aeronaves, conexões atrasadas, excesso de tráfego aéreo, voos com baixa ocupação e principalmente problemas meteorológicos. Segundo a ANAC (2017) devem ser realizados uma série de procedimentos para averiguar as condições meteorológicas dos aeródromos de origem e destino:

- Consultar METAR, SPECI, TAF, SIGMET, AIRMET, GAMET, verificar condições de vento, visibilidade, teto, nebulosidade, temperaturas e precipitação;
- Atentar para os mínimos meteorológicos para operação VFR e IFR estabelecidos nos regulamentos e nas cartas dos aeródromos;
- Verificar se o vento presente e o vento previsto enquadram-se dentro das limitações do manual da aeronave;
- Verificar registros e previsões do tempo para a rota a ser voada, consultando imagens de satélites, cartas SIGWX e cartas de vento em altitude;
- Verificar se há condições meteorológicas adversas previstas (formação de gelo, nevoeiro, chuva forte, tesoura de vento, rajadas de vento, turbulência);

- Verificar condições para retorno ou pouso de precaução em caso de ambiente visual degradado (velocidade e altura mínima para helicópteros voando VFR).

Analisar todas essas informações pode se tornar uma tarefa cansativa e demorada. Desta forma pretende-se criar uma ferramenta de auxílio, que analisa a possibilidade de cancelamentos ou atrasos de voos baseando-se em um histórico dados meteorológicos e de voos. Com conhecimentos de *machine learning* é possível automatizar o processo de análise dos dados facilitando a tomada de decisão. Existem diversos fatores além das condições climáticas que podem causar os cancelamentos e atrasos nos voos (STERNBERG et al., 2017). Esse projeto estará focado apenas na relação entre as previsões do tempo e os cancelamentos/atrasos dos voos.

### Descrição do problema

O objetivo deste trabalho é prever se um voo pode ser cancelado ou sofrer atrasos baseando-se nos históricos de dados meteorológicos e de voos de aeroportos do Brasil. O conjunto de dados foi rotulado em duas categorias: uma contendo voos que sofreram cancelamentos/atrasos e outra com os voos que ocorreram normalmente. Foram analisados os desempenhos de diversos algoritmos de classificação (SVM, Métodos *Ensemble*, etc).

## 1 Base de Dados

### 1.1 Conjuntos de dados e entradas

Neste trabalho foram utilizadas informações sobre os voos e as previsões do tempo durante o ano de 2018 no Brasil. Do site da ANAC (2016) foram obtidos dados sobre os voos de janeiro a dezembro de todos os aeroportos do país. A base de dados consiste em uma planilha com dados em formato *.csv* (Figura 1). Dela foram extraídos os registros de voos nos 10 maiores aeroportos do Brasil, indicados na Tabela 1.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 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 |
| 2 | AAL | 213 | 0 | I | KMIA | SBBR | 01/12/2018 02:15 | 01/12/2018 03:11 | 01/12/2018 09:50 | 01/12/2018 10:35 | REALIZADO | MX |
| 3 | AAL | 213 | 0 | I | KMIA | SBBR | 02/12/2018 02:15 | 02/12/2018 02:15 | 02/12/2018 09:50 | 02/12/2018 09:50 | REALIZADO | |
| 4 | AAL | 213 | 0 | I | KMIA | SBBR | 03/12/2018 02:15 | 03/12/2018 02:15 | 03/12/2018 09:50 | 03/12/2018 09:50 | REALIZADO | |
| 5 | AAL | 213 | 0 | I | KMIA | SBBR | 04/12/2018 02:15 | 04/12/2018 02:15 | 04/12/2018 09:50 | 04/12/2018 09:50 | REALIZADO | |
| 6 | AAL | 213 | 0 | I | KMIA | SBBR | 05/12/2018 02:15 | 05/12/2018 02:15 | 05/12/2018 09:50 | 05/12/2018 09:50 | REALIZADO | |
| 7 | AAL | 213 | 0 | I | KMIA | SBBR | 06/12/2018 02:15 | 06/12/2018 02:15 | 06/12/2018 09:50 | 06/12/2018 09:50 | REALIZADO | |
| 8 | AAL | 213 | 0 | I | KMIA | SBBR | 07/12/2018 02:15 | 07/12/2018 03:08 | 07/12/2018 09:50 | 07/12/2018 10:38 | REALIZADO | MX |
| 9 | AAL | 213 | 0 | I | KMIA | SBBR | 08/12/2018 02:15 | 08/12/2018 02:15 | 08/12/2018 09:50 | 08/12/2018 09:50 | REALIZADO | |
| 10 | AAL | 213 | 0 | I | KMIA | SBBR | 09/12/2018 02:15 | 09/12/2018 02:15 | 09/12/2018 09:50 | 09/12/2018 09:50 | REALIZADO | |
| 11 | AAL | 213 | 0 | I | KMIA | SBBR | 10/12/2018 02:15 | 10/12/2018 02:15 | 10/12/2018 09:50 | 10/12/2018 09:50 | REALIZADO | |
| 12 | AAL | 213 | 0 | I | KMIA | SBBR | 11/12/2018 02:15 | 11/12/2018 02:15 | 11/12/2018 09:50 | 11/12/2018 09:50 | REALIZADO | |
| 13 | AAL | 213 | 0 | I | KMIA | SBBR | 12/12/2018 02:15 | 12/12/2018 02:15 | 12/12/2018 09:50 | 12/12/2018 09:50 | REALIZADO | |
| 14 | AAL | 213 | 0 | I | KMIA | SBBR | 13/12/2018 02:15 | 13/12/2018 02:15 | 13/12/2018 09:50 | 13/12/2018 09:50 | REALIZADO | |
| 15 | AAL | 213 | 0 | I | KMIA | SBBR | 14/12/2018 02:15 | 14/12/2018 02:15 | 14/12/2018 09:50 | 14/12/2018 09:50 | REALIZADO | |
| 16 | AAL | 213 | 0 | I | KMIA | SBBR | 15/12/2018 02:15 | 15/12/2018 02:15 | 15/12/2018 09:50 | 15/12/2018 09:50 | REALIZADO | |
| 17 | AAL | 213 | 0 | I | KMIA | SBBR | 16/12/2018 02:15 | 16/12/2018 02:01 | 16/12/2018 09:50 | 16/12/2018 09:08 | REALIZADO | AT |
| 18 | AAL | 213 | 0 | I | KMIA | SBBR | 17/12/2018 02:15 | 17/12/2018 02:13 | 17/12/2018 09:50 | 17/12/2018 09:13 | REALIZADO | AT |
| 19 | AAL | 213 | 0 | I | KMIA | SBBR | 18/12/2018 02:15 | 18/12/2018 02:12 | 18/12/2018 09:50 | 18/12/2018 09:18 | REALIZADO | AT |

Figura 1 – Exemplo de planilha contendo o histórico de voos

Foram utilizados os dados das seguintes colunas:

- ICAO Aeródromo de origem: o ICAO é um código de 4 letras que identifica os aeroportos;
- Código de justificativa: identifica situações associadas aos voos segundo a Instrução de Aviação Civil (DAC, 2000). Os códigos associados a problemas meteorológicos são apresentados na Tabela 2. Caso este campo esteja em branco significa que o voo ocorreu sem imprevistos;

- Partida Prevista: horário previsto para os voos.

| ICAO | Descrição |
|---|---|
| SBBR | Aeroporto Internacional de Brasília / Presidente Jucelino Kubitschek |
| SBCF | Aeroporto Internacional de Minas Gerais / Confins - Tancredo Neves |
| SBGL | Aeroporto Internacional do Rio de Janeiro / Galeão - Antônio carlos Jobim |
| SBGR | Aeroporto Internacional de São Paulo / Guarulhos - Governador André Franco Motoro |
| SBKP | Aeroporto Internacional de Viracopos / Campinas |
| SBPA | Aeroporto Internacional de Porto Alegre / Salgado Filho |
| SBRF | Aeroporto Internacional do Recife / Guararapes - Gilberto Freyre |
| SBRJ | Aeroporto Santos Dumont |
| SBSP | Aeroporto de São Paulo / Congonhas |
| SBSV | Aeroporto Internacional de Salvador / Deputado Luis Eduardo Magalhães |

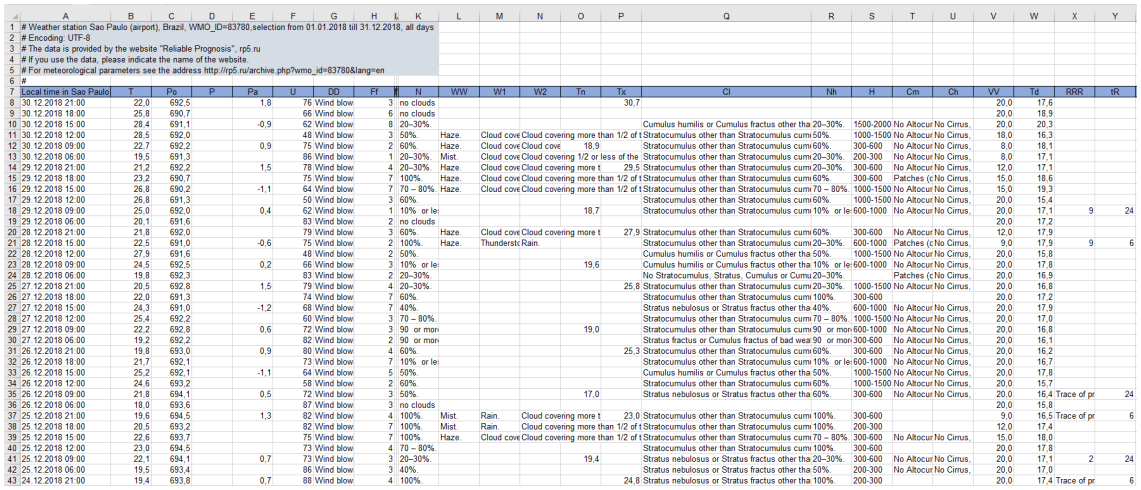
Tabela 1 – Código ICAO dos aeroportos

| Código | Justificativa |
|---|---|
| AM | ATRASO AEROPORTO DE ALTERNATIVA – CONDIÇÕES METEOROLÓGICAS |
| RM | CONEXÃO AERONAVE/VOLTA – VÔO DE IDA NÃO PENALIZADO CONDIÇÕES METEOROLÓGICAS |
| WR | ATRASO DEVIDO RETORNO – CONDIÇÕES METEOROLÓGICAS |
| XS | CANCELAMENTO – CONEXÃO AERONAVE/VOLTA – VÔO DE IDA CANCELADO – COND. METEOR. |

Tabela 2 – Código de justificativa dos voos

Como a base de dados fornecida pela ANAC não continha informações meteorológicas associadas aos voos, foi necessário buscar o histórico de previsão do tempo nos aeroportos (POGODI, 2019). Com a data e o horário previstos dos voos foi possível consultar o histórico de previsão do tempo para cada voo. Os dados foram obtidos, para cada aeroporto, no formato *.csv*, conforme a Figura 2. Foram utilizados os dados das seguintes colunas:

- Data / Hora local: data e hora no formato dd.mm.aaaa hh:mm;
- T: temperatura em graus Celsius;
- Po: pressão atmosférica em milímetros de mercúrio;
- U: percentual de umidade relativa do ar;
- DD: direção do vento (norte, nordeste, sul, etc);
- Ff: velocidade do vento em metros por segundo;
- N: percentual de nebulosidade geral;
- VV: distância de visibilidade horizontal;
- Td: temperatura de condensação em graus Celsius.



| | A | B | C | D | E | F | G | H | I | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | # Weather station Sao Paulo (airport), Brazil, WMO_ID=83780,selection from 01.01.2018 till 31.12.2018, all days | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | # Encoding: UTF-8 | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | # The data is provided by the website "Reliable Prognosis", rp5.ru | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | # If you use the data, please indicate the name of the website. | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | # For meteorological parameters see the address http://rp5.ru/archive.php?wmo_id=83780&lang=en | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | # | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | Local time in Sao Paulo | T | Po | P | Pa | U | DD | Ff | ff | N | WW | W1 | W2 | Tn | Tx | Cl | Nh | H | Cm | Ch | VV | Td | RRR | tR |
| 8 | 30.12.2018 21:00 | 22,0 | 692,5 | | 1,8 | 76 | Wind blow | 3 | | no clouds | | | | | 30,7 | | | | | | 20,0 | 17,6 | | |
| 9 | 30.12.2018 18:00 | 25,8 | 690,7 | | | 66 | Wind blow | 6 | | no clouds | | | | | | | | | | | 20,0 | 18,9 | | |
| 10 | 30.12.2018 15:00 | 28,4 | 691,1 | | -0,9 | 62 | Wind blow | 8 | | 20–30% | | | | | | Cumulus humilis or Cumulus fractus other tha | 20–30% | 1500-2000 | No Altocur | No Cirrus, | 20,0 | 20,3 | | |
| 11 | 30.12.2018 12:00 | 28,5 | 692,0 | | | 48 | Wind blow | 3 | | 50% | Haze. | Cloud cove | Cloud covering more than 1/2 of t | | | Stratocumulus other than Stratocumulus cumi | 50% | 1000-1500 | No Altocur | No Cirrus, | 18,0 | 16,3 | | |
| 12 | 30.12.2018 09:00 | 22,7 | 692,2 | | 0,9 | 75 | Wind blow | 2 | | 60% | Haze. | Cloud cove | Cloud cove | 18,9 | | Stratocumulus other than Stratocumulus cumi | 60% | 300-600 | No Altocur | No Cirrus, | 8,0 | 18,1 | | |
| 13 | 30.12.2018 06:00 | 19,5 | 691,3 | | | 86 | Wind blow | 1 | | 20–30% | Mist. | Cloud cove | Cloud covering 1/2 or less of the | | | Stratocumulus other than Stratocumulus cumi | 20–30% | 200-300 | No Altocur | No Cirrus, | 8,0 | 17,1 | | |
| 14 | 29.12.2018 21:00 | 21,2 | 692,2 | | 1,5 | 78 | Wind blow | 4 | | 20–30% | Haze. | Cloud cove | Cloud covering more t | | 29,5 | Stratocumulus other than Stratocumulus cumi | 20–30% | 300-600 | No Altocur | No Cirrus, | 12,0 | 17,1 | | |
| 15 | 29.12.2018 18:00 | 23,2 | 690,7 | | | 75 | Wind blow | 7 | | 100% | Haze. | Cloud cove | Cloud covering more than 1/2 of t | | | Stratocumulus other than Stratocumulus cumi | 60% | 300-600 | Patches (c | No Cirrus, | 15,0 | 18,6 | | |
| 16 | 29.12.2018 15:00 | 26,8 | 690,2 | | -1,1 | 64 | Wind blow | 7 | | 70 – 80% | Haze. | Cloud cove | Cloud covering more than 1/2 of t | | | Stratocumulus other than Stratocumulus cumi | 70 – 80% | 1000-1500 | No Altocur | No Cirrus, | 15,0 | 19,3 | | |
| 17 | 29.12.2018 12:00 | 26,8 | 691,3 | | | 50 | Wind blow | 3 | | 60% | | | | | | Stratocumulus other than Stratocumulus cumi | 60% | 1000-1500 | No Altocur | No Cirrus, | 20,0 | 15,4 | | |
| 18 | 29.12.2018 09:00 | 25,0 | 692,0 | | 0,4 | 62 | Wind blow | 1 | | 10% or le | | | | 18,7 | | Stratocumulus other than Stratocumulus cumi | 10% or le | 600-1000 | No Altocur | No Cirrus, | 20,0 | 17,1 | 9 | 24 |
| 19 | 29.12.2018 06:00 | 20,1 | 691,6 | | | 83 | Wind blow | 2 | | no clouds | | | | | | | | | | | 20,0 | 17,2 | | |
| 20 | 28.12.2018 21:00 | 21,8 | 692,0 | | | 79 | Wind blow | 3 | | 60% | Haze. | Cloud cove | Cloud covering more t | | 27,9 | Stratocumulus other than Stratocumulus cumi | 60% | 300-600 | No Altocur | No Cirrus, | 12,0 | 17,9 | | |
| 21 | 28.12.2018 15:00 | 22,5 | 691,0 | | -0,6 | 75 | Wind blow | 2 | | 100% | Haze. | Thunderstc | Rain. | | | Stratocumulus other than Stratocumulus cumi | 20–30% | 600-1000 | Patches (c | No Cirrus, | 9,0 | 17,9 | 9 | 6 |
| 22 | 28.12.2018 12:00 | 27,9 | 691,6 | | | 48 | Wind blow | 2 | | 50% | | | | | | Cumulus humilis or Cumulus fractus other tha | 50% | 1000-1500 | No Altocur | No Cirrus, | 20,0 | 15,8 | | |
| 23 | 28.12.2018 09:00 | 24,5 | 692,5 | | 0,2 | 66 | Wind blow | 3 | | 10% or le | | | | 19,6 | | Cumulus humilis or Cumulus fractus other tha | 10% or le | 600-1000 | No Altocur | No Cirrus, | 20,0 | 17,8 | | |
| 24 | 28.12.2018 06:00 | 19,8 | 692,3 | | | 83 | Wind blow | 2 | | 20–30% | | | | | | No Stratocumulus, Stratus, Cumulus or Cumu | 20–30% | | Patches (c | No Cirrus, | 20,0 | 16,9 | | |
| 25 | 27.12.2018 21:00 | 20,5 | 692,8 | | 1,5 | 79 | Wind blow | 4 | | 20–30% | | | | | 25,8 | Stratocumulus other than Stratocumulus cumi | 20–30% | 1000-1500 | No Altocur | No Cirrus, | 20,0 | 16,8 | | |
| 26 | 27.12.2018 18:00 | 22,0 | 691,3 | | | 74 | Wind blow | 7 | | 60% | | | | | | Stratocumulus other than Stratocumulus cumi | 100% | 300-600 | | | 20,0 | 17,2 | | |
| 27 | 27.12.2018 15:00 | 24,3 | 691,0 | | -1,2 | 68 | Wind blow | 7 | | 40% | | | | | | Stratus nebulosus or Stratus fractus other tha | 40% | 600-1000 | No Altocur | No Cirrus, | 20,0 | 17,9 | | |
| 28 | 27.12.2018 12:00 | 25,4 | 692,2 | | | 60 | Wind blow | 3 | | 70 – 80% | | | | | | Stratocumulus other than Stratocumulus cumi | 70 – 80% | 1000-1500 | No Altocur | No Cirrus, | 20,0 | 17,0 | | |
| 29 | 27.12.2018 09:00 | 22,2 | 692,8 | | 0,6 | 72 | Wind blow | 3 | | 90 or mor | | | | 19,0 | | Stratocumulus other than Stratocumulus cumi | 90 or mor | 600-1000 | No Altocur | No Cirrus, | 20,0 | 16,8 | | |
| 30 | 27.12.2018 06:00 | 19,2 | 692,2 | | | 82 | Wind blow | 2 | | 90 or mor | | | | | | Stratus fractus or Cumulus fractus of bad weat | 90 or mor | 300-600 | No Altocur | No Cirrus, | 20,0 | 16,1 | | |
| 31 | 26.12.2018 21:00 | 19,8 | 693,0 | | 0,9 | 80 | Wind blow | 4 | | 60% | | | | | 25,3 | Stratocumulus other than Stratocumulus cumi | 60% | 300-600 | No Altocur | No Cirrus, | 20,0 | 16,2 | | |
| 32 | 26.12.2018 18:00 | 21,7 | 692,1 | | | 73 | Wind blow | 7 | | 10% or le | | | | | | Stratocumulus other than Stratocumulus cumi | 10% or le | 600-1000 | No Altocur | No Cirrus, | 20,0 | 16,7 | | |
| 33 | 26.12.2018 15:00 | 25,2 | 692,1 | | -1,1 | 64 | Wind blow | 5 | | 50% | | | | | | Cumulus humilis or Cumulus fractus other tha | 50% | 1000-1500 | No Altocur | No Cirrus, | 20,0 | 17,8 | | |
| 34 | 26.12.2018 12:00 | 24,6 | 693,2 | | | 58 | Wind blow | 2 | | 60% | | | | | | Stratocumulus other than Stratocumulus cumi | 60% | 1000-1500 | No Altocur | No Cirrus, | 20,0 | 15,7 | | |
| 35 | 26.12.2018 09:00 | 21,8 | 694,1 | | 0,5 | 72 | Wind blow | 3 | | 50% | | | | 17,0 | | Stratus nebulosus or Stratus fractus other tha | 60% | 300-600 | No Altocur | No Cirrus, | 20,0 | 16,4 | Trace of pr | 24 |
| 36 | 26.12.2018 06:00 | 18,0 | 693,6 | | | 87 | Wind blow | 3 | | no clouds | | | | | | | | | | | 20,0 | 15,8 | | |
| 37 | 25.12.2018 21:00 | 19,6 | 694,5 | | 1,3 | 82 | Wind blow | 4 | | 100% | Mist. | Rain. | Cloud covering more t | | 23,0 | Stratocumulus other than Stratocumulus cumi | 100% | 300-600 | | | 9,0 | 16,5 | Trace of pr | 6 |
| 38 | 25.12.2018 18:00 | 20,5 | 693,2 | | | 82 | Wind blow | 7 | | 100% | Mist. | Rain. | Cloud covering more than 1/2 of t | | | Stratocumulus other than Stratocumulus cumi | 100% | 200-300 | | | 12,0 | 17,4 | | |
| 39 | 25.12.2018 15:00 | 22,6 | 693,7 | | | 75 | Wind blow | 7 | | 100% | Haze. | Cloud cove | Cloud covering more than 1/2 of t | | | Stratocumulus other than Stratocumulus cumi | 70 – 80% | 300-600 | No Altocur | No Cirrus, | 15,0 | 18,0 | | |
| 40 | 25.12.2018 12:00 | 23,0 | 694,5 | | | 73 | Wind blow | 4 | | 70 – 80% | | | | | | Stratocumulus other than Stratocumulus cumi | 100% | 300-600 | | | 20,0 | 17,8 | | |
| 41 | 25.12.2018 09:00 | 22,1 | 694,1 | | 0,7 | 73 | Wind blow | 3 | | 20–30% | | | | 19,4 | | Stratus nebulosus or Stratus fractus other tha | 20–30% | 300-600 | No Altocur | No Cirrus, | 20,0 | 17,1 | 2 | 24 |
| 42 | 25.12.2018 06:00 | 19,5 | 693,4 | | | 86 | Wind blow | 3 | | 40% | | | | | | Stratus nebulosus or Stratus fractus other tha | 50% | 200-300 | No Altocur | No Cirrus, | 20,0 | 17,0 | | |
| 43 | 24.12.2018 21:00 | 19,4 | 693,8 | | 0,7 | 88 | Wind blow | 4 | | 100% | | | | | 24,8 | Stratus nebulosus or Stratus fractus other tha | 100% | 200-300 | | | 20,0 | 17,4 | Trace of pr | 6 |

Figura 2 – Histórico de previsão do tempo para uma localidade

## 1.2 Preparação de Dados

Para gerar o conjunto de dados utilizado neste projeto foram realizados diversos tratamentos nos dados, filtragens e mesclagem de informação. A primeira tarefa realizada foi obter as informações apenas dos voos realizados entre os aeroportos da Tabela 1. Com a data dos voos e os ICAO de origem e destino (Figura 1) buscou-se a previsão do tempo (Figura 2) para cada voo. A coluna "Código de justificativa" teve os valores "AM", "RM", "WR" e "XS" substituídos pelo valor "Anormal" e os campos em branco receberam o valor "Normal". Além disso a coluna passou a se chamar "Previsão do Tempo". A coluna "DD" teve seus valores substituídos por zero (indicando que não há vento) ou um (indicando que há vento). Após os ajustes foi gerada uma tabela contendo colunas com detalhes do voo e previsão do tempo dos aeroportos de origem e destino. Foram utilizadas as colunas a seguir:

- Previsão do Tempo: indica se o voo ocorreu normalmente ou teve anormalidades devido a problemas no tempo. Contém os rótulos "Normal"ou "Anormal".
- T_origem: temperatura em graus Celsius no aeroporto de origem;
- Po_origem: pressão atmosférica em milímetros de mercúrio no aeroporto de origem;
- U_origem: percentual de umidade relativa do ar no aeroporto de origem;
- DD_origem: indica se há vento ou não no aeroporto de origem;
- Ff_origem: velocidade do vento em metros por segundo no aeroporto de origem;
- N_origem: percentual de nebulosidade geral no aeroporto de origem;
- VV_origem: distância de visibilidade horizontal no aeroporto de origem;
- Td_origem: temperatura de condensação em graus Celsius no aeroporto de origem;
- T_destino: temperatura em graus Celsius no aeroporto de destino;
- Po_destino: pressão atmosférica em milímetros de mercúrio no aeroporto de destino;
- U_destino: percentual de umidade relativa do ar no aeroporto de destino;

- DD_destino: indica se há vento ou não no aeroporto de destino;
- Ff_destino: velocidade do vento em metros por segundo no aeroporto de destino;
- N_destino: percentual de nebulosidade geral no aeroporto de destino;
- VV_destino: distância de visibilidade horizontal no aeroporto de destino;
- Td_destino: temperatura de condensação em graus Celsius no aeroporto de destino;

# 2 Análise dos Dados

Este capítulo apresenta toda a parte de processamento e análise dos dados. Foram detalhadas todas as etapas do processo de análise dos dados.

## 2.1 Estatística Descritiva

A base de dados teve os seus valores organizados em histogramas, com o objetivo de analisar a sua distribuição. Foi encontrada uma distribuição de dados conforme a Figura 3, que mostra os atributos meteorológicos dos aeroportos de origem dos voos. Os *outliers* encontrados foram considerados relevantes para a análise e não foram removidos. Atributos como a temperatura (T), umidade (U), velocidade do vento (Ff) e a temperatura de condensação (Td) apresentaram distribuições dos dados bem definidas, com os dados agrupados em torno de um ponto central. Para os demais atributos foi identificada uma descontinuidade na dispersão e a dominância nos valores em alguns atributos como por exemplo no caso da distância de visibilidade (VV).

Foi observada uma grande similaridade na distribuição dos dados entre os atributos metrológicos dos aeroportos de origem e destino conforme a Figura 4.

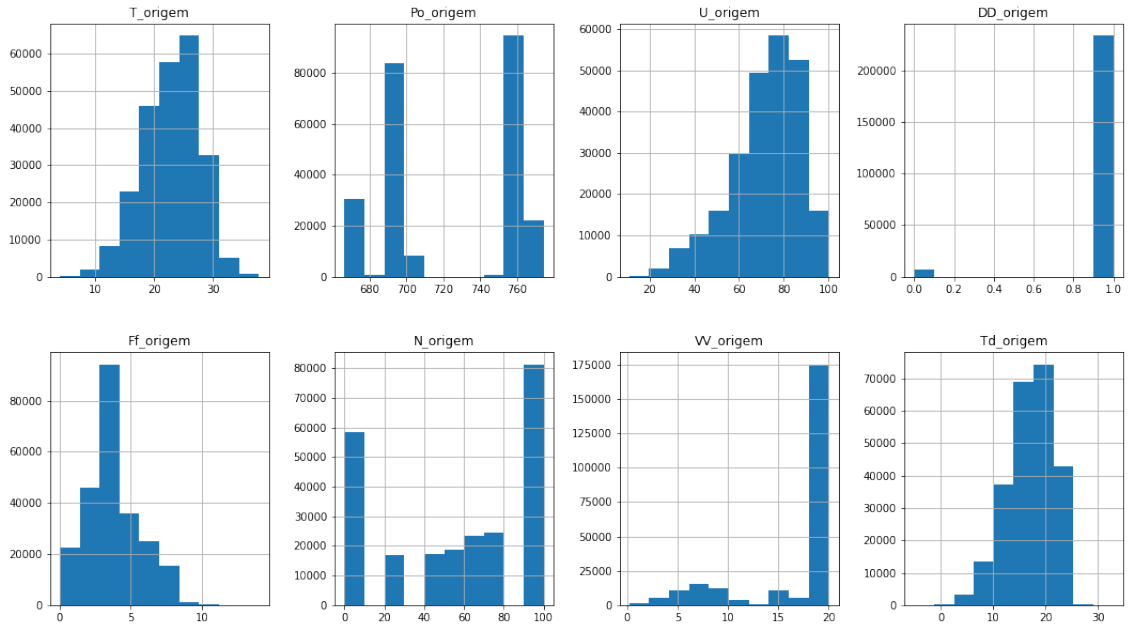



Figura 3 – Distribuição dos dados meteorológicos do aeroporto de origem

```
T_origem => MIN: 3.90 | MÁX: 37.80 | MÉDIA: 22.88 | DESV. PADRÃO: 4.79
Po_origem => MIN: 666.40 | MÁX: 774.30 | MÉDIA: 724.05 | DESV. PADRÃO: 36.41
U_origem => MIN: 11.00 | MÁX: 100.00 | MÉDIA: 72.10 | DESV. PADRÃO: 15.51
DD_origem => MIN: 0.00 | MÁX: 1.00 | MÉDIA: 0.97 | DESV. PADRÃO: 0.16
Ff_origem => MIN: 0.00 | MÁX: 14.00 | MÉDIA: 3.70 | DESV. PADRÃO: 1.83
N_origem => MIN: 0.00 | MÁX: 100.00 | MÉDIA: 55.44 | DESV. PADRÃO: 38.13
VV_origem => MIN: 0.20 | MÁX: 20.00 | MÉDIA: 17.14 | DESV. PADRÃO: 5.20
Td_origem => MIN: -5.10 | MÁX: 32.90 | MÉDIA: 17.11 | DESV. PADRÃO: 4.30

T_destino => MIN: 3.90 | MÁX: 37.80 | MÉDIA: 22.88 | DESV. PADRÃO: 4.82
Po_destino => MIN: 666.40 | MÁX: 774.30 | MÉDIA: 724.11 | DESV. PADRÃO: 36.42
U_destino => MIN: 11.00 | MÁX: 100.00 | MÉDIA: 72.14 | DESV. PADRÃO: 15.44
DD_destino => MIN: 0.00 | MÁX: 1.00 | MÉDIA: 0.97 | DESV. PADRÃO: 0.16
Ff_destino => MIN: 0.00 | MÁX: 14.00 | MÉDIA: 3.75 | DESV. PADRÃO: 1.84
N_destino => MIN: 0.00 | MÁX: 100.00 | MÉDIA: 55.43 | DESV. PADRÃO: 38.17
VV_destino => MIN: 0.20 | MÁX: 20.00 | MÉDIA: 17.15 | DESV. PADRÃO: 5.19
Td_destino => MIN: -5.10 | MÁX: 32.90 | MÉDIA: 17.12 | DESV. PADRÃO: 4.30
```

Figura 4 – Informações sobre os atributos dos aeroportos de origem e destino

## 2.2 Pré-processamento

Os valores presentes em cada atributo foram normalizados com o objetivo de limitar os valores numéricos e melhorar a performance dos classificadores. A normalização garante que cada atributo será tratado com o mesmo peso durante a aplicação do aprendizado supervisionado.

Os atributos da coluna "Previsão do Tempo" foram modificados para valores numéricos para o correto funcionamento do algoritmo de classificação. O rótulo "Normal" recebeu o valor zero e o rótulo "Anormal" recebeu o valor um. Apenas as linhas contendo informações em todas as colunas foram utilizadas. A Figura 5 mostra um trecho da base de dados após as alterações.

| Previsão do Tempo | T_origem | Po_origem | U_origem | DD_origem | Ff_origem | N_origem | VV_origem | Td_origem | T_destino | Po_destino | U_destino | DD_destino | Ff_destino |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.719764 | 0.909175 | 0.584270 | 1.0 | 0.214286 | 0.25 | 1.000000 | 0.673684 | 0.557522 | 0.289157 | 0.696629 | 1.0 | 0.285714 |
| 1 | 0.775811 | 0.857275 | 0.337079 | 1.0 | 0.285714 | 0.00 | 1.000000 | 0.539474 | 0.651917 | 0.863763 | 0.752809 | 1.0 | 0.214286 |
| 1 | 0.522124 | 0.268767 | 0.775281 | 1.0 | 0.142857 | 0.95 | 1.000000 | 0.605263 | 0.595870 | 0.863763 | 0.797753 | 1.0 | 0.357143 |
| 1 | 0.545723 | 0.251158 | 0.651685 | 1.0 | 0.142857 | 0.95 | 1.000000 | 0.568421 | 0.613569 | 0.805375 | 0.910112 | 1.0 | 0.214286 |
| 1 | 0.522124 | 0.264133 | 0.797753 | 1.0 | 0.142857 | 0.95 | 0.747475 | 0.615789 | 0.584071 | 0.877665 | 0.719101 | 1.0 | 0.500000 |
| 1 | 0.660767 | 0.882298 | 0.584270 | 1.0 | 0.214286 | 0.60 | 1.000000 | 0.628947 | 0.557522 | 0.279889 | 0.820225 | 1.0 | 0.285714 |
| 1 | 0.471976 | 0.229842 | 0.865169 | 1.0 | 0.428571 | 1.00 | 0.141414 | 0.605263 | 0.569322 | 0.848934 | 0.651685 | 1.0 | 0.214286 |
| 1 | 0.743363 | 0.231696 | 0.393258 | 1.0 | 0.428571 | 0.50 | 1.000000 | 0.560526 | 0.790560 | 0.806302 | 0.584270 | 1.0 | 0.214286 |
| 1 | 0.445428 | 0.275255 | 0.887640 | 1.0 | 0.214286 | 1.00 | 0.343434 | 0.589474 | 0.584071 | 0.886006 | 0.876404 | 1.0 | 0.142857 |
| 1 | 0.566372 | 0.248378 | 0.696629 | 1.0 | 0.500000 | 1.00 | 1.000000 | 0.605263 | 0.690265 | 0.843373 | 0.696629 | 1.0 | 0.357143 |

Figura 5 – Trecho da base de dados após o pré-processamento

## 2.3 Embaralhamento e Divisão do Dados

Os dados foram separados em um conjunto de treinamento e de testes. 80% foi utilizado para treinar os classificadores e 20% para teste.

Uma das preocupações durante a avaliação dos dados foi em relação ao seu desbalanceamento. Como a maioria dos voos ocorre sem problemas, há mais informações sobre voos da categoria "Normal" (Figura 6) . A base de dados continha 240531 registros, sendo que cerca de 1 % deles pertenciam à classe "Anormal" e 99 % pertenciam à classe "Normal". Para

evitar erros na avaliação dos classificadores, devido ao desbalanceamento dos dados entre as classes, foram utilizadas métricas de avaliação como a matriz de confusão, $F_\beta$-score, precisão, *recall* e a curva ROC.

```
Número de registros: 240531
Número de voos sem atrasos/cancelamentos devido a problemas meteorológicos: 238175
Número de voos com atrasos/cancelamentos devido a problemas meteorológicos: 2356
Porcentagem de voos com problemas: 0.99%
```

Figura 6 – Número de registros de voo

Com o objetivo de balancear os dados entre as duas classes foram utilizados menos registros da classe "Normal" conforme a Figura 7. Os dados foram separados em 3790 amostras de treinamento e 948 amostras de teste.

```
Número de registros: 4738
Número de voos sem atrasos/cancelamentos devido a problemas meteorológicos: 2382
Número de voos com atrasos/cancelamentos devido a problemas meteorológicos: 2356
Porcentagem de voos com problemas: 49.73%
```

Figura 7 – Número de registros de voo (reduzido)

## 2.4 Embasamento Teórico

Nesta seção são apresentados os principais algoritmos e técnicas utilizadas no projeto.

### 2.4.1 Métodos *Ensemble*

Os métodos *ensemble* são responsáveis por realizar predições baseando-se nos resultados obtidos em outros modelos, ou seja, toma-se uma decisão após agrupar e avaliar o desempenho de diversos modelos "fracos". Por combinar diversos modelos esses métodos tendem a ser mais flexíveis (menor *bias*) e apresentar menor sensibilidade aos dados (menor variância). O *bias* é a diferença entre o valor esperado da predição do nosso modelo (média das predições) e o valor real que queremos predizer. Já a variância é a variabilidade das predições (COUTO, 2013).

Normalmente, à medida que a complexidade do modelo aumenta, há uma redução no erro de previsão devido ao *bias* ser mais baixo no modelo. À medida que o modelo se torna mais complexo ele começará a sofrer com a variância. Um modelo ótimo deve manter o equilíbrio entre estes dois tipos de erros. Isto é conhecido como a gestão de "trade-off" entre erros de variância e *bias* (TEAM, 2016). Aprendizagem por *ensemble* é uma maneira de analisar esse "trade-off" (Figura 8).

Dois métodos *ensemble* populares são o *Bagging* e *Boosting*:

- Bagging: treina vários modelos individualmente de forma paralela. Cada modelo é treinado com um subconjunto aleatório de dados;
- Boosting: treina vários modelos individualmente de forma sequencial. Cada modelo aprende com os erros do modelo anterior a ele.

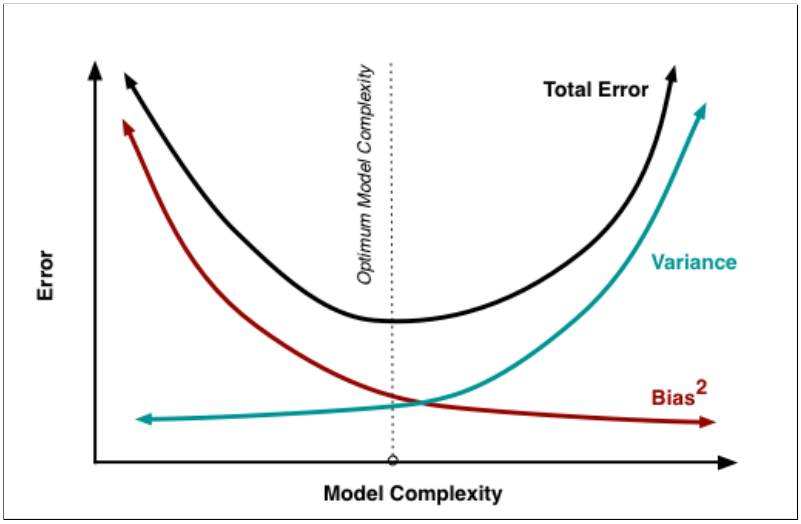



Figura 8 – Comportamento da variância e *bias* de acordo com a complexidade do modelo

### 2.4.2 Floresta Aleatória

Floresta aleatória é um modelo que geralmente combina o *Bagging* (método *ensemble*) com o modelo de árvore de decisão. Ele funciona através dos passos a seguir (CHEN, 2019):

- Passo 1: seleciona aleatoriamente $n$ subconjuntos de um conjunto de treinamento;
- Passo 2: treina $n$ árvores de decisão, sendo que cada subconjunto é utilizado para treinar uma árvore de decisão. A separação ótima ("split") para cada árvore de decisão é baseada em um subconjunto aleatório de características (por exemplo, de um conjunto total de 10 características seleciona aleatoriamente 5);
- Passo 3: cada árvore realiza as predições utilizando um conjunto de teste de forma independente;
- Passo 4: realiza a predição final. Para cada candidato do conjunto de teste, a floresta aleatória escolhe a classe com maioria dos votos, obtidos de cada uma das árvores de decisão.

Vantagens da Floresta Aleatória: permite resolver problemas de classificação e regressão. Lida bem com grande volume de dados e com muitas dimensões. Pode ser utilizado para identificar dentre as variáveis de entrada, as mais significativas, portanto, pode ser considerado como um método de redução de dimensões. Possui métodos para equilibrar erros em conjuntos de dados onde as classes são desequilibradas. Tem menos chances de sofrer com *overfitting*.

Desvantagens da Floresta Aleatória: não apresenta um bom funcionamento em problemas de regressão, por não fornecer previsões precisas para variáveis contínuas. Sua maior desvantagem é a sua complexidade, já que uma grande quantidade de árvores pode tornar o algoritmo lento e ineficiente para predições em tempo real.

### 2.4.3 Métricas de Avaliação

Antes de avaliar o desempenho de um classificador é necessário conhecer os seguintes conceitos:

- Acurácia: mede com que frequência o classificador faz a predição correta. É a proporção entre o número de predições corretas e o número total de predições;
- Precisão: é o número de vezes que uma classe foi predita corretamente dividido pelo número de vezes que a classe foi predita;
- Revocação: é o número de vezes que uma classe foi predita corretamente dividido pelo número total de elementos que pertencem realmente a classe.

Foi utilizada o F-beta score (Equação 1) como uma métrica que considera ambos: precisão e revocação.

$$F_\beta = (1 + \beta^2) \cdot \frac{precisão \cdot revocação}{(\beta^2 \cdot precisão) + revocação} \tag{1}$$

Quando $0 \leq \beta < 1$ é dado mais peso para a acurácia. Se $\beta = 1$ a acurácia e a revocação têm o mesmo peso. Para $\beta > 1$ a revocação se torna mais relevante.

Para escolher o valor de $\beta$ foi preciso entender melhor o problema, que no caso é a predição de anormalidades nos voos. Na Figura 9 foi analisada a pior situação, que ocorre quando os elementos pertencentes a classe "Anormal" são classificados como sendo da classe "Normal". Nesse caso a revocação torna-se importante para a análise e deve ter um peso um pouco maior que a precisão. Foi adotado o valor de $\beta = 2$.

| | | Previsto | |
|---|---|---|---|
| | | Anormal | Normal |
| Classe Real | Anormal | Verdadeiro Positivo | Falso Negativo |
| | Normal | Falso Positivo | Verdadeiro Negativo |

Figura 9 – Matriz de confusão do problema

### 2.4.4 *Benchmark*

Para avaliar a performance do classificador foi criado um outro classificador de referência ("burro"), que utiliza critérios de classificação simples e apresenta resultados inferiores aos desejados. Para criar esse tipo de classificador foi utilizada o pacote "DummyClassifier" presente na biblioteca *Sklearn*. Foi utilizada a estratégia *stratified*, que gera predições respeitando a distribuição dos dados entre as classes. Foi obtida uma acurácia de 51,90 % e um $F_\beta$-score de 52,43 %.

## 2.5 Resultados

### 2.5.1 Classificação

Foram utilizados quatro modelos de classificadores (Figura 10) e o melhor deles foi escolhido e ajustado posteriormente. O primeiro deles, o SVC (*Support Vector Classifier*) teve o

menor *score* e um tempo de treinamento e predição bem superior em relação aos demais classificadores. O modelo de classificação por Regressão Logística apresentou os menores tempos de treinamento e predição. Os métodos *Ensemble* (*Adaboost* e *Random Forest*) tiveram os melhores resultados tanto em acurácia quanto no *f1-Score*, além de demandarem de um tempo reduzido tanto para treino quanto para teste. O método de florestas aleatórias (*random forest*) apresentou o melhor resultado (Tabela 3).

| Classificador | Train Time (s) | Pred Time (s) | Acc Train (%) | Acc Test (%) | f1-score Train (%) | f1-score Test (%) |
|---|---|---|---|---|---|---|
| SVC | 0.4998 | 0.1093 | 60.00 | 59.28 | 49.15 | 47.41 |
| Regr.Logística | 0.0164 | 0.0009 | 61.33 | 59.91 | 57.35 | 55.29 |
| Adaboost | 0.1718 | 0.0155 | 66.66 | 62.97 | 65.51 | 62.29 |
| RandomForest | 0.0486 | 0.0156 | 95.66 | 66.87 | 95.50 | 65.49 |

Tabela 3 – Comparação do desempenho dos classificadores

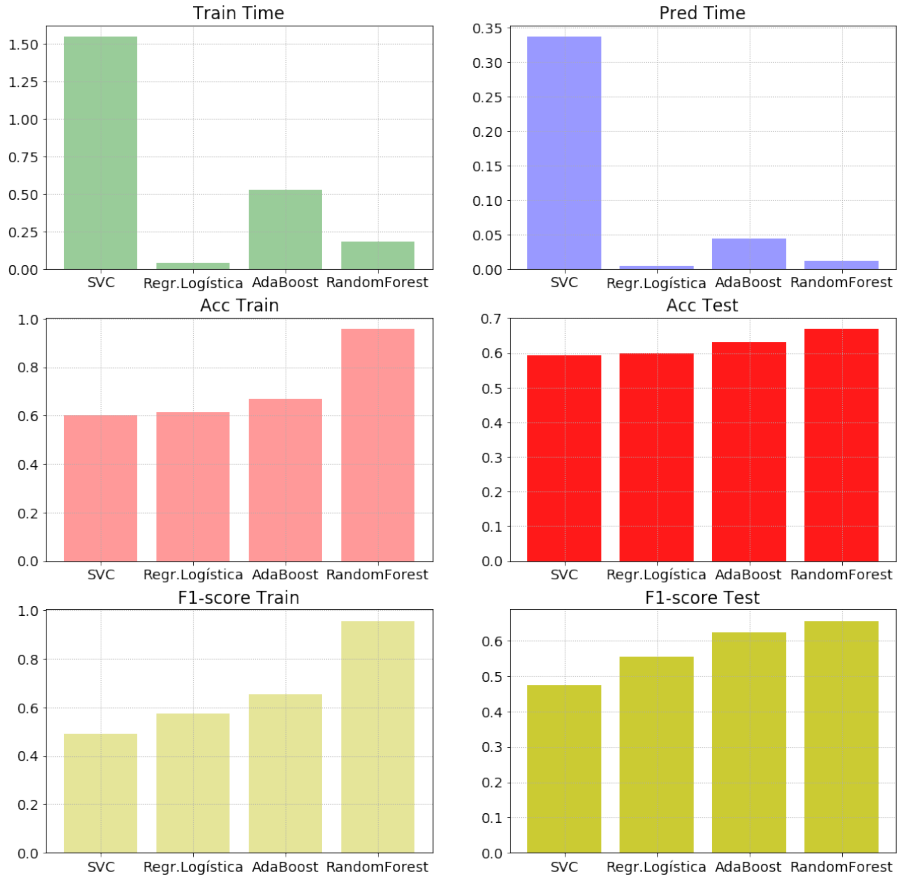



Figura 10 – Desempenho dos classificadores

### 2.5.2 Otimização do Classificador

O modelo de classificador adotado, de floresta aleatória, é uma combinação (*ensemble*) de árvores de decisão. Na maioria dos vezes ele é treinado com o método de *bagging*, que combina diversos modelos de aprendizado, com o objetivo de alcançar melhores resultados. São criadas várias árvores de decisão utilizando subconjuntos aleatórios das características, o que resulta na criação de modelos melhores (DONGES, 2018). A floresta aleatória tem como vantagens o fato de poder ser utilizada tanto na classificação quanto na regressão, permite

visualizar a importância dos atributos na classificação e possui poucos hiperparâmetros. Além disso ela apresenta menores chances de sofrer com sobreajuste (*overfitting*) caso haja uma quantidade de árvores suficiente na floresta. A maior limitação do modelo é que uma quantidade grande de árvores pode tornar o algoritmo lento e ineficiente para predições em tempo real.

Para aumentar o poder de predição do modelo foram realizados os ajustes dos hiperparâmeros a seguir, disponíveis na biblioteca *Sklearn*:

- n_estimators: indica o número de árvores construídas pelo algoritmo antes de tomar uma votação ou fazer uma média de predições. Em geral, uma quantidade elevada de árvores aumenta a performance e torna as predições mais estáveis, mas também torna a computação mais lenta;
- max_depth: indica a profundidade a árvore;
- max_features: indica o número máximo de características a serem utilizadas pela floresta aleatória na construção de uma dada árvore;
- min_sample_leaf: indica o número mínimo de folhas que devem existir em uma árvore;
- Criterion: Função que mede a qualidade das divisões;

Os hiperparâmetros do modelo de floresta aleatória foram otimizados através do método de *Grid Search* (PEDREGOSA et al., 2011) que utilizou os dados da Figura 11. Este método realiza uma busca pelos valores dos hiperparâmetros que proporcionam um melhor desempenho do classificador. O classificador com melhor desempenho apresentou as características da Figura 12. Percebeu-se melhores resultados quando os valores do atributo "max_depth", ou seja, a profundidade da árvore aumentava. Foi preciso adotar um limite para os valores para evitar problemas de *overfitting*, ou seja, quando o modelo se ajusta tanto aos dados de treino que não apresenta bons resultados nos dados de teste.

```
parameters = {
    'n_estimators': [10,20,30,50],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [8,9,10,12,15,20],
    'criterion' :['gini', 'entropy']

}
```

Figura 11 – Valores dos hiperâmetros utilizados no procedimento de *Grid Search*

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='entropy',
            max_depth=20, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=50, n_jobs=None,
            oob_score=False, random_state=42, verbose=0, warm_start=False)
```

Figura 12 – Hiperâmetros otimizados do classificador

### 2.5.3 Classificação utilizando o modelo otimizado

A Tabela 4 apresenta o desempenho do classificador após sua otimização. Houve uma pequena melhoria nos resultados. Observe como o valor do $F_\beta$-score foi inferior ao valor da acurácia. Foi obtida uma acurácia de 68,57 % e um $F_\beta$-score de 64,59 %.

| Métrica | Modelo Sem Otimização | Modelo Otimizado |
|---|---|---|
| Acurácia (%) | 66.03 | 68.57 |
| $F_\beta$-score ($\beta = 2$)(%) | 61.33 | 64.59 |

Tabela 4 – Performance do classificador

Para melhor visualização, as predições foram organizadas em uma matriz de confusão (Figura 13). Das 948 predições 173 foram de falsos negativos, ou seja, de elementos que pertencem à classe "Anormal" mas que foram consideradas pelo classificador como sendo da classe "Normal".

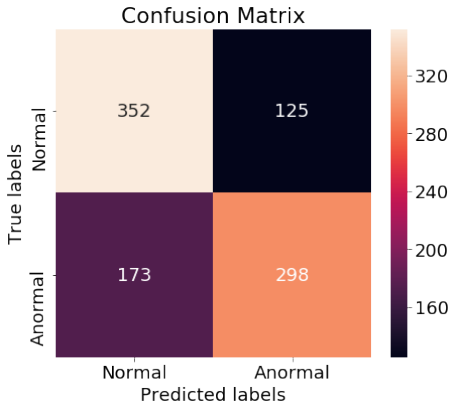



Figura 13 – Matriz de confusão dos dados de teste

A curva ROC relaciona a taxa verdadeiros positivos com a taxa de falsos positivos e é utilizada para avaliar o desempenho dos classificadores. Quanto maior a área sob a curva melhor é o classificador, portanto, quando mais o gráfico se assemelha com um quadrado, maior a área sob ele. Para este experimento foi obtida uma AUC (*Area Under the Curve*) de 0,74 (Figura 14).

### 2.5.4 Desempenho do modelo com novos dados

Foi realizado um experimento para avaliar o desempenho do classificador com novos dados de voos de outros aeroportos. Foram obtidos registros de voo e previsão do tempo de quatro aeroportos mostrados na Tabela 5.

| ICAO | Descrição |
|---|---|
| SBCT | Aeroporto Internacional de Curitiba / PR - Afonso Pena |
| SBCF | Aeroporto Internacional de Florianópolis / SC – Hercílio Luz |
| SBVT | Aeroporto de Vitória / ES - Eurico de Aguiar Salles |
| SBGR | Aeroporto Internacional de São Paulo / Guarulhos - Governador André Franco Motoro |

Tabela 5 – Código ICAO dos quatro aeroportos

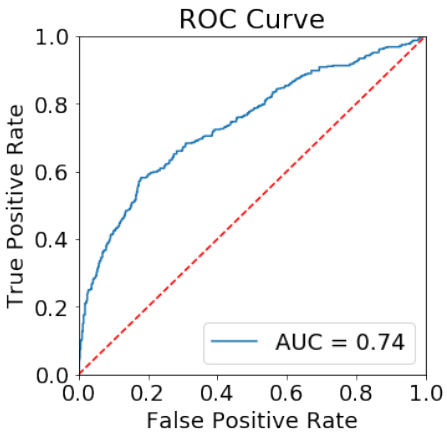



Figura 14 – Curva ROC

Os dados passaram pelo pré-processamento mostrado na Seção 2.2 e foram classificados utilizando o modelo otimizado do classificador. Foram analisados 566 registros, sendo 307 pertencentes a classe "Normal" e 259 da classe "Anormal". Os resultados obtidos após a classificação foram organizados na matriz de confusão da Figura 15. Foi obtida uma acurácia de 64,13 % e um $F_\beta$-score de 52,42 %.

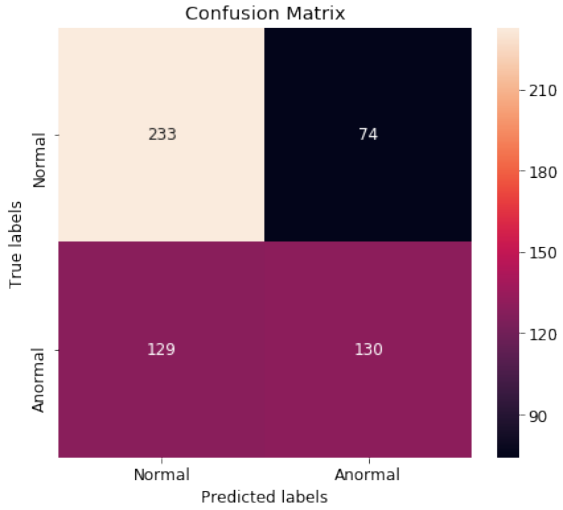



Figura 15 – Matriz de confusão obtida após experimento com os novos dados

### 2.5.5 Análise da Relevância dos Atributos

Reduzir a dimensionalidade dos dados é um processo importante para a otimização e redução dos custo computacional envolvido na classificação. Dependendo do problema a ser resolvido alguns atributos apresentam maior relevância que os demais, e podem ser utilizados para simplificar o modelo do classificador. O *Sklearn* tem uma excelente ferramenta para isto chamado de *feature_importances*, que mede a importância dos atributos. Ele calcula este valor automaticamente para cada atributo após o treinamento e normaliza os resultados para que a soma de todas as importâncias seja igual a 1.

Na Figura 16 foram analisados os pesos de cada atributo, ou seja, a importância de cada um para o modelo. Os atributos "Temperatura" (T), "Temperatura de Condensação" (Td), "Pressão Atmosférica" (Po) e a "Umidade" (U) tiveram maior importância. Observou-se que o atributo Direção do Vento (DD) apresentou pouca importância em relação aos demais atributos e portanto poderia ser removido. Reduzir o número de atributos simplifica o modelo e reduz as chances de sobre-ajuste (*overfitting*).

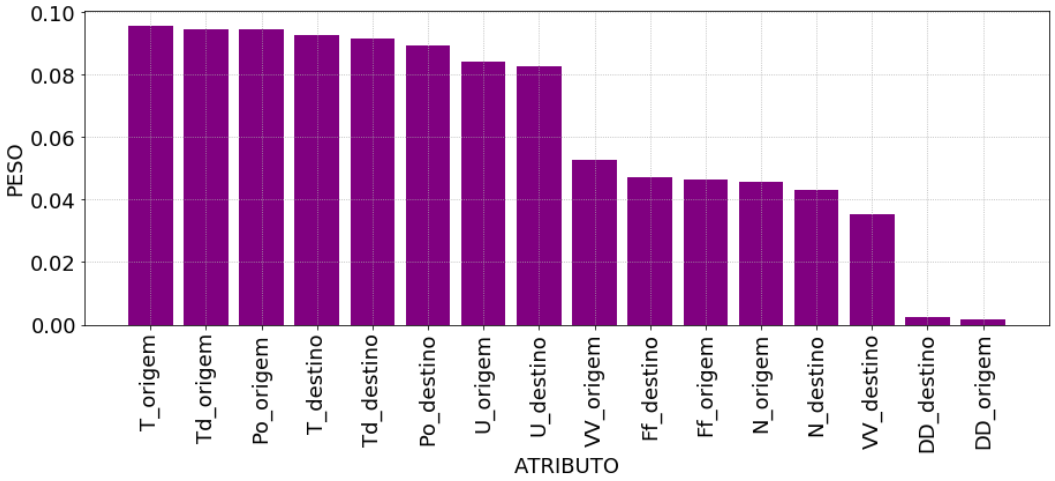



Figura 16 – Importância dos atributos para a classificação

Realizou-se um novo experimento, no qual foi criado um novo modelo de classificador que não utilizava os atributos "DD_origem" e "DD_destino". Foi obtida uma acurácia de 68,04 % e um $F_\beta$-score de 64,12 %, valores ligeiramente inferiores aos obtidos anteriormente (Tabela 4). Entretanto, o modelo apresentou uma performance melhor com novos dados (Seção 2.5.4). Nesse caso, o modelo apresentou uma acurácia de 66,25 % e um $F_\beta$-score de 55,15 %.

# 3 Conclusão

Neste trabalho foi utilizado o aprendizado supervisionado para criar um modelo de predição de anormalidades em voos baseando-se no histórico de voos e previsão do tempo dos dez principais aeroportos do Brasil.

Durante os experimentos foi identificado que o desbalanceamento entre as duas classes (Normal e Anormal) era muito elevado, portanto, decidiu-se equilibrar o número de dados relacionados a cada uma. Da base de dados consolidada foram utilizadas as informações de 4738 registros de voo. O Classificador de florestas aleatórias apresentou o melhor resultado dentre os classificadores analisados. Foi obtida uma acurácia de 68,57 % e um $F_\beta$-score de 64,59 %, valores superiores ao modelo de referência (*benchmark*) adotado.

Após criar o modelo e otimizá-lo o classificador foi avaliado com novos dados, ou seja, dados de voos de outros quatro aeroportos. Foi obtida uma acurácia de 64,13 % e um $F_\beta$-score de 52,42 %. Os resultados foram melhorados após realizar-se a redução do número de atributos do modelo, que obteve para esses dados uma acurácia de 66,25 % e um $F_\beta$-score de 55,15 %.

O modelo apresentou um desempenho acima do *benchmark* definido, contudo é preciso melhorá-lo. Deve-se utilizar uma quantidade de dados maior para a criação do modelo e

estudar o uso de outros tipos de atributo.

Como sugestão para trabalhos futuros, recomenda-se utilizar um conjunto de atributos mais relevantes, ou seja, que permitem distinguir com maior facilidade as amostras das duas classes. Seria interessante também analisar o desempenho de redes neurais nesse tipo de problema (SAUVESTRE ROMAIN, 2016).

## Referências


ANAC. *Histórico de Voos*. 2016. Disponível em: <http://www.anac.gov.br/assuntos/dados-e-estatisticas/historico-de-voos>. Citado na página 2.

ANAC. *Meteorologia e Planejamento de Voo*. 2017. Disponível em: <http://www.anac.gov.br/assuntos/setor-regulado/profissionais-da-aviacao-civil/meteorologia-aeronautica/veja-mais/meteorologia-e-o-planejamento-de-voo>. Citado na página 1.

CHEN, L. *Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)-Step by Step Explained*. 2019. Disponível em: <https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725>. Citado na página 8.

COUTO, E. *Bias vs. Variância (Parte 1)*. 2013. Disponível em: <https://ericcouto.wordpress.com/2013/06/29/bias-vs-variancia-parte-1/>. Citado na página 7.

DAC. Iac 1504. *Procedimentos para o Registro de Alterações em Voos de Empresas de Transporte Aéreo Regular*, p. 12–14, 2000. Disponível em: <http://www.anac.gov.br/assuntos/legislacao/legislacao-1/iac-e-is/iac/iac-1504/@@display-file/arquivo_norma/IAC1504.pdf>. Citado na página 2.

DONGES, N. *The Random Forest Algorithm*. 2018. Disponível em: <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>. Citado na página 10.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 11.

POGODI, R. *Weather archive in Sao Paulo (airport)*. 2019. Disponível em: <https://rp5.ru/Weather_archive_in_Sao_Paulo_(airport)>. Citado na página 3.

SAUVESTRE ROMAIN, D. L. e. L. J. *Predicting flight delays and cancellations using weather as a feature*. [S.l.], 2016. Disponível em: <http://cs229.stanford.edu/proj2016/report/DuperierSauvestreLeaf-ModelingFlightDelays-report.pdf>. Citado na página 15.

STERNBERG, A. et al. *A Review on Flight Delay Prediction*. 2017. Disponível em: <https://arxiv.org/pdf/1703.06118.pdf>. Citado na página 2.

TEAM, A. V. C. *A Complete Tutorial on Tree Based Modeling from Scratch (in R Python)*. 2016. Disponível em: <https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/>. Citado na página 7.