# Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays

IEEE WCCI 2018
Rio de Janeiro, Brazil

Leonardo Moreira, Christofer Dantas, Leonardo Oliveira, Jorge Soares, Eduardo Ogasawara
CEFET/RJ - Federal Center for Technological Education of Rio de Janeiro

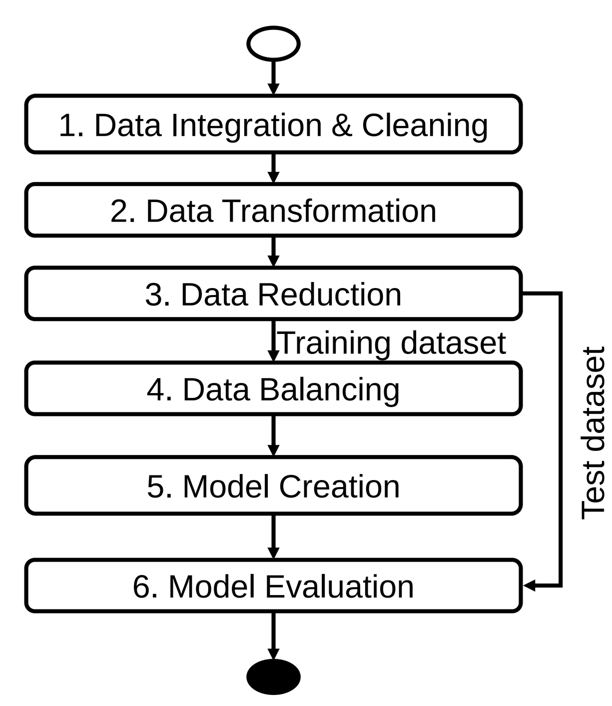
## Introduction

- Flight delays cause various inconveniences for airlines, airports, and passengers.
- According to the Brazilian National Civil Aviation Agency (ANAC), between 2009 and 2015, 22% Brazilian flights were delayed by more than 15 minutes.
- Airlines, airports, and users may be more interested in when delays are likely to occur (sensitivity) than the correct prediction of an absence of delays (accuracy).
- Building machine learning models under such unbalanced distribution is challenging.
- Few works explore different preprocessing methods for the development of machine-learning flight delay classification models.

## Problem Statement

- Which preprocessing methods may aid in solving the sensitivity while preserving good accuracy under such unbalanced distribution.
- This paper focuses on the unbalanced distribution of the classes of delay (presence and absence) by performing an experimental evaluation of several preprocessing methods for the development of machine-learning flight delay classification models.

## Data Mining Process




## Data integration & Transformation

- This work builds a dataset that integrates a database containing flight operations data provided by the Brazilian National Civil Aviation Agency (ANAC) (http://www.anac.gov.br) and airport weather data provided by Weather Underground (http://www.wunderground.com).


| Dimension | Original attribute | Original range of values | Indexed attribute | Indexed values | Indexing technique |
|---|---|---|---|---|---|
| Temporal | Scheduled departure (date/time) | from 01/01/2009 0:00 to 28/02/2015 23:59 | Year<br>Month<br>Day of the week<br>Working day<br>Time of the day | 2009–2015<br>1–12<br>Sunday to Saturday<br>True or false<br>Early morning: 5:00–8:59<br>Mid-morning: 9:00–10:59<br>Late morning: 11:00–12:59<br>Afternoon: 13:00–16:59<br>Early evening: 17:00–19:59<br>Late evening: 20:00–22:59<br>Night: 23:00–4:49 | Concept hierarchy<br>Concept hierarchy<br>Concept hierarchy<br>Concept hierarchy<br>Concept hierarchy |
| Meteorological | Temperature (C) | −3 to 41.4 | Temperature | Low: −3 to 20<br>Medium: 21–25<br>High: 26–41 | Binning |
| Meteorological | Pressure (hPa) | 996–1036 | Pressure | Low: 996–1013<br>Medium: 1014–1017<br>High: 1018–1036 | Binning |
| Meteorological | Humidity (%) | 4–100 | Humidity | Low: 4–64<br>Medium: 65–81<br>High: 82–100 | Binning |
| Meteorological | Wind speed (km/h) | 0–213 | Wind speed | Calm: 0–1.852<br>Light air: 1.853–5.556<br>Light breeze: 5.557–11.112<br>Gentle breeze: 11.113–18.520<br>Moderate breeze: 18.521–29.632<br>Fresh breeze: 29.633–38.892<br>Strong breeze: 38.893–50.004<br>Near gale: 50.005–61.116<br>Gale: 61.117–74.080<br>Strong gale: 74.081–87.044<br>Storm: 87.045–101.860<br>Violent storm: 101.861–116.676<br>Hurricane: 116.677–213 | Concept hierarchy |
| Meteorological | Wind direction (degrees) | 0–360 | Wind direction | N: 0–11 or 349–360, NNE: 12–33<br>NE: 34–56, ENE: 57–78<br>E: 79–101, ESE: 102–123<br>SE: 124–146, SSE: 147–168<br>S: 169–191, SSW: 192–213<br>SW: 214–236, WSW: 237–258<br>W: 259–281, WNW: 282–303<br>NW: 304–326, NNW: 327–348 | Concept hierarchy |
| Meteorological | Visibility (km) | 0–28 | Visibility | IFR: 0–4.82<br>VFR: 4.83–28 | Concept hierarchy |
| State of the system | Percentage of delays | 0–100% | Delay level | Low: 0–11.76%<br>Medium: 11.77–30.00%<br>High: 30.00–100% | Temporal aggregation and binning |

## Data Reduction

- The data reduction activity aims to create a reduced representation of the dataset (either by filtering attributes or tuples) to improve performance during analytical result. Many approaches exist for attribute selection, such as (i) Absolute Minimum Shrinkage and LASSO, (ii) Information Gain, (iii) Attribute Selection based on Correlation (CFS), (iv) Principal Component Analysis (PCA).

## Data Balancing

- Sampling is a direct approach to the problem of class balancing in a dataset. From the use of balancing methods, it is possible to change the distribution of classes aiming at obtaining a more balanced distribution of the data and improve the performance of the data classification models. The data balancing strategies used in this study are Random Sub-Sampling (RS) and the Synthetic Minority Oversampling Technique (SMOTE).

## Preliminary experimental evaluation

- Choosing machine learning methods (using LASSO):

TABLE III: Analysis of Machine Learning Methods

| Method | Accuracy (MSE) | Elapsed time (hours) | Parameter combinations |
|---|---|---|---|
| NN | 78.02 | 00:02 | 28 |
| RF | 77.94 | 00:01 | 28 |
| $SVM_{rbf}$ | 77.99 | 05:01 | 14 |
| $SVM_{tanh}$ | 77.99 | 03:09 | 14 |
| NB | 74.81 | 00:03 | - |
| kNN | 67.80 | 00:23 | 28 |

- Exploring data balancing methods:

TABLE IV: Balancing Results

| Balancing Method | Number of Records With Delay | Without Delay | Total |
|---|---|---|---|
| None | 184.094 | 52.156 | 236.250 |
| RS | 52.156 | 52.156 | 104.312 |
| SMOTE | 104.312 | 104.312 | 208.624 |

- Evaluation of preprocessing methods (using Neural Networks):

TABLE VII: Accuracy (in %) for NN according to the method of selection of attributes and balancing

| Selection Method | Balancing Method RS | SMOTE |
|---|---|---|
| None | 61.44 | **73.81** |
| LASSO | 59.14 | 59.04 |
| CFS | 60.20 | 60.32 |
| INFOGAIN | 60.23 | 64.65 |
| PCA | 60.52 | 67.24 |

TABLE VI: Sensitivity (in %) for NN according to the method of selection of attributes and balancing

| Selection Method | Balancing Method None | RS | SMOTE |
|---|---|---|---|
| None | 5.93 | 58.41 | 26.03 |
| LASSO | 1.81 | 58.75 | 58.89 |
| CFS | 1.88 | 56.46 | 56.08 |
| INFOGAIN | 3.57 | 58.47 | 49.10 |
| PCA | 5.17 | **60.13** | 43.47 |

## Conclusions

- An experimental evaluation using different data preprocessing and machine learning models was carried out over a Brazilian national commercial flight dataset with the objective of building a classification model with higher sensitivity to the occurrences of flight delays.
- A broader spectrum analysis of different data preprocessing methods was evaluated when compared to the literature review, with a particular focus on the unbalanced distribution of the classes of delay.
- Future work will focus on a more in-depth exploratory analysis of the data, and an extensive combining of data preprocessing methods with machine learning methods, particularly the deep-learning ones. Finally, a clustering analysis is also intended to analyze data mining process effectiveness and the quality of prediction, to improve the results obtained by the classifier.

CEFET/RJ

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

CAPES

FAPERJ Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro