CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

EVALUATION OF DATA PREPROCESSING METHODS FOR PREDICTING BRAZILIAN FLIGHT DELAYS

Leonardo da Silva Moreira

Dissertation submitted to the Postgraduate Program in of the Federal Center for Technological Education of Rio de Janeiro, CEFET/RJ, as partial fulfillment of the requirements for the degree of master.

Advisor: Jorge de Abreu Soares
Co-advisor: Eduardo Soares Ogasawara

Rio de Janeiro,
November 13, 2019

EVALUATION OF DATA PREPROCESSING METHODS FOR PREDICTING BRAZILIAN FLIGHT DELAYS

Dissertation submitted to the Postgraduate Program in of the Federal Center for Technological Education of Rio de Janeiro, CEFET/RJ, as partial fulfillment of the requirements for the degree of master.

Leonardo da Silva Moreira

Examining jury:

President, Prof. D.Sc Jorge de Abreu Soares (CEFET/RJ) (Advisor)

Prof. D.Sc Eduardo Soares Ogasawara (CEFET/RJ) (Co-advisor)

Prof. D.Sc. Eduardo Bezerra da Silva(CEFET/RJ)

Prof. D.Sc. Leonardo Gresta Paulino Murta (UFF/RJ - Universidade Federal Fluminense)

Rio de Janeiro,

November 13, 2019

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M838 Moreira, Leonardo da Silva.
Evaluation of data preprocessing methods for predicting brazilian flight delays / Leonardo da Silva Moreira – 2019.
123f. + apêndices : il.color. , grafs. ; enc.

Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2019.
Bibliografia : f. 115-123.
Orientadores: Jorge de Abreu Soares [e] Eduardo Soares Ogasawara.

1. Mineração de dados. 2. Redes neurais (Computação). 3. Análise de redes (Planejamento). 4. Teoria da Análise de sistemas. 5. Predição, Teoria da. 6. Modelos matemáticos. I. Soares, Jorge de Abreu (Orient.). II. Ogasawara, Eduardo Soares. III. Título.

CDD 006.312

Elaborada pelo bibliotecário Leandro Mota de Menezes CRB-7/5281

DEDICATION

Gratidão é uma palavra que felizmente tem sido reutilizada nos últimos tempos, no lugar de ”dedico”, ”obrigado” ou ”agradeço”.

Gratidão à Deus e à minha família que me ajudaram, apoiaram e guiaram ao longo de toda a minha jornada.

Gratidão a todas as pessoas e condições que de alguma forma me estimularam e me provocaram a perguntar, questionar, pesquisar, aprender, apreender, ensinar.

ACKNOWLEDGMENTS

O presente trabalho foi desenvolvido com o apoio da Coordenação de Aperfeicionamento de Pessoal de Nível Superior - Brasil(CAPES), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ)

RESUMO

Evaluation of Data Preprocessing Methods for Predicting Brazilian Flight Delays

Em 2016, as receitas do setor de serviços aéreos do Brasil alcançaram um recorde histórico de receita de R$ $35,59$ bilhões, transportando 109,6 milhões de passageiros de acordo com levantamento da Agência Nacional de Aviação Civil (ANAC). Considerando esse cenário, atrasos nos voos causam vários inconvenientes para as companhias aéreas, aeroportos e passageiros como ocorreram entre 2009 e 2015, onde cerca de 22% dos voos domésticos realizados no Brasil sofreram atrasos superiores a 15 minutos. A previsão desses atrasos é fundamental para mitigar sua ocorrência e otimizar o processo de tomada de decisão de um sistema de transporte aéreo. Particularmente, companhias aéreas, aeroportos e usuários podem estar mais interessados em saber quando é provável que ocorram atrasos do que a previsão precisa de quando não ocorrerão. Neste contexto, esta pesquisa apresenta uma avaliação experimental de métodos de pré-processamento de dados para modelos de classificação de aprendizado de máquina para a predição dos atrasos aéreos, de forma a identificar quais métodos e combinações destes métodos podem auxiliar na melhora da predição e dos resultados do classificador sob uma distribuição desequilibrada de classes de atraso. Para isto a metodologia utilizada inclui a integração de dados aéreos e meteorológicos, etapas de pré-processamento (limpeza, transformação, redução e balanceamento) e finalmente a comparação da predição de dados a partir destes diferentes métodos de pré-processamento. Particularmente, esta pesquisa contribui com a análise de um espectro de métodos de pré-processamento de dados quando comparado à revisão bibliográfica, focando especialmente a distribuição das classes de atraso. Incluem-se entre os objetivos deste trabalho a verificação mais detalhada em relação aos atributos do classificador, a normalização, discretização e seleção e extração de recursos, principalmente no que diz respeito à faixa de parâmetros do filtro. Em comparação aos trabalhos relacionados, com o uso de uma comparação normalizada das melhorias, foram obtidos resultados até 54,70% superiores em termos de Acurácia; 4,58% superiores em termos de Precisão; até 63,33% superiores em termos de Recall; e cerca de 25,38% superiores em termos de F1-Score.

Palavras-chave: Predição;Atrasos Aéreos;Pré-Processamento

ABSTRACT

Evaluation of Data Preprocessing Methods for Predicting Brazilian Flight Delays

In 2016, revenues from Brazil's air services industry reached a historical record revenue of R$ 35.59 billion, carrying 109.6 million passengers according to a survey by the National Civil Aviation Agency (ANAC). Given this scenario, flight delays cause a number of inconveniences for airlines, airports and passengers, as occurred between 2009 and 2015, where about 22 % of domestic flights performed in Brazil were delayed by more than 15 minutes. Predicting these delays is critical to mitigating their occurrence and optimizing the decision-making process of an air transport system. In particular, airlines, airports and passengers may be more interested in knowing when delays are likely to occur than the precise forecast of when they will not occur. In this context, this research presents an experimental evaluation of data preprocessing methods for machine learning classification models for the prediction of flight delays, in order to identify which methods and combinations of these methods can assist in improving prediction and classifier results under an unbalanced distribution of delay classes. For this the methodology used includes the integration of flight and weather data, preprocessing steps (cleaning, transformation, reduction and balancing) and finally the comparison of data prediction from these different preprocessing methods. In particular, this research contributes to the analysis of a spectrum of data preprocessing methods as compared to the literature review, focusing in particular on the distribution of delay classes. The objectives of this work include more detailed verification regarding the attributes of the classifier, normalization, discretization, selection and extraction of resources, especially with regard to the filter parameter range. In comparison to the related studies, with the use of a normalized improvement, results were obtained up to 54.70% superior in terms of Accuracy; up to 4.58% higher in Precision terms; up to 63.33% higher in terms of Recall; and results about 25.38% higher in terms of F1-Score.

Keywords: Prediction; Flight Delays; Preprocessing

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IG	Information Gain
KDD	Knowledge Discovery In Databases
LASSO	Least Absolute Shrinkage And Selection Operator
NN	Neural Network
PCA	Principal Component Analysis
RF	Random Forests
SMOTE	Synthetic Minority Over-sampling Technique

CONTENTS

Introduction

Brazilian aviation is administered and regulated by the Civil Aviation Secretary through INFRAERO (Brazilian Airport Infrastructure Company) and ANAC (National Civil Aviation Agency) bodies responsible for the administration and regulation of flights in Brazilian airspace. Besides, they concentrate most of the data on civil aviation, providing an appropriate database to build a panorama of the Brazilian air sector concerning the number of passengers, aircraft, and other aspects of the main airport units in the country.

In 2016, revenues from Brazil's air services sector reached a historical record of revenue of R$ 35.59 billion [ANAC, 2016]. Considering domestic and international flights, Brazilian and foreigners companies carried 109.6 million passengers paid in 2016 [ANAC, 2016, 2017]. It is estimated that in the medium and long-term (up to 2030), given the projected growth, the total number of passengers will increase from 130 million to 310 million passengers a year [BNDES, 2010].

Delay is one of the key performance indicators of any transportation system. A flight delay shall be represented by the difference between the programmed time and the actual time of departure or arrival of a flight [ANAC, 2016, 2012]. Given the uncertainty of its occurrence, many passengers are forced to reschedule their travels in order to arrive at the destination on time, which often leads to increased travel costs [Britto et al., 2012]. Thus, methods of predicting flight delays are fundamental to mitigate their occurrence, and therefore reduce the costs generated.

In the commercial aviation scenario, delays have a high financial impact on airlines, such as fines, additional operating costs, and declining customer loyalty. Also, given the uncertainty of their occurrence, many passengers are forced to reschedule their travels to arrive at the destination on time, which often leads to increased travel costs.

In Brazil, in 2016, the percentage of cancellations was 10.5% of total scheduled flights, while 6.2% of flights performed were delayed by 30 minutes or more, and 2.5% were delayed by 60 minutes or more. Thus, 11.8% of scheduled domestic flights were canceled, 5.9% of the flights were delayed by more than 30 minutes, and 2.2% were delayed by more than 60 minutes.

A large volume of data has been collected in databases of public and private

institutions for studying and understanding the operations of the air transport system. Analysis of this large amount of data, such as a big data problem, allows us to gain the knowledge needed to detect and predict delays. This process of prediction, which presentation of a repertoire of data, makes it possible to perform in different types of the prediction, especially that refers to the label [Han et al., 2011]. In this context, there are several analyzes, which involve domain comprehension, the relationship between the data, and the application of models to solve the problem [Sternberg et al., 2016; Dhar, 2013; Jagadish et al., 2014; Matsudaira, 2015].

Although there are these public databases and regulations regarding the disclosure of cases of delays and cancellations, there are few studies that aim to analyze the conditions that generate delays and cancellations regarding data science. The lack of jobs in the area causes many airlines, airports, and investors to make decisions that may not consider all the factors associated with delays [Sternberg et al., 2016].

In this scenario, the objective of this work is to perform an experimental evaluation of data preprocessing methods, especially normalization, categorical mapping, and discretization, with the objective of optimizing the accuracy, sensitivity, and F1-Score of the prediction models, considering all the factors involved and collected by the dataset. Any improvement on this topic can be beneficial to airlines, airports, and passengers. For this, a preprocessing methodology will be used for data mining, including the evaluation of classification in two types of machine learning.

This research contributes by exploring a broader spectrum of data preprocessing methods for building machine learning models. Although flight delay prediction is an open problem, our results indicated the need for balanced training data. Workflows were assembled with different combinations of data transformation techniques to perform the tests to generate classification models. Given the imbalance of the data, balancing techniques and classification threshold definitions were applied. These workflows were observed for the accomplishment of numerous experiments that verified the effectiveness of the techniques.

The experimental evaluation was conducted using a dataset that flight operations information from ANAC (2016), and weather information from The Weather Company (2016). The dataset was made available by Ogasawara (2018), containing aviation data from January 2009 to December 2017. Many data preprocessing methods were applied in combination with machine learning classification models. Their performance evaluation

regarding delay prediction was preliminarily analyzed.

Compared with the original workflow, without any transformation, results were obtained with much higher performance, with improvements of 130% in F1-Score and more than 40% improvement in the results obtained in Accuracy and Sensitivity.

Besides this introduction, this work is structured as follows. Chapter 1 refers to the theoretical foundation. In this chapter, the concepts related to preprocessing and machine learning are presented. Chapter 2 presents a literature overview of prediction and classification in flight delays that uses preprocessing techniques and model classification.

Chapter 3 presents the methodology used to carry out the work, pointing out each stage of preprocessing steps applied to the dataset and workflow experiments performed for the creation, training, testing, and evaluation of the models. Chapter 4 presents the results of testing and comparison of data obtained on experiments. On the final, the Conclusion presenting some considerations with the proposed future step for this work.

1- Background

The process of Knowledge Discovery in Databases (KDD) was created in 1989 as a reference to the domain of knowledge in data mining (Data Mining), referring to any process of finding useful knowledge of data, while Data Mining refers to the application of algorithms to extract models of the data[Fayyad et al., 1996; Macedo and Matos, 2010].

Six stages represent KDD Process, organized From Data to Knowledge, as depicted in Figure 1: Problem Specification, Problem Understanding, Data Preprocessing, Data Mining, Evaluation, Result Exploitation [García et al., 2016].

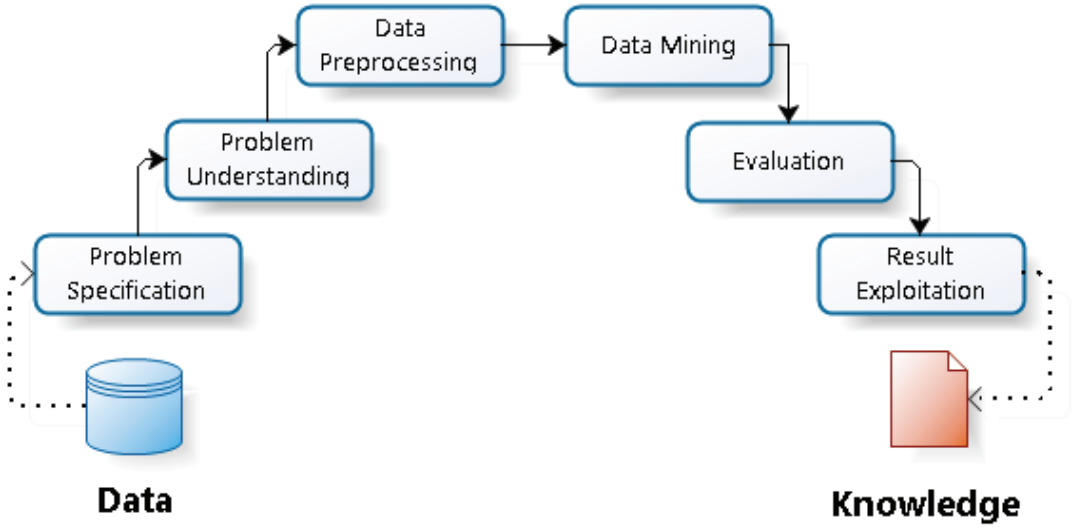

Figure 1 – KDD Process. Adapted From: [García et al., 2016]

The project starts specifying the problem stage, where it is necessary to understand the problem, the definition of objectives (mining problems), and the requirements of a business perspective [Fayyad et al., 1996]. This stage, in addition to the selected data understanding, also comprises an associated specialized knowledge to achieve a somewhat higher degree of reliability, with a rework reduction.

During preprocessing problem, a number of tasks such as data cleaning (how to deal with noise removal and inconsistent data), data integration (where multiple data sources can be combined in one), data selection (giving higher relevance to data), data transformation (how to deal with the format of the data, adjusting it), data reduction

(decreasing dimensionality from data making them more significant), and data balancing (how to deal with class data distribution) are applied [García et al., 2016].

Often confused with KDD, which refers to the whole process of knowledge discovery, data mining is a stage of this process, combining statistical analysis, machine learning, and data management to extract information from datasets [Thuraisingham, 2000]. Each data mining technique serves, depending on the modeling objective, to a different purpose [Han et al., 2011].

Finally, we have the last two stages of this process: evaluation, which includes the estimation and interpretation of the standards mined in the previous stage; and result exploitation, which from the evaluation performed, can extract the knowledge.

1.1- Data Preprocessing

The presence of noise, redundant data, missing data, inconsistencies, and data in large sizes and dimensions usually influences datasets. Such factors can dramatically reduce the performance of data mining. These threats reinforce that the Data Preprocessing step is one of the most important in the KDD process. Typically, this stage requires more than 60 % of the total project time [Press, 2016]. It exemplifies the importance of data quality to meet the desired requirements, including factors such as accuracy, integrity, consistency, timeliness, credibility, sensitivity, and interpretability [Aggarwal, 2015; Han et al., 2011]. After data collection, this phase includes integration, selection, cleaning, transformation, reduction, and balancing procedures, as exhibited in figure 2.

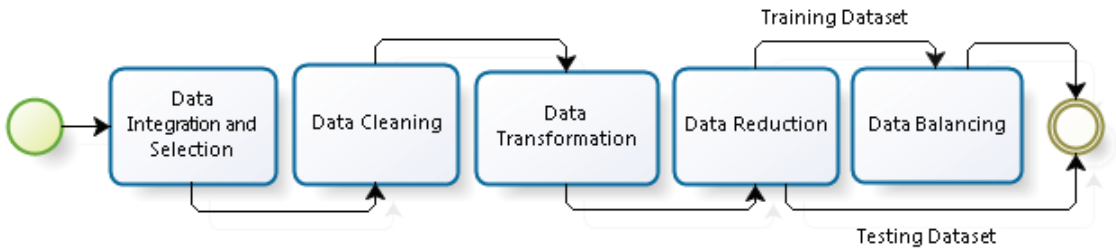

Figure 2 – Data Preprocessing.

1.1.1 Integration and Selection

The Integration stage, as demonstrated in Figure 25, permits the integration of different databases (multiple autonomous systems and heterogeneous data sources) in order to create a unified dataset for a complete data analysis [Han et al., 2011; Lenzerini, 2002; Halevy et al., 2006].

If the data integration phase disregards semantic heterogeneity and structure of data, redundancies and inconsistencies might appear, resulting in accuracy and speed decrease [Han et al., 2011; García et al., 2016].

A practical example of data integration, as shown in Figure 25, refers to the flight data integration. For each flight record (arrival and departure), the process looks for historical data concerning the climatic conditions through the web service of Weather Underground (WU). To each of flights and climatic conditions, this stage uses data of departure (date and time expected) and arrival (date and time expected), resulting in integrated base VRA-WU.

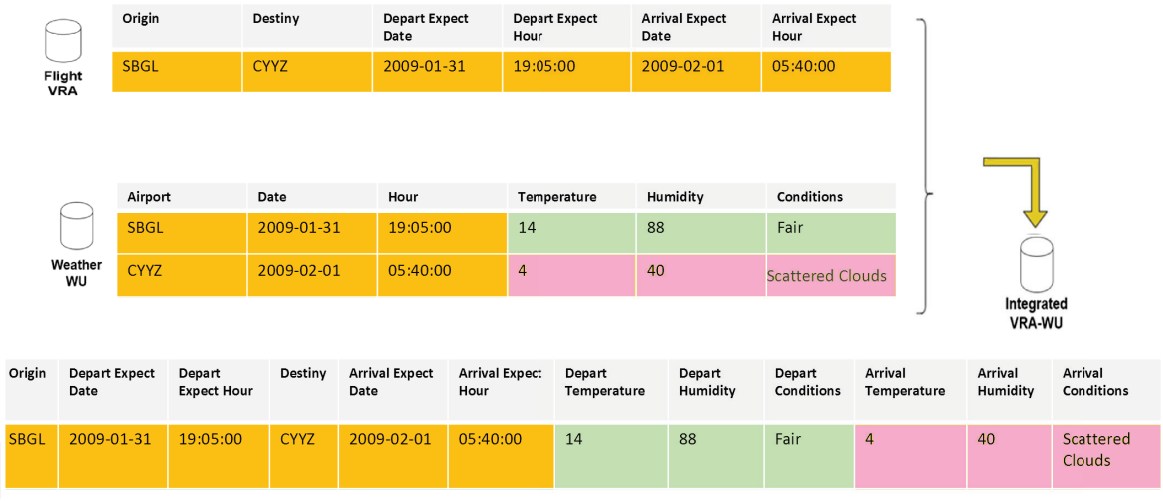

Figure 3 – Data Integration Example

Properly performing the data integration phase, dismissing semantic heterogeneity, and structure of data, the probability of redundancies and inconsistencies reduces substantially, resulting in accuracy and speed decreasing [Han et al., 2011; García et al., 2016].

Data Selection

Sometimes it may be necessary when performing the integration in a large set of data and attributes, the use of selection technique as a form of redundancy reduction, or even for the relevance concentration [Han et al., 2011; Rahm and Do, 2000].

Commercial aviation problems commonly employ data selection techniques; for example, from the flight frequency at a given airport [Sternberg et al., 2016].

In the context of commercial aviation (main airports, for example), sampling techniques can reduce the number of tuples, using methods as stratified samplings, quartiles, histograms, and Pareto [Xiong and Hansen, 2013; Rebollo and Balakrishnan, 2014].

1.1.2 Cleaning

Even after the data integration phase, errors may still exist in the dataset. This cleanup routine aims to correct these types of problems by filling in missing values, smoothing noisy data, identifying or removing outliers, and solving inconsistencies [García et al., 2016]. Hence, the data cleansing process is crucial, filtering incorrect data for the data set and processing, because dirty data may noise the mining procedure, resulting in unreliable output. It also can overload subsequent mining routines, although most mining routines have some procedures to deal with incomplete or noisy data [Han et al., 2011; García et al., 2016].

Inconsistent Data Entries and Duplication

Essential methods can remove or correct inconsistent entries. It includes inconsistency detection (when the data is available from different sources in different formats); domain knowledge (a significant amount of knowledge referring to attributes or rules ranges - that specify the relationships across different attributes); or data-centric methods

(when the statistical behavior of the data is used to detect outliers) [Han et al., 2011; Rahm and Do, 2000].

Figure 4 presents flight data discrepancies concerning values referring to the flight spent time. For example, negative time (-830) and duration over $24h$ denotes inconsistency.

Flight	Duration
1	5
2	-830
3	40
4	138
5	45
6	114
7	46
8	52
9	41
10	137
11	1578
12	450
13	58
14	60
15	39

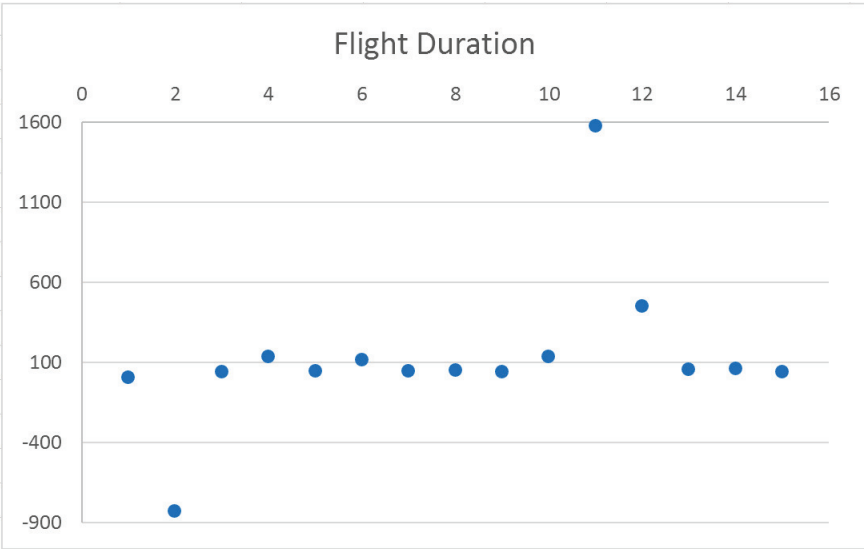

Figure 4 – Flight Duration Example - Inconsistency and Outlier

Redundancy can lead to duplication, especially with denormalized data, commonly used to accelerate processes involving join operations. This problematic duplication can be a source of inconsistency [Silberschatz et al., 2016; García et al., 2016].

The following step after data set integration is to check data errors. Data collected on a day-to-day basis tend to be incomplete, containing noises and inconsistencies. As the purpose is to use the data to generate classification models, data must be complete, correct, and compatible with reality to prevent the classifier's performance from being adversely affected. Hence, the cleanup step prepares data to use, either by identifying and removing outliers, by smoothing noisy data, or by filling in lost values [Han et al., 2011; Rahm and Do, 2000]. For this, there are methods to remove or fix as missing and inconsistent data entries.

Missing Data Entries

Missing data is a common problem faced by preprocessing data step. Various reasons may cause this situation, such as user response lack, storage problems, among others [Schafer and Graham, 2002].

Missing data mechanisms are classified as missing completely at random (MCAR), missing at random (MAR), and missing in non-Random (NMAR)[Little and Rubin, 1989]. In MCAR, the reason that caused the problem is the loss of power of the analysis. In MAR, the completed variables can explain the missing values. Regarding NMAR, the missing non-measurable data depends only on the missing attribute, revealing as the most severe form of missing data Soares [2007].

Some of the methods used to treat such missing data involve the removal or replacement of missing data.

There are methods to deal with lost data, as shown in figure 5. These can be divided into categories such as the exclusion of selected cases of variables, consisting of the straightforward deletion of data that contains missing values (such as listwise and pairwise, with a disadvantage of lead to possible biased parameter estimates); and data imputation(which aims to replace missing values with more plausible values in order to infer the information of the variables in that incomplete cases) where methods fill in missing values given the others the available data. A consequence of imputation is that the deviation in mean imputation decreases the variation in the data set. Mainstream imputation methods include case substitution, mean substitution, hot deck, regression, multiple imputations, and closest neighborhood imputation [Andridge and Little, 2010; Schafer and Graham, 2002].

Appropriate methods to process lost data tend to present satisfactory results even if the missing data are not originated by purely measurable factors [Brick and Kalton, 1996; Little and Schenker, 1995].

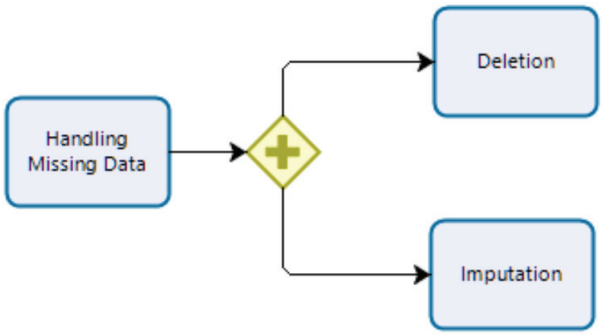

Figure 5 – Handling Missing Data.

Data with Missing

A	B	C
1	-	1
1	2	3

With Deletion Method

A	B	C
~~1~~	-	~~1~~
1	2	3

A	B	C
1	2	3

With Imputation Method

A	B	C
1	2	1
1	2	3

A	B	C
1	2	1
1	2	3

Figure 6 – Missing Data Example.

Redundancy

Redundancy caused by an attribute derived from another one should be an avoided problem because it can increase the size of the resulting dataset [Han et al., 2011]. Correlation analysis can detect redundancies. Given two attributes, the implication between them is calculated based on availability. The chi-square test is appropriated to use in nominal data. Numerical attributes demand correlation coefficient and co-variance, to evaluate variations of the value.

1.1.3 Transformation

The Transformation stage is responsible for transforming and consolidating data in an appropriate format, facilitating the data mining process and the understanding of hidden data patterns [Han et al., 2011]. The data transformation strategies used in this study are the Min-Max and Z-score normalization, Conceptual Hierarchy, Smoothing, and Categorical Mapping.

Normalization: Min-Max and Z-score

Normalization transforms the scale of the values of an attribute so that they fit into a new range. For example, because the unit of measurement used can affect data analysis, changing the unit of measurement from miles to miles can lead to different results. Therefore, to avoid using the unit of measures, the data should be normalized [Rissanen, 2001].

Normalization transforms data to a range, usually [0.0, 1.0] or [-1.0, 1.0], a very important process for classification algorithms such as neural networks or k-neighbors. It increases the speed of the learning phase and prevents attributes with distorted (very high or small) initial values, such as income, overlap attributes, and binary attributes [García et al., 2016].

Min-Max normalization is one of the standardization methods that apply a linear transformation in the original data, where the minimum value, min_A, and the maximum value, max_A, are used to transform each value v_i of an attribute A to a value v_i', in the new interval [$newMin_A, newMax_A$], as shown in the following equation 1 [Ogasawara et al., 2009], and exemplified in figure 7.

Min-max normalization is not useful or cannot be applied if minimum or maximum values of attribute A are not known. Although the minimum and maximum values are available, the presence of outliers can influence the min-max normalization by grouping them and limiting the digital precision available to represent the values [García et al., 2016]. In normalization Z-score (or zero-mean normalization), the values for an attribute,

A, are normalized based on the mean and standard deviation of A. A value, vi, of $\overline{A}$ is normalized to vi' by computing, as shown in equation 2 [Al Shalabi and Shaaban, 2006] and exemplified in figure 8.

$$v\prime_i = \frac{v_i - Min_A}{Max_A - Min_A} \cdot (NewMax_A - NewMin_A) + NewMin_A \quad (1)$$

$$v\prime_i = (v_i - \overline{A})/(\partial a) \quad (2)$$

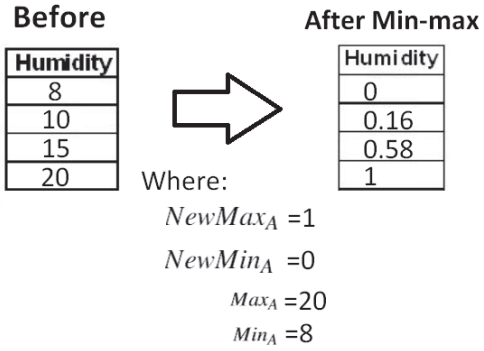

Figure 7 – Min-Max

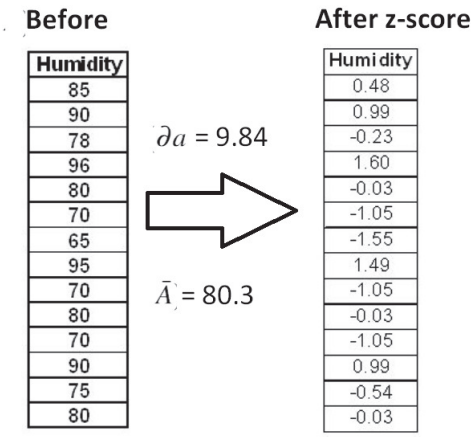

Figure 8 – z-score

Conceptual Hierarchy

The Conceptual Hierarchy is a preprocessing technique of the transformation stage. The objective is to transform an attribute in n other attributes, exploiting a hierarchical division among them.

This technique does not fit into one procedure step. It can also be used for the reduction step, replacing the original data by a smaller number of intervals and concepts representing them. It also simplifies the original data and makes the data mining process more efficient [Chan, 1998].

Nominal attributes have a finite number of distinct values without order between them, with many hierarchies implicit in the database schema. Concept hierarchies can be used to transform data at various levels of granularity, as demonstrated in Figure 10.

Discretization

The Smoothing technique is used to correct noises in data, generated by some random error or an unusual variation obtained in the variable measurement. These methods soft a data sample noise by querying the closest values and distributing them in several ”buckets” or boxes. Because smoothing methods query neighboring values for noisy values, they do a local smoothing on the data [Han et al., 2011]. Figure 9 demonstrates this technique.

Temperature Data: -3, 15, 18, 21, 24, 27, 29, 31, 36, 38, 39, 41.

Low: -3 to 21
Medium: 21.1 to 26
High: 27 to 41

Figure 9 – Smoothing

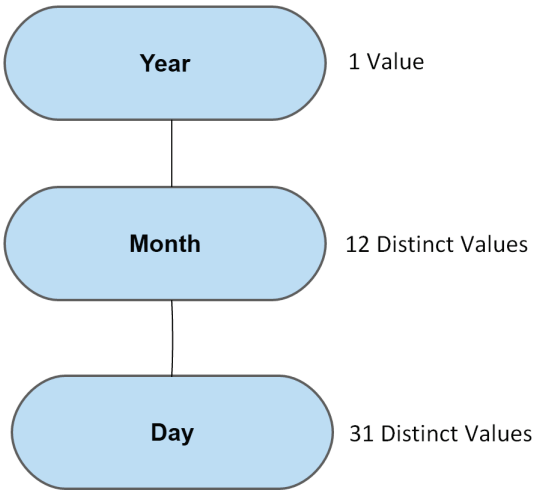

Figure 10 – Conceptual Hierarchy.

Categorical Mapping

Some machine learning methods need to use categorical attributes before usage. Assume a categorical attribute with n distinct values. The basic idea is to produce n derived binary attributes. Since n can be higher, many advanced approaches reduce the original 1-to-n mapping problem to a 1-to-k mapping problem with $k \ll n$. For this purpose, the cardinality of the data is first reduced by grouping individual values into k sets of values. Then each set is represented by a binary derived input, identifying the group the value belongs, and then the corresponding definition in the numerical representation. Original attribute needs to be grouped by values that present similar target statistics to become effective [Micci-Barreca, 2001]. Figure 11 demonstrates the use of this method.

ID	STATUS
A1	Light Wind
A2	Calm Wind
A3	Moderate Wind
A3	Strong Wind

ID	Light Wind	Calm Wind	Moderate Wind	Strong Wind
A1	1	0	0	0
A2	0	1	0	0
A3	0	0	1	0
A4	0	0	0	1

Figure 11 – Categorical Mapping.

1.1.4 Feature Selection and Extraction

Dimensionality impacts data differently depending on the following DM task or algorithm. The reduction step can create a shortened representation of the data set and still produce the same analytical result. A set of data to be analyzed can contain several attributes. However, some of them may be irrelevant during the mining process, or even redundant [Han et al., 2011]. For example, to classify airline companies based on flight delays, attributes such as flight numbers tend to be irrelevant, unlike starting delay time or real departure time, which are attributes that can add value to the analysis.

Some tasks to data reduction step are known as Feature Extraction (when a function calculates new features based on the original ones); and Feature Selection (where chooses an optimal subset according to a criterion).

The data reduction strategies used in this study for Feature Selection are LASSO, Information Gain (IG), Attribute Selection based on Correlation (CFS), and Principal Component Analysis (PCA) for Feature Extraction.

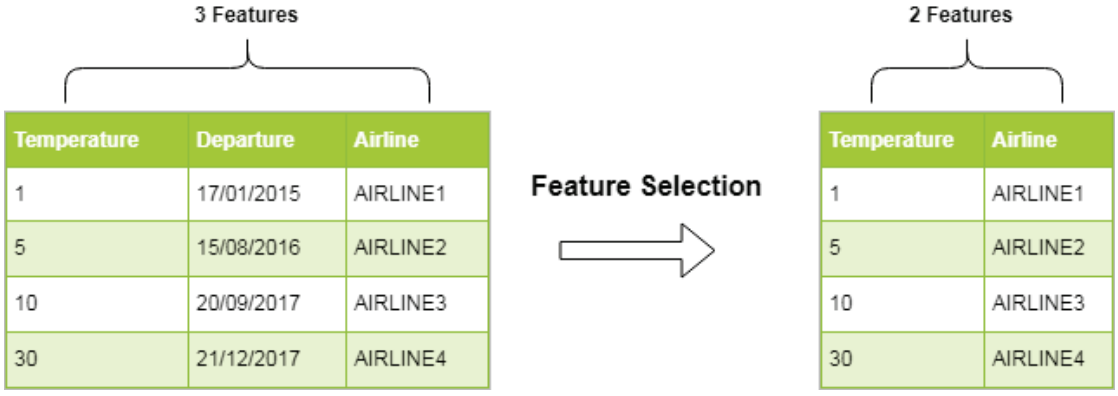

Figure 12 – Feature Selection

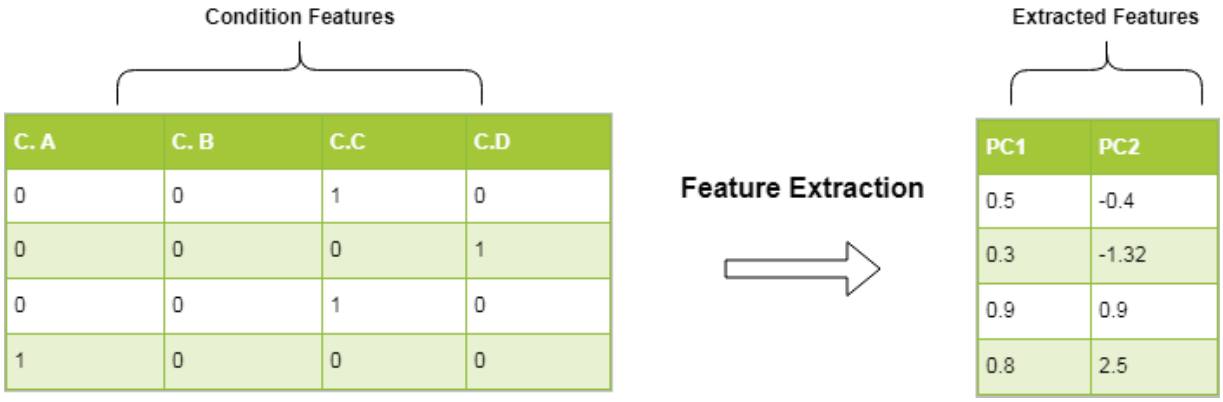

Figure 13 – Feature Extraction

LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) is a powerful two-tasks method (regularization and selection) which involves penalizing the absolute size of the regression coefficients and having as its primary objective to minimize the prediction error [Hastie et al., 2009].

This method performs a restriction on the sum of the absolute values of the model parameters, with the sum smaller than an upper limit. Hence, the method applies a shrinking process, also called regularization, where regression coefficients penalized, reducing some of them to zero. It also holds some of the favorable selection properties of both subsets, revealing the boundary regression stability [Tibshirani, 2011].

Supposing the data $(x^i, y_i), i = 1, 2, ...N$, where $x^i = (x_{i1}, ...x_{i1})^T$ are predictor variables and y_i is the response. Assume x_{ij} as standardized so that $\sum_i x_{ij}/N = 0, \sum_i x_{ij}^2/N = 1$.

Letting $\widehat{\beta} = (\widehat{\beta}_1 ... \widehat{\beta}_p)^p$, the lasso estimate $(\alpha, \widehat{\beta})$ is, according to Tibshirani [2011], defined by:

$$(\widehat{\alpha}, \widehat{\beta}) = argmin \sum_{i=1}^{N} (y_i - \alpha - \sum \beta_j x_{ij})^2 \text{ subject to } \sum_j |\beta_j| \leq t \tag{3}$$

where $t \geq 0$ is a tuning parameter.

Information Gain

Information Gain is a method that individually evaluates the accretion of each attribute. For an attribute, it is defined as the difference between entropy before and after the distribution of the data [Witten et al., 2011].

The use of this method makes it possible to obtain attributes that minimize the amount of information needed to classify the data. It is used to select the essential attributes, i.e., those that have the least entropy. It allows the treatment of the missing values separately, or to distribute counts to each other in proportion to their frequency [Quinlan, 1986].

To calculate the Information Gain (IG), firstly is necessary to calculate the entropy, computing the quality of a single (sub)set of examples corresponds to a single value, as shown in Equation 4.

$$E(D) = -\sum_{i=1}^{m} p_i log(p_i) \tag{4}$$

where p_i represents the probability that an object in D belongs, and E(D) represents the average amount of information needed to find out the class label of an object in partition D.

After that, it is necessary to compute the weighted average over all sets resulting from the split. I(D) represents the simplification of computation of average entropy (information), represented in equation 5.

$$I(D) = \sum_{j=1}^{v} \frac{D_j}{D} E(D_j) \tag{5}$$

where $\frac{D_j}{D}$ represents the weight of the j^{th} partition.

Finally, IG is the difference between the original information before splitting or partitioning, E(D); and the new Information, I(D), obtained after partitioning on A, as given in the equation below.

$$IG(A) = E(D) - I(D) \tag{6}$$

In other words, equation 6 calculates how much would be gained by branching on

A. It represents the expected reduction in the information requirement caused by knowing the value of feature A.

CFS

The CFS is a simple filter algorithm that classifies subsets of attributes according to a heuristic evaluation function based on the correlation. The bias of this function is for subsets that contain attributes that are highly correlated with the class and uncorrelated to each other [Hall, 1998].

Irrelevant attributes should be ignored because they will have a low-class correlation. The strong correlation with one or more remaining attributes recommends avoiding the redundant attributes. Accepting an attribute will depend on the extent to which it predicts classes in areas of space not already provided by other attributes [Hall, 1998]. CFS is given by equation 1.1.4:

$$r_{zc} = \frac{k\bar{r_{zi}}}{\sqrt{k + k(k-1)\bar{r_{ii}}}} \tag{7}$$

where the number of features, $\bar{r_{zi}}$ is the average of the correlation between feature-class and $\bar{r_{ii}}$ is the average inter-correlation between each pair of features [Hall, 1998].

PCA

The Principal Component Analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data, preserving most of the variation in the dataset [Jolliffe, 2002]. PCA combines the essence of attributes by creating a smaller set of variables [Han et al., 2011]. Moreover, from points in n-dimensional space, it presents patterns of similarity between observations and variables.

Its purpose is to extract the critical information from the data and express it as a set of new orthonormal variables, called Principal Components (PC), as shown in Figure 14. For this, the following process is performed: (i) normalization of the input data; (ii)

the calculation of the main components; (iii) the ordering of the significant components in descending order of significance or force; (iv) reducing the size of the data from the elimination of the more ineffective components, i.e., those with the smallest variance [Han et al., 2011].

$$C_{ij} = \frac{1}{n-1} \sum_{m-1}^{n} (X_{im} - \overline{X_i})(X_{jm} - \overline{X_j}) \tag{8}$$

where

C_{ij} covariance of the variable i and j

$\sum_{m-1}^{n}$ the sum of all n objects

X_{im} value of the variable i in object m

$\overline{X_i}$ means of variable i

X_{jm} value of variable j in object m

$\overline{X_j}$means of variable j

All eigenvectors are orthogonal (perpendicular). Hence, data is a linear combination of these vectors. The factor (value) multiplied to each vector is known as an eigenvalue.

The significant components are obtained by eigenvalues of the covariance matrix C, as presented in equation 9.

$$C_{vi} = \lambda_i v_i \tag{9}$$

The covariance matrix of the original data vectors X (represented by C and λ_i), refers to the eigenvalues of matrix C,v_i and corresponds to eigenvectors.

Considering $E_k = [v_1, v_2, v_3....v_k]$ and $\Lambda = [\lambda_1, \lambda_2, \lambda_3....\lambda_k]$, having $CE_k = E_k\Lambda$ obtains:

$$X^{PCA} = E_K^T X \tag{10}$$

The number of characteristics of the data matrix original X is reduced by multiplication with the matrix E_K^T, which has eigenvectors k corresponding to the highest eigenvalues k. The result of the array is X^{PCA}.

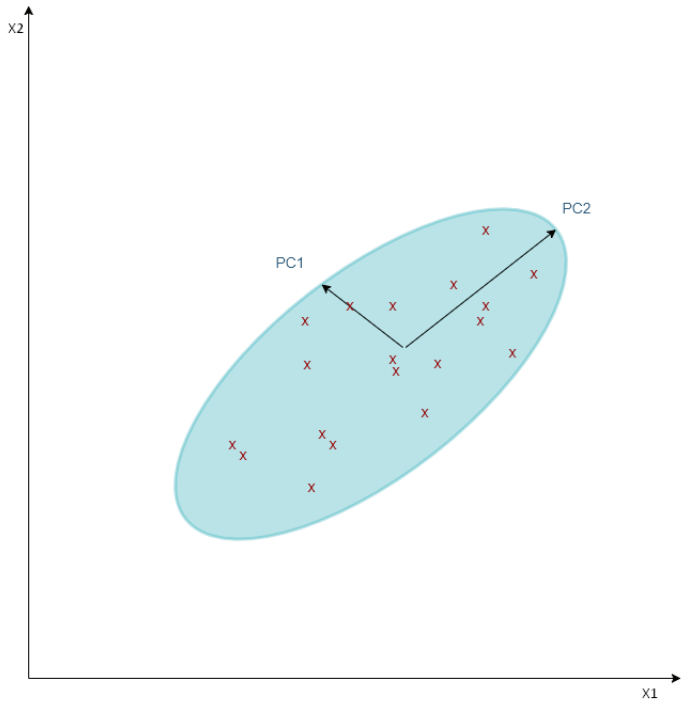

Figure 14 – PCA Example.

1.1.5 Data Sampling

Sampling can be used as a data reduction technique because it allows us to represent a large dataset using a sample of random data, or subsets, much smaller [Lantz, 2013]. The sample formation depends on the type of approach adopted.

Random Sampling consists of creating a subset where each tuple belonging to a dataset has the same probability of being selected to compose it. The Stratified Sampling consists of separating the dataset into mutually disjoint parts, called strata, extracting then a sample from each stratum generated [Han et al., 2011]. Thus, Stratified Sampling creates a reduced set of data that attempts to maintain the same ratio between the existing classes in the original dataset.

1.1.6 Balancing

A widespread problem in data mining is the class assignment in the dataset since inadequate distribution can induce the result of the classifiers. In several applications, the number of records of a particular class is much larger than the number of records belonging to another [Prati et al., 2009]. Some examples are the detection of credit card fraud, where the number of fraudulent transactions is much less than the number of legal transactions, and air delays, in which only about 25% of the flights show more delay than 15 minutes.

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 to obtain a more balanced distribution of the data and improve the performance of the data classification models [Prati et al., 2009]. The data-balancing strategies used in this study are Random Sub-Sampling and the Synthetic Minority Oversampling Technique (SMOTE).

Random Subsampling & Synthetic Minority Over-sampling Technique

It is a non-heuristic method that aims to balance the distribution of classes in the data from a random deletion of the tuples of the majority class, that is, the class more frequently in the original data set [Prati et al., 2009]. This random elimination can generate information loss about the majority classes[More, 2016].

Synthetic Minority Over-sampling Technique (SMOTE) is a data-balancing method that aims at generating synthetic tuples of the minority class in the data set. The minority class tuples oversampling accomplishes introducing synthetic tuples from a less frequent class tuple and its nearest neighboring k-tuples. The difference between the tuple attributes of the chosen minority class and the attributes of their neighbors generates the synthetic tuples. This difference is then multiplied by a random number between 0 and 1 and added to the chosen minority class tuple. Depending on the number of synthetic tuples needed, the nearest k neighbors are chosen randomly [Chawla et al., 2002].

As shown in Figure 15, the distribution of the original database was firstly presented,

naming the majority class (with the highest number of tuples) as 'condition 1' and the minority class as 'condition2'. Then, in Random subsampling, the balancing occurred from the deletion of tuples of the majority class, equaling with the minority. In SMOTE, tuples were introduced to the minority class until the number was equal to the majority class.

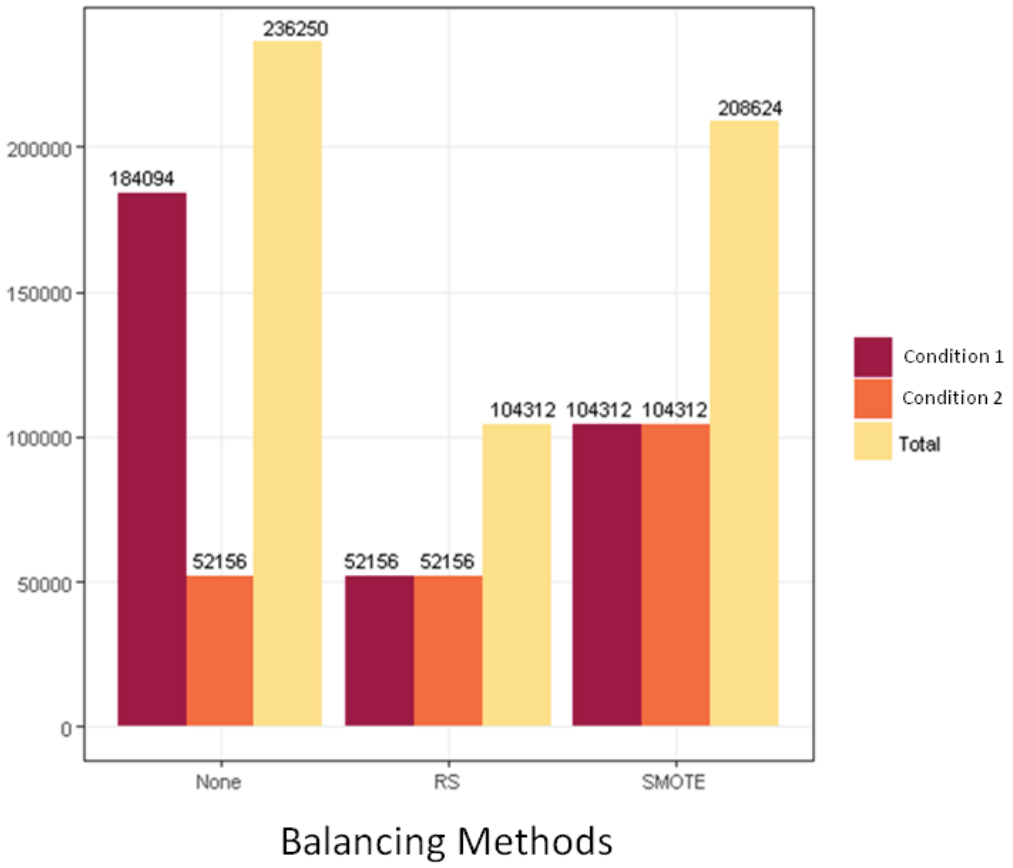

Figure 15 – Balancing Example.

1.2- Data Mining and Machine Learning

Machine Learning investigates how computers can learn (or improve their performance) based on the data [Han et al., 2011]. It is useful for automating complex pattern recognition processes and making smart decisions based on data.

These methods subdivide into supervised, unsupervised, semi-supervised, and active. The first ones, supervised, are represented by the classification methods since

supervision in learning is given by the known labels of the class in the training data set. Clustering methods represent unsupervised methods. The learning process in the input data does not have the class labels. The semi-supervised uses both labeled and unlabeled examples when learning to model. Finally, active learning is the approach that lets the user play an active role in the learning process because the user can be asked to label an example optimizing the goal by knowledge [Han et al., 2011].

1.2.1 Models

The machine learning methods were evaluated according to their hyper-parameters configurations targeting better accuracy during cross-validation [Bergstra et al., 2011]. Table 1 presented the general performance of machine learning methods for the dataset produced using LASSO. This table shows the approximate execution time, their ranking according to accuracy and the number of combinations of parameters to be explored for each method. This number was fixed into 28 different setups for each machine learning method. In the case of SVM, the kernel itself is a parameter. Thus, we had 14 parameters for each kernel. Also, it is worth mentioning that NB is a parameter-free method.

Table 1 – Analysis of Machine Learning Methods

Method	Accuracy(%)	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

Therefore, based on the analysis performed, considering their accuracy ranking, their execution time, and the number of combinations of parameters to be explored, the machine learning method chosen for analysis of the preprocessing methods were random forest (RF) and neural network with back-propagation (NN).

Neural Networks (NN)

Neural Network (NN) is an information processing system like biological neural networks as a generalization of the mathematical model of human cognition or neural biology. It comprises a computational approach performing information processing in basic units called neurons. Signals pass by these neurons through the connecting links.

Each connection link has an associated weight, which, in a typical neural network, multiplies the transmitted signal. Each neuron applies an activation function to its network input to determine its output signal [Fausett and others, 1994; Haykin et al., 2009].

As shown in Figure 16, a set of synapses, or connection links, each of which is characterized by a weight associated. A An signal at the synapse input n connects to the neuron t is multiplied by the synaptic weight wtn.

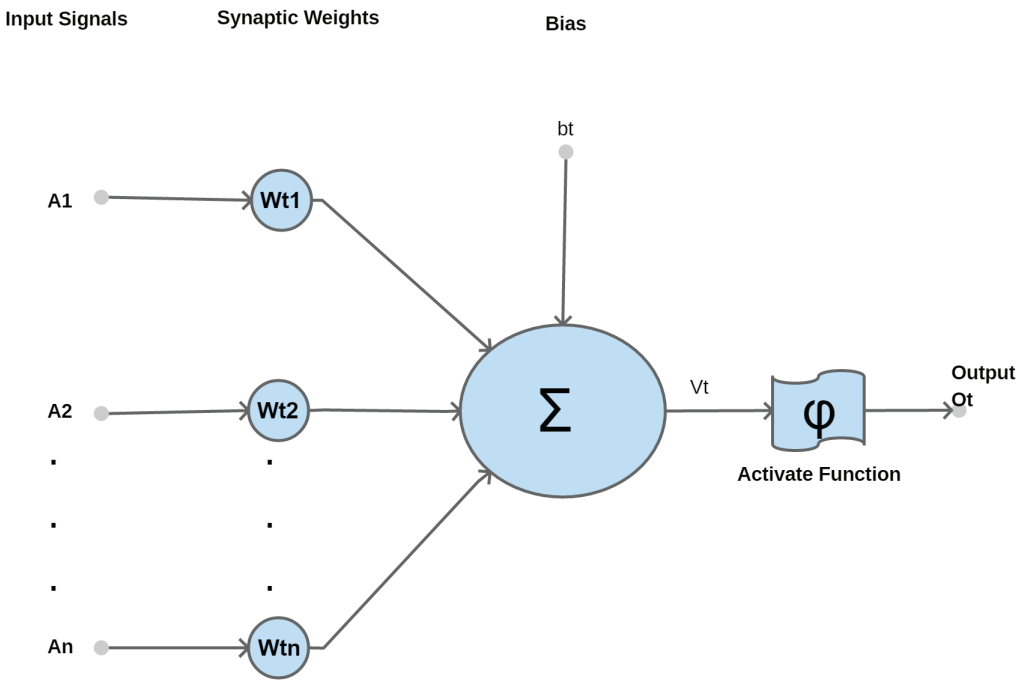

Figure 16 – Model of Neuron. Adapted from [Haykin et al., 2009]

The most common neural network is the multilayer perceptron Haykin et al. [2009].). In this model, each neuron in the network includes a non-linear activation function, which may contain one or more hidden layers from the input and outgoing nodes. The

network has a distributed presence of non-linearity and high connectivity that tenders to a more sophisticated theoretical analysis. Hidden neurons make it challenging to visualize the learning process, with the research focused on a much larger space of possible functions, and a choice has to be between alternative representations of the input pattern, as depicted in figure 17. One popular method for the training of multilayer perceptron is a backpropagation algorithm Fausett and others [1994].

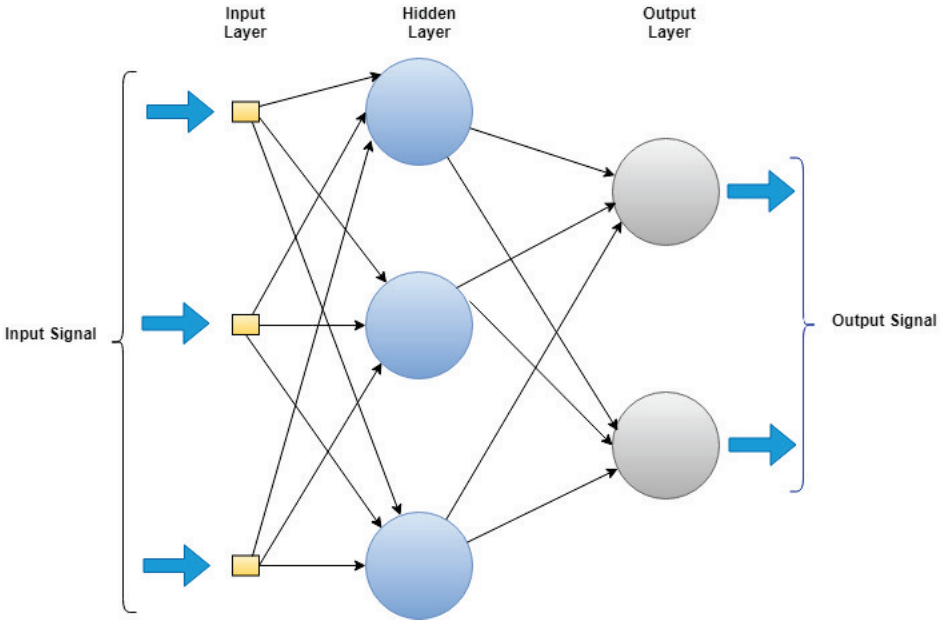

Figure 17 – Multi-Layer Perceptron.

Random Forest

Breiman [2001] proposed Random Forests (RF) combining decision trees so that each tree depends on the values of a vector sampled randomly independently and with the same distribution for all trees in the forest. Each generated decision tree is a result of an attributes random selection, done at each node, to determine the division [Han et al., 2011].

After the forest formation, the model uses the vote to combine the predictions of each tree. The most voted class is returned as a result of the forecast [Lantz, 2013] . Its precision depends on the strength of each tree and the dependence between them, as shown in Figure 18. The idea is to preserve the strength of each tree without increasing

its correlations [Han et al., 2011]. The generalization error for a forest converges while the number of trees in the forest is large, which makes the overfitting not being a problem [Han et al., 2011]. Besides, it can handle a large dataset since the set uses only a small random part of the original data set [Lantz, 2013].

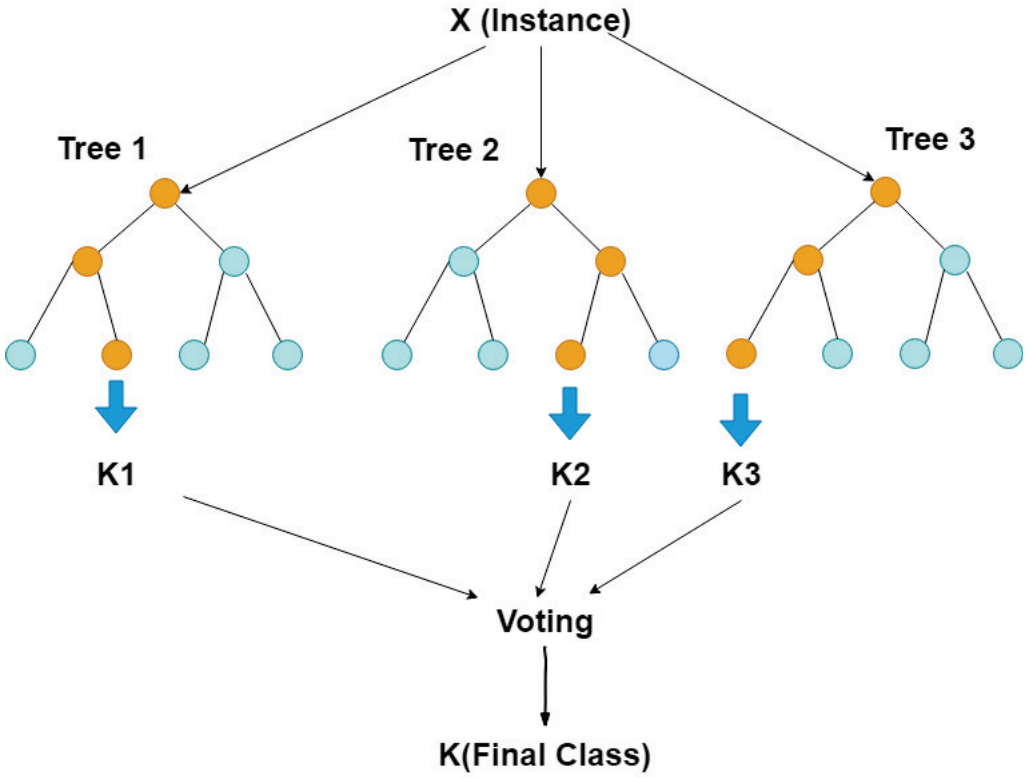

Figure 18 – Random Forest Architecture.

1.3- Model Evaluation

After the construction of the classification models, it is necessary to quantify its accuracy in a given data set, that is, to evaluate the effectiveness of the classifier comparing the different models, selecting the best one for the dataset, adjusting and tuning parameters [Aggarwal, 2015; Han et al., 2011].

1.3.1 Cross-Validation

Cross-validation is a statistical method for evaluating and comparing machine learning algorithms where data sets are divided into training and other test and/or model validation segments with the possibility of training and validation sets passing by crossing in successive rounds. The most commonly used methods are k-fold and holdout, however there are other methods like Leave-One-Out cross-validation, Re-substitution Validation [Refaeilzadeh et al., 2009].

K-Fold

The K-Fold cross-validation method consists of dividing the dataset into numerous k groups (usually between 5 and 10 groups) with random selection of groups for validation or testing. In this method there is repetition of training and validation iterations until all groups have been contemplated as represented in Figure 19. At the end of the validation group, a percentage of error is consolidated by representing the mean error of the model [Kohavi et al., 1995]. As an advantage of this method is the accurate performance estimation Small samples of performance estimation; But with disadvantage of overlapped training data, causing low performance and underestimated performance variance or overestimated degree of freedom for comparison.

Hold-out

Holdout method randomly divides labeled data into two disjoint sets: training and test. The most used proportion for this division is about two-thirds for training and the remainder for testing and derivation of the model, as represented in Figure 20. An essential feature of this method is that because the model training process does not use the set of tests. It is an indicator of how well the model performs in unseen data. Some

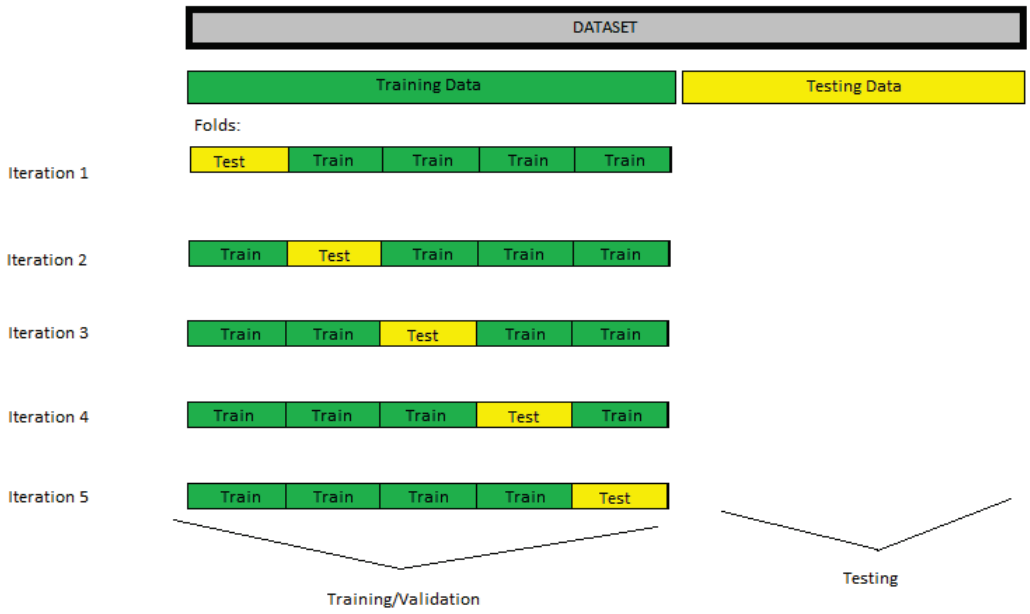

Figure 19 – 5-Fold Example

of the problems of this method are over-representation of classes in training, especially when the original class reveals an imbalance in its distribution [Aggarwal, 2015]. Still, considering that the data are independent with a need for a single execution, there is a lower computational cost. As a advantage of this method is the independent training and test with reduced data for training and testing, improving performance; but has as an disadvantage the large variance that can cause bias [Refaeilzadeh et al., 2009].

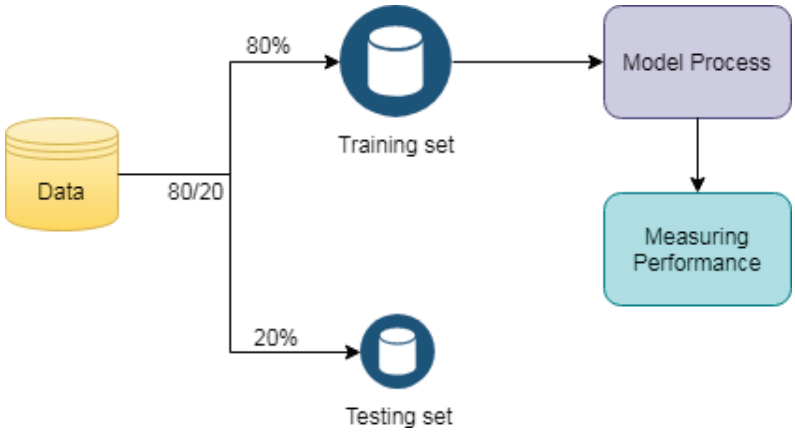

Figure 20 – Holdout - Measuring Performance. From Aggarwal [2015]

1.3.2 Measuring and Classification Performance Metrics

The main goal in the classification of learning algorithms is the construction of a classifier, which from a training set, can predict the test samples satisfactorily. Therefore, it is necessary to measure the predictive capacity of the classification algorithm, either by precision, accuracy, or other methods. However, in some cases, precision does not consider prediction probability, either by the classifier or unbalance of the data set (usually the class with the highest probability estimate is the same as the target). A more precise classification may surpass other methods.

Confusion Matrix

A confusion matrix is a tool used to measure the performance of the classification problem in machine learning, where the output can be of two or more classes, serving as the basis for calculating many other measures of performance [Kelleher et al., 2015]. Each cell represented in the confusion matrix represents one of four results (TP, FP, TN, FN) in binary classification, counting the number of occurrences of the result when presented to the test set as represented in Figure 21.

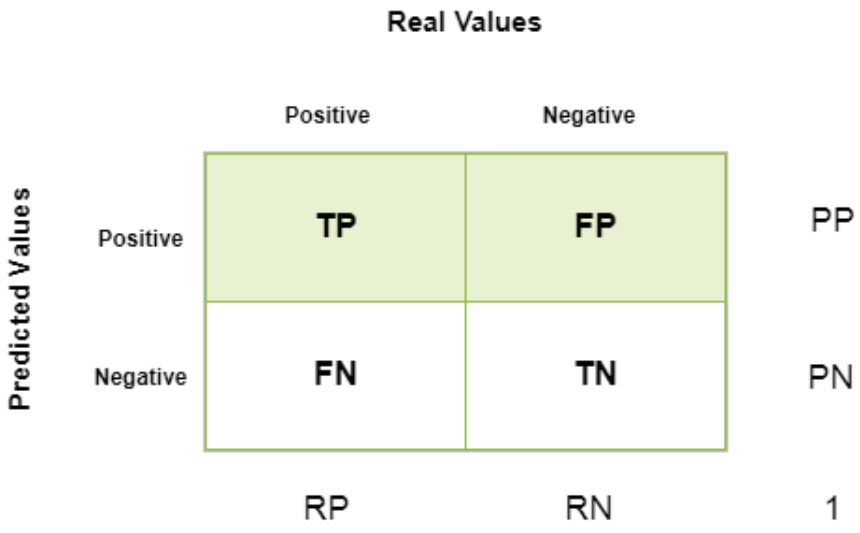

Figure 21 – Confusion Matrix

The upper-left cell of the results in the confusion matrix, TP (True Positive), represents the quantification of the total of instances in the test set with predicted value as positive and that they are positive values. Still in the upper cell, on the right, we have FP (False Positive) that represents the quantification of the total number of instances of the test set with the value predicted to be positive, but which were, in fact, negative values. In the lower left-hand cell, we have FN (False Negative) that represents the quantification of the total number of instances of the test set with predicted value as unfavorable. However, they were negative values. Still, in the lower part of the array, but in the right cell, we have TN (True Negative) that represents the quantification of the total instances of the test set with predicted value as unfavorable, and that was negative. As shown in the Figure 21, the columns of the table are labeled Real Value, both positive and negative, with the total positive value (RP) represented by equation 11, and the total negative value (RN) is represented by equation 12. The table lines represent positive and negative predictive values, with the positive predicted value (PP) represented by equation 13, and the total negative value (PN) represented by equation 14.

$$RP = TP + FN \tag{11}$$

$$RN = FP + TN \tag{12}$$

$$PP = TP + FP \tag{13}$$

$$PN = FN + TN \tag{14}$$

Rate measurement (TPR, TNR, FPR, FNR) is one of the measurement methods for verifying the actual prediction results. The equation 15 represents TPR (True Positive Rate), which means the proportion of instances predicted as, and that was positive concerning the total of really positive instances. TNR (True Negative Rate), represented by the equation 16, means the proportion of instances predicted, as and that was negative about the total of really negative instances.

$$TPR = \frac{TP}{(TP + FN)} \tag{15}$$

$$TNR = \frac{TN}{(TN + FP)} \tag{16}$$

Rate measurement (TPR, TNR, FPR, FNR) is one of the measurement methods for verifying the actual prediction results. The equation 15 represents TPR (True Positive Rate), which means the proportion of instances predicted as, and that was positive concerning the total of really positive instances. TNR (True Negative Rate), represented by the equation 16, means the proportion of instances predicted as, and that was negative about the total of really negative instances. FPR (False Positive Rate), represented by the equation 17, means the proportion of instances predicted to be positive. However, in reality, they did not concern the total of truly positive instances. FNR (False Negative Rate), represented by the equation 18, means the proportion of negative, but in truth positive, about the total number of really positive cases.

$$FPR = \frac{FP}{(TN + FP)} = 1 - TNR \tag{17}$$

$$FNR = \frac{FN}{(TP + FN)} = 1 - TPR \tag{18}$$

Accuracy

The accuracy of a classifier is a performance metric that, from a given set of tests, independent of the number of examples it has, represents the percentage of instances in the set of tests correctly sorted by the classifier [Han et al., 2011; Aggarwal, 2015]. It is represented by the equation 19 below.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{19}$$

Precision and Recall

Precision corresponds to a measure of exactness that reports the total percentage of correctly classified instances, while Recall, or Sensibility, corresponds to the percentage of true instances that have been correctly classified [Han et al., 2011; Aggarwal, 2015]. Precision and Recall are represented, respectively by the equation 20 and 21.

$$Precision = \frac{TP}{(TP + FP)} \tag{20}$$

$$Sensitivity = Recall = \frac{TP}{(TP + FN)} \tag{21}$$

F1 Score

F1 Score, also called F-Measure or F Score, is an accuracy measure that uses the weighted harmonic mean of the test's precision and recalls, as shown in Equation 22. This measure is widely used for the classification evaluation of unbalanced data, reflecting how good the classifier is in the presence of rare class [Han et al., 2011; Davis and Goadrich, 2006].

$$F1_{score} = 2\frac{(PrecisionXRecall)}{(Precision + Recall)} \tag{22}$$

Threshold

In classification, it is necessary to define a decision limit for mapping the values of binary categories. For this purpose, this definition is called a classification threshold (also called a decision threshold).

Often the use of conventional threshold, with the 50/50 ratio, is not enough. It

frequently happens when applied over unbalanced data. For this, there are many ways to maximize F1-Score in the context of binary classification [Krawczyk and Woźniak, 2015]. One way to calibrate the output is thresholding classifiers, reflecting the proportion of data, i.e., respecting the proportion of the majority and minority classes called the majority threshold [Lever et al., 2016]. Another possible way is to use cost-sensitive learning, which can produce probability estimates on training and test examples [Sheng and Ling, 2006; Lipton et al., 2014].

2- Related Works

This chapter will present a two-step analysis of existing work on preprocessing methods. Firstly, we will have a more general preliminary analysis of the techniques and methods used for classification in machine learning, then this research will be refined with a focus on preprocessing methods in relation to the context of this research that refers to delays. on flights.

In this first step, to establish a literature review map that shows similarities and differences when compared to the work presented, the string *("preprocessing" or "preprocessing method") and ("classification" or "prediction") and "machine learning "* was used to search for publications in the Science Direct database in June 2019. The search yielded 75 articles, out of a total of 455 results related to the topic. From these 75 articles, a verification was performed from the analysis of the introduction, methodology and results, which directly dealt with the use of processing methods, resulting in a complete reading of a subset of 22 articles, which supported this first stage of the process review.

As shown in Chapter 1, there are numerous preprocessing techniques. The related works, quantitatively, used the techniques and models according to Tables 2 and 3.

Table 2 – Comparison of the techniques used in the related works for Pre-Processing

Preprocessing Techniques				
Integration	Cleaning	Reduction	Transformation	Balancing
5	11	17	17	8

Table 3 – Comparison of the Models used in the related works for Classification

Supervised Models					
NB	SVM	MLP	RF	KNN	Others
5	8	7	3	8	10

Torunoğlu et al. [2011]; Uysal and Gunal [2014] performed analysis on data preprocessing techniques in text mining, aiming at classification in the text using a wide range of datasets. Extensive experiments use stop-word, stemming, and weighting of words, and report their effect on classification performance. For this, classifiers such as Naive Bayes, Support Vector Machines, and K-Nearest Neighbor.

Nikulin et al. [1998] addressed the extraction (data reduction) method explicitly designed for preprocessing magnetic resonance spectra of biomedical origin to search for and select optimal spectral subregions. This research demonstrates the method in two biomedical examples: discrimination between meningioma and astrocytoma in biopsies of brain tissue, and a colorectal classification biopsy in normal and tumor classes. Both preprocessing methods lead to classification accuracies greater than 97% for both examples. A similar approach was dealt with in AlMuhaideb and Menai [2016]; Bilski [2014]; Luypaert et al. [2004].

García et al. [2012] evaluated how learning is affected when different resampling algorithms transform the originally unbalanced date into artificially balanced class distributions, mainly on the influence of the imbalance ratio and the classifier on the effectiveness of the most popular resampling strategies. Classification uses the rule of k-neighbors closer (1,7,13-NN), a multilayer perceptron (MLP), a support vector machine (SVM), Naive Bayes classifier (NBC), a decision tree (J48), and a base function network (RBF). Iliou et al. [2017]; López et al. [2012]; Tsoi and Back [1995]; Marques et al. [2011] applies a similar approach.

Majidi and Oskuoee [2015] propose new methods for data preprocessing based on the first, second, and infinite signal norm. It also uses the autocorrelation function (ACF), performing resource extraction and data compression in a single step, as well as the fractional resources extraction. The neural network pattern recognition toolbox (nprtool) supported the standards classification.

Li et al. [2008]; Xiang-wei and Yian-fang [2012] addressed the effects of noise, distortion, observational environment, and other factors that make preprocessing adequate before automatic sorting. For this purpose, preprocessing may include, for example, noise elimination, calibration, flow standardization, continuous normalization, wild point removal, skyline subtraction, feature extraction to improve the quality of spectral data, suppression of unnecessary distortion to raise specific spectral characteristics of automatic processing. The classification method used several methods, including k-nearest neighbor classifier (KNN), the Fisher linear discriminant analysis (FDA), and the vector support machine (SVM).

Kamiran and Calders [2012]; Lavangnananda and Waiwing [2015]; Luypaert et al. [2004] target to improve the accuracy of the classifier, but without discrimination of its predictions. Existing data experienced the use of preprocessing techniques coupled with

the suppression of the sensitive attribute, modifying the dataset by changing the class labels or replacing/resampling the data to remove discrimination without reclassifying the instances.

Kotsiantis et al. [2006] demonstrate the data preparation and filtering steps processing time in machine learning problems, with data preprocessing including various methods of cleaning, normalization, transformation, extraction of characteristics, and selection. Huang et al. [2015]; García et al. [2016] applies a similar approach.

Dara et al. [2008] try to determine whether major preprocessing complaints before automatically classifying them improves classification performance. It uses preprocessed master complaints using two preprocessors (CCP and EMT-P) and evaluating whether classification performance for a probabilistic classifier (CoCo) or a classifier based on keywords (modification of the New York Department of Health) and Mental Hygiene coder chief of complaints (KC)).

Hoshyar et al. [2014]; Xu et al. [2016] involves preprocessing of images for improvement in pattern recognition and classification.

Table 4 shows the related works selected after the search.

Table 4 – Related Publications

Pub	Reference
1	Torunoğlu et al. [2011]
2	Nikulin et al. [1998]
3	García et al. [2012]
4	Majidi and Oskuoee [2015]
5	Li et al. [2008]
6	Kamiran and Calders [2012]
7	Kotsiantis et al. [2006]
8	AlMuhaideb and Menai [2016]
9	Huang et al. [2015]
10	Iliou et al. [2017]
11	Xiang-wei and Yian-fang [2012]
12	García et al. [2016]
13	Tsoi and Back [1995]
14	Bilski [2014]
15	Luypaert et al. [2004]
16	Lavangnananda and Waiwing [2015]
17	Luypaert et al. [2002]
18	Xu et al. [2016]
19	Dara et al. [2008]
20	Uysal and Gunal [2014]
21	Marques et al. [2011]
22	López et al. [2012]

Tables 5 and 6 summarize the number of citations and lists related jobs that use the data preprocessing and machine learning techniques described in the background sections. The end of each table exhibits the percentages of use of each of the preprocessing techniques and classification models.

Table 5 – Publications on preprocessing methods for classification

Pub.	Preprocessing Techniques				
	Integration	Cleaning	Reduction	Transformation	Balancing
1			X		
2			X		
3					X
4			X	X	
5			X	X	
6			X	X	X
7	X	X	X	X	
8		X	X	X	
9			X	X	
10	X	X	X	X	
11			X	X	
12	X	X	X	X	X
13				X	
14	X				
15		X	X	X	
16		X		X	X
17		X	X	X	
18		X	X	X	
19		X	X	X	X
20		X	X	X	X
21	X	X	X	X	X
22					X
Total	22.7%	50%	77.2%	77.2%	36.3%

Table 6 – Publications on preprocessing methods for classification

Pub.	Classification Models					
	NB	SVM	MLP	RF	KNN	Others[1]
1	X	X			X	
2						X
3	X	X	X	X	X	X
4		X		X		X
5		X			X	
6	X					
7						
8						
9			X			X
10	X	X	X		X	X
11						X
12						
13			X			
14		X	X			X
15						X
16			X			
17					X	
18					X	
19						X
20	X	X		X	X	
21			X			
22		X			X	X
Total	22.7%	36.3%	31.8%	13.6%	36.3%	45.4%

From the overview of the preprocessing methods obtained through the survey, a second step of this literature review was performed in order to establish a view that demonstrates the similarities and differences compared to the present work. It wishes to demonstrate the use of preprocessing methods in the context of flight delays. Hence, a

[1] `LDA, JR48, RBF, CBR, CART, KeywordClassifier, SOM, RoughSets, Fuzzy, BayesNet.`

search was carried out in Science Direct Database publications on October 2019, using the search string (”classification” or ”prediction”) and (”flight” or ”air”) and ”machine learning” and ”delay.” This search resulted in approximately 44 results within 276 items returned from this set of articles related to the topic, after an abstract observation. From that 44 results, nine were selected for a complete read because of a direct relationship with this research, after a complete reading.The other works not selected to a complete read was retired as example in lecture of introduction for presenting themes like mapping causal, ticket price and Arrival time prediction, that are not the guideline of this work. Tables 7 and 8 presents the selected related works and data used in each work after the search. Table 9 presents the better results achieved in those related works.

Table 7 – Selected Related Works

Pub.	**Reference**	**Main Target**
1	Rebollo and Balakrishnan [2014]	Prediction of air traffic delays
2	Cao and Fang [2012]	Airport Flight Departure Delay
3	Khaksar and Sheikholeslami [2019]	Predict Delay Occurence and Magnitude
4	Chakrabarty et al. [2019]	Delay prediction of individual flight
5	Choi et al. [2016]	Prediction of Weather-induced Airline Delays
6	Nigam and Govinda [2017]	Flight Delay Prediction
7	Choi et al. [2017]	Cost-sensitive Prediction of Airline Delays
8	Saadat and Moniruzzaman [2019]	Airlines Delay Prediction
9	Belcastro et al. [2016]	Predicting Flight Delays
10	Henriques and Feiteira [2018]	Predict arrival delays of individual flight

Table 8 – Data used in Selected Related Works

Pub.	**Related Data**
1	Historical Flight Data and weather information
2	Historical Flight data
3	Historical Flight Data and weather information
4	Historical Flight data
5	Historical Flight Data and weather information
6	Historical Flight Data and weather information
7	Historical Flight Data and weather information
8	Historical Flight Data
9	Historical Flight data and weather information
10	Historical Flight data; Weather data; Airplane info; Delay Propagation information

Rebollo and Balakrishnan [2014] use Random Forest to predict Flight Delays with data from Historical flight and forecast weather data, considering both temporal and spatial delay states. The authors were able to achieve 81% accuracy and 76.40% Recall of the

model in best predictions.

In Cao and Fang [2012], based on the flight data created a BN model and Experiments show that parameters learning can reflect departure delay, achieving 88.33% accuracy of the model in best predictions.

Khaksar and Sheikholeslami [2019] objective is to predict flight delay with different machine learning algorithms approachs as bayesian modeling, decision tree, cluster classification, random forest and hybrid method to estimate the occurrences and magnitude of delay in network, using Us flight and Iranian airline datasets (specially visibility, wind and departure time). The authors were able to achieve 76.44% accuracy and 60% Recall of the model in best predictions.

Chakrabarty et al. [2019] analyse arrival delay of the flights using data mining and supervised machine learning algorithms [random forest, Support Vector Machine (SVM), Gradient Boosting Classifier (GBC) and k-nearest neighbour algorithm(KNN)] to obtain the best performing classifier with data collected from BTS, United States Department of Transportation(flights operated by American Airlines, connecting the top five busiest airports of United States in the years 2015 and 2016). Some of features utilized: Year, Quarter, Month, Day of Month, Day of Week, Flight Num, Origin Airport ID, Origin World Area Code, Destination Airport ID, Destination World Area Code, CRS Departure Time, CRS Arrival Time, Arr Del 15. With use of Gradient boosting, the authors were able to achieve 79.72% accuracy, 76% Precision, 80% Recall and 74% F-Score of the model in best predictions.

Choi et al. [2016] predict airline delays caused by inclement weather conditions using data mining and supervised machine learning algorithms[Decision Trees(DT), Random Forest(RF), AdaBoost, k-Nearest-Neighbors Classifier (kNN)]. The data used refers to US domestic flight and the weather conditions from 2005 to 2015. Some example of data used is: Quarter of Year, Month, Day of Month, Day of Week, Departure and Arrival Schedule in Local Time, Arrival Delay Indicator, Wind Direction Angle [deg], Wind Speed Rate [m/s], Visibility [m], Precipitation [mm], Snow Depth [cm], Snow Accumulation [cm] and others. With use of Random Forest, the authors were able to achieve 83.4% accuracy of the model in best predictions.

Nigam and Govinda [2017] forecast Flight delay logist regression supervised learning method, using historical flight data and weather data such as temperature, humidity, precipitation and dew point. The authors were able to achieve 80.6% accuracy,

32.1% Precision, 11.5% Recall and 20.9% F-Score of the model in best predictions.

Choi et al. [2017] use a combining of the sampling method called costing and supervised machine learning algorithms to predict individual flight delays. The costing method converts cost-insensitive classifiers to cost-sensitive ones by subsampling examples from the original training dataset according to their misclassification costs. This study uses flight and weather data (Destination, Quarter of Year, Month, Day of Month, Day of Week, Scheduled Departure Time in Local Time, Scheduled Arrival Time in Local Time, Arrival Delay Indicator,Wind Direction Angle [deg], Wind Speed Rate [m/s], Visibility [m], Precipitation [mm] , Snow Depth [cm] and others). The authors were able to achieve 83.07% accuracy of the model in best predictions.

Saadat and Moniruzzaman [2019] predict airlines flight delays by analyzing flight data, especially, for the domestic Airlines those moves around the United States of America. In order to transform the high dimension data into a low dimension Principal component analysis is used. This work uses Deep learning algorithms [Recursive Neural Network (RNN), Deep Neural Network (DNN), Convolutional neural network (CNN), Deep belief network (DBN) and more. With this deep learning approach the authors were able to achieve 82.1% accuracy of the model in best predictions.

Belcastro et al. [2016] predict arrival delay uses flight information (origin airport, destination airport, scheduled departure and arrival time) and weather forecast at origin airport and destination airport according to the flight timetable and as supervised machine learning the Random Forest(RF) algorithm, achieving 85.8% accuracy and 86.9% Recall of the model in best predictions.

Henriques and Feiteira [2018] objective to predict the occurrence of delays in arrivals at the international airport of Hartsfield-Jackson using several data mining techniques and historical flight and weather data, besides delay propagation. This work uses Decision Trees, Random Forest and Multilayer Perceptron as machine learning classifiers. The authors were able to achieve 85.63% accuracy of the model in best predictions.

Table 9 – Results achieved in Selected Related Works

Pub.	Classifier	Threshold	Results			
			Accuracy	**Precision**	**Recall**	**F1-score**
1	Random Forest	Conventional	81.00%	-	76.40%	-
2	Bayesian network	Conventional	88.33%	-	-	-
3	Hybrid(decision tree combined with cluster classification)	Conventional	76.44%	-	60.00%	-
4	Gradient boosting	Conventional	79.72%	76.00%	80.00%	74.00%
5	Random Forest	Conventional	83.40%	-	-	-
6	Logistic Regression	Conventional	80.60%	32.10%	11.50%	20.90%
7	Adaboost	Conventional	83.07%	-	-	-
8	Deep Learning	Conventional	82.10%	-	-	-
9	Random Forest	Conventional	85.80%	-	86.90%	-
10	MLP	Conventional	85.63%	-	-	-

Taking into consideration the description presented in related works, this work intends to through the massive use of transformation techniques, as well as other pre-processing techniques such as integration, cleaning, reduction and balancing in order to understand and improve the performance of models already presented with this approach for the prediction of flight delays.

3- Methodology

This work aims at evaluating data preprocessing methods for predicting Brazilian flight delays optimizing the accuracy, sensitivity and F1-Score of the prediction models, considering all the factors involved and collected by the dataset.Those measure meters were considered basicaly because is important to show if accuracy is higher than the proportion of the majority class and together with the sensitivity, which is directly linked to the true positives, that is, how many of these delays were correctly classified, in addition to the F1- Score that will show in a balanced way the set of the classifier in terms of sensitivity and precision.

Specifically, this work focus on transformation strategies as demonstrated on Subsection 1.1.3: 1) **normalization**, that transforms the scale of the values of an attribute so that they fit into a new range; 2) **categorical mapping**, when the original attribute needs to be grouped by values that present similar target statistics to become effective [Micci-Barreca, 2001]; and 3) **discretization**, replacing raw values of numerical attributes by interval or conceptual labels [Han et al., 2011], in addition to the tuning of hyperparameters in the machine learning process. A proposed methodology helps to evaluate these preprocessing methods.

The methodology used to analyze flight delays is composed of six steps, as presented in the workflow of Figure 22. Step 1 (Section 3.1, 3.2 and 3.3) consists on the databases integration, selection of relevant data and cleaning and removal of the outliers. Step 2 (Section 3.4) addresses data processing and consolidation. Step 3 (Section 3.5) is responsible for performing training and test sampling. Step 4 (Section 3.6) applies balancing methods to the training data. Step 5 (Section 3.7) is responsible for treating the curse of dimensionality (a term coined by Bellman and Dreyfus [2015], which refers to the phenomena caused by the exponential increase in data volume associated with a large number of dimensions in a mathematical space). It also conducts different attribute selections in the data. Step 6 (Section 3.8) consists of applying machine learning methods, performing hold-out validation, optimizing hyper-parameters and evaluation.

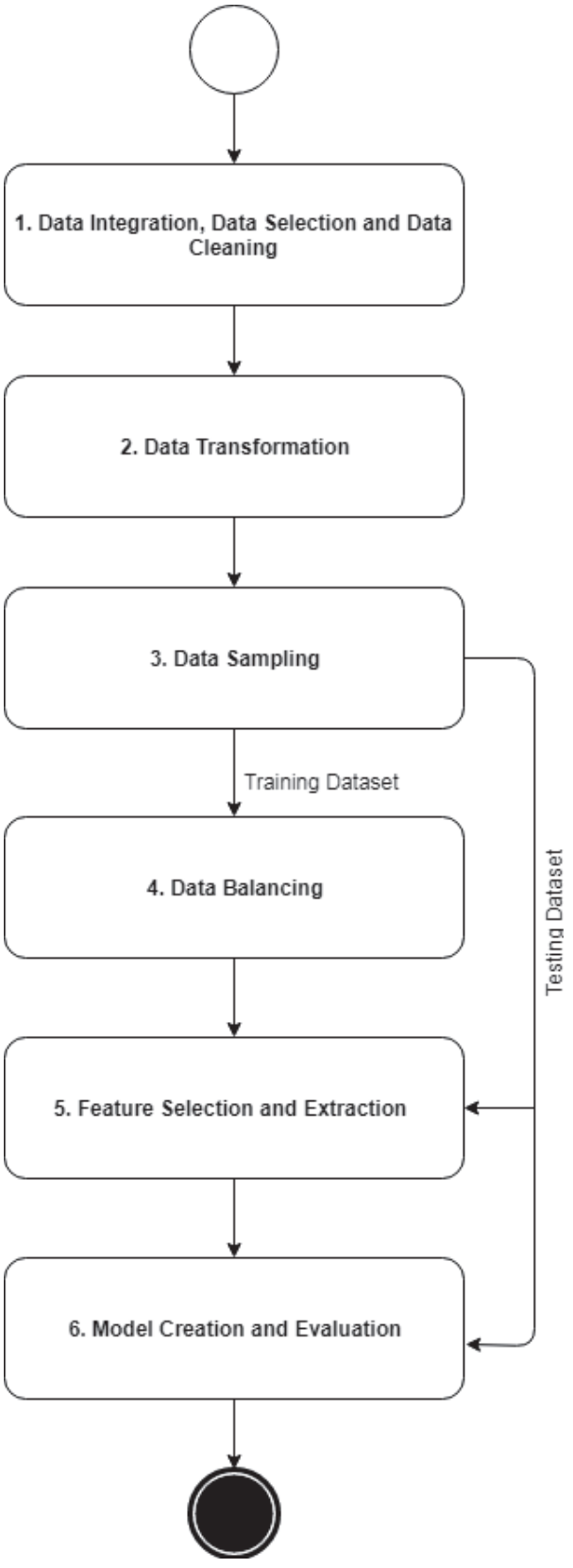

Figure 22 – Workflow

3.1- Integration

ANAC is responsible for regulating and supervising civil aviation activities in Brazil, which includes all commercial flight takeoffs and landings on Brazilian airports. The agency provides a public dataset named VRA [ANAC, 2016], updated monthly. It contains information about flight operations, such as the scheduled time for departure and arrival in each airport, as well as actual departure and arrival times.

VRA dataset has no weather information. Thus, to create a complete flight dataset, an integrated dataset was formed using data from the weather service provider Weather Underground (WU). WU provides hourly information about weather conditions (as temperature, pressure, humidity) for each airport. The integration process considered weather conditions closest to the scheduled departure and arrival times of each airport for all flights.

Tables 10 and 11 contains the source data descriptions, regarding the type of variable and description of VRA and WU to the integration process.

The data integration (Figure 23) illustrates the concept of the data integration process from data extraction of two bases, one of the realized flights (VRA) and one of the meteorological data (WU - Weather Underground) that converges, after several selections, to VRA-WU integrated database.

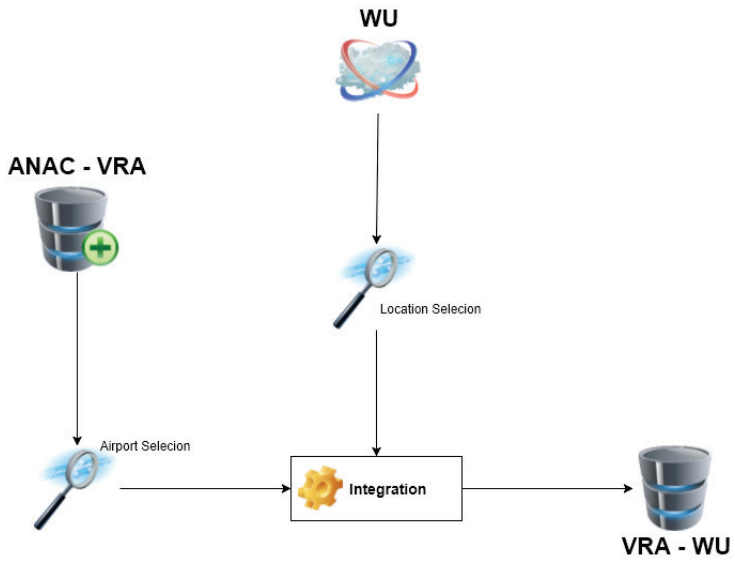

Figure 23 – Data Integration Concept.

Table 10 – VRA Attributes

Attribute	Type	Description
destiny	char(4)	Destiny of Flight
origin	char(4)	Origin of Flight
airlines	char(3)	Name of the company that provides air transport services
flight	Seq	Number of Flight
autho_code	char(1)	Identification of flight Authorization
line_type	char(1)	Identification of Type of Line
depart_expect	Datetime	Date and Time of Expected Depart
depart	Datetime	Date and Time of Real Depart
arrival_expect	Datetime	Date and Time of Expected Arrival
arrival	Datetime	Date and Time of Real Arrival
status	char	Status of Flight
observation	char(2)	Observations about Flight
depart_expect_date	Date	Date of Expected Depart
depart_expect_hour	Time	Time of Expected Depart
arrival_expect_date	Date	Date of Expected Arrival
arrival_expect_hour	Time	Time of Expected Arrival
departure_delay	Integer	Delay of Departure
arrival_delay	Integer	Delay of Arrival
duration_expect	Integer	Expected Duration of Flight
duration	Integer	Real Duration of Flight
duration_delta	Integer	Difference Between Real and Expected Duration
name.x	String	Complete Name of Origin
city.x	String	City of Origin
state.x	String	State of Origin
name.y	String	Complete Name of Destiny
city.y	String	City of Destiny
state.y	String	State of Destiny

Table 11 – Weather Attributes

Attribute	Type	Description
data.airport	char(4)	Airport Code
data.date	Date	Date of Weather Forecast
data.hour	Time	Time of Weather Forecast
data.temperature	Integer	Temperature
data.dewpoint	Integer	Dew Point
data.humidity	Integer	Humidity
data.pressure	Integer	Pressure
data.visibility	Integer	Visibility
data.events	String	Events
data.conditions	String	Description of Conditions

Using the conceptual model that demonstrates the integrated data schema from the VRA and WU databases is another way to represent this integration. It describes

the structure of the system, presenting its classes, relations, and attributes. Figure 24 represents this model, by the classes Airports, Flight and Weather and their respective attributes; and the relations between flight and airport that represent the origin and destiny of flights, beyond the relationship between flight and weather that represents the weather conditions on departure and arrival.

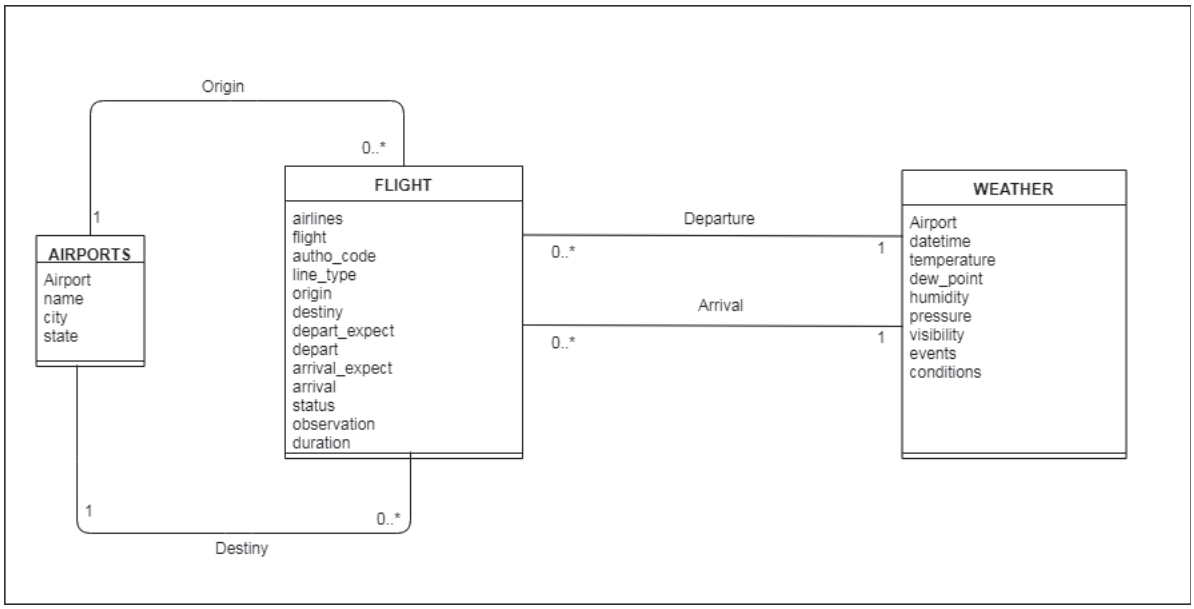

Figure 24 – Data Integration Schema.

Figure 25 exemplifies a way of practically data integration VRA and Weather WU databases. In the first tuple of the flight data table (marked in gold color), there are data from a flight that departed from SBGL to CYYZ, with their respective date and hour.

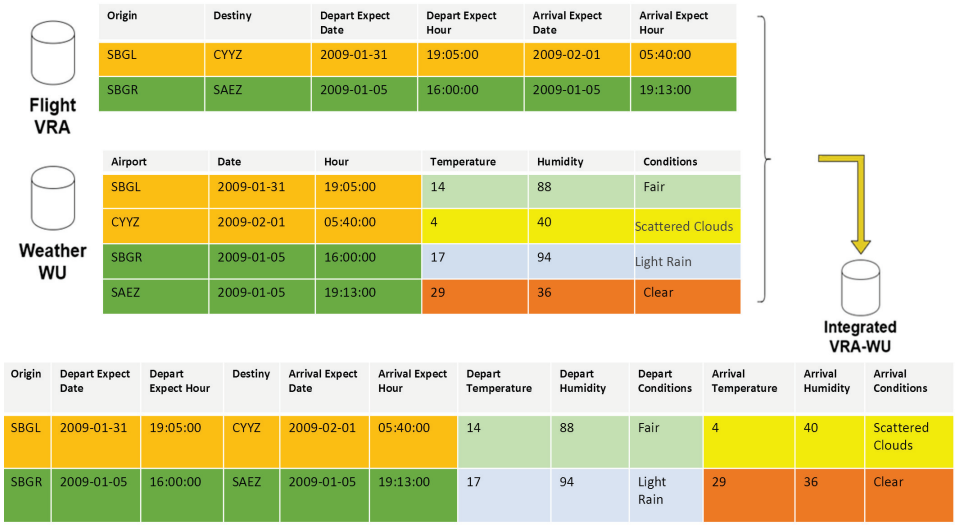

Figure 25 – Data Integration Example.

3.2- Selection

Firstly, the airport presentation attended ANAC database, which contains more than 200 airports. Based on these data, the analysis of the number of flights per airport was carried out, resulting in the choice that concentrates 94% of the trips made in the country, as shown Distribution Diagram in Figure 26.

That analysis returned the major 62 airports shown in Figure 27, that interact with Brazilian flight mesh. They correspond to 94% of all monitored flights by ANAC. Ten of these are foreign airports (KMIA, SAEZ, SABE, SCEL, MPTO, LPPT, SUMU, SPJC, SKBO, KJFK).

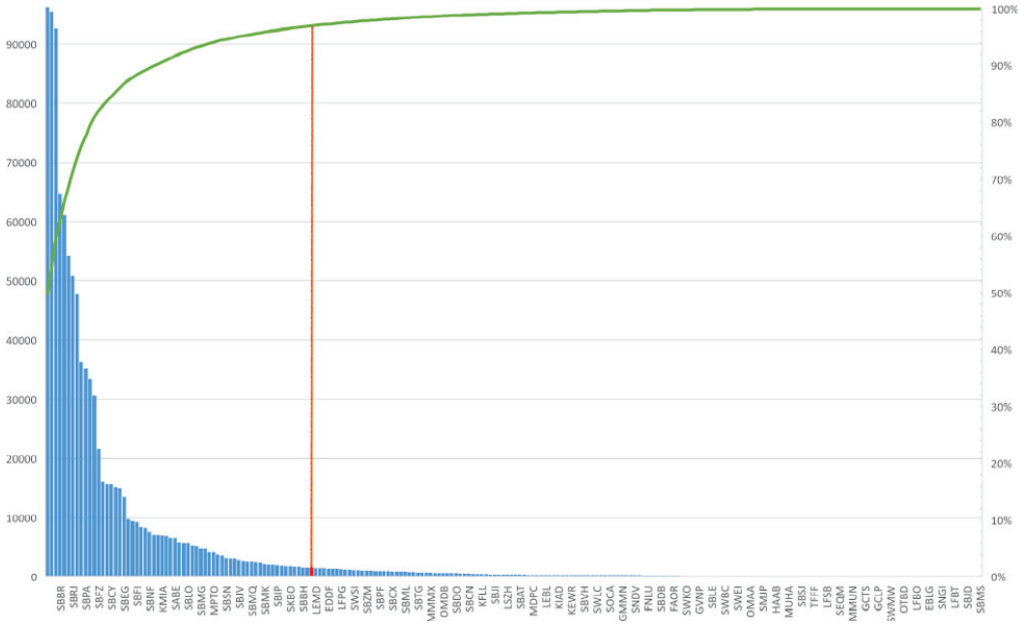

Figure 26 – Distribution Diagram

Selected airports guided the selection of the travel data (arrivals and departures) from the ANAC database (VRA), as shown in Table 12 and Figure 28. This list summarizes the number of original records and the resulting value of the integration and selection of flight data. 24/5000 Even considering the selection task, one can notice a reduction of the number of records to approximately $82,5\%$ at the end of this step.

Aviation data ranging from January 2009 to December 2017 Ogasawara [2018] [1] marked the integrated database construction, followed by a data cleansing process.

[1] https://github.com/eogasawara/flight-data

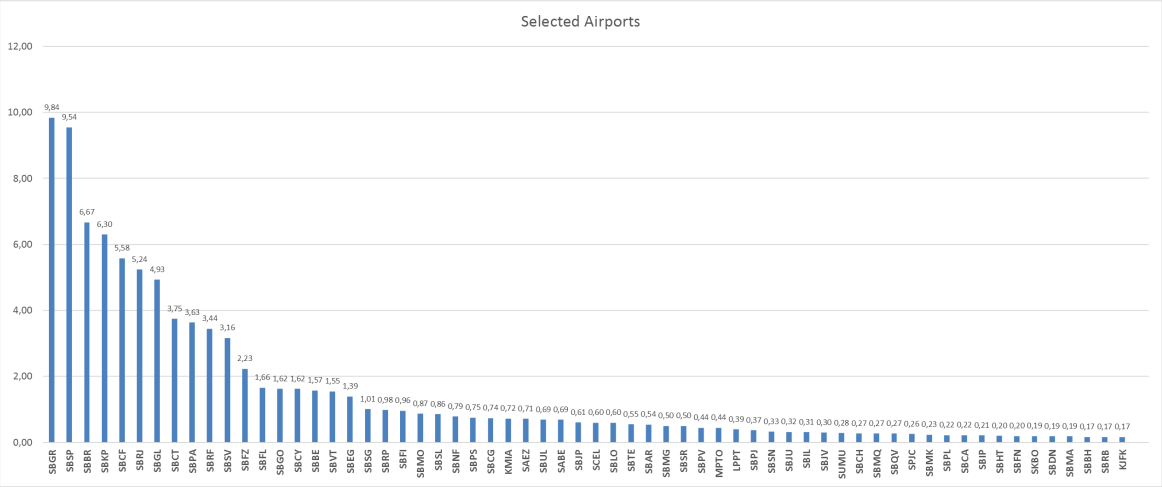

Figure 27 – Selected Airports.

Table 12 – Number of Records after Selection Step

Initial Dataset	Dataset with Airport Selection
10517228	8683195

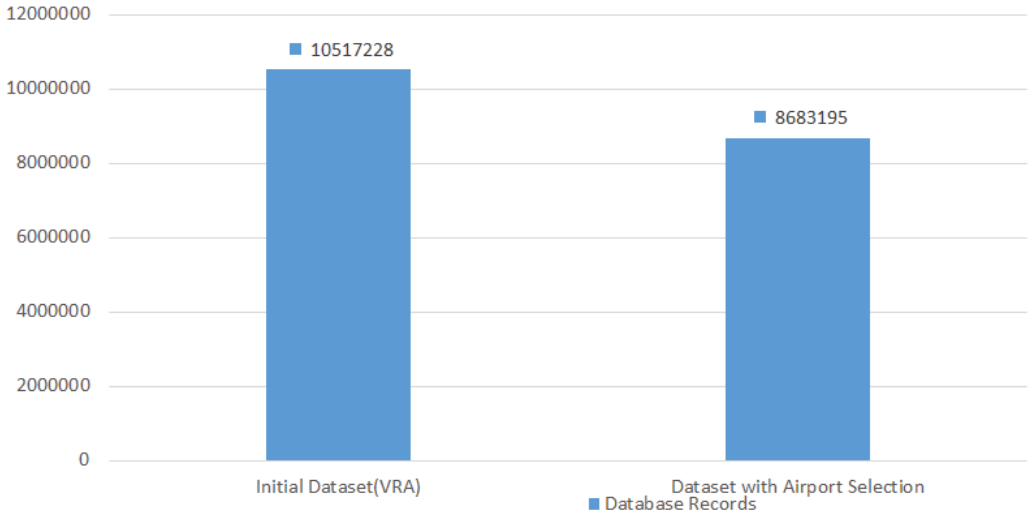

Figure 28 – The number of flight records after Airport Selection.

3.3- Cleaning

Two steps define the cleaning process. The first one refers to the verification of inconsistencies (identification and treatment of upper and lower cases and removing of empty spaces; identification and correcting of outliers; and another inconsistencies filtering). The second considered verification and treatment of missing data, as shown in

Figure 29.

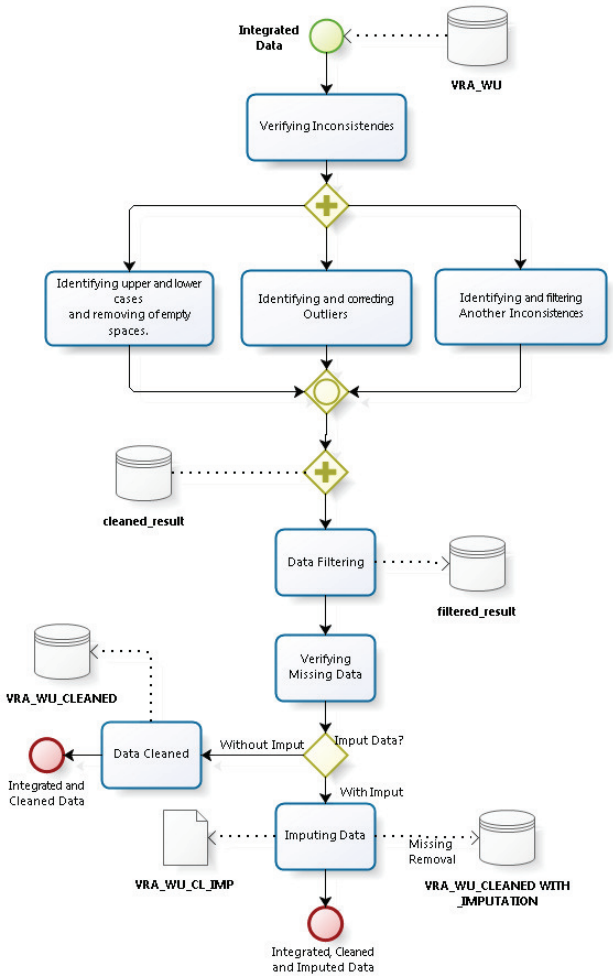

Figure 29 – The workflow of Data Integration and Cleaning

3.3.1 Verifying Inconsistencies

Table 13 exposes different motivations and conditions of data cleaning caused by outlier identification and inconsistency filtering. Subsequently, there is a more detailed description of some domain inconsistencies in the definition of ANAC for delays.

According to Brazilian regulations, a flight with a balance greater than 24 hours is considered canceled. Other types of inconsistencies are related to arrivals of flights occurring before departure, or flights of negative duration, as well as flights of the same origin and destination.

The relative humidity is the ratio of the partial pressure of the water vapor in the air to the vapor pressure of the water at room temperature [Perry et al., 2015]. It is usually expressed as a percentage, on a scale of zero to one hundred, where a more significant percentage means that a moister air-water mixture. In an aspect of the motivation of cleaning, this item defines humidity over 100%.

Another analyzed factor refers to the minimum and maximum temperature. Measurement failures can occur in the sensors, generating a non-consistent data. For this, there were temperatures in the dataset that already recorded on an inhabited planet area - the highest temperature, 56.7^oCelsius; and as lowest, -68^oCelsius, already recorded on an inhabited planet area, as available in Weather Underground and Guinness[Organization, 2018; Records, 2018a,b].

Dew point is the temperature to which air must be cooled to become saturated, called *dew*. The higher dew point registered in the planet was 84^oC and this [Underground, 2011].

Barometric pressure, also known as atmospheric pressure, is the mass of an entire air column in a unit of sea-level surface area, usually expressed in millibars (mbar). This measure is widely used in meteorological observations on the movement of fronts and meteorological systems [noa, 2011b]. The pattern values are situated between 860 mbar and 1080 mbar [noa, 2010, 2011a].

Visibility is a weather measure that indicates how far away the air can perceive an object or light. This measure is dependent on the transparency of the air and affects all forms of traffic, from the road to air and sea, being expressed in miles or kilometers [Seinfeld and Pandis, 2016]. The highest visibility in the cleanest possible atmosphere is limited about 184 miles or 296km.

Considering the cleaning tasks performed for each attribute described in Table 13, there are a set of tuples that satisfied these conditions, according to the Table 14 and Figure 30. The dataset comprises a high number of flights canceled (in the range of 830 thousand records), and the number of records with destination equal to the origin (something like 15 thousand records).

The cleaning process reduced the data in 10% . Compared to the airport selection, there was a decrease of 18% in the total number of records, as shown in Figure 31.

Table 13 – Motivation for Cleaning

Motivation	Feature	Value Condition
depature time occuring after arrival time	arrival/depart	arrival<=depart
depature equals arrival	orgin/destination	origin==destiny
flight with a negative duration	duration	duration<0
flight lasting more than 24 hours	duration	duration>1440
canceled flight	status	status!="DONE"
humidity over 100 %	arrival_humidity	arrival_humidity>100
	depart_humidity	departure_humidity>100
temperature Over 57^oC	arrival_temperature	arrival_temperature>57
	depart_temperature	depart_temperature>57
dew Point over 84^oC	arrival_dew_point	arrival_dew_point>84
	depart_dew_point	depart_dew_point>84
pressure under 860 mbar	arrival_pressure	arrival_pressure<860
	depart_pressure	depart_pressure<860
pressure over 1084 mbar	arrival_pressure	arrival_pressure>1084
	depart_pressure	depart_pressure>1084
visibility over 184 miles	arrival_visibility	arrival_visibility>184
	depart_visibility	depart_visibility>184

Table 14 – Numbers of Cleaning

Condition	Quantity
canceled flights	830436
departure time occurring after arrival time	1758
departure equals arrival	15775
negative flight duration	1758
flight duration greater than one day	2349
too low or too high-temperature values	0
too high dew point	0
invalid humidity range	21
invalid pressure range	139
invalid visibility range	2

3.3.2 Data Filtering

A major remodeling of the infrastructure of Brazilian airports occurred due to the great World Cup Soccer 2014 and Olympics 2016, in order to meet the high demand of tourists who traveled to the country. Before data cleaning, data were selected for the analysis since 2015 to represent the current infrastructure of Brazilian airports, considering all improvements completed, excluding data for 2018. They were not yet available for

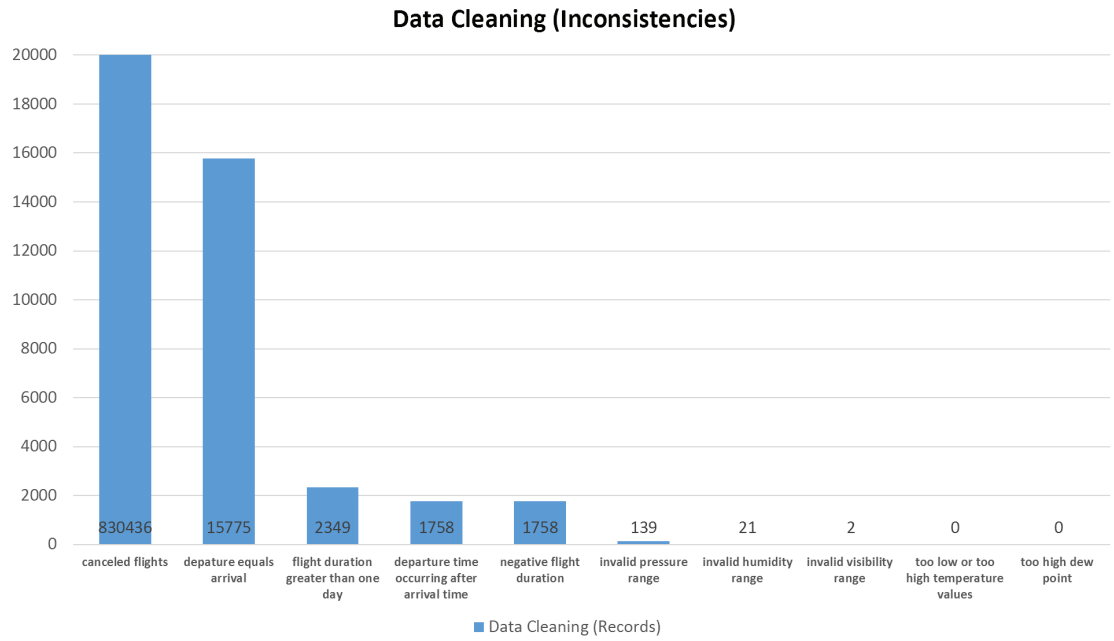

Figure 30 – Data Cleaning - Inconsistencies

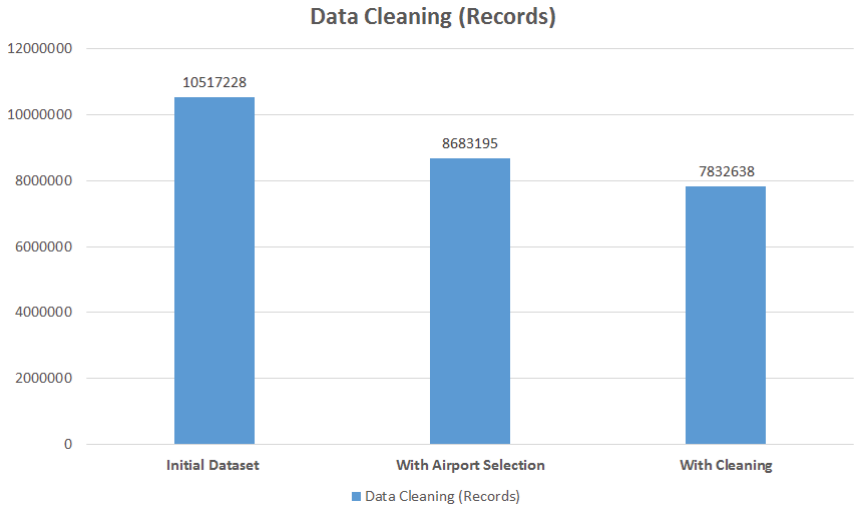

Figure 31 – Result of Data Cleaning

download at the time of data acquisition.

After the execution of this data filtering, we obtain 1.652.941 records, a subset that represents 21% of the original data, in order to give a greater representativity of that data.

Initial Dataset	With Airport Selection	With Cleaning	After Filtering(2015-2017)
10517228	8683195	783638	1652941

Figure 32 – Data Filtering

3.3.3 Verifying Missing Data

Listwise deletion (removal of lines that contain null values on their attributes) was the first method chosen to treat missing data. This missing data treatment does not require any previous verification. It solves the data lack the problem quickly. Thus, after selection and simplified withdrawal, the reduction of the records to 35% of the observed in the selection step was observed, as shown in Figure 33.

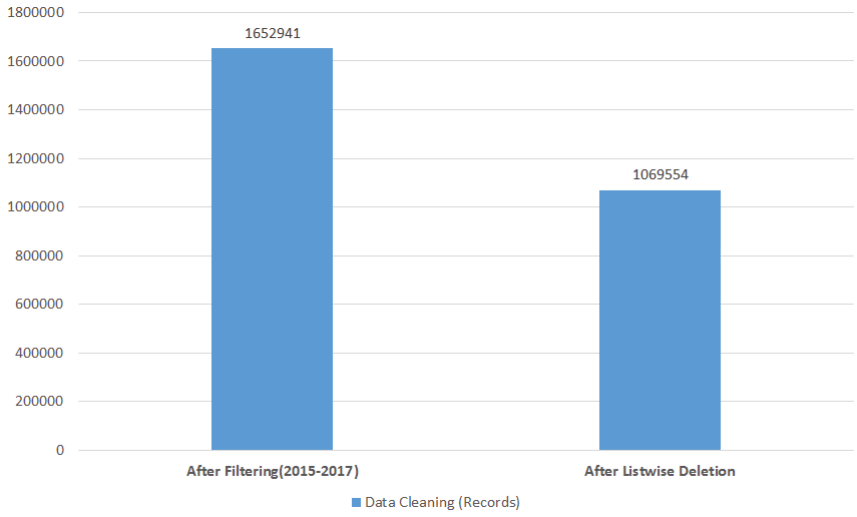

Figure 33 – Data After Filtering and Listwise Deletion.

Figure 34 presents features with the most percentage of missing data are visibility (departure) - in the range of $30\%(28.77\%$ exactly) - followed by departure pressure and dew point.

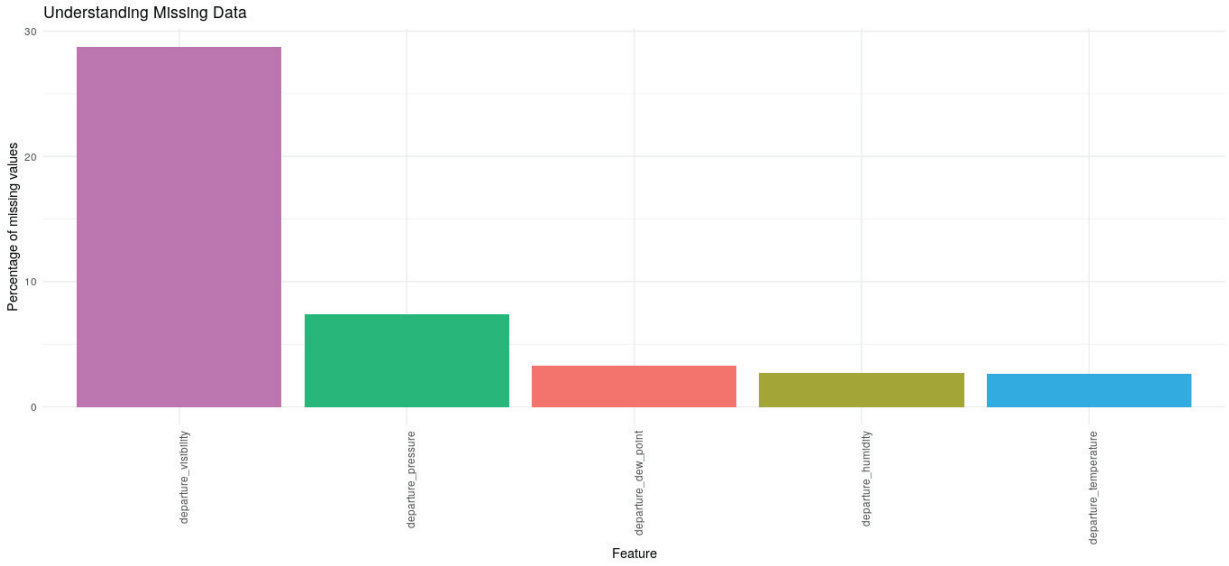

Figure 34 – Understanding Missing Data - After Filtering

Table 15 – Fields with missing data

Field	Percentual Missing
departure_visibility	28,77
departure_pressure	7,43
departure_dew_point	3,27
departure_humidity	2,71
departure_temperature	2,60
other fields	0,00

Imputation

Listwise deletion revealed a total of 30% of missing data. Hence, a necessary data imputation to avoid analysis distortions that could be generated by the simple withdrawal of these values.

Therefore, the hot-deck imputation method substituted missing values of one or more variables of a non-respondent, replacing it to similar observed values of a respondent

or donor.

Figure 35 presents the application of the imputation method organized by the departure. It exhibits existing records in blue and imputations one in orange.

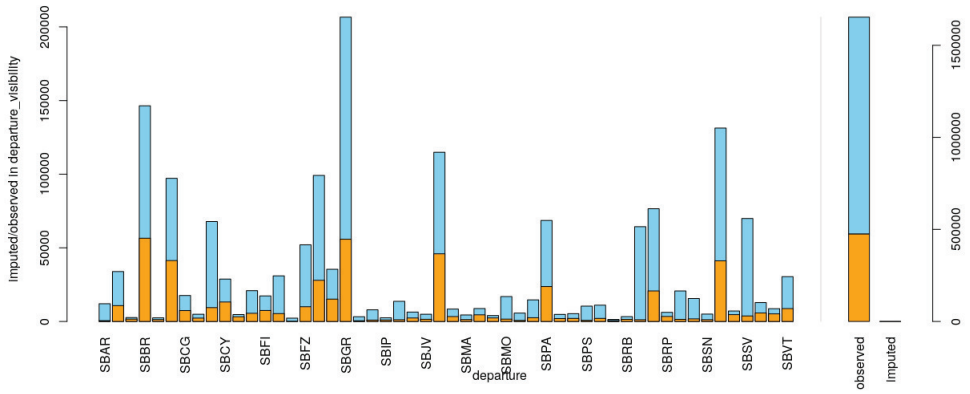

Figure 35 – Hot-Deck Imputation before Filtering

Imputation has restored the original number of records, softening the listwise deletion effects, as presented in Figure 36.

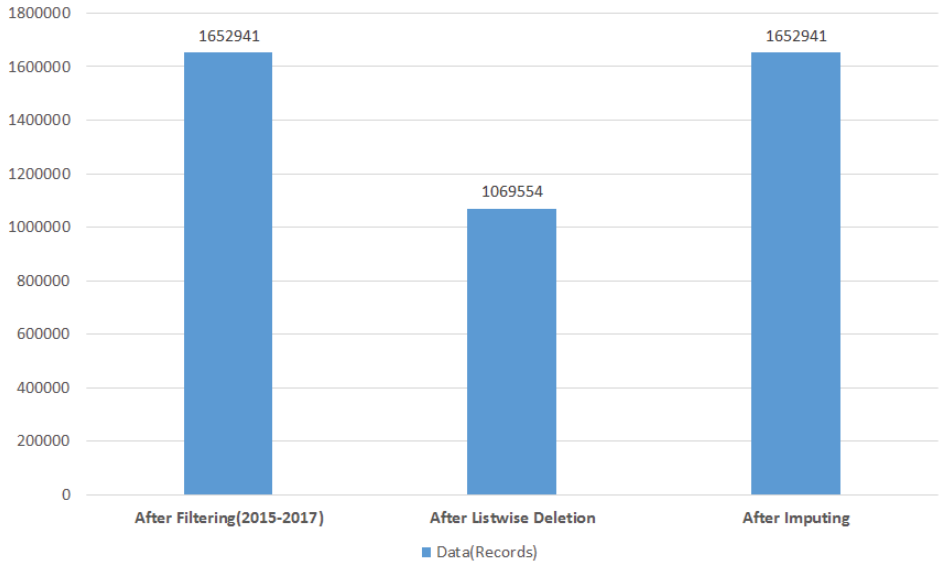

Figure 36 – Data After Imputing Comparison.

3.4- Transformation

Data transformation follows the integration, cleaning, and data filtering steps, as presented in Figure 37. Discretization (Binning), Categorical Mapping, and Conceptual Hierarchy compose this step. Its primary purpose is to improve data quality, increasing the significant classifier predictions probability. Data normalization should also follow these transformations.

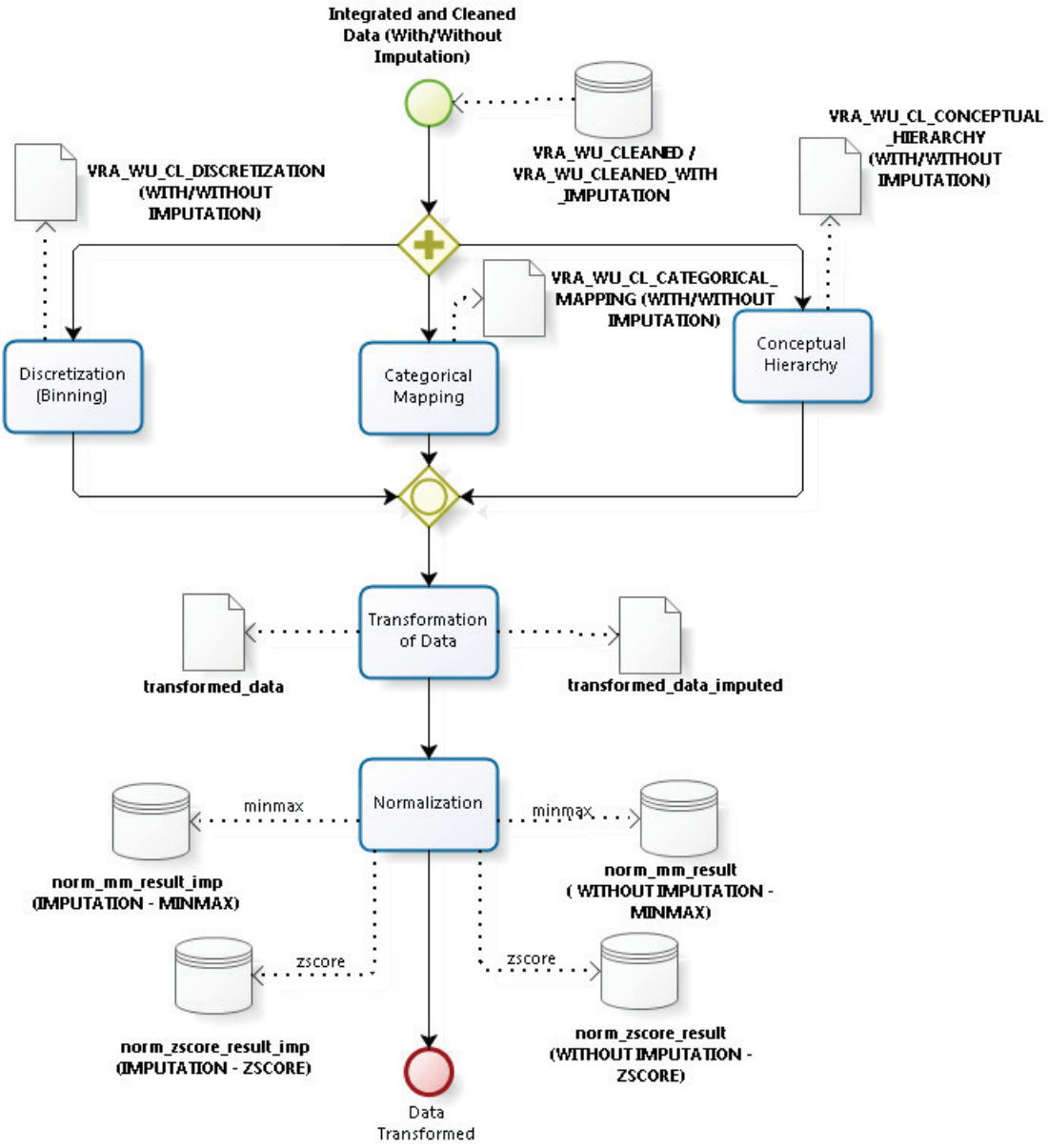

Figure 37 – The workflow of Data Transformation

Listwise deletion revealed a total of 30% of missing data. Hence, a necessary data

imputation to avoid analysis distortions that could be generated by the simple withdrawal of these values. Transformation encompasses Discretization-Binning (bin), Conceptual Hierarchy (CH), Categorical Mapping (CM), and unused data (removed). Table 16 presents data used in the other steps with descriptions of the transformations made.The delay feature was demarked with class one to delayed flights and zero to flights on time, according the difference over 15 minutes of flight depart time expected and real appured. As described in id line one, the departure and arrival date and time feature as well as the flight duration, were removed after that delay check and establishment of the target class delayed. More specific details about transformations are provided in Appendix A.

Table 16 – Transformed Data Dictionary

ID	Description	Attributes	Type
0	remove	flight;departure_expect;arrival_expect;duration	Factor; Datetime (POSIXct); Datetime (POSIXct) Integer
1	basic	airline;departure;arrival;	Factor
2	CH	departure_year;departure_month; departure_day;departure_hour	Numeric
3	bin	departure_hour_bin	Numeric
4	original	departure_temperature;departure_dew_point; departure_humidity;departure_pressure; departure_visibility	Numeric
5	bin	departure_temperature_bin;departure_dew_point_bin; departure_humidity_bin;departure_pressure_bin; departure_visibility_bin	Numeric
6	original	departure_events	Factor
7	CM	departure_events	Numeric
8	original	departure_conditions	Factor
9	CM	departure_conditions	Numeric
10	Target	delayed	Factor
11	CM	airline	Numeric

Table 16 – Transformed Data Dictionary

ID	Description	Attributes	Type
12	CH	departure_time	Time(ITime)
13	CM	departure	Numeric
14	CM	arrival	Numeric
15	CH	departure_weekday	Numeric
99	original	departure_time_original	Datetime (POSIXct)

3.4.1 Discretization

Discretization was applied in several attributes, for example, in temperature. For the binning, an unsupervised method was used, transforming the numerical variables into categorical variables, dividing the data into ranges of values, according to their frequency. Values were mapped to one of six possible values (ranging from 1 to 6), as exhibited in Figure 38.

```
> arrival_temperature.bin$binning
       1        2        3        4        5        6
14.20531 18.75328 22.59854 26.40415 30.05091 34.19500
```

Figure 38 – Example of Binning Discretization

Figure 39 presents the discretization effects, contrasting the original temperature (arrival_temperature and departure_temperature columns) and the respectively discretized one (columns arrival_temperature_bin and depart_temperature_bin).

arrival_temperature	arrival_temperature_bin	depart_temperature	depart_temperature_bin
25	4	23	3
19	2	21	3
17	2	13	1
32	5	29	5
22	3	26	4
25	4	26	4
19	2	19	2
19	2	24	3
25	4	19	2
26	4	27	4

Figure 39 – Example of Discretization

3.4.2 Conceptual Hierarchy

The attribute depart_expect received conceptual hierarchy technique application, slicing it in the attributes year, month, day, day of the week and time, as represented in figure 41.

depart_expect	depart_expect_year	depart_expect_month	depart_expect_day	depart_expect_weekday	depart_expect_hr
2009-01-01 08:45:00	2009	1	1	5	8
2009-01-01 22:30:00	2009	1	1	5	22
2009-01-01 22:15:00	2009	1	1	5	22
2009-01-01 23:55:00	2009	1	1	5	23
2009-01-01 23:25:00	2009	1	1	5	23
2009-01-01 23:40:00	2009	1	1	5	23
2009-01-01 23:15:00	2009	1	1	5	23
2009-01-01 23:10:00	2009	1	1	5	23
2009-01-02 08:45:00	2009	1	2	6	8
2009-01-02 22:30:00	2009	1	2	6	22

Figure 40 – Example of Conceptual Hierarchy

3.4.3 Categorical Mapping

For the categorical mapping technique, the event attribute was chosen, in cases referring to the events recorded in the flight departure: Gentle Breeze, Light Breeze, Moderate Breeze, None, Strong Breeze. These records were transformed into columns, using ”0” when they did not occur, and ”1” when they occurred (Figure 41).

depart_eventsGENTLE BREEZE	depart_eventsLIGHT BREEZE	depart_eventsMODERATE BREEZE	depart_eventsNONE	depart_eventsSTRONG BREEZE
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0

Figure 41 – Example of Categorical Mapping

At the end of the application of these three techniques (Discretization, Conceptual Hierarchy, and Categorical Mapping), new attributes were generated, as exhibited in Table 16. These changes are quantified in Table 17, where it is verified the high number of attributes generated with the categorical mapping. Table 18 shows the consolidated numbers of transformation techniques compared to original data.

Table 17 – Numbers of Transformation

Transformation Technique	**Number of Original**	**Number after Technique**
Discretization (Binning- bin)	6	12
Conceptual Hierarchy (CH)	1	7
Categorical Mapping (CM)	5	162

Table 18 – Consolidated Numbers of Transformation

Original	**After Transformation**
12	181

3.4.4 Normalization

Typically, machine learning algorithms try to find trends in the data by comparing resources of the data points [Han et al., 2011]. In the given flight delay data situation, some features are at drastically different scales, such as departure_dew_point and depart_temperature, which can generate dominance of one of these. Normalization aims at

softening this kind of noise. Initially, the min-max technique was applied; however, this technique does not work very well with outliers. Therefore, Z-score data normalization was also used to avoid this problem. Those techniques were applied only in training set.

Figures 42, 43 and 44 show, respectively, the data without normalization and after the application of the min-max and Z-score.

	departure_temperature	departure_dew_point	departure_humidity	departure_pressure	departure_visibility
54	22	19	83	1020	10
55	22	21	94	1015	10
56	24	22	89	1016	6
57	23	19	78	1019	10

Figure 42 – Data Without Normalization

	departure_temperature	departure_dew_point	departure_humidity	departure_pressure	departure_visibility
54	0.5365854	0.3725490	0.83	0.9435185	0.10
55	0.5365854	0.4117647	0.94	0.9388889	0.10
56	0.5853659	0.4313725	0.89	0.9398148	0.06
57	0.5609756	0.3725490	0.78	0.9425926	0.10

Figure 43 – Data with Min-max Normalization

	departure_temperature	departure_dew_point	departure_humidity	departure_pressure	departure_visibility
54	-0.25081973	0.3063952	0.6257735	0.1949837	-0.09479544
55	-0.25081973	0.7004870	1.1558525	0.1660201	-0.09479544
56	0.10430457	0.8975329	0.9149075	0.1718128	-0.58710767
57	-0.07325758	0.3063952	0.3848284	0.1891910	-0.09479544

Figure 44 – Data with Z-score Normalization

3.5- Sampling

After transformation and normalization, the dataset was divided into training and test sets in the 80:20 ratio, as shown in Figures 45 and 46.

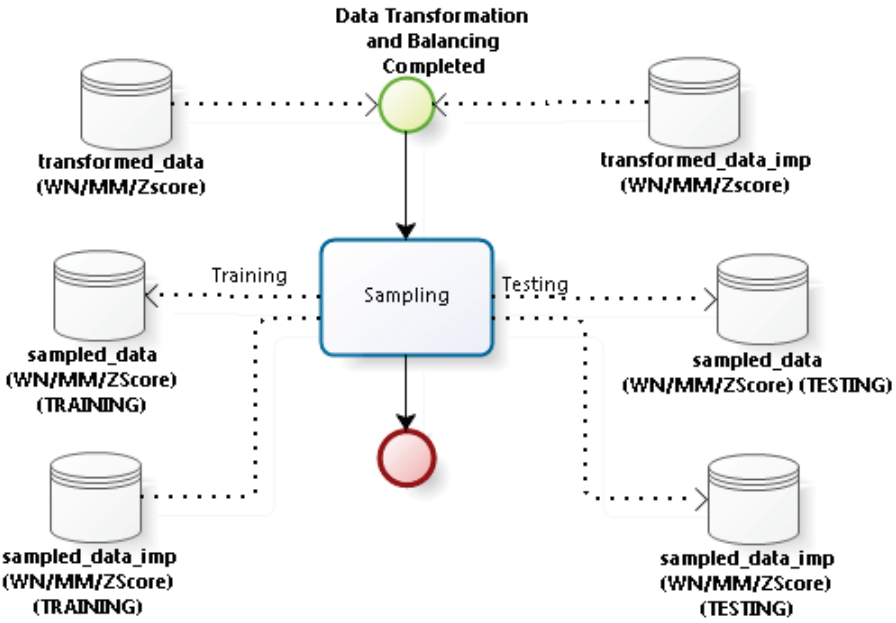

Figure 45 – Data Sampling

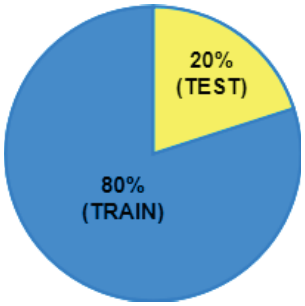

Figure 46 – Train and Testing

3.6- Balancing

Data unbalance verification is the step following data sampling. Random Sub-sampling (RS) and Synthetic Minority Over-sampling Technique (SMOTE) balancing techniques were applied to the training sets, maintaining the original unbalance of the test sets.

The unbalance of the data set is 85,69% without delay and 14,31% delayed (ratio 86:14, approximately), and balanced with the application of the balancing techniques, as indicated in Figure 48.

Figure 49 and Table 19 presents a quantity comparison using balancing techniques

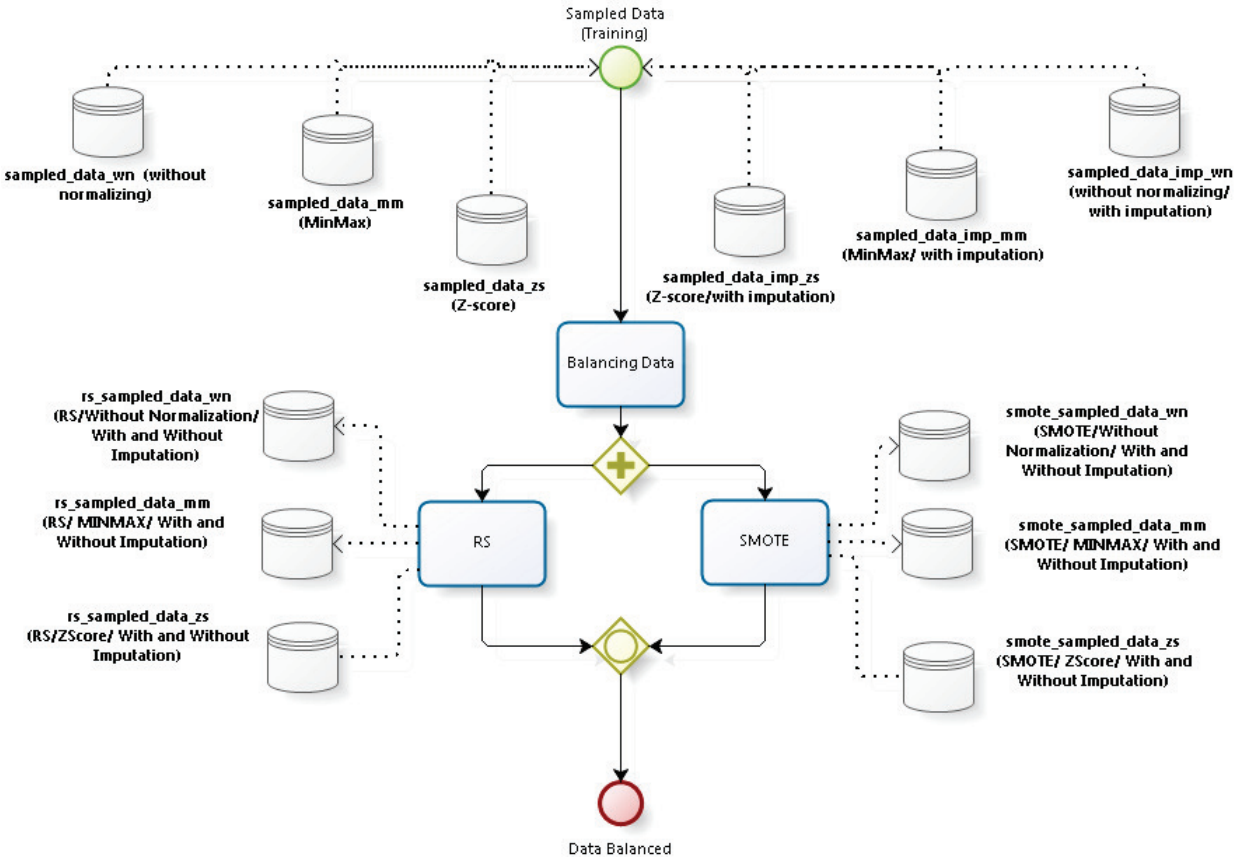

Figure 47 – The workflow of Data Balancing

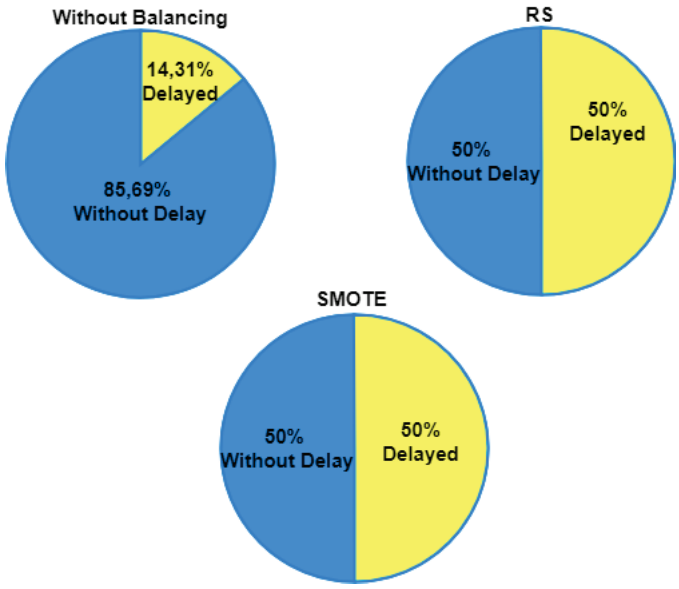

Figure 48 – Balancing Distribution - Original/RS/SMOTE

over the training set. RS reduces the number of records without delay (minority class), matching the same number of records with delay. Still, SMOTE includes random records in the minority class to match the number of records without delay (dominant class).

Table 19 – Data Balancing Numbers

	Without Balancing	With Balancing (RS)	With Balancing (SMOTE)
With Delay	189418	189418	1133050
Without Delay	1132935	189418	1133050
Total	1322353	378836	2266100

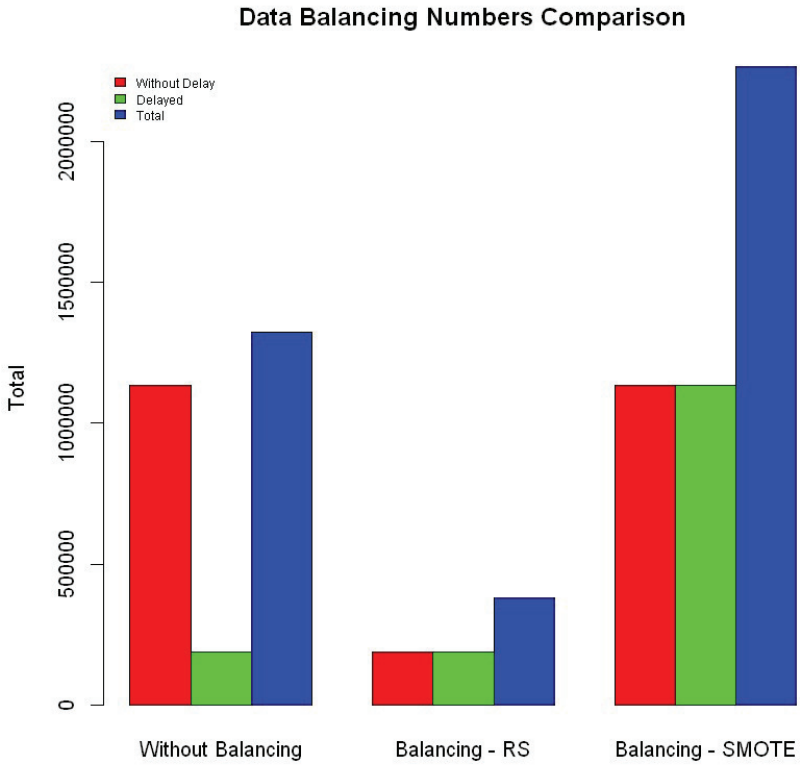

Figure 49 – Data Balancing Comparison

3.7- Feature Selection and Extraction

After finished balancing, the next step to perform the step focused on data reduction, more specifically feature selection and extraction represented in Figure 50.

For the application of the feature selection strategies (LASSO, CFS, IG) and feature extraction (PCA), the attributes that best fit this technique applied and gathered in sets, called workflows, are chosen, as exhibited in Figures 50 and 51.

Table 20 reveals all the workflows assembled for evaluation in this work.

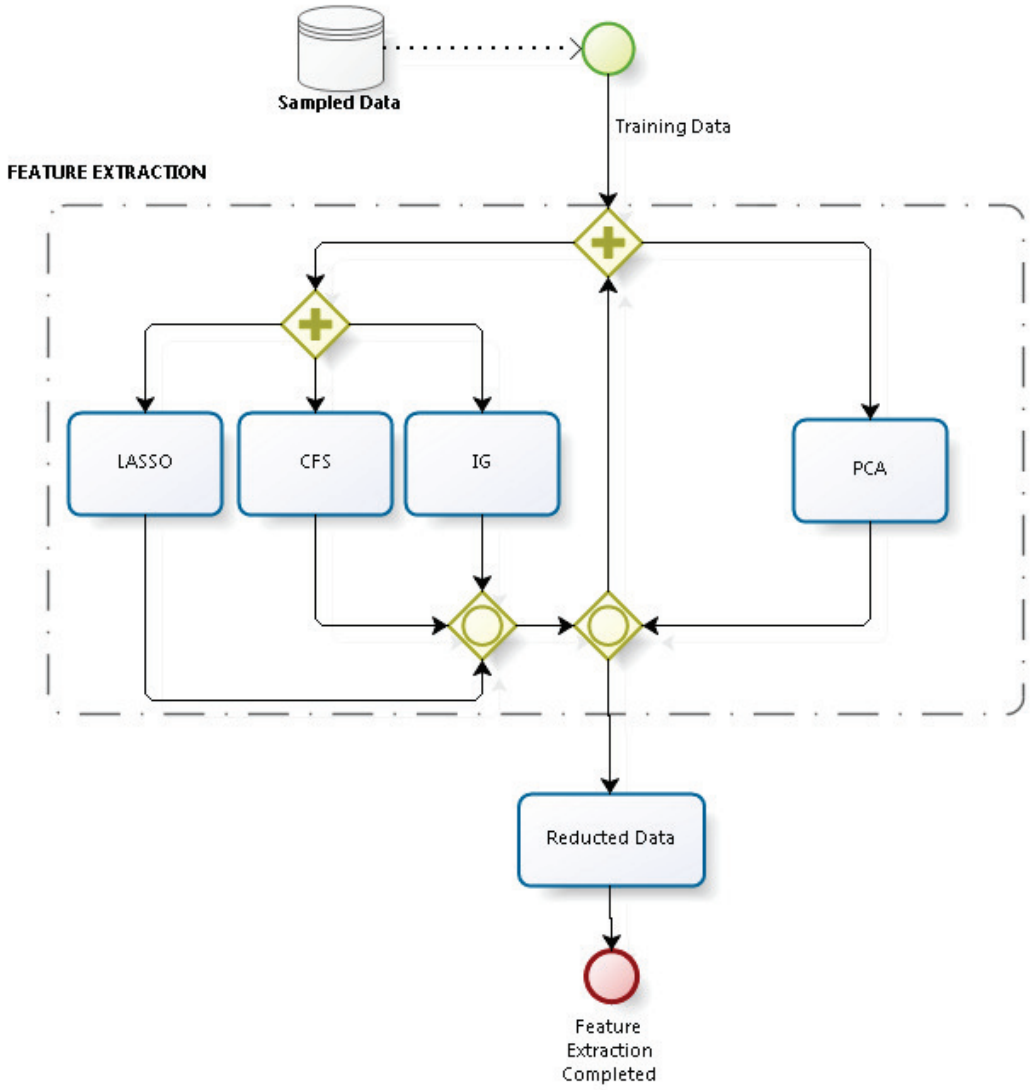

Figure 50 – The workflow of Feature Extraction

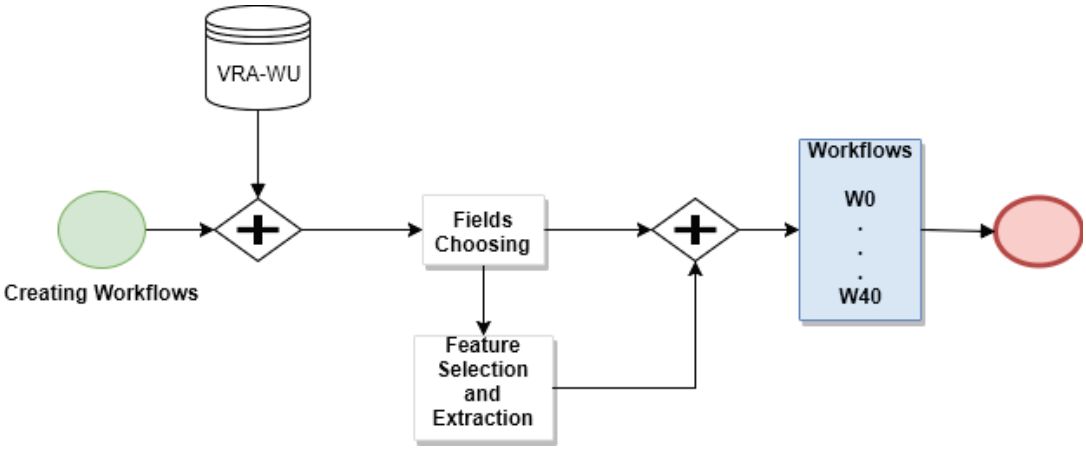

Figure 51 – Workflow Mounting

Table 20 – Workflow Description

Workflow	Description	ID(Dictionary)
0	Attributes[1,99]	1,2,99,10
1	Attributes[1,2,12]	1,2,12,10
2	Attributes[1,2,12,15]	1,2,12,15,10
3	Attributes[1,12]	1,12,10
4	Attributes[1,2,12] + Conditions[8]	1,2,12,8,10
5	Attributes[1,2,12] + Conditions[9] (PCA)	1,2,12,9,10
6	Attributes[1,2,12] + Conditions[9] (PCA+LASSO)	1,2,12,9,10
7	Attributes[1,2,12] + Conditions[9] (PCA+INFOGAIN)	1,2,12,9,10
8	Attributes[1,2,12] + Conditions[9] (PCA+CFS)	1,2,12,9,10
9	Attributes[1,2,12] + Conditions[9] (LASSO)	1,2,12,9,10
10	Attributes[1,2,12] + Conditions[9] (INFOGAIN)	1,2,12,9,10
11	Attributes[1,2,12] + Conditions[9] (CFS)	1,2,12,9,10
12	Attributes[1,2,12] + Events[6]	1,2,12,6,10
13	Attributes[1,2,12] + Events[7] (PCA+LASSO)	1,2,12,7,10
14	Attributes[1,2,12]+ Events[7] (CFS)	1,2,7,12,10
15	Attributes[1,2,12] + Events[7](LASSO)+ Conditions[9](CFS)	1,2,7,9,12,10
16	Attributes[1,2,12] + Events[7](PCA+LASSO)+ Conditions[9] (PCA+CFS)	1,2,7,9,12,10
17	Attributes[1,6,8,99]	1,6,8,99,10
18	Attributes[2,12]	2,12,10
19	Attributes[2,12]+ Airlines[11](PCA+INFOGAIN)	2,10,11,12
20	Attributes[2,12]+ Departure[13](PCA+INFOGAIN)	2,12,10,13
21	Attributes[2,12]+ Arrival[14](PCA+CFS)	2,12,10,14
22	Attributes[2,12]+ Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	2,12,11,13,14,10
23	Attributes[2,12]+ Airlines[11](CFS)+Departure[13](CFS)+ Arrival[14](CFS)	2,12,11,13,14,10

Table 20 continued from the previous page

Workflow	Description	ID(Dictionary)
24	Attributes[2,12]+ Events[7](PCA+CFS)+ Airlines[11](PCA+CFS)+Departure[13](PCA+CFS)+ Arrival[14](PCA+CFS)	2,12,7,11,13,14,10
25	Attributes[2,12]+ Conditions[9](CFS)+ Airlines[11](CFS)+ Departure[13](CFS)+Arrival[14](CFS)	2,12,9,11,13,14,10
26	Attributes[2,12]+Conditions[8]+Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	2,8,11,13,14,12,10
27	Attributes[2,12]+ Events[7](CFS)+ Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	2,12,7,11,13,14,10
28	Attributes[2,12]+ Conditions[9](CFS)+ Airlines[11](PCA+CFS)+Departure[13](PCA+CFS)+ Arrival[14](PCA+CFS)	2,12,9,11,13,14,10
29	Attributes[2,12]+ [Events[7]+ Airlines[11]+Departure[13]+ Arrival[14]](PCA+CFS)	2,12,7,11,13,14,10
30	Attributes[1,3,12]	1,3,12,10
31	Attributes[3,12]+ Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	3,12,11,13,14,10
32	Attributes[1,4,12]	1,4,12,10
33	Attributes[4,12]+ Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	4,12,11,13,14,10
34	Attributes[1,5,12]	1,5,12,10
35	Attributes[5,12]+ Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	5,12,11,13,14,10
36	Attributes[1,2,4,12]	1,2,4,12,10
37	Attributes[2,4,12]+ Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+Arrival[14](PCA+CFS)	2,4,12,11,13,14,10
38	Attributes[1,3,5,12]	1,3,5,12,10
39	Attributes[3,5,12]+Airlines[11](PCA+CFS)+ Departure[13](PCA+CFS)+ Arrival[14](PCA+CFS)	2,4,12,11,13,14,10
40	Attributes[1,4,99]	1,4,6,99,10

Table 21 presents, for each workflow assembled, the types of machine learning applied, as well as the reduction strategies applied.

Table 21 – Workflow Machine Learning and Reduction Strategies

			Reduction Strategies			
	ML		Feature Selection			Feature Extraction
Workflow	RF	NN	LASSO(LS)	INFOGAIN(IG)	CFS	PCA
0	X					
1	X					
2	X	X				
3	X					
4	X					
5	X					X
6	X		X			X
7	X			X		X
8	X				X	X
9	X		X			
10	X			X		
11	X				X	
12	X					
13	X		X			X
14	X				X	
15	X		X		X	
16	X		X		X	X
17	X					
18	X	X				
19	X	X		X		X
20	X	X		X	X	X
21	X	X			X	X
22	X	X			X	X
23	X	X			X	
24	X	X			X	
25	X	X			X	X

Table 21 continued from the previous page

	Reduction Strategies					
	ML		Feature Selection			Feature Extraction
Workflow	RF	NN	LASSO(LS)	INFOGAIN(IG)	CFS	PCA
26	X	X			X	X
27	X	X		X	X	X
28	X	X			X	X
29	X	X			X	X
30	X					X
31	X	X			X	X
32	X					
33	X	X			X	X
34	X					
35	X	X			X	X
36	X					
37	X	X			X	X
38	X					
39	X	X			X	X
40	X					

3.8- Model Creation, Evaluation and Implementation

The classification task is binary which assigning an individual values to one of two categories, by measuring a series of attributes Parmigiani [2001] as described in transformation step. After all these preprocessing steps permitted the model conception, applying optimized settings for Random Forest (RF) and Neural Networks (MLP) algorithms. Therefore, the output model evaluation tests data sets. This process is demonstrated in Figure 52. The experiments were conducted on a computer with a Xeon processor with 32GB of RAM and using Ubuntu 16.04 operating system and implemented in R for both preprocessing and machine learning methods [Lantz, 2013]. The source code, dataset,

and a Jupyter sample are made available at GitHub [2].

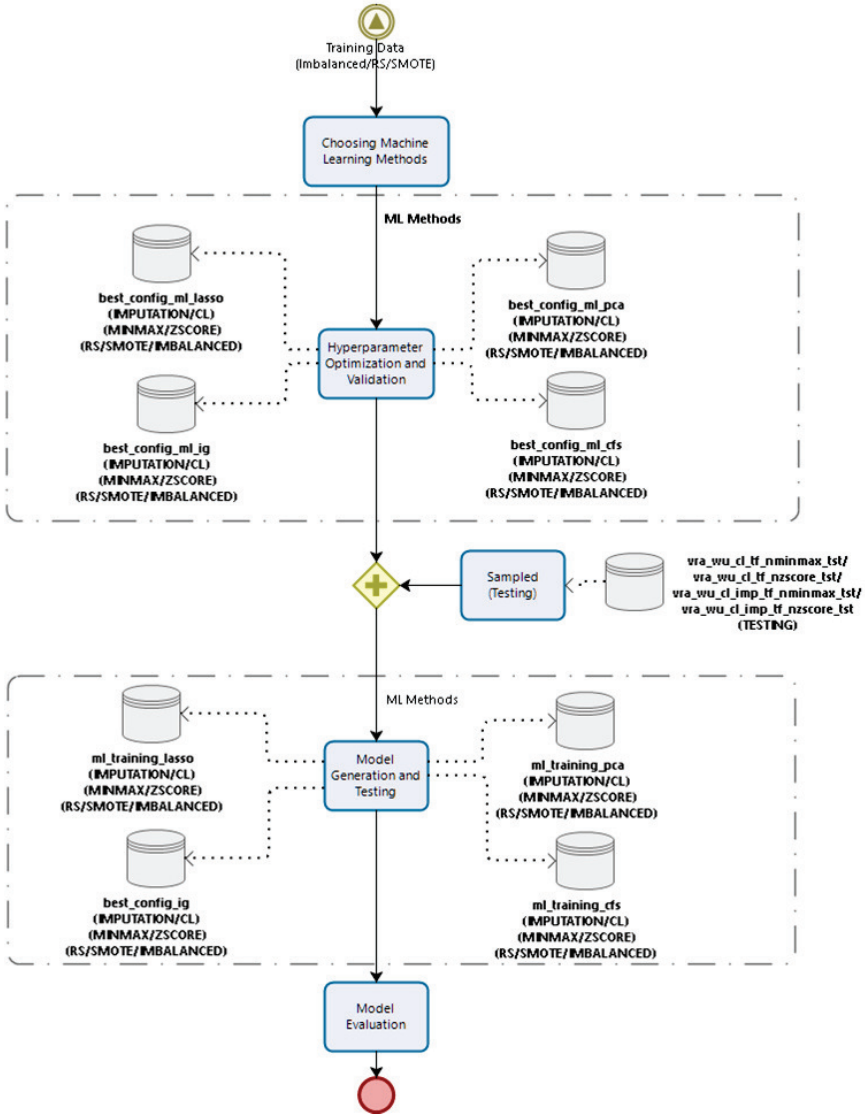

Figure 52 – The workflow of Model Creation and Evaluation

Due to the lower computational cost, the model training used the holdout method. This process proceeded with the realization of Workflow Analysis and Results, as detailed in Chapter 4.

[2] https://github.com/leonardosminfo/presentation

4- Workflow Analysis and Results

Each experiment workflow, that was executed one time, considers a combination of training datasets and machine learning methods. Model generation demands a reliable way to evaluate all experiments through the test data sets, verifying the predictive capacity of the classification algorithm, either by precision, accuracy, or other methods. Hence, experiments were performed for this evaluation, as listed in Table 22, and their corresponding workflows.

Table 22 – Realized Experiments

Code	Experiment Description	Workflow
1	Comparison Between Original and Transformed Time	0,1,22
2	Comparison of Conceptual Hierarchy (Departure Time Original)	0,1,2,3
3	Comparison of Feature Selection Techniques(CFS, PCA,LASSO,IG) on Conditions	4;5:11;25,28
4	Comparison of Feature Selection Techniques(LASSO, CFS,PCA) on Events	13,14;24,27,29
5	Comparison of Feature Selection Techniques(LASSO, CFS,PCA) on Events and Conditions	15,16,17
6	Comparison of data with and without Airline, Departure and Arrival fields	1,18
7	Comparison of impacts of separately and together with data of Airline, Departure and Arrival	19,20,21,22
8	Comparison of Feature Selection to completely data (Airline,Arrival,Departure)	22,23
9	Comparison of Best {Basics+Events} and CM (Airline,Arrival,Departure)+ Events{7}	12,14,27
10	Comparison Between CM(Airline,Arrival, Departure) + Conditions{8,9}	4,9,26,28

Table 22 continued from the previous page

Code	Experiment Description	Workflow
11	Comparison between Original {2} and Discretized{3}, with and without CM(Airline,Departure,Arrival)	1,30,31
12	Comparison between Original {4} and Discretized{5}, with and without CM(Airline,Departure,Arrival)	1,32,33,34,35
13	Comparsion between include Discretized{4} on Basic with and without CM(Airline, Departure,Arrival)	1,36,37,40
14	Comparsion between include Discretized{3,5} and Basic	1,38,39
15	Comparison of Normalization Methods	All
16	Comparison of Balancing Methods	All
17	Comparison of Threshold Approach	All
18	Random Forest X Neural Networks	All
19	Time Elapsed	All
20	Accuracy, Sensibility and F1-Score	All

All experiments are organized and presented according to the methods, techniques, and strategies of transformation, normalization, data reduction, balancing, classifier limit approach, learning, and elapsed time in a table format.

Consider the following legend:

Threshold - The approach used for the classifier limit

WF - Workflow Tested (0 to 40)

ML - Machine Learning Method (RF - Random Forest / NN - Neural Networks)

BL - Balancing Technique (IMB - Imbalanced / SUB-Random Subsampling / SMOTE - Synthetic Minority Over-sampling Technique)

NM - Standardization Technique (WN- Without Normalization / MM - Min-max / ZS - Z-score)

AC - Accuracy

SS / R - Sensibility / Recall

SP - Specificity

P - Precision

F1-Score - Accuracy that uses the weighted harmonic mean of the test's precision and recall

Time - Time Elapsed in test

Next, the results of the experiments will be presented, evaluated and discussed through the respective application of methods of Transformation (Conceptual Hierarchy, Discretization, Categorical Mapping), Normalization, Feature Selection and Extraction, Balancing, Threshold Approach and Machine Learning. At the end are also evaluated aspects regarding the time elapsed to perform each experiment, better results (in aspect of Accuracy, Sensitivity/Recall and F1-Score), as well as a comparison of the best results achieved in this experiment and the best results achieved in related works.

4.1- Transformation

Transformation experiments focused on normalization, conceptual hierarchy, discretization, and categorical mapping. The objective is to analyze the impacts on results.

4.1.1 Comparison of Conceptual Hierarchy

Experiments 2 and 6 analyze the impacts of the conceptual hierarchy, as shown in Tables 23 and 24.

Experiment 2

Experiment 2 (Table 23) compares the application of the conceptual hierarchy through workflows 0, 1, 2, and 3. Workflow 0 uses the original attributes 1 (Airline, Departure, Arrival) and 99 (depart date and time); workflow 1 uses of attributes 1, 2 (departure_year, departure_month, departure_day, departure_hour), and 12 (departure_time); workflow 2, with attributes 1, 2, 12 and 15 (depart_weekday); and workflow 3, has the attribute sets 1 and 12 only.

Table 23 – Comparison of Conceptual Hierarchy (Departure Time Original)

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Conventional	0	RF	SUB	MM	0,647	0,644	0,648	0,235	0,344	1,385
Majority	1	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,481
Majority	2	RF	IMB	WN	0,922	0,929	0,920	0,660	0,772	7,317
Majority	3	RF	IMB	MM	0,937	0,826	0,956	0,756	0,790	4,495

Results reveal that workflow 0 with original data produced a better result with the sub-sampling method and normalization min-max, with a considerable low accuracy compared to the other workflows from 1 to 3. On the scenario that uses the workflows and conceptual hierarchy, the experiments revealed that the addition of the day of the week (used in workflow 2) raised the sensitivity, but reduced all the other aspects faced to workflow 1. This configuration revealed the best F1-Score. Workflow 3 obtained the best accuracy of the set of workflows with a slightly lower sensitivity to workflows 1 and 2.

Experiment 6

Experiment 6 presented in Table 24 compares workflows 1 and 18, that is, between workflow 1 that uses the set of attributes (1, 2 and 12), where attribute 1 corresponding to airline, departure and arrival data; attribute 2, corresponding to the date of departure with the application of the conceptual hierarchy, and attribute 12, the attribute of the departure time. Workflow 18 uses only the set of attributes (2 and 12), which helps to evaluate the impact on the use of attribute set 1.

Table 24 – Comparison of data with and without Airline,Departure and Arrival fields

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	1	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,481
	18	RF	IMB	WN	0,908	0,508	0,974	0,765	0,610	3,878
Conventional	18	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	3,033

Now the scenario considers the conceptual hierarchy, but this time with the removal of attribute 1 (referring to airline, departure, and arrival) in the workflow 18. Workflow 1 (using the attributes referring to airline, departure, and arrival) obtains better results than workflow 18 both in accuracy and sensitivity, losing only in specificity and precision.

4.1.2 Comparison of Discretization

To evaluate the impacts of the Discretization, experiments 11, 12, 13 and 14 were performed as shown in Tables 25, 26 and 27.

Experiment 11

Table 25 refers to an evaluation comparing workflow 1 and workflow scenarios 30 and 31, applying the discretization transformation technique to an item of the attribute set 2, generating the attribute 3. Workflow 30 also has attributes 1 and 12. Workflow 31 has attribute 12 and derived attributes produced by conceptual hierarchy with the application of extraction and selection of PCA and CFS characteristics (11, 13 and 14).

Table 25 – Comparison between Original {2} and Discretized{3}

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
	1	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,481
Majority	30	RF	IMB	MM	0,940	0,818	0,961	0,776	**0,797**	8,000
	31	RF	IMB	ZS	0,927	0,660	0,971	0,793	0,720	7,771
Conventional	31	NN	SUB	MM	0,639	0,567	0,651	0,215	0,312	49,773

This experiment compares workflows 1, 30, and 3. The base workflow 1 obtained better results insensitivity. However, according to the produced results of workflow 30, the use of discretization on the set of attributes 2 (which generated the set of attributes 3) increased accuracy, specificity, and precision, slightly improving F1-Score. However, workflow 31 did not offer any improvement in terms of accuracy and sensitivity, both with machine learning RF and NN techniques, with small improvements in specificity and precision.

Experiment 12

Experiment 12 examines the basic workflow 1 and workflows 32, 33, 34, and 35. While Workflows 32 and 33 counts with the original set of climate attributes (4), Workflows 34 and 35 also have these climate attributes (5) discretized. Workflows 32 and 34 have, in addition to their climate data sets, the attribute sets 1 and 12. Workflows 33 and 35, in addition to the climate data sets, have the attribute 12 and the attribute set 1 transformed by conceptual hierarchy (11, 13, and 14).

Table 26 – Comparison between Original $\{4\}$ and Discretized$\{5\}$

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
	1	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,481
Majority	32	RF	IMB	WN	0,834	0,882	0,826	0,457	0,602	9,968
	33	RF	IMB	WN	0,878	0,766	0,897	0,552	0,642	13,23
Conventional	33	NN	SUB	ZS	0,474	0,729	0,432	0,176	0,283	59,082
	34	RF	IMB	WN	0,893	0,830	0,903	0,588	0,688	8,767
Majority	35	RF	IMB	ZS	0,907	0,742	0,935	0,655	0,695	11,845
	35	NN	IMB	MM	0,523	0,649	0,501	0,179	0,280	2,829

Table 26 presents the results obtained with workflows 32 (which included the time data set 4) and workflow 34 (which included the discretization of the time - result data set in the data set 5). The use of discretization raised the results obtained in specificity, accuracy, precision, and F1-Score in workflow 34 concerning workflow 32. Comparison between data sets 33 and 35 revealed better results in specificity, accuracy, precision, and F1-Score in workflow 35 concerning workflow 33 (according to the result obtained between workflows 32 and 34). None of the results obtained in workflows 32,33, 34, and 35 were higher than those achieved by the base workflow 1, which did not contain the time data set.

Experiments 13 and 14

Experiment 13 aims to compare workflows involving the basic attribute set (workflow 1) with sets containing the addition of time attribute set 4, with and without the

categorical mapping relationship in attribute 1 (airline, departure, and arrival), especially in workflows 36, 37, and 40. Conversely, experiment 14, aims at the comparison of the basic workflow 1, with the workflows 38 and 39 that have the use of sets of attributes 3, 5 and 12, besides the set of attributes 1, with or without the application of the technique of transformation of hierarchy in workflows 39 and 38 respectively. Results are listed in Table 27.

Table 27 – Comparsion between include orginal{4} and discretized{5} on Basic with and without CM(Airline, Departure,Arrival)

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	1	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,481
Conventional	36	RF	IMB	ZS	0,922	0,504	0,991	0,907	0,648	11,118
Majority	37	RF	IMB	WN	0,864	0,896	0,858	0,512	0,652	16,077
Conventional	37	NN	IMB	WN	0,907	0,409	0,990	0,876	0,557	6,359
Majority	38	RF	IMB	WN	0,905	0,846	0,914	0,621	0,716	9,670
	39	RF	IMB	WN	0,917	0,758	0,944	0,691	0,723	13,286
	40	RF	IMB	MM	0,666	0,655	0,668	0,248	0,359	8,767

Regarding the results obtained in workflows 36 and 37, workflow 1 obtained better results only in specificity and precision, with a significant decrease in the sensitivity in workflow 36. The overall results measured in F1-Score in workflow 40 were considerably inferior compared to the other workflows tested, revealing the difference that the application of the transformation techniques generated in the results. The workflows (38,39) using discretization of attribute sets 2 and 4 (derived to the attribute sets 3 and 5), were not superior to workflow 1, considering essential transformation. It loses in the overall result as measured by F1-Score. However, it surpassed workflows 36 and 37.

4.1.3 Comparison of Categorical Mapping

Experiments 1, 7, 9, and 10 evaluate the impact of the categorical mapping. Tables 28, 29, 30 and 31 presents the obtained results.

Experiment 1

Aiming at comparing categorical mapping, exhibited in Table 28, this experiment examines the differences between workflow 0 (with the primary data and without transformation) and workflow 1. It obtained the best result with the conceptual hierarchy applied in workflow 22, which uses the same attributes of workflow 1. However, attribute 1 (composed of airline, departure, and arrival) was categorically mapped, aiming both this comparison and the application in neural networks.

Table 28 – Comparison Between Original and Transformed Time

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Conventional	0	RF	SUB	MM	0,647	0,644	0,648	0,235	0,344	1,385
Majority	1	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,481
Majority	22	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,296

Comparing workflow 1 (original attribute 1) and workflow 22 (attribute 1 transformed in attributes 11,13 and 14), it occurred by part of workflow 22 a slight improvement in accuracy and slight worsening in the sensitivity. Results were also superior in specificity and precision. In the overall comparison, F1-Score obtained better results, albeit with an overcome elapsed time.

Experiment 7

Table 29 – Comparison of impacts of separate data of Airline, Departure, and Arrival

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	19	RF	IMB	WN	0,915	0,513	0,982	0,822	0,631	5,450
Majority	20	RF	IMB	WN	0,916	0,665	0,958	0,724	0,693	5,318
Majority	21	RF	IMB	ZS	0,912	0,728	0,942	0,678	0,702	5,076
Majority	22	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,296
Conventional	19	NN	IMB	ZS	0,903	0,371	0,992	0,889	0,523	3,862
Conventional	20	NN	IMB	WN	0,909	0,423	0,990	0,872	0,570	3,088
Conventional	21	NN	IMB	WN	0,909	0,422	0,990	0,876	0,570	5,219
Conventional	22	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	6,115

Next, as presented in Table 29, the categorical mapping tests considered sets of attributes for airlines(workflow 19), departure(workflow 20), arrival(workflow 21) separately; in addition to the set composed by the attributes of the arrival and departure airlines (workflow 22), to verify which one could generate the greatest impact on the prediction. Results using the Random Forest (RF) machine learning method, in terms of accuracy, were very close between the separately arranged attributes (workflows 19,20,21) and a slightly higher result when the attributes are arranged together (workflow 22). For the Neural Network (NN) method, none of the workflows had a significant difference in terms of accuracy. In terms of recall, in the RF machine learning method, there was an improved highlight for workflow 22, with a more significant difference in workflow 21 sensitivity (match), which was the best followed by the flow of work 20 (arrival) and 19 (airlines). For the NN machine learning method, the results in terms of sensitivity were higher in workflows 20 (categorized arrival attribute) and 21 (categorized departure attribute).

Experiment 9

This experiment, as presented in Table 30 has workflow 12 deals with the use of the attributes related to sets (1,2,12) commonly with the original event attribute set (6). Workflow 14 owns sets of attributes (1,2,12), besides the event attribute transformed by categorical mapping (attribute set 7). Workflow 27 has the attributes (2,12), the event transformed by categorical mapping (attribute set 7). It also has the categorical mapping transformation of attribute 1, resulting in the attribute sets (11, Airlines; 13, Departure and 14, Arrival).

Table 30 – Comparison of Best {Basics+Events} and CM(Airline,Arrival,Departure)+ Events[7]

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Conventional	12	RF	SUB	MM	0,656	0,703	0,648	0,252	0,371	1,789
Majority	14	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,695
Majority	27	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,415

Relatively to workflow 12, the workflows 14 and 27 show performance far superior in all aspects, except on the elapsed time. It shows that the use of categorical mapping

contributed to the increase in accuracy and sensitivity. The comparison between workflows 14 and 27 obtains a higher total workflow performance, mostly because of the F1-score. Categorical mapping performs better not only on the attribute set 6 but also on the set of attributes.

Experiment 10

Experiment 10 compares workflows 4, 9, 26, and 28. Workflow 4 has sets of attributes (1,2,12) and original conditions (attribute 8). In workflow 9, we have the same set of attributes (1,2,12) and the conditions transformed through the categorical mapping (represented by the set of attributes 9). Workflow 28 has, as well in workflow 9, the transformed conditions (set of attributes 9). However, it also presents attributes (2,12) together with the transformation through categorical mapping of the attribute set 1, resulting in sets of attributes 11 (Airlines), 13 (Departure), and 14 (Arrival). Workflow 26 presents transformations by categorical mapping of the attribute set 1 (resulting in sets of attributes 11,13 and 14), but with the original set of the condition attribute (8). The results are presented in Table 31.

Table 31 – Comparison Between CM(Airline,Arrival,Departure) + Conditions[8,9]

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Conventional	4	RF	IMB	MM	0,943	0,674	0,988	0,902	0,772	8,252
Majority	9	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,874
Majority	26	RF	IMB	WN	0,914	0,878	0,920	0,645	0,744	11,816
Majority	28	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,508

Test results were very close between workflows. Comparing to the workflow 4, workflow 9 (with the categorical mapping of the set of condition attributes) obtained a slightly lower accuracy, but with a substantial increase of the sensitivity and reduction of the elapsed time. However, workflow 26, which had the categorical mapping only concerning the attribute set 1, obtained the worst result in all the queries tested, achieving the most significant time elapsed. However, workflow 28, which had the categorical mapping both relative to the set of attributes of conditions, and relative to the set of attributes 1 (airline, departure, arrival), obtained the best performance of this experiment, with better F1-Score.

4.1.4 Comparison of Normalization Methods

In order to evaluate normalization impacts, experiment 15 considered mean, maximum and minimum values (of the values obtained with all workflows [0 to 40]) and the use of the two machine learning techniques, as shown in Tables 32 and 33.

Table 32 – Normalization - Random Forest

Machine Learning: Random Forest(RF)								
C.R	**NM**	**Threshold**	**AC**	**SS/R**	**SP**	**P**	**F1-Score**	**Time**
MED	WN	Majority	0,521	0,884	0,460	0,334	0,424	10,002
MIN			0,143	0,203	0,000	0,139	0,201	1,102
MAX			0,940	1,000	0,982	0,822	0,806	39,522
MED	MM	Majority	0,520	0,885	0,460	0,334	0,423	9,891
MIN			0,143	0,211	0,000	0,139	0,201	1,164
MAX			0,940	1,000	0,988	0,845	0,801	39,029
MED	ZS	Majority	0,498	0,910	0,430	0,331	0,422	9,644
MIN			0,143	0,212	0,000	0,143	0,249	1,158
MAX			0,940	1,000	0,989	0,854	0,804	38,899
MED	WN	Conventional	0,722	0,625	0,738	0,503	0,450	10,002
MIN			0,193	0,039	0,064	0,101	0,074	1,102
MAX			0,946	0,969	0,999	0,948	0,786	39,522
MED	MM	Conventional	0,719	0,632	0,734	0,508	0,453	9,891
MIN			0,193	0,040	0,064	0,103	0,075	1,164
MAX			0,947	0,969	0,999	0,949	0,789	39,029
MED	ZS	Conventional	0,629	0,738	0,611	0,497	0,458	9,644
MIN			0,143	0,040	0,000	0,142	0,076	1,158
MAX			0,947	1,000	0,999	0,937	0,792	38,899

Results obtained with the Random Forest (RF) machine learning technique, pre-

sented in Table 32, reveal a reduction of the elapsed time of the Min-Max (MM) and Z-Score (ZS) techniques, comparing to the non-standardization method (WN), progressively. There were no changes found on the maximum and minimum accuracy values due to normalization. Nevertheless, there was a quality debase when using the Z-score application comparing to Mix-Max method, or even when normalization was not used. Regarding Sensibility/Recall, in the majority threshold approach, there was a progressive increase in mean sensitivity, relative to non-use of the normalization method, and Min-Max, and Z-score approaches, respectively. In the conventional threshold strategy, there was an increase of the sensibility/recall when applying the Z-score technique concerning Min-Max method and the non-use of normalization techniques. In both the majority and the conventional threshold approach, there was no significant change in the maximum and minimum values of the sensibility/recall. The use of Z-score reduced Specificity. There were no significant variations between Precision and F1-Score.

Table 33 – Normalization - Neural Networks(NN)

Machine Learning: Neural Network(NN)								
C.R	**NM**	**Threshold**	**AC**	**SS/R**	**SP**	**P**	**F1-Score**	**Time**
MED	WN	Majority	0,335	0,836	0,251	0,215	0,279	15,654
MIN			0,143	0,000	0,000	0,000	0,000	1,049
MAX			0,859	1,000	1,000	0,878	0,521	58,617
MED	MM	Majority	0,354	0,835	0,273	0,188	0,279	14,602
MIN			0,141	0,022	0,000	0,141	0,039	1,060
MAX			0,849	1,000	0,987	0,341	0,445	52,736
MED	ZS	Majority	0,362	0,773	0,294	0,194	0,246	13,977
MIN			0,142	0,016	0,000	0,142	0,031	1,245
MAX			0,857	1,000	0,998	0,539	0,453	59,082
MED	WN	Conventional	0,614	0,566	0,622	0,355	0,303	15,654
MIN			0,141	0,000	0,000	0,000	0,000	1,049
MAX			0,909	1,000	1,000	0,899	0,570	58,617
MED	MM	Conventional	0,603	0,538	0,614	0,402	0,283	14,602
MIN			0,141	0,000	0,000	0,000	0,000	1,060

Table 33 – Normalization - Neural Networks(NN)

Machine Learning: Neural Network(NN)								
MAX			0,903	1,000	1,000	0,948	0,516	52,736
MED	ZS	Conventional	0,628	0,474	0,653	0,451	0,226	13,977
MIN			0,143	0,005	0,000	0,143	0,009	1,245
MAX			0,904	1,000	1,000	0,935	0,524	59,082

Results obtained with Neural Network (NN) machine learning technique, presented in Table 33, reveal a reduction in the elapsed time of the Min-Max (MM) techniques concerning the non-use of normalization technique (WN), progressively.

Normalization caused no significant variations of the minimum and maximum numbers in Accuracy. Still, on the average value measured in workflows, the accuracy raised using Z-score standardization application instead of Min-Max and the non-use of techniques of normalization, both in the majority and conventional threshold approach.

Regarding the Sensibility / Recall, in the majority threshold approach, there was a decrease in the mean sensitivity about the use of Z-score normalization. In the conventional threshold approach, there was a reduction of sensibility/recall when applying the Min-max and Z-score techniques, progressively. In both the majority and the traditional threshold approach, there was no significant change in the maximum and minimum values of the sensibility/recall.

Results concerning Specificity revealed a progressive improvement in the majority threshold approach when using Min-Max and Z-score methods, progressively. In the conventional threshold approach, growth only occurred when the Z-score technique was applied. Precision using majority threshold approach worsened in the application of Min-Max and Z-score techniques. However, improvements occurred using the conventional threshold approach applying normalization technique, instead of non-application. Z-score outperformed the Min-Max method.

In the context of the conventional threshold approach, there was a progressive worsening on the F1-Score, non-application of the normalization technique outperforming the Min-Max, and Z-score techniques, respectively.

4.2- Feature Selection and Extraction

Experiments 3, 4, 5, and 8 evaluate the impacts of the feature selection and extraction, as shown in Tables 34, 35, 36, 37.

Experiment 3

This experiment evaluates workflows that applied feature selection and extraction techniques. Workflow 4 consists of sets of attributes (1,2,12) and the original conditions (8).

Transformation on workflows 5 and 11, composed by the attributes (1,2,12) and the set of attributes transformed conditions (9), experienced numerous feature selection techniques like LASSO, CFS, and INFOGAIN, as well as feature extraction method called PCA.

There also have workflows 25 and 28. They present sets of attributes (2.12), the transformed condition attribute set (9), and the set of transformed attributes 11, 13 and 14 (Airline, Departure, Arrival), resulting from the transformation of the attribute set 1. On these two workflows, the feature selection (CFS) reduction occurred. One case applied the feature extraction feature PCA. Results are presented in Table 34.

Workflows of 5 to 8 (respectively presenting the PCA and LASSO/PCA and INFOGAIN/PCA and CFS techniques) revealed inferior results comparing to workflow 4 (composed of the original condition attributes). Still, the workflows from 9 to 11, which had only the application of the feature selection techniques (LASSO, INFOGAIN, and CFS, respectively), obtained the same result, being superior to both workflows of 4 and numbered from 5 to 8.

Considering machine learning method RF, workflows 25 and 28 revealed the best result. It significantly increased the sensitivity and the overall result in workflow 28, using features selection CFS technique and extraction PCA. Related to the machine learning method NN, workflow 25 manifested the best result (especially the one that refers to the sensitivity), using only the feature selection CFS method. Contrasting to workflow 4 and

Table 34 – Comparison of Feature Selection Techniques(CFS,PCA,LASSO,IG) on Conditions

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Conventional	4	RF	IMB	MM	0,943	0,674	0,988	0,902	0,772	8,252
	5	RF	IMB	MM	0,910	0,786	0,930	0,653	0,713	14,264
	6	RF	IMB	ZS	0,898	0,857	0,905	0,601	0,707	12,564
	7	RF	IMB	WN	0,891	0,914	0,888	0,575	0,706	9,163
Majority	8	RF	IMB	WN	0,894	0,866	0,898	0,586	0,699	9,844
	9	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,874
	10	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	6,091
	11	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	6,080
	25	RF	IMB	WN	0,908	0,508	0,974	0,765	0,610	3,947
Conventional	25	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	2,992
Majority	28	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,508
Conventional	28	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	5,918

workflows 5 to 11, workflow 28 with the machine learning method RF also showed higher performance.

Experiment 4

In experiment 4 experience, workflows try to evaluate different aspects of feature selection. Workflows 13 and 14 compare the transformed set of event attributes (7), applying feature selection and extraction PCA and LASSO algorithms in workflow 13, and feature selection technique CFS in workflow 14. Attributes (1, 2, and 12) integrates both of them.

Workflows 24, 27, and 29 have in common the use of sets of attributes (2,12). There were applied feature extraction and selection methods PCA and CFS on the sets of attributes (7,11,13,14) of workflow 24. In workflow 27, there were used CFS feature selection techniques on the set of event attributes (7) and individually the application of the techniques of feature extraction and selection, PCA and CFS. Finally, workflow 29 experienced the use of feature extraction and selection PCA and CFS methods, considering the complete set of attributes (7,11,13,14).

Table 35 illustrates the effects obtained using workflows 13 and 14. There was no difference in the results regarding the application of the feature selection technique CFS

Table 35 – Comparison of Feature Selection Techniques(LASSO,CFS,PCA) on Events

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	13	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	6,069
Majority	14	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,695
Majority	24	RF	IMB	WN	0,939	0,881	0,949	0,740	0,805	10,732
Conventional	24	NN	IMB	ZS	0,896	0,375	0,983	0,786	0,507	6,430
Majority	27	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,415
Conventional	27	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	6,047
Majority	29	RF	IMB	MM	0,937	0,883	0,946	0,733	0,801	13,638
Conventional	29	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	4,468

(workflow 14) and feature extraction and selection, PCA, and LASSO (workflow 13).

On workflows 24, 27, and 29, evaluation befalls in two parts. Firstly, results regarding machine learning RF algorithms were very close, highlighting better F1-Score on workflow 27, achieving the best overall effect of this experiment. Regarding the neural networks, there was a slight difference, mainly in what concerns the accuracy, with workflow 29 standing out among the three workflows tested with higher F1-Score and shorter time elapsed.

Experiment 5

Experiment 5 embraces tests regarding workflows 15,16 and 17, for sets of event attributes and conditions (original and transformed). Both workflows 15 and 16 use the basic set of attributes (1,2,12). Besides, workflow 15 also has the application of the LASSO feature selection technique, on the set of events (7) and the CFS technique on the set of conditions (9). The set of attributes (1,2,12) of workflow 16 held feature extraction and selection techniques PCA and LASSO for the set of event attributes (7), and PCA and CFS methods for the set of condition attributes (9). In workflow 17, the goal is to analyze the set with original data for Airline, Departure, Arrival (1); Events (6); Conditions (8); and the original data set of the departure date and time (99).

Table 36 highlights results on workflow 15, except for referring to the elapsed time, revealing that the application of feature extraction technique before feature selection may not be the best choice in this situation. Still, a significantly lower result occurred in

Table 36 – Comparison of Feature Selection Techniques(LASSO,CFS,PCA) on Events and Conditions

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	15	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,620
	16	RF	IMB	WN	0,894	0,866	0,898	0,586	0,699	9,764
Conventional	17	RF	SUB	MM	0,655	0,637	0,658	0,238	0,347	1,549

almost all the questions, is the best only in the time elapsed, in workflow 17, presenting inadequate performance using original attributes.

Experiment 8

One goal in experiment 8 is to evaluate the application of feature selection techniques on workflows 1, 22, and 23. Workflow 1 has the underlying attribute schema, considering the transformation of the data only of the departure date (2.12) in addition to original set 1 (Airline, Departure, Arrival). Workflows 22 and 23 have the basic attribute sets (2.12). Besides the set of primary attributes, workflow 22 has the data sets (11,13,14) with the individual application of feature extraction and selection PCA and CFS algorithms. On workflow 23, the experiment considered only the individual application of the feature selection method CFS.

Table 37 – Comparison of Feature Selection to completely data(Airline,Arrival,Departure)

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	1	RF	IMB	WN	0,934	0,895	0,941	0,716	0,795	5,481
Majority	22	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,296
Conventional	22	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	6,115
Majority	23	RF	IMB	WN	0,908	0,508	0,974	0,765	0,610	3,953
Conventional	23	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	2,951

Results related to workflows 22 and 23, presented in Table 37, referring to workflows 22 and 23, considered the use of the machine learning RF technique initially. In this context, workflow 22 presented a far superior result in practically all aspects tested, losing only in the time elapsed. This scenario reveals that the use of the techniques of feature extraction and selection proved very productive in this situation. Results involving machine

learning technique NN on workflow 23 was the opposite - superior to that obtained in workflow 22. Comparing to base workflow 1, the result obtained in workflow 22 using machine learning RF method was superior, revealing itself as the best of this experiment.

4.3- Comparison of Balancing Methods

In order to evaluate the impacts of the application of balancing techniques, experiment 16 presents the consolidated results in their mean and minimum and maximum values with the Tables 38 and 39.

Table 38 – Balancing Results - Random Forest

Machine Learning Method: Random Forest(RF)								
C.R	**BL**	**Threshold**	**AC**	**SS/R**	**SP**	**P**	**F1-Score**	**Time**
MED	IMB	Majority	0,900	0,756	0,923	0,646	0,677	8,362
MIN			0,666	0,203	0,668	0,246	0,267	3,149
MAX			0,940	0,933	0,989	0,854	0,806	16,597
MED	SUB	Majority	0,446	0,973	0,358	0,206	0,339	2,273
MIN			0,194	0,893	0,062	0,149	0,259	1,102
MAX			0,691	0,995	0,647	0,310	0,469	3,879
MED	SMOTE	Majority	0,193	0,950	0,067	0,147	0,252	18,902
MIN			0,143	0,362	0,000	0,139	0,201	9,044
MAX			0,729	1,000	0,773	0,254	0,329	39,522
MED	IMB	Conventional	0,908	0,402	0,993	0,886	0,532	8,362
MIN			0,859	0,039	0,981	0,524	0,074	3,149
MAX			0,947	0,700	0,999	0,949	0,792	16,597
MED	SUB	Conventional	0,826	0,809	0,829	0,463	0,585	2,273
MIN			0,634	0,633	0,627	0,231	0,341	1,102
MAX			0,912	0,896	0,915	0,637	0,745	3,879
MED	SMOTE	Conventional	0,335	0,783	0,261	0,158	0,243	18,902
MIN			0,143	0,068	0,000	0,101	0,087	9,044
MAX			0,854	1,000	0,982	0,433	0,273	39,522

The application of the Random Forest balancing method reveals that the use of unbalanced data produces the best average accuracy results (considering both conventional and majority threshold approach). Table 38 presents the related results. In the majority approach, the SUB and SMOTE methods had incredibly lower results regarding accuracy. In the conventional approach, the subsampling technique was slightly inferior to the use of

unbalanced data. In this case, the SMOTE technique presented a significantly lower result comparing to the two other states previously mentioned (unbalanced and SUB). Regarding sensibility/recall, the use of balancing techniques produced a significant increase in the values obtained, both in the conventional and in the majority threshold approach.

In the majority threshold approach, the Specificity and Precision, as well as observed regarding the accuracy, had much lower results in sets that had the balancing techniques applied. In the Conventional approach, concerning the average values, the same phenomenon occurred; however, for the maximum values of Specificity, Precision, and F1-Score were very close in the different balancing techniques.

Table 39 – Balancing Results - Neural Networks

Machine Learning Method: Neural Networks(NN)								
C.R	**BL**	**Threshold**	**AC**	**SS/R**	**SP**	**P**	**F1-Score**	**Time**
MED	IMB	Majority	0,741	0,450	0,790	0,308	0,300	4,533
MIN			0,305	0,000	0,204	0,000	0,000	1,060
MAX			0,859	0,913	1,000	0,878	0,521	50,456
MED	SUB	Majority	0,163	0,997	0,024	0,146	0,254	32,937
MIN			0,142	0,989	0,000	0,142	0,249	1,049
MAX			0,222	1,000	0,095	0,155	0,268	59,082
MED	SMOTE	Majority	0,146	0,997	0,005	0,143	0,250	6,423
MIN			0,141	0,974	0,000	0,141	0,247	1,084
MAX			0,177	1,000	0,043	0,146	0,253	18,068
MED	IMB	Conventional	0,880	0,182	0,996	0,759	0,260	4,533
MIN			0,857	0,000	0,983	0,000	0,000	1,060
MAX			0,909	0,423	1,000	0,948	0,570	50,456
MED	SUB	Conventional	0,739	0,468	0,784	0,300	0,298	32,937
MIN			0,352	0,014	0,272	0,100	0,027	1,049
MAX			0,859	0,836	0,999	0,659	0,470	59,082
MED	SMOTE	Conventional	0,226	0,928	0,109	0,149	0,255	6,763
MIN			0,141	0,300	0,000	0,141	0,220	1,084
MAX			0,697	1,000	0,763	0,174	0,271	18,068

Results considering the use of Neural Networks method, as shown in Table 39), reveals the best average accuracy results considering unbalanced data. It considers conventional and with the majority threshold approach. It was a similar scenario to the RF machine learning method. In the majority threshold approach, the SUB and SMOTE methods had incredibly lower results regarding accuracy. The subsampling and smote techniques were slightly inferior to the use of the unbalanced data in the conventional threshold approach. Regarding Sensibility/Recall, the use of balancing techniques caused

a significant increase in the values obtained, both in the conventional and in the majority threshold approach.

The Specificity and Precision, in the majority threshold approach, as well as observed regarding the accuracy, had much lower results in sets that had the applied balancing techniques. All balancing techniques revealed similar values of F1-Score, considering the difference between the maximum values found in the unbalanced data sets. It happened in the majority approach and with the use of the conventional threshold.

4.4- Comparison of Threshold Approach

Experiment 17 evaluates the impacts of the application of the threshold approach (Conventional and Majority). Table 40 and Figure 53 present consolidated results (mean, minimum, and maximum values).

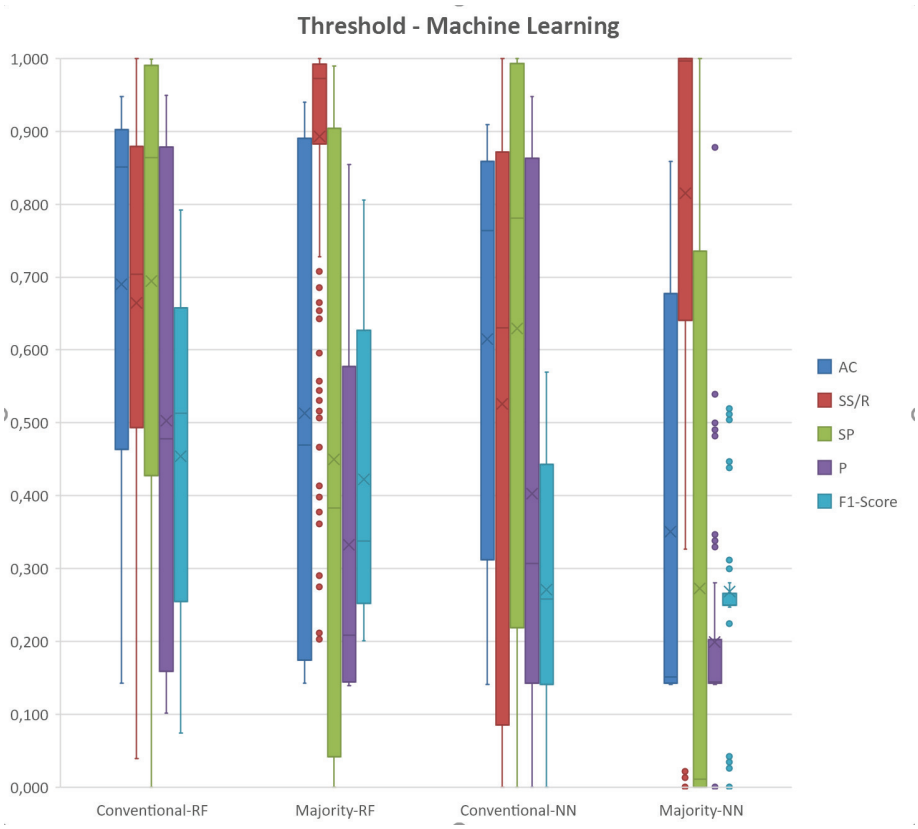

Figure 53 – Threshold Results

Accuracy results produced by the application of the Random Forest (RF) learning

method presents that the conventional approach shows a much higher average result. The minimum results in both approaches are very similar. Regarding Sensibility/Recall, the opposite occurs: the average value is higher in the majority approach compared to the conventional approach. On the specificity, precision, and F1-Score, the average values of the conventional approach surpass those of the majority approach. However, evaluating the maximum values found, the majority approach reveals slightly higher results.

Concerning the Neural Networks (NN) machine learning method, results resemble similar to that found in the RF learning method. However, evaluating the maximum values found, the conventional approach reveals slightly higher overall results, unlike the other method analyzed.

Table 40 – Threshold Result

C.R	Threshold	ML	AC	SS/R	SP	P	F1-Score	Time
MED	Conventional	RF	0,690	0,665	0,694	0,503	0,454	9,846
MIN			0,143	0,039	0,000	0,101	0,074	1,102
MAX			0,947	1,000	0,999	0,949	0,792	39,522
MED	Majority	RF	0,513	0,893	0,450	0,333	0,423	9,846
MIN			0,143	0,203	0,000	0,139	0,201	1,102
MAX			0,940	1,000	0,989	0,854	0,806	39,522
MED	Conventional	NN	0,632	0,564	0,643	0,443	0,320	13,201
MIN			0,141	0,000	0,000	0,000	0,000	1,049
MAX			0,909	1,000	1,000	0,948	0,570	59,082
MED	Majority	NN	0,371	0,810	0,298	0,227	0,278	13,201
MIN			0,141	0,000	0,000	0,000	0,000	1,049
MAX			0,859	1,000	1,000	0,878	0,521	59,082

4.5- Machine Learning: Random Forest X Neural Networks

Experiment 18 evaluates the impacts of the Machine Learning RF / NN methods, with best results presented in Table 41.

Table 41 – Machine Learning

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
Majority	22	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,296
	18	NN	IMB	WN	0,859	0,537	0,913	0,505	0,521	3,033
Conventional	2	RF	IMB	ZS	0,947	0,700	0,989	0,912	0,792	7,638
	20	NN	IMB	WN	0,909	0,423	0,990	0,872	0,570	3,088

This scenario presents the best results obtained using RF/NN methods combined with each threshold approach (Majority/Conventional). Random Forest machine learning method was superior in the two threshold approaches in all analyzed factors, such as accuracy, sensibility, F1-Score, except for the elapsed time. Neural Networks machine learning method obtained its best results with a much shorter time.

Figure 54 – Machine Learning

4.6- Evaluation of Time Elapsed

In order to evaluate the impacts on the Elapsed Times, especially concerning the balancing and normalization, experiment 19 exhibits the best and worst results in Table 42.

Table 42 – Time Elapsed - Balancing and Normalization

Machine Learning : RF			
	Time - Balancing		
CR	IMB	SUB	SMOTE
MED	8,362	2,273	18,902
MIN	3,149	1,102	9,044
MAX	16,597	3,879	39,522
	Time - Normalization		
CR	WN	MM	ZS
MED	10,002	9,782	9,644
MIN	1,102	1,164	1,158
MAX	39,522	39,029	38,899

Machine Learning : NN			
	Time - Balancing		
CR	IMB	SUB	SMOTE
MED	4,533	32,937	6,763
MIN	1,060	1,049	1,084
MAX	50,456	58,617	18,068
	Time - Normalization		
CR	WN	MM	ZS
MED	15,654	14,602	13,977
MIN	1,049	1,060	1,245
MAX	57,399	49,953	57,384

Random Forest presented inferior average times using balancing SUB, followed by Imbalanced, and with superior time considering SMOTE technique. Table 42 consolidates these experiment results.

For the Neural Networks machine learning method, SUB had the worst result, followed by SMOTE, and the best result was not to use balance techniques. Normalization results were very close in all techniques in the two methods of machine learning.

Considering the aspect of better and worse elapsed time results, as shown in Figures 55 and 56, Random Forest produced the worst results (higher times) with the SMOTE balancing technique. Neural Networks method obtained all higher times results using the SUB balancing technique.

The use of the SUB balancing technique unveiled the best results (lower times) for the Random Forest method. Considering Neural Networks, several methods achieved the fleetest times, as, for example, SUB (workflows 31, 21 and 19), SMOTE (workflow 24), and IMB (workflow 24).

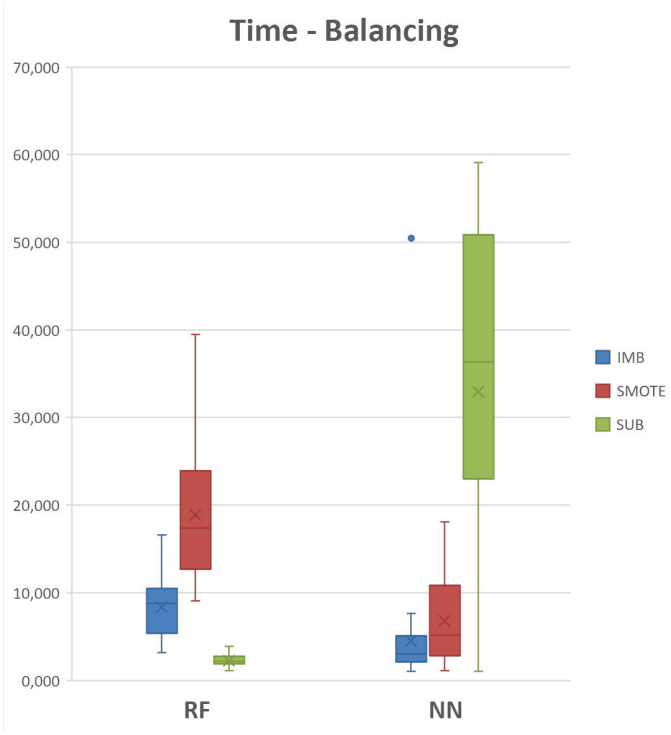

Figure 55 – Time - Balancing

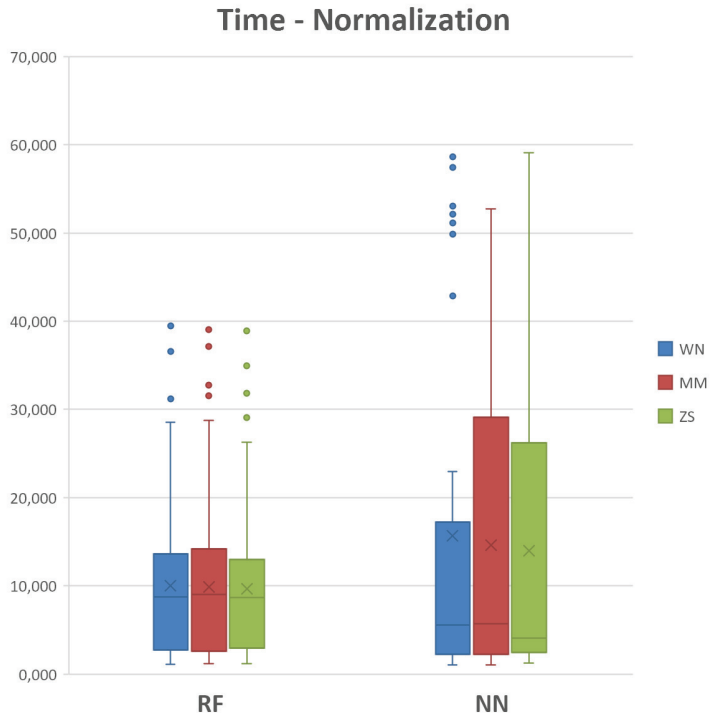

Figure 56 – Time - Normalization

4.7- Accuracy, Sensibility/Recall, and F1-Score

Experiment 20 evaluates the best results of Accuracy, Sensibility/Recall, and F1-Score, with the best results presented in table 43.

Table 43 – Better F1-Score

Threshold	WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
	2	RF	IMB	ZS	0,947	0,700	0,989	0,912	**0,792**	7,638
	2	RF	IMB	MM	0,947	0,695	0,989	0,911	**0,789**	7,855
Conventional	2	RF	IMB	WN	0,946	0,691	0,989	0,911	**0,786**	7,317
	4	RF	IMB	MM	0,943	0,674	0,988	0,902	**0,772**	8,252
	4	RF	IMB	ZS	0,942	0,674	0,987	0,897	**0,770**	7,981
	22	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,296
	27	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,415
Majority	28	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,508
	24	RF	IMB	WN	0,939	0,881	0,949	0,740	**0,805**	10,732
	22	RF	IMB	ZS	0,938	0,896	0,945	0,729	**0,804**	9,777

This experiment selected the five best F1-Score results for both the conventional threshold approach and the Majority approach. In this selection, the learning method of random forest machine had all the results of this ranking. The results where the normalization technique did not occur were the most frequent among those ranked with five items among the ten most, being a vast majority in the approach threshold majority. Then Z-score with three items and Min-Max with two items.

The Accuracy of this coming in all the best-ranked results, getting in the range of 94%.

Sensibility/Recall presented its best results in the majority threshold approach, ranging from 87.5% to 89.6%. Meanwhile, in the conventional approach, the results were between 67.4% and 70%.

The F1-Score showed, in general, in this listing results always higher than 77%, reaching results of 80.6%.

4.8- Comparison With Related Works

The two best results obtained in this work (workflow 2 with Conventional Threshold and Worflow 22 with Majority Threshold) were then selected for comparison and evaluation with the best results obtained in the related works as present in Table 44. That comparison with these related works used normalized improvement, so as not to overestimate any improvement obtained.

Figure 57 – Accuracy, Sensibility/Recall, Specificity and F1-Score

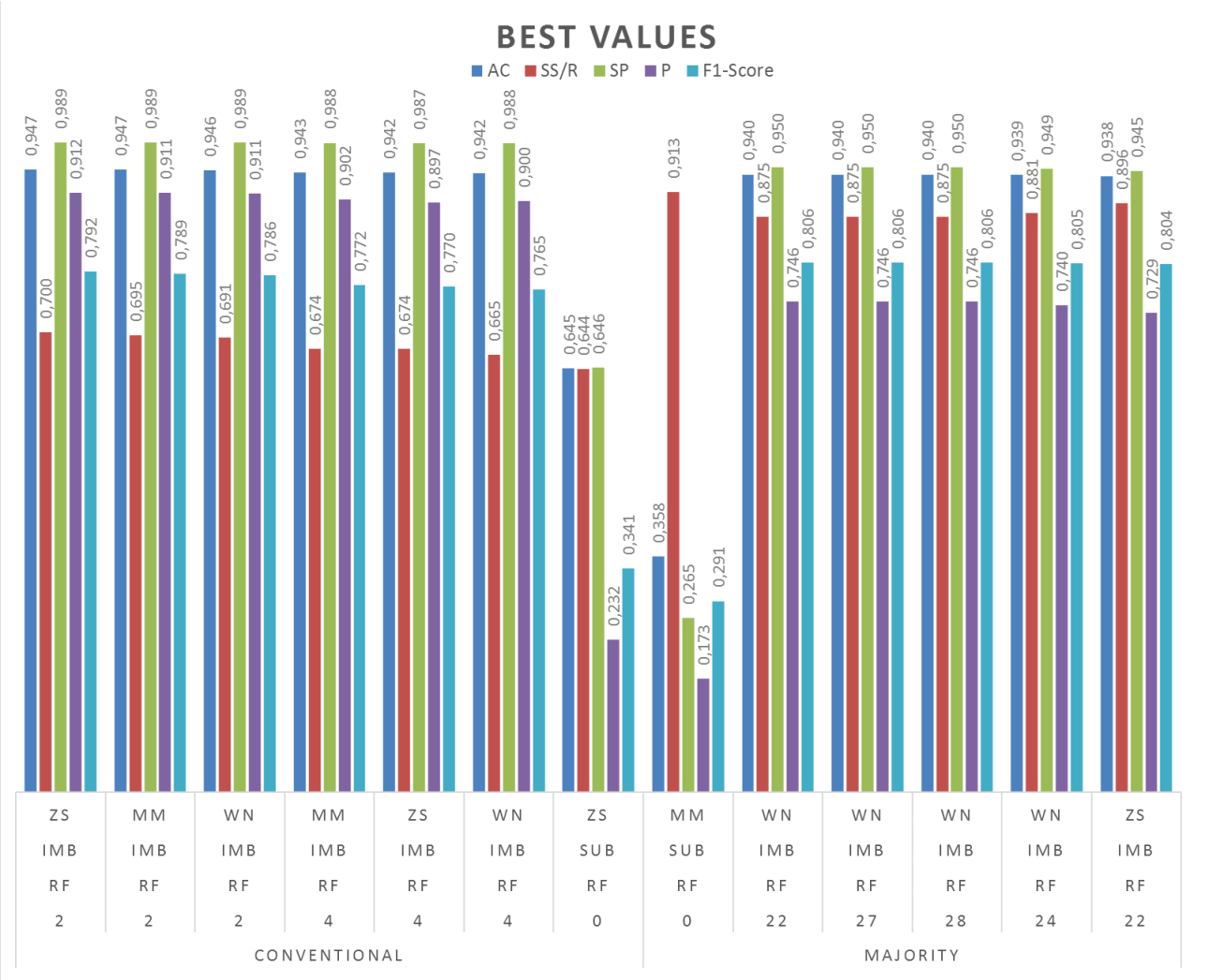

Figure 58 – Accuracy, Sensibility/Recall, Specificity and F1-Score

Comparing the best results achieved in the work related to the results achieved in this work in terms of Accuracy, Precision, Recall, F1-Score, promising results were found as expressed in Figure 59.

Regarding Accuracy, in relation to the work related to the best results presented by Cao and Fang [2012] which obtained 0.883; This work obtained 0.947 (with conventional Threshold) and 0.940 (with Majority Threshold). Comparatively, considering a normalized improvement, the improvement rate was 54.70% and 48.72% respectively in this study.

In aspect of Precision, the related work that presented the best results was that of Belcastro et al. [2016] with 0.869; In relation to this research was not achieved a significant improvement, having obtained the best result, in its version with Majority Threshold, the

Table 44 – Better Results achieved in in this work and related works.

Pub.	**Threshold**	Results			
		Accuracy	Precision	Recall	F1-score
Rebollo and Balakrishnan [2014]	Conventional	0.810	-	0.764	-
Cao and Fang [2012]	Conventional	**0.883**	-	-	-
Khaksar and Sheikholeslami [2019]	Conventional	0.764	-	0.600	-
Chakrabarty et al. [2019]	Conventional	0.797	**0.760**	0.800	**0.740**
Choi et al. [2016]	Conventional	0.834	-	-	-
Nigam and Govinda [2017]	Conventional	0806	0.321	0.1150	0.209
Choi et al. [2017]	Conventional	0.831	-	-	-
Saadat and Moniruzzaman [2019]	Conventional	0.821	-	-	-
Belcastro et al. [2016]	Conventional	0.858	-	**0.869**	-
Henriques and Feiteira [2018]	Conventional	0.856	-	-	-
This Work	Conventional	**0.947**	**0.912**	0.700	0.792
This Work	Majority	0.940	0.746	**0.875**	**0.806**

value 0.875. That is, a 4.58% normalized improvement rate over the best result observed in the related works.

Regarding Recall, the best related work was in Chakrabarty et al. [2019] with 0.760. Already in this research was obtained as best result 0.912 (Conventional Threshold). Comparing the results achieved in this research with those of the best related work, a normalized improvement rate of 63.33% was achieved.

Regarding F1-Score, the work related to best result was also Chakrabarty et al. [2019] with 0.740. In this current work we obtained results of 0.792 (Conventional Threshold) and 0.806 (Majority Threshold). These results were in terms of normalized improved 20% and 25.38% higher, respectively, than the best result achieved in related work.

Figure 59 – Normalized Improvement

Conclusions

In this work, we performed an experimental evaluation of data preprocessing methods, especially normalization, categorical mapping and discretization to optimize the accuracy and sensitivity of the prediction models.

To produce this assessment, a database was built with the integration of multiple data sources, such as flight data (called VRA [ANAC, 2016]) and weather data (Weather Underground (WU), collected for each airport. From there, a selection of the airports with the highest number of flights was carried out, besides a data cleaning segmented in two steps: first was a verification of the inconsistencies; and second, verification and processing of missing data. However, as a remodeling of the infrastructure of Brazilian airports took place due to the great events of the World Cup and Olympics, it was also necessary to select data for analysis from 2015, in order to represent the current infrastructure of the Brazilian airports with the improvements already completed.

The data transformation process was performed after the proper verification of each field, in the application of techniques such as Binning, Categorical Mapping, and Conceptual Hierarchy were applied in this data, increasing the probability of significant predictions by the classifier. In addition to these techniques, the data normalization was also applied, generating three distinct data sets: one without normalization, and another two with the application of the Min-max and Z-score techniques, respectively.

With the transformation adequately applied, these data sets were divided into training and test sets, in the ratio 80:20, allowing the continuity of the tests. Then, a new round of data was applied to the training data sets with three levels of distinct sets: one without imbalanced and two with Random Subsampling (RS) and Synthetic Minority Over-sampling Technique (SMOTE), respectively.

To optimize the experiments with the selection of more representative features, data reduction was also performed, more specifically feature selection (LASSO, CFS, IG) and feature extraction (PCA).

Workflows have been developed thinking for each training dataset and machine learning methods from the creation of an evaluation model. With model generation, it was possible to carry out the appropriate analysis of the test datasets and verification of the

predictive capacity of the classification algorithm.

Based on these workflows, and as a way of evaluating the impact transformations caused on reaching results, twenty experiments were carried out.

Experiment analysis points out that the application of the transformation techniques allowed a considerable improvement of the results obtained in the prediction models, optimizing Accuracy, Sensitivity, and F1-Score. It is necessary to emphasize that the transformation of the data (conceptual hierarchy, categorical mapping, and application of the data type ITime) and data reduction (feature selection and extraction) generated a absolute difference that increased in more than 130% the results of F1-Score, 40% the results of Accuracy, 35% the sensitivity results, compared with the original data. Compared to the related studies, with the use of a normalized improvement, results were obtained up to 54.70% superior in terms of Accuracy; up to 4.58% higher in Precision terms; up to 63.33% higher in terms of Recall; and results about 25.38% higher in terms of F1-Score.

As suggestions for improvements and future work, it may be considered the application of other transformation methods and strategies to the dataset and balancing. Another possibility of improvement can be considered in the application of other limit modalities, cost-sensitive learning, other approaches to tunning, application of hyperparameter search techniques and machine learning methods, such as deep learning. Another interesting new study consists of the analysis of the imputation methods applied to original missing data using different techniques. Other opportunities for future improvements also include comparing the periods before and after the refurbishment of airports to see if a classifier trained in the period prior to the improvements would be able to make good recommendations later. It is also worth checking the discretization of data that went through the conceptual hierarchy semantically, such as grouping of months in season and hours in shifts.

Bibliography

(2010). *Arctic Blast Sets Record High Pressure, url=https://weather.com/science/weather-explainers/news/arctic-blast-record-high-pressure-plains, author=The Weather Channel*.

(2011a). *Barometric Pressure Records, url=https://wu.com/blog/weatherhistorian/, author=The Weather Channel*.

(2011b). *Fishing by the Barometer*. The Weather Channel.

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Al Shalabi, L. and Shaaban, Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. In *Dependability of Computer Systems, 2006. DepCos-RELCOMEX'06. International Conference on*, pages 207–214. IEEE.

AlMuhaideb, S. and Menai, M. E. B. (2016). An individualized preprocessing for medical data classification. *Procedia Computer Science*, 82:35–42.

ANAC (2012). Agência Nacional de Aviação Civil. Technical report, http://www.anac.gov.br/assuntos/legislacao.

ANAC (2016). Agência Nacional de Aviação Civil.

ANAC (2017). Agência Nacional de Aviação Civil. Technical report, http://www.anac.gov.br/assuntos/dados-e-estatisticas/mercado-de-transporte-aereo/anuario-do-transporte-aereo/dados-do-anuario-do-transporte-aereo.

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64.

Belcastro, L., Marozzo, F., Talia, D., and Trunfio, P. (2016). Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):5.

Bellman, R. E. and Dreyfus, S. E. (2015). *Applied dynamic programming*, volume 2050. Princeton university press.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In J. Shawe-Taylor, R.S. Zemel, P. B. F. P. K. W., editor, *25th Annual Conference on Neural Information Processing Systems (NIPS 2011)*, volume 24 of *Advances in Neural Information Processing Systems*, Granada, Spain. Neural Information Processing Systems Foundation.

Bilski, P. (2014). Data set preprocessing methods for the artificial intelligence-based diagnostic module. *Measurement*, 54:180–190.

BNDES (2010). Banco Nacional de Desenvolvimento Econômico e Social. Technical report, https://www.bndes.gov.br/Arquivos/empresa/pesquisa/chamada3/relatorioconsolidado.pdf.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical methods in medical research*, 5(3):215–238.

Britto, R., Dresner, M., and Voltes, A. (2012). The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review*, 48(2):460 – 469.

Cao, W. and Fang, X. (2012). Airport flight departure delay model on improved bn structure learning. *Physics Procedia*, 33:597–603.

Chakrabarty, N., Kundu, T., Dandapat, S., Sarkar, A., and Kole, D. K. (2019). Flight arrival delay prediction using gradient boosting classifier. In *Emerging Technologies in Data Mining and Information Security*, pages 651–659. Springer.

Chan, C.-C. (1998). A rough set approach to attribute generalization in data mining. *Information Sciences*, 107(1-4):169–176.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357.

Choi, S., Kim, Y. J., Briceno, S., and Mavris, D. (2016). Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE.

Choi, S., Kim, Y. J., Briceno, S., and Mavris, D. (2017). Cost-sensitive prediction of airline delays using machine learning. In *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, pages 1–8. IEEE.

Dara, J., Dowling, J. N., Travers, D., Cooper, G. F., and Chapman, W. W. (2008). Evaluation of preprocessing techniques for chief complaint classification. *Journal of Biomedical Informatics*, 41(4):613–623.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Fausett, L. V. and others (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*, volume 3. Prentice-Hall Englewood Cliffs.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

García, S., Luengo, J., and Herrera, F. (2016). *Data preprocessing in data mining*. Springer.

García, V., Sánchez, J. S., and Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21.

Halevy, A., Rajaraman, A., and Ordille, J. (2006). Data integration: the teenage years. In *Proceedings of the 32nd international conference on Very large data bases*, pages 9–16. VLDB Endowment.

Hall, M. A. (1998). Correlation-based feature selection for machine learning. Technical report.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Haykin, S. S., Haykin, S. S., Haykin, S. S., and Haykin, S. S. (2009). *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA:.

Henriques, R. and Feiteira, I. (2018). Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport. *Procedia computer science*, 138:638–645.

Hoshyar, A. N., Al-Jumaily, A., and Hoshyar, A. N. (2014). The beneficial techniques in preprocessing step of skin cancer detection system comparing. *Procedia Computer Science*, 42:25–31.

Huang, J., Li, Y.-F., and Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, 67:108–127.

Iliou, T., Anagnostopoulos, C.-N., Stephanakis, I. M., and Anastassopoulos, G. (2017). A novel data preprocessing method for boosting neural network performance: a case study in osteoporosis prediction. *Information Sciences*, 380:92–100.

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., and Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94.

Jolliffe, I. (2002). *Principal component analysis*. Springer Verlag, New York.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kelleher, J. D., Mac Namee, B., and D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

Khaksar, H. and Sheikholeslami, A. (2019). Airline delay prediction by machine learning algorithms. *Scientia Iranica*, 26(5):2689–2702.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117.

Krawczyk, B. and Woźniak, M. (2015). Cost-sensitive neural network with roc-based moving threshold for imbalanced classification. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 45–52. Springer.

Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.

Lavangnananda, K. and Waiwing, S. (2015). Effectiveness of Different Preprocessing Techniques on Classification of Various Lengths of Control Charts Patterns. *Procedia Computer Science*, 69:44–54.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM.

Lever, J., Krzywinski, M., and Altman, N. (2016). *Points of significance: classification evaluation*. Nature Publishing Group.

Li, X.-r., Hu, Z.-y., Zhao, Y.-h., and Liu, Z.-t. (2008). Spectral Preprocessing and Its Effect on Galaxy/Quasar Classification. *Chinese Astronomy and Astrophysics*, 32(1):13–22.

Lipton, Z. C., Elkan, C., and Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326.

Little, R. J. and Schenker, N. (1995). Missing data. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 39–75. Springer.

Luypaert, J., Heuerding, S., De Jong, S., and Massart, D. (2002). An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. *Journal of pharmaceutical and biomedical analysis*, 30(3):453–466.

Luypaert, J., Heuerding, S., Vander Heyden, Y., and Massart, D. (2004). The effect of pre-processing methods in reducing interfering variability from near-infrared measurements of creams. *Journal of pharmaceutical and biomedical analysis*, 36(3):495–503.

López, V., Fernández, A., Moreno-Torres, J. G., and Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608.

Macedo, D. C. and Matos, S. N. (2010). Extração de conhecimento através da mineração de dados. *Revista de Engenharia e Tecnologia*, 2(2):Páginas–22.

Majidi, M. and Oskuoee, M. (2015). Improving pattern recognition accuracy of partial discharges by new data preprocessing methods. *Electric Power Systems Research*, 119:100–110.

Marques, F. J., Moutinho, A., Vieira, S. M., and Sousa, J. M. (2011). Preprocessing of clinical databases to improve classification accuracy of patient diagnosis. *IFAC Proceedings Volumes*, 44(1):14121–14126.

Matsudaira, K. (2015). The science of managing data science. *Communications of the ACM*, 58(6):44–47.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1):27–32.

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Nigam, R. and Govinda, K. (2017). Cloud based flight delay prediction using logistic regression. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 662–667. IEEE.

Nikulin, A. E., Dolenko, B., Bezabeh, T., and Somorjai, R. L. (1998). Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra. *NMR in Biomedicine*, 11(45):209–216.

Ogasawara, E. (2018). *Brazilian flight datasets*.

Ogasawara, E., Murta, L., Zimbrão, G., and Mattoso, M. (2009). Neural networks cartridges for data mining on time series. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 2302–2309. IEEE.

Organization, W. M. (2018). *July sees extreme weather with high impacts*.

Parmigiani, G. (2001). Decision theory: Bayesian.

Perry, R. H., Green, D. W., and Maloney, J. O. (2015). *Perry's chemical engineers' handbook*. McGraw-Hill New York.

Prati, R. C., Batista, G. E., and Monard, M. C. (2009). Data mining with imbalanced class distributions: concepts and methods. In *Indian International Conference Artificial Intelligence*, pages 359–376.

Press, G. (2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. Forbes Magazine.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13.

Rebollo, J. J. and Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 44:231 – 241.

Records, G. W. (2018a). *Highest recorded temperature*. Guinness World Records.

Records, G. W. (2018b). *Lowest recorded temperature*. Guinness World Records.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, pages 532–538.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717.

Saadat, M. N. and Moniruzzaman, M. (2019). Enhancing airlines delay prediction by implementing classification based deep learning algorithms. In *International Conference on Ubiquitous Information Management and Communication*, pages 886–896. Springer.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Seinfeld, J. H. and Pandis, S. N. (2016). *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons.

Sheng, V. S. and Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In *AAAI*, pages 476–481.

Silberschatz, A., Korth, H. F., and Sudarshan, S. (2016). *Introduction to Data base Management System*. Tata McGraw Hill, New Delhi.

Soares, J. A. (2007). Pré-processamento em mineração de dados: Um estudo comparativo em complementação. *Rio de Janeiro, RJ*.

Sternberg, A., Carvalho, D., Murta, L., Soares, J., and Ogasawara, E. (2016). An analysis of Brazilian flight delays based on frequent patterns. *Transportation Research Part E: Logistics and Transportation Review*, 95:282 – 298.

Thuraisingham, B. (2000). A primer for understanding and applying data mining. *It Professional*, 2(1):28–31.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Torunoğlu, D., Çakirman, E., Ganiz, M. C., Akyokuş, S., and Gürbüz, M. Z. (2011). Analysis of preprocessing methods on classification of Turkish texts. In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on*, pages 112–117. IEEE.

Tsoi, A. C. and Back, A. (1995). Static and dynamic preprocessing methods in neural networks. In *Safety, Reliability and Applications of Emerging Intelligent Control Technologies*, pages 153–160. Elsevier.

Underground, W. (2011). *Record Dew Point Temperatures*. Weather Underground.

Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Xiang-wei, L. and Yian-fang, Q. (2012). A data preprocessing algorithm for classification model based on Rough sets. *Physics Procedia*, 25:2025–2029.

Xiong, J. and Hansen, M. (2013). Modelling airline flight cancellation decisions. *Transportation Research Part E: Logistics and Transportation Review*, 56:64–80.

Xu, Y., Zhang, Z., Lu, G., and Yang, J. (2016). Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification. *Pattern Recognition*, 54:68–82.

A- Detailed Transformations

Table 45 – Data Transformation detailing

Dimension	Original Attribute	Transformed Attribute	Transformed Values	Technique
Metereologic	Depart Temperature	Depart Temperature	1: below 14.2 2: 14.2 to 18.7 3: 18.8 to 22.6 4: 22.7 to 26.4 5: 26.5 to 30.0 6: above 30.0	Discretization Binning (Interval)
Metereologic	Depart Dew Point	Depart Dew Point	1: below 7.8 2: 7.8 to 12.7 3: 12.8 to 16.6 4: 16.7 to 20.5 5: 20.6 to 23.6 6: 23.7 to 27.4 7: above 27.4	Discretization Binning (Interval)
Metereologic	Depart Humidity	Depart Humidity	1: below 26.5 2: 26.5 to 41.6 3: 41.7 to 54.4 4: 54.5 to 67.3 5: 67.4 to 79.4 6: 79.5 to 92.5 6: above 92.5	Discretization Binning (Interval)

Table 45 continued from the previous page

Dimension	Original Attribute	Transformed Attribute	Transformed Values	Technique
Metereologic	Depart Pressure	Depart Pressure	1: below 1006.9 2: 1006.9 to 1010.8 3: 1010.9 to 1014.5 4: 1014.6 to 1018.4 5: 1018.5 to 1022.2 6: 1022.3 to 1026.3 7: above 1026.3	Discretization Binning (Interval)
Metereologic	Depart Visibility	Depart Visibility	1: below 2.3 2: 2.3 to 6.7 3: 6.8 to 10.0 4: 10.1 to 14.9 5: 15.0 to 20.0 6: 21.1 to 37.2 7: 37.3 to 54.1 8: 54.2 to 89.2 9: above 89.2	Discretization Binning (Interval)
Metereologic	Events	Events		Categorical Mapping
Metereologic	Conditions	Conditions		Categorical Mapping
Temporal	Departure Expected	Year Month Day WeekDay Hour	2015 to 2017 1 to 12 1 to 30 1 to 7 00:00 to 23:59	Conceptual Hierarchy

Table 45 continued from the previous page

Dimension	Original Attribute	Transformed Attribute	Transformed Values	Technique
Temporal	Hour	Departure_Hour_bin	1: 00:00 to 01:42 2: 01:43 to 06:18 3: 06:19 to 09:24 4: 09:25 to 13.24 5: 13:25 to 17:36 6: 17:37 to 21:19 7: 21:20 to 23:59	Discretization Binning (Interval)
Spatial	Airline	Airline		Categorical Mapping
Spatial	Arrival	Arrival		Categorical Mapping
Spatial	Departure	Departure		Categorical Mapping

B- Complete Test Results

B.1- Random Forest

B.1.1 Conventional Threshold

Table 46 – Results of Random Forest Workflow Tests - Conventional Threshold

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
0	RF	IMB	WN	0,860	0,039	0,997	0,662	**0,074**	3,149
0	RF	IMB	MM	0,860	0,040	0,997	0,659	**0,075**	3,276
0	RF	IMB	ZS	0,860	0,040	0,997	0,669	**0,076**	3,265
0	RF	SUB	WN	0,645	0,645	0,645	0,232	**0,341**	1,427
0	RF	SUB	MM	0,647	0,644	0,648	0,235	**0,344**	1,385
0	RF	SUB	ZS	0,645	0,644	0,646	0,232	**0,341**	1,427
0	RF	SMOTE	WN	0,214	0,923	0,096	0,145	**0,251**	9,650
0	RF	SMOTE	MM	0,214	0,923	0,096	0,145	**0,251**	9,352
0	RF	SMOTE	ZS	0,214	0,923	0,096	0,145	**0,251**	9,275
1	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	5,481
1	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,731
1	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,635
1	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	1,999
1	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	2,053
1	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	1,957
1	RF	SMOTE	WN	0,264	0,918	0,155	0,153	**0,262**	14,009
1	RF	SMOTE	MM	0,268	0,914	0,161	0,153	**0,263**	13,385
1	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,700

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
2	RF	IMB	WN	0,946	0,691	0,989	0,911	**0,786**	7,317
2	RF	IMB	MM	0,947	0,695	0,989	0,911	**0,789**	7,855
2	RF	IMB	ZS	0,947	0,700	0,989	0,912	**0,792**	7,638
2	RF	SUB	WN	0,908	0,891	0,911	0,624	**0,734**	2,566
2	RF	SUB	MM	0,909	0,893	0,911	0,628	**0,738**	2,620
2	RF	SUB	ZS	0,912	0,896	0,915	0,637	**0,745**	2,505
2	RF	SMOTE	WN	0,294	0,903	0,192	0,157	**0,267**	18,362
2	RF	SMOTE	MM	0,285	0,909	0,181	0,156	**0,266**	17,543
2	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,311
3	RF	IMB	WN	0,929	0,581	0,987	0,878	**0,699**	4,030
3	RF	IMB	MM	0,928	0,576	0,987	0,879	**0,696**	4,495
3	RF	IMB	ZS	0,929	0,584	0,986	0,876	**0,701**	4,029
3	RF	SUB	WN	0,889	0,862	0,894	0,574	**0,689**	1,316
3	RF	SUB	MM	0,889	0,863	0,893	0,575	**0,690**	1,356
3	RF	SUB	ZS	0,892	0,862	0,896	0,580	**0,694**	1,263
3	RF	SMOTE	WN	0,262	0,876	0,160	0,148	**0,253**	9,635
3	RF	SMOTE	MM	0,262	0,876	0,160	0,148	**0,253**	9,163
3	RF	SMOTE	ZS	0,262	0,876	0,160	0,148	**0,253**	9,044
4	RF	IMB	WN	0,942	0,665	0,988	0,900	**0,765**	8,018
4	RF	IMB	MM	0,943	0,674	0,988	0,902	**0,772**	8,252
4	RF	IMB	ZS	0,942	0,674	0,987	0,897	**0,770**	7,981
4	RF	SUB	WN	0,884	0,891	0,883	0,559	**0,687**	2,617
4	RF	SUB	MM	0,884	0,892	0,882	0,561	**0,689**	2,558
4	RF	SUB	ZS	0,887	0,894	0,885	0,564	**0,691**	2,578
4	RF	SMOTE	WN	0,245	0,931	0,131	0,151	**0,260**	18,237
4	RF	SMOTE	MM	0,239	0,937	0,122	0,151	**0,260**	17,220
4	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,293
5	RF	IMB	WN	0,884	0,198	0,998	0,937	**0,327**	13,697
5	RF	IMB	MM	0,885	0,212	0,998	0,935	**0,346**	14,264
5	RF	IMB	ZS	0,884	0,200	0,998	0,930	**0,329**	12,917

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
5	RF	SUB	WN	0,795	0,814	0,791	0,393	**0,530**	3,641
5	RF	SUB	MM	0,789	0,816	0,785	0,389	**0,527**	3,866
5	RF	SUB	ZS	0,792	0,822	0,787	0,391	**0,530**	3,739
5	RF	SMOTE	WN	0,827	0,073	0,952	0,201	**0,107**	19,747
5	RF	SMOTE	MM	0,838	0,068	0,966	0,247	**0,106**	18,532
5	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	18,952
6	RF	IMB	WN	0,927	0,543	0,991	0,908	**0,680**	9,144
6	RF	IMB	MM	0,927	0,541	0,991	0,909	**0,678**	9,232
6	RF	IMB	ZS	0,903	0,346	0,995	0,926	**0,504**	12,564
6	RF	SUB	WN	0,821	0,846	0,817	0,435	**0,574**	2,957
6	RF	SUB	MM	0,816	0,844	0,812	0,430	**0,570**	2,771
6	RF	SUB	ZS	0,815	0,841	0,810	0,424	**0,564**	3,592
6	RF	SMOTE	WN	0,742	0,187	0,835	0,158	**0,171**	15,098
6	RF	SMOTE	MM	0,720	0,219	0,804	0,156	**0,182**	15,631
6	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	15,250
7	RF	IMB	WN	0,931	0,571	0,990	0,907	**0,701**	9,163
7	RF	IMB	MM	0,931	0,576	0,990	0,907	**0,704**	9,265
7	RF	IMB	ZS	0,885	0,211	0,997	0,930	**0,344**	9,898
7	RF	SUB	WN	0,838	0,860	0,834	0,463	**0,602**	2,937
7	RF	SUB	MM	0,845	0,865	0,841	0,478	**0,616**	2,859
7	RF	SUB	ZS	0,811	0,839	0,807	0,419	**0,559**	2,926
7	RF	SMOTE	WN	0,683	0,310	0,745	0,168	**0,218**	19,010
7	RF	SMOTE	MM	0,651	0,369	0,697	0,169	**0,231**	19,099
7	RF	SMOTE	ZS	0,144	0,999	0,002	0,143	**0,250**	15,364
8	RF	IMB	WN	0,905	0,363	0,995	0,923	**0,521**	9,844
8	RF	IMB	MM	0,905	0,367	0,995	0,921	**0,525**	9,598
8	RF	IMB	ZS	0,906	0,376	0,995	0,921	**0,534**	9,916
8	RF	SUB	WN	0,841	0,862	0,838	0,469	**0,608**	2,826
8	RF	SUB	MM	0,871	0,879	0,870	0,532	**0,663**	2,619
8	RF	SUB	ZS	0,883	0,885	0,882	0,555	**0,682**	2,724

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
8	RF	SMOTE	WN	0,275	0,913	0,168	0,154	**0,264**	17,862
8	RF	SMOTE	MM	0,269	0,916	0,161	0,154	**0,263**	17,377
8	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	18,156
9	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	5,874
9	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,461
9	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,652
9	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	2,066
9	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	1,973
9	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	2,002
9	RF	SMOTE	WN	0,264	0,918	0,155	0,153	**0,262**	13,320
9	RF	SMOTE	MM	0,268	0,914	0,161	0,153	**0,263**	13,254
9	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,711
10	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	6,091
10	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,421
10	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,644
10	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	2,006
10	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	2,012
10	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	1,981
10	RF	SMOTE	WN	0,264	0,918	0,155	0,153	**0,262**	13,368
10	RF	SMOTE	MM	0,268	0,914	0,161	0,153	**0,263**	13,748
10	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,768
11	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	6,080
11	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,357
11	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,713
11	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	1,969
11	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	1,984
11	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	2,138
11	RF	SMOTE	WN	0,264	0,918	0,155	0,153	**0,262**	13,306
11	RF	SMOTE	MM	0,268	0,914	0,161	0,153	**0,263**	13,893
11	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,970

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
12	RF	IMB	WN	0,862	0,045	0,998	0,789	**0,086**	4,945
12	RF	IMB	MM	0,862	0,049	0,998	0,778	**0,091**	4,440
12	RF	IMB	ZS	0,862	0,048	0,998	0,765	**0,091**	4,528
12	RF	SUB	WN	0,655	0,704	0,647	0,249	**0,368**	1,864
12	RF	SUB	MM	0,656	0,703	0,648	0,252	**0,371**	1,789
12	RF	SUB	ZS	0,660	0,702	0,653	0,251	**0,370**	2,000
12	RF	SMOTE	WN	0,853	0,081	0,982	0,424	**0,136**	12,382
12	RF	SMOTE	MM	0,854	0,083	0,982	0,433	**0,139**	11,579
12	RF	SMOTE	ZS	0,143	1,000	0,001	0,143	**0,250**	11,692
13	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	6,069
13	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,418
13	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,554
13	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	2,085
13	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	1,961
13	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	2,155
13	RF	SMOTE	WN	0,229	0,946	0,110	0,150	**0,259**	17,005
13	RF	SMOTE	MM	0,224	0,951	0,104	0,150	**0,259**	16,785
13	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	16,319
14	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	5,695
14	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,372
14	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,637
14	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	2,040
14	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	1,994
14	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	2,157
14	RF	SMOTE	WN	0,264	0,918	0,155	0,153	**0,262**	13,645
14	RF	SMOTE	MM	0,268	0,914	0,161	0,153	**0,263**	13,457
14	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,555
15	RF	IMB	WN	0,926	0,518	0,994	0,936	**0,667**	5,620
15	RF	IMB	MM	0,925	0,515	0,994	0,935	**0,664**	5,344
15	RF	IMB	ZS	0,926	0,516	0,994	0,935	**0,665**	5,876

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
15	RF	SUB	WN	0,886	0,876	0,888	0,565	**0,687**	1,949
15	RF	SUB	MM	0,884	0,876	0,886	0,563	**0,685**	1,969
15	RF	SUB	ZS	0,887	0,879	0,888	0,565	**0,688**	2,199
15	RF	SMOTE	WN	0,264	0,918	0,155	0,153	**0,262**	13,606
15	RF	SMOTE	MM	0,268	0,914	0,161	0,153	**0,263**	13,311
15	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,720
16	RF	IMB	WN	0,905	0,363	0,995	0,923	**0,521**	9,764
16	RF	IMB	MM	0,905	0,367	0,995	0,921	**0,525**	9,096
16	RF	IMB	ZS	0,906	0,376	0,995	0,921	**0,534**	9,404
16	RF	SUB	WN	0,841	0,862	0,838	0,469	**0,608**	2,799
16	RF	SUB	MM	0,871	0,879	0,870	0,532	**0,663**	2,624
16	RF	SUB	ZS	0,883	0,885	0,882	0,555	**0,682**	2,941
16	RF	SMOTE	WN	0,302	0,889	0,204	0,157	**0,266**	18,373
16	RF	SMOTE	MM	0,289	0,901	0,188	0,156	**0,265**	18,219
16	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,558
17	RF	IMB	WN	0,860	0,041	0,996	0,659	**0,077**	3,638
17	RF	IMB	MM	0,860	0,043	0,996	0,653	**0,080**	3,770
17	RF	IMB	ZS	0,860	0,045	0,996	0,643	**0,083**	3,778
17	RF	SUB	WN	0,655	0,633	0,659	0,236	**0,344**	1,550
17	RF	SUB	MM	0,655	0,637	0,658	0,238	**0,347**	1,549
17	RF	SUB	ZS	0,658	0,636	0,661	0,237	**0,346**	1,558
17	RF	SMOTE	WN	0,247	0,884	0,141	0,146	**0,251**	11,041
17	RF	SMOTE	MM	0,247	0,884	0,141	0,146	**0,251**	11,255
17	RF	SMOTE	ZS	0,247	0,884	0,141	0,146	**0,251**	11,179
18	RF	IMB	WN	0,907	0,407	0,990	0,875	**0,556**	3,878
18	RF	IMB	MM	0,906	0,398	0,991	0,876	**0,548**	3,964
18	RF	IMB	ZS	0,907	0,409	0,990	0,869	**0,556**	4,179
18	RF	SUB	WN	0,812	0,692	0,832	0,407	**0,513**	1,232
18	RF	SUB	MM	0,813	0,689	0,834	0,412	**0,515**	1,188
18	RF	SUB	ZS	0,815	0,689	0,836	0,411	**0,515**	1,300

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
18	RF	SMOTE	WN	0,208	0,941	0,086	0,146	**0,253**	10,892
18	RF	SMOTE	MM	0,217	0,936	0,097	0,147	**0,254**	10,645
18	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,641
19	RF	IMB	WN	0,886	0,221	0,997	0,924	**0,357**	5,450
19	RF	IMB	MM	0,886	0,222	0,997	0,921	**0,358**	5,237
19	RF	IMB	ZS	0,866	0,067	0,999	0,926	**0,124**	3,975
19	RF	SUB	WN	0,702	0,660	0,709	0,274	**0,387**	1,220
19	RF	SUB	MM	0,689	0,676	0,691	0,269	**0,385**	1,164
19	RF	SUB	ZS	0,690	0,666	0,694	0,265	**0,379**	1,158
19	RF	SMOTE	WN	0,669	0,299	0,731	0,156	**0,205**	10,268
19	RF	SMOTE	MM	0,666	0,282	0,730	0,148	**0,194**	10,050
19	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,112
20	RF	IMB	WN	0,902	0,346	0,995	0,920	**0,502**	5,318
20	RF	IMB	MM	0,866	0,066	0,999	0,949	**0,123**	15,046
20	RF	IMB	ZS	0,866	0,069	0,999	0,937	**0,128**	16,023
20	RF	SUB	WN	0,730	0,729	0,730	0,310	**0,435**	2,599
20	RF	SUB	MM	0,712	0,716	0,711	0,294	**0,417**	2,145
20	RF	SUB	ZS	0,743	0,750	0,742	0,325	**0,454**	2,587
20	RF	SMOTE	WN	0,772	0,076	0,888	0,101	**0,087**	31,175
20	RF	SMOTE	MM	0,760	0,088	0,872	0,103	**0,095**	31,527
20	RF	SMOTE	ZS	0,144	0,997	0,002	0,142	**0,249**	31,830
21	RF	IMB	WN	0,877	0,148	0,999	0,948	**0,255**	5,373
21	RF	IMB	MM	0,878	0,157	0,999	0,947	**0,269**	5,480
21	RF	IMB	ZS	0,914	0,449	0,991	0,896	**0,598**	5,076
21	RF	SUB	WN	0,802	0,742	0,812	0,397	**0,517**	1,568
21	RF	SUB	MM	0,852	0,782	0,864	0,491	**0,604**	1,913
21	RF	SUB	ZS	0,837	0,770	0,848	0,456	**0,573**	2,027
21	RF	SMOTE	WN	0,364	0,760	0,299	0,153	**0,254**	21,857
21	RF	SMOTE	MM	0,365	0,762	0,299	0,153	**0,255**	21,882
21	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	16,544

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
22	RF	IMB	WN	0,923	0,494	0,994	0,933	**0,646**	10,296
22	RF	IMB	MM	0,921	0,480	0,994	0,932	**0,634**	9,430
22	RF	IMB	ZS	0,935	0,597	0,992	0,925	**0,726**	9,777
22	RF	SUB	WN	0,883	0,833	0,891	0,561	**0,670**	2,031
22	RF	SUB	MM	0,898	0,867	0,903	0,601	**0,710**	2,168
22	RF	SUB	ZS	0,876	0,858	0,879	0,541	**0,663**	2,694
22	RF	SMOTE	WN	0,463	0,680	0,427	0,165	**0,265**	24,156
22	RF	SMOTE	MM	0,457	0,684	0,420	0,164	**0,264**	24,444
22	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	21,719
23	RF	IMB	WN	0,907	0,407	0,990	0,875	**0,556**	3,953
23	RF	IMB	MM	0,906	0,398	0,991	0,876	**0,548**	3,988
23	RF	IMB	ZS	0,907	0,409	0,990	0,869	**0,556**	3,904
23	RF	SUB	WN	0,812	0,692	0,832	0,407	**0,513**	1,190
23	RF	SUB	MM	0,813	0,689	0,834	0,412	**0,515**	1,304
23	RF	SUB	ZS	0,815	0,689	0,836	0,411	**0,515**	1,210
23	RF	SMOTE	WN	0,208	0,941	0,086	0,146	**0,253**	10,853
23	RF	SMOTE	MM	0,217	0,936	0,097	0,147	**0,254**	10,817
23	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,556
24	RF	IMB	WN	0,925	0,510	0,994	0,930	**0,659**	10,732
24	RF	IMB	MM	0,925	0,515	0,994	0,931	**0,663**	9,336
24	RF	IMB	ZS	0,934	0,589	0,992	0,924	**0,719**	9,790
24	RF	SUB	WN	0,865	0,827	0,871	0,517	**0,636**	2,110
24	RF	SUB	MM	0,900	0,870	0,905	0,607	**0,715**	2,167
24	RF	SUB	ZS	0,844	0,825	0,847	0,473	**0,601**	2,581
24	RF	SMOTE	WN	0,471	0,661	0,440	0,164	**0,263**	25,640
24	RF	SMOTE	MM	0,469	0,669	0,436	0,165	**0,264**	25,116
24	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	23,259
25	RF	IMB	WN	0,907	0,407	0,990	0,875	**0,556**	3,947
25	RF	IMB	MM	0,906	0,398	0,991	0,876	**0,548**	3,965
25	RF	IMB	ZS	0,907	0,409	0,990	0,869	**0,556**	4,131

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
25	RF	SUB	WN	0,812	0,692	0,832	0,407	**0,513**	1,252
25	RF	SUB	MM	0,813	0,689	0,834	0,412	**0,515**	1,267
25	RF	SUB	ZS	0,815	0,689	0,836	0,411	**0,515**	1,291
25	RF	SMOTE	WN	0,208	0,941	0,086	0,146	**0,253**	10,706
25	RF	SMOTE	MM	0,217	0,936	0,097	0,147	**0,254**	10,774
25	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,626
26	RF	IMB	WN	0,916	0,450	0,994	0,923	**0,605**	11,816
26	RF	IMB	MM	0,916	0,451	0,994	0,926	**0,607**	10,639
26	RF	IMB	ZS	0,929	0,554	0,991	0,914	**0,690**	11,517
26	RF	SUB	WN	0,851	0,827	0,855	0,487	**0,613**	2,332
26	RF	SUB	MM	0,865	0,857	0,866	0,518	**0,645**	2,578
26	RF	SUB	ZS	0,850	0,850	0,850	0,485	**0,618**	3,010
26	RF	SMOTE	WN	0,460	0,690	0,422	0,166	**0,267**	26,069
26	RF	SMOTE	MM	0,468	0,681	0,432	0,166	**0,267**	25,927
26	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	24,651
27	RF	IMB	WN	0,923	0,494	0,994	0,933	**0,646**	10,415
27	RF	IMB	MM	0,921	0,480	0,994	0,932	**0,634**	9,009
27	RF	IMB	ZS	0,935	0,597	0,992	0,925	**0,726**	10,324
27	RF	SUB	WN	0,883	0,833	0,891	0,561	**0,670**	2,081
27	RF	SUB	MM	0,898	0,867	0,903	0,601	**0,710**	2,260
27	RF	SUB	ZS	0,876	0,858	0,879	0,541	**0,663**	2,507
27	RF	SMOTE	WN	0,463	0,680	0,427	0,165	**0,265**	24,276
27	RF	SMOTE	MM	0,457	0,684	0,420	0,164	**0,264**	23,913
27	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	21,715
28	RF	IMB	WN	0,923	0,494	0,994	0,933	**0,646**	10,508
28	RF	IMB	MM	0,921	0,480	0,994	0,932	**0,634**	9,401
28	RF	IMB	ZS	0,935	0,597	0,992	0,925	**0,726**	10,274
28	RF	SUB	WN	0,883	0,833	0,891	0,561	**0,670**	1,963
28	RF	SUB	MM	0,898	0,867	0,903	0,601	**0,710**	2,174
28	RF	SUB	ZS	0,876	0,858	0,879	0,541	**0,663**	2,410

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
28	RF	SMOTE	WN	0,463	0,680	0,427	0,165	**0,265**	24,222
28	RF	SMOTE	MM	0,457	0,684	0,420	0,164	**0,264**	23,843
28	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	21,504
29	RF	IMB	WN	0,920	0,479	0,993	0,924	**0,631**	14,362
29	RF	IMB	MM	0,928	0,545	0,992	0,920	**0,684**	13,638
29	RF	IMB	ZS	0,909	0,398	0,995	0,924	**0,556**	16,597
29	RF	SUB	WN	0,899	0,880	0,902	0,600	**0,713**	2,366
29	RF	SUB	MM	0,895	0,875	0,898	0,591	**0,706**	2,581
29	RF	SUB	ZS	0,899	0,882	0,902	0,598	**0,713**	3,103
29	RF	SMOTE	WN	0,496	0,621	0,475	0,164	**0,260**	36,602
29	RF	SMOTE	MM	0,494	0,620	0,474	0,164	**0,259**	37,142
29	RF	SMOTE	ZS	0,143	1,000	0,001	0,143	**0,250**	38,899
30	RF	IMB	WN	0,922	0,512	0,991	0,900	**0,653**	4,860
30	RF	IMB	MM	0,923	0,516	0,991	0,901	**0,656**	8,000
30	RF	IMB	ZS	0,922	0,511	0,990	0,899	**0,652**	4,843
30	RF	SUB	WN	0,883	0,852	0,889	0,560	**0,676**	1,482
30	RF	SUB	MM	0,885	0,854	0,890	0,566	**0,681**	1,414
30	RF	SUB	ZS	0,887	0,857	0,892	0,569	**0,684**	1,438
30	RF	SMOTE	WN	0,193	0,969	0,064	0,147	**0,255**	12,165
30	RF	SMOTE	MM	0,193	0,969	0,064	0,147	**0,255**	9,988
30	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,004
31	RF	IMB	WN	0,880	0,178	0,997	0,911	**0,298**	10,364
31	RF	IMB	MM	0,879	0,168	0,998	0,919	**0,283**	11,394
31	RF	IMB	ZS	0,892	0,268	0,996	0,913	**0,414**	7,771
31	RF	SUB	WN	0,770	0,685	0,784	0,346	**0,459**	1,102
31	RF	SUB	MM	0,785	0,730	0,794	0,373	**0,494**	1,281
31	RF	SUB	ZS	0,839	0,810	0,843	0,462	**0,588**	1,868
31	RF	SMOTE	WN	0,322	0,856	0,233	0,156	**0,265**	20,066
31	RF	SMOTE	MM	0,322	0,856	0,233	0,156	**0,265**	19,989
31	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,739

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
32	RF	IMB	WN	0,909	0,441	0,987	0,851	**0,581**	9,968
32	RF	IMB	MM	0,910	0,447	0,987	0,850	**0,586**	16,000
32	RF	IMB	ZS	0,909	0,445	0,986	0,844	**0,583**	10,130
32	RF	SUB	WN	0,811	0,827	0,808	0,417	**0,555**	2,947
32	RF	SUB	MM	0,811	0,830	0,808	0,421	**0,559**	2,934
32	RF	SUB	ZS	0,812	0,833	0,808	0,419	**0,557**	2,987
32	RF	SMOTE	WN	0,731	0,312	0,800	0,206	**0,248**	22,567
32	RF	SMOTE	MM	0,499	0,587	0,485	0,159	**0,251**	22,879
32	RF	SMOTE	ZS	0,144	0,999	0,001	0,143	**0,250**	22,875
33	RF	IMB	WN	0,891	0,274	0,994	0,879	**0,418**	13,230
33	RF	IMB	MM	0,886	0,237	0,995	0,883	**0,374**	14,212
33	RF	IMB	ZS	0,898	0,337	0,992	0,870	**0,486**	12,982
33	RF	SUB	WN	0,750	0,747	0,751	0,333	**0,460**	2,438
33	RF	SUB	MM	0,792	0,784	0,794	0,390	**0,521**	2,574
33	RF	SUB	ZS	0,788	0,771	0,790	0,379	**0,508**	2,848
33	RF	SMOTE	WN	0,777	0,209	0,871	0,212	**0,211**	39,454
33	RF	SMOTE	MM	0,689	0,325	0,749	0,177	**0,229**	32,771
33	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	29,051
34	RF	IMB	WN	0,909	0,430	0,989	0,869	**0,575**	8,767
34	RF	IMB	MM	0,910	0,441	0,989	0,866	**0,584**	8,998
34	RF	IMB	ZS	0,910	0,442	0,988	0,864	**0,585**	9,055
34	RF	SUB	WN	0,828	0,826	0,829	0,445	**0,578**	2,721
34	RF	SUB	MM	0,827	0,828	0,827	0,445	**0,579**	2,705
34	RF	SUB	ZS	0,829	0,828	0,829	0,445	**0,579**	2,693
34	RF	SMOTE	WN	0,757	0,233	0,844	0,199	**0,215**	22,765
34	RF	SMOTE	MM	0,754	0,238	0,840	0,198	**0,216**	15,326
34	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	15,273
35	RF	IMB	WN	0,888	0,247	0,995	0,892	**0,387**	12,017
35	RF	IMB	MM	0,883	0,204	0,996	0,894	**0,332**	10,569
35	RF	IMB	ZS	0,897	0,319	0,993	0,883	**0,469**	11,845

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
35	RF	SUB	WN	0,752	0,735	0,755	0,333	**0,458**	2,070
35	RF	SUB	MM	0,794	0,778	0,796	0,391	**0,521**	2,341
35	RF	SUB	ZS	0,783	0,763	0,786	0,372	**0,500**	2,589
35	RF	SMOTE	WN	0,805	0,131	0,917	0,208	**0,160**	26,593
35	RF	SMOTE	MM	0,803	0,137	0,914	0,209	**0,166**	26,144
35	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	24,320
36	RF	IMB	WN	0,920	0,485	0,992	0,911	**0,633**	10,833
36	RF	IMB	MM	0,921	0,496	0,992	0,911	**0,642**	11,063
36	RF	IMB	ZS	0,922	0,504	0,991	0,907	**0,648**	11,118
36	RF	SUB	WN	0,838	0,862	0,834	0,463	**0,602**	3,317
36	RF	SUB	MM	0,837	0,864	0,833	0,465	**0,604**	3,313
36	RF	SUB	ZS	0,838	0,868	0,833	0,463	**0,604**	3,270
36	RF	SMOTE	WN	0,368	0,814	0,294	0,161	**0,269**	26,601
36	RF	SMOTE	MM	0,280	0,889	0,179	0,153	**0,260**	26,331
36	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	26,267
37	RF	IMB	WN	0,911	0,409	0,994	0,920	**0,566**	16,077
37	RF	IMB	MM	0,909	0,395	0,995	0,923	**0,553**	14,176
37	RF	IMB	ZS	0,917	0,461	0,993	0,913	**0,612**	15,550
37	RF	SUB	WN	0,824	0,814	0,826	0,437	**0,569**	2,994
37	RF	SUB	MM	0,858	0,851	0,859	0,504	**0,633**	3,462
37	RF	SUB	ZS	0,853	0,853	0,853	0,490	**0,622**	3,879
37	RF	SMOTE	WN	0,491	0,670	0,461	0,171	**0,273**	39,522
37	RF	SMOTE	MM	0,448	0,716	0,403	0,166	**0,270**	39,029
37	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	34,956
38	RF	IMB	WN	0,920	0,504	0,989	0,887	**0,643**	9,670
38	RF	IMB	MM	0,921	0,512	0,989	0,883	**0,648**	9,493
38	RF	IMB	ZS	0,920	0,508	0,988	0,878	**0,644**	9,570
38	RF	SUB	WN	0,854	0,838	0,857	0,493	**0,621**	2,759
38	RF	SUB	MM	0,854	0,845	0,856	0,497	**0,626**	2,750
38	RF	SUB	ZS	0,855	0,845	0,856	0,494	**0,623**	2,762

Table 46 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
38	RF	SMOTE	WN	0,310	0,874	0,216	0,156	**0,265**	16,664
38	RF	SMOTE	MM	0,313	0,871	0,220	0,157	**0,265**	16,842
38	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	16,685
39	RF	IMB	WN	0,905	0,371	0,994	0,907	**0,527**	13,286
39	RF	IMB	MM	0,900	0,331	0,994	0,907	**0,485**	11,705
39	RF	IMB	ZS	0,913	0,441	0,992	0,898	**0,591**	12,988
39	RF	SUB	WN	0,796	0,757	0,802	0,389	**0,514**	2,260
39	RF	SUB	MM	0,832	0,802	0,837	0,453	**0,579**	2,522
39	RF	SUB	ZS	0,855	0,830	0,859	0,494	**0,620**	3,151
39	RF	SMOTE	WN	0,475	0,635	0,449	0,161	**0,257**	28,555
39	RF	SMOTE	MM	0,486	0,625	0,462	0,162	**0,257**	28,769
39	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	26,085
40	RF	IMB	WN	0,860	0,124	0,982	0,532	**0,202**	8,984
40	RF	IMB	MM	0,859	0,129	0,981	0,531	**0,208**	8,767
40	RF	IMB	ZS	0,859	0,128	0,981	0,524	**0,206**	8,658
40	RF	SUB	WN	0,635	0,673	0,628	0,231	**0,344**	3,189
40	RF	SUB	MM	0,634	0,678	0,627	0,234	**0,348**	3,188
40	RF	SUB	ZS	0,634	0,675	0,627	0,231	**0,344**	3,200
40	RF	SMOTE	WN	0,630	0,347	0,677	0,152	**0,211**	23,395
40	RF	SMOTE	MM	0,522	0,480	0,529	0,145	**0,223**	23,145
40	RF	SMOTE	ZS	0,144	0,998	0,002	0,143	**0,250**	23,009

B.1.2 Majority Threshold

Table 47 – Results of Random Forest Workflow Tests - Majority Threshold

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
0	RF	IMB	WN	0,841	0,203	0,948	0,392	**0,267**	3,149

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
0	RF	IMB	MM	0,840	0,211	0,945	0,389	**0,274**	3,276
0	RF	IMB	ZS	0,839	0,212	0,943	0,383	**0,273**	3,265
0	RF	SUB	WN	0,353	0,913	0,260	0,170	**0,287**	1,427
0	RF	SUB	MM	0,358	0,913	0,265	0,173	**0,291**	1,385
0	RF	SUB	ZS	0,355	0,915	0,262	0,171	**0,288**	1,427
0	RF	SMOTE	WN	0,171	0,974	0,037	0,144	**0,251**	9,650
0	RF	SMOTE	MM	0,171	0,974	0,037	0,144	**0,251**	9,352
0	RF	SMOTE	ZS	0,171	0,974	0,037	0,144	**0,251**	9,275
1	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,481
1	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,731
1	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,635
1	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	1,999
1	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	2,053
1	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	1,957
1	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	14,009
1	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	13,385
1	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,700
2	RF	IMB	WN	0,922	0,929	0,920	0,660	**0,772**	7,317
2	RF	IMB	MM	0,920	0,932	0,918	0,654	**0,769**	7,855
2	RF	IMB	ZS	0,920	0,933	0,918	0,654	**0,769**	7,638
2	RF	SUB	WN	0,506	0,990	0,426	0,223	**0,364**	2,566
2	RF	SUB	MM	0,507	0,990	0,426	0,225	**0,366**	2,620
2	RF	SUB	ZS	0,518	0,989	0,440	0,227	**0,369**	2,505
2	RF	SMOTE	WN	0,143	1,000	0,001	0,143	**0,250**	18,362
2	RF	SMOTE	MM	0,143	1,000	0,001	0,143	**0,250**	17,543
2	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,311
3	RF	IMB	WN	0,937	0,824	0,956	0,757	**0,789**	4,030
3	RF	IMB	MM	0,937	0,826	0,956	0,756	**0,790**	4,495
3	RF	IMB	ZS	0,937	0,825	0,955	0,754	**0,788**	4,029
3	RF	SUB	WN	0,682	0,961	0,636	0,305	**0,463**	1,316

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
3	RF	SUB	MM	0,682	0,960	0,635	0,307	**0,465**	1,356
3	RF	SUB	ZS	0,691	0,959	0,647	0,310	**0,469**	1,263
3	RF	SMOTE	WN	0,178	0,964	0,047	0,144	**0,251**	9,635
3	RF	SMOTE	MM	0,178	0,964	0,047	0,144	**0,251**	9,163
3	RF	SMOTE	ZS	0,178	0,964	0,047	0,144	**0,251**	9,044
4	RF	IMB	WN	0,902	0,926	0,897	0,600	**0,728**	8,018
4	RF	IMB	MM	0,900	0,928	0,895	0,596	**0,726**	8,252
4	RF	IMB	ZS	0,900	0,928	0,895	0,597	**0,726**	7,981
4	RF	SUB	WN	0,496	0,989	0,414	0,219	**0,359**	2,617
4	RF	SUB	MM	0,498	0,990	0,415	0,221	**0,362**	2,558
4	RF	SUB	ZS	0,504	0,989	0,424	0,222	**0,362**	2,578
4	RF	SMOTE	WN	0,143	1,000	0,001	0,143	**0,250**	18,237
4	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	17,220
4	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,293
5	RF	IMB	WN	0,910	0,773	0,932	0,655	**0,709**	13,697
5	RF	IMB	MM	0,910	0,786	0,930	0,653	**0,713**	14,264
5	RF	IMB	ZS	0,908	0,779	0,929	0,648	**0,707**	12,917
5	RF	SUB	WN	0,348	0,991	0,241	0,178	**0,302**	3,641
5	RF	SUB	MM	0,344	0,991	0,235	0,179	**0,303**	3,866
5	RF	SUB	ZS	0,346	0,991	0,239	0,178	**0,301**	3,739
5	RF	SMOTE	WN	0,391	0,767	0,329	0,160	**0,264**	19,747
5	RF	SMOTE	MM	0,422	0,733	0,371	0,162	**0,266**	18,532
5	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	18,952
6	RF	IMB	WN	0,887	0,910	0,883	0,564	**0,697**	9,144
6	RF	IMB	MM	0,883	0,911	0,878	0,554	**0,689**	9,232
6	RF	IMB	ZS	0,898	0,857	0,905	0,601	**0,707**	12,564
6	RF	SUB	WN	0,358	0,992	0,253	0,181	**0,306**	2,957
6	RF	SUB	MM	0,357	0,993	0,250	0,182	**0,308**	2,771
6	RF	SUB	ZS	0,362	0,992	0,258	0,181	**0,307**	3,592
6	RF	SMOTE	WN	0,269	0,901	0,164	0,152	**0,260**	15,098

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
6	RF	SMOTE	MM	0,258	0,913	0,149	0,151	**0,260**	15,631
6	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	15,250
7	RF	IMB	WN	0,891	0,914	0,888	0,575	**0,706**	9,163
7	RF	IMB	MM	0,888	0,916	0,883	0,566	**0,700**	9,265
7	RF	IMB	ZS	0,904	0,790	0,923	0,632	**0,702**	9,898
7	RF	SUB	WN	0,389	0,992	0,288	0,188	**0,316**	2,937
7	RF	SUB	MM	0,399	0,992	0,299	0,192	**0,322**	2,859
7	RF	SUB	ZS	0,349	0,992	0,243	0,179	**0,303**	2,926
7	RF	SMOTE	WN	0,206	0,962	0,080	0,148	**0,257**	19,010
7	RF	SMOTE	MM	0,204	0,961	0,078	0,148	**0,256**	19,099
7	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	15,364
8	RF	IMB	WN	0,894	0,866	0,898	0,586	**0,699**	9,844
8	RF	IMB	MM	0,888	0,869	0,892	0,572	**0,690**	9,598
8	RF	IMB	ZS	0,892	0,868	0,895	0,581	**0,696**	9,916
8	RF	SUB	WN	0,394	0,992	0,295	0,190	**0,318**	2,826
8	RF	SUB	MM	0,445	0,991	0,353	0,205	**0,339**	2,619
8	RF	SUB	ZS	0,470	0,990	0,383	0,210	**0,347**	2,724
8	RF	SMOTE	WN	0,143	1,000	0,001	0,143	**0,250**	17,862
8	RF	SMOTE	MM	0,143	1,000	0,001	0,143	**0,250**	17,377
8	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	18,156
9	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,874
9	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,461
9	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,652
9	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	2,066
9	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	1,973
9	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	2,002
9	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	13,320
9	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	13,254
9	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,711
10	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	6,091

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
10	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,421
10	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,644
10	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	2,006
10	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	2,012
10	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	1,981
10	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	13,368
10	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	13,748
10	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,768
11	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	6,080
11	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,357
11	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,713
11	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	1,969
11	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	1,984
11	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	2,138
11	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	13,306
11	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	13,893
11	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,970
12	RF	IMB	WN	0,851	0,275	0,947	0,463	**0,345**	4,945
12	RF	IMB	MM	0,850	0,295	0,942	0,461	**0,359**	4,440
12	RF	IMB	ZS	0,847	0,290	0,940	0,446	**0,352**	4,528
12	RF	SUB	WN	0,254	0,975	0,135	0,158	**0,272**	1,864
12	RF	SUB	MM	0,259	0,974	0,138	0,160	**0,274**	1,789
12	RF	SUB	ZS	0,257	0,976	0,138	0,158	**0,272**	2,000
12	RF	SMOTE	WN	0,729	0,466	0,773	0,254	**0,329**	12,382
12	RF	SMOTE	MM	0,727	0,467	0,770	0,252	**0,327**	11,579
12	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	11,692
13	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	6,069
13	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,418
13	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,554
13	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	2,085

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
13	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	1,961
13	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	2,155
13	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	17,005
13	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	16,785
13	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	16,319
14	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,695
14	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,372
14	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,637
14	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	2,040
14	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	1,994
14	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	2,157
14	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	13,645
14	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	13,457
14	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,555
15	RF	IMB	WN	0,934	0,895	0,941	0,716	**0,795**	5,620
15	RF	IMB	MM	0,931	0,896	0,937	0,705	**0,789**	5,344
15	RF	IMB	ZS	0,930	0,899	0,935	0,698	**0,786**	5,876
15	RF	SUB	WN	0,462	0,990	0,374	0,208	**0,344**	1,949
15	RF	SUB	MM	0,463	0,990	0,374	0,210	**0,347**	1,969
15	RF	SUB	ZS	0,469	0,990	0,383	0,210	**0,347**	2,199
15	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	13,606
15	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	13,311
15	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	12,720
16	RF	IMB	WN	0,894	0,866	0,898	0,586	**0,699**	9,764
16	RF	IMB	MM	0,888	0,869	0,892	0,572	**0,690**	9,096
16	RF	IMB	ZS	0,892	0,868	0,895	0,581	**0,696**	9,404
16	RF	SUB	WN	0,394	0,992	0,295	0,190	**0,318**	2,799
16	RF	SUB	MM	0,445	0,991	0,353	0,205	**0,339**	2,624
16	RF	SUB	ZS	0,470	0,990	0,383	0,210	**0,347**	2,941
16	RF	SMOTE	WN	0,143	1,000	0,001	0,143	**0,250**	18,373

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
16	RF	SMOTE	MM	0,143	1,000	0,001	0,143	**0,250**	18,219
16	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,558
17	RF	IMB	WN	0,842	0,207	0,948	0,397	**0,272**	3,638
17	RF	IMB	MM	0,840	0,215	0,944	0,391	**0,278**	3,770
17	RF	IMB	ZS	0,839	0,218	0,943	0,390	**0,279**	3,778
17	RF	SUB	WN	0,332	0,924	0,234	0,167	**0,283**	1,550
17	RF	SUB	MM	0,334	0,923	0,235	0,169	**0,285**	1,549
17	RF	SUB	ZS	0,336	0,923	0,239	0,167	**0,284**	1,558
17	RF	SMOTE	WN	0,171	0,976	0,037	0,144	**0,251**	11,041
17	RF	SMOTE	MM	0,171	0,976	0,037	0,144	**0,251**	11,255
17	RF	SMOTE	ZS	0,171	0,976	0,037	0,144	**0,251**	11,179
18	RF	IMB	WN	0,908	0,508	0,974	0,765	**0,610**	3,878
18	RF	IMB	MM	0,906	0,507	0,973	0,757	**0,607**	3,964
18	RF	IMB	ZS	0,906	0,507	0,972	0,754	**0,607**	4,179
18	RF	SUB	WN	0,475	0,928	0,399	0,204	**0,335**	1,232
18	RF	SUB	MM	0,475	0,928	0,399	0,206	**0,337**	1,188
18	RF	SUB	ZS	0,473	0,929	0,397	0,204	**0,334**	1,300
18	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	10,892
18	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	10,645
18	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,641
19	RF	IMB	WN	0,915	0,513	0,982	0,822	**0,631**	5,450
19	RF	IMB	MM	0,914	0,512	0,981	0,818	**0,630**	5,237
19	RF	IMB	ZS	0,902	0,377	0,989	0,854	**0,523**	3,975
19	RF	SUB	WN	0,441	0,909	0,363	0,192	**0,317**	1,220
19	RF	SUB	MM	0,422	0,916	0,339	0,189	**0,313**	1,164
19	RF	SUB	ZS	0,435	0,913	0,355	0,190	**0,315**	1,158
19	RF	SMOTE	WN	0,345	0,797	0,270	0,154	**0,258**	10,268
19	RF	SMOTE	MM	0,351	0,780	0,280	0,153	**0,255**	10,050
19	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,112
20	RF	IMB	WN	0,916	0,665	0,958	0,724	**0,693**	5,318

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
20	RF	IMB	MM	0,903	0,398	0,988	0,845	**0,541**	15,046
20	RF	IMB	ZS	0,904	0,413	0,986	0,834	**0,553**	16,023
20	RF	SUB	WN	0,410	0,956	0,319	0,189	**0,316**	2,599
20	RF	SUB	MM	0,408	0,948	0,317	0,189	**0,316**	2,145
20	RF	SUB	ZS	0,419	0,959	0,329	0,192	**0,320**	2,587
20	RF	SMOTE	WN	0,589	0,362	0,627	0,139	**0,201**	31,175
20	RF	SMOTE	MM	0,591	0,362	0,629	0,139	**0,201**	31,527
20	RF	SMOTE	ZS	0,143	0,999	0,001	0,143	**0,249**	31,830
21	RF	IMB	WN	0,914	0,531	0,978	0,797	**0,637**	5,373
21	RF	IMB	MM	0,914	0,544	0,976	0,792	**0,645**	5,480
21	RF	IMB	ZS	0,912	0,728	0,942	0,678	**0,702**	5,076
21	RF	SUB	WN	0,414	0,964	0,323	0,191	**0,319**	1,568
21	RF	SUB	MM	0,478	0,968	0,395	0,212	**0,348**	1,913
21	RF	SUB	ZS	0,473	0,962	0,391	0,208	**0,342**	2,027
21	RF	SMOTE	WN	0,150	0,996	0,010	0,143	**0,250**	21,857
21	RF	SMOTE	MM	0,153	0,995	0,013	0,143	**0,251**	21,882
21	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	16,544
22	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,296
22	RF	IMB	MM	0,938	0,864	0,951	0,745	**0,800**	9,430
22	RF	IMB	ZS	0,938	0,896	0,945	0,729	**0,804**	9,777
22	RF	SUB	WN	0,525	0,978	0,450	0,228	**0,370**	2,031
22	RF	SUB	MM	0,528	0,986	0,451	0,232	**0,375**	2,168
22	RF	SUB	ZS	0,501	0,985	0,421	0,220	**0,360**	2,694
22	RF	SMOTE	WN	0,154	0,997	0,014	0,144	**0,252**	24,156
22	RF	SMOTE	MM	0,155	0,997	0,015	0,144	**0,252**	24,444
22	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	21,719
23	RF	IMB	WN	0,908	0,508	0,974	0,765	**0,610**	3,953
23	RF	IMB	MM	0,906	0,507	0,973	0,757	**0,607**	3,988
23	RF	IMB	ZS	0,906	0,507	0,972	0,754	**0,607**	3,904
23	RF	SUB	WN	0,475	0,928	0,399	0,204	**0,335**	1,190

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
23	RF	SUB	MM	0,475	0,928	0,399	0,206	**0,337**	1,304
23	RF	SUB	ZS	0,473	0,929	0,397	0,204	**0,334**	1,210
23	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	10,853
23	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	10,817
23	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,556
24	RF	IMB	WN	0,939	0,881	0,949	0,740	**0,805**	10,732
24	RF	IMB	MM	0,937	0,879	0,947	0,734	**0,800**	9,336
24	RF	IMB	ZS	0,937	0,898	0,943	0,724	**0,802**	9,790
24	RF	SUB	WN	0,462	0,982	0,375	0,207	**0,342**	2,110
24	RF	SUB	MM	0,526	0,986	0,448	0,231	**0,375**	2,167
24	RF	SUB	ZS	0,450	0,984	0,361	0,204	**0,338**	2,581
24	RF	SMOTE	WN	0,155	0,996	0,015	0,144	**0,252**	25,640
24	RF	SMOTE	MM	0,158	0,995	0,018	0,144	**0,252**	25,116
24	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	23,259
25	RF	IMB	WN	0,908	0,508	0,974	0,765	**0,610**	3,947
25	RF	IMB	MM	0,906	0,507	0,973	0,757	**0,607**	3,965
25	RF	IMB	ZS	0,906	0,507	0,972	0,754	**0,607**	4,131
25	RF	SUB	WN	0,475	0,928	0,399	0,204	**0,335**	1,252
25	RF	SUB	MM	0,475	0,928	0,399	0,206	**0,337**	1,267
25	RF	SUB	ZS	0,473	0,929	0,397	0,204	**0,334**	1,291
25	RF	SMOTE	WN	0,143	1,000	0,000	0,143	**0,250**	10,706
25	RF	SMOTE	MM	0,143	1,000	0,000	0,143	**0,250**	10,774
25	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,626
26	RF	IMB	WN	0,914	0,878	0,920	0,645	**0,744**	11,816
26	RF	IMB	MM	0,911	0,872	0,918	0,639	**0,737**	10,639
26	RF	IMB	ZS	0,910	0,901	0,912	0,630	**0,742**	11,517
26	RF	SUB	WN	0,418	0,986	0,323	0,195	**0,326**	2,332
26	RF	SUB	MM	0,434	0,990	0,340	0,201	**0,335**	2,578
26	RF	SUB	ZS	0,427	0,990	0,334	0,198	**0,330**	3,010
26	RF	SMOTE	WN	0,154	0,997	0,014	0,144	**0,252**	26,069

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
26	RF	SMOTE	MM	0,156	0,997	0,016	0,144	**0,252**	25,927
26	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	24,651
27	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,415
27	RF	IMB	MM	0,938	0,864	0,951	0,745	**0,800**	9,009
27	RF	IMB	ZS	0,938	0,896	0,945	0,729	**0,804**	10,324
27	RF	SUB	WN	0,525	0,978	0,450	0,228	**0,370**	2,081
27	RF	SUB	MM	0,528	0,986	0,451	0,232	**0,375**	2,260
27	RF	SUB	ZS	0,501	0,985	0,421	0,220	**0,360**	2,507
27	RF	SMOTE	WN	0,154	0,997	0,014	0,144	**0,252**	24,276
27	RF	SMOTE	MM	0,155	0,997	0,015	0,144	**0,252**	23,913
27	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	21,715
28	RF	IMB	WN	0,940	0,875	0,950	0,746	**0,806**	10,508
28	RF	IMB	MM	0,938	0,864	0,951	0,745	**0,800**	9,401
28	RF	IMB	ZS	0,938	0,896	0,945	0,729	**0,804**	10,274
28	RF	SUB	WN	0,525	0,978	0,450	0,228	**0,370**	1,963
28	RF	SUB	MM	0,528	0,986	0,451	0,232	**0,375**	2,174
28	RF	SUB	ZS	0,501	0,985	0,421	0,220	**0,360**	2,410
28	RF	SMOTE	WN	0,154	0,997	0,014	0,144	**0,252**	24,222
28	RF	SMOTE	MM	0,155	0,997	0,015	0,144	**0,252**	23,843
28	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	21,504
29	RF	IMB	WN	0,937	0,868	0,948	0,736	**0,797**	14,362
29	RF	IMB	MM	0,937	0,883	0,946	0,733	**0,801**	13,638
29	RF	IMB	ZS	0,936	0,848	0,950	0,739	**0,790**	16,597
29	RF	SUB	WN	0,573	0,983	0,505	0,248	**0,396**	2,366
29	RF	SUB	MM	0,560	0,983	0,489	0,244	**0,392**	2,581
29	RF	SUB	ZS	0,591	0,981	0,527	0,256	**0,406**	3,103
29	RF	SMOTE	WN	0,190	0,979	0,059	0,147	**0,256**	36,602
29	RF	SMOTE	MM	0,189	0,979	0,057	0,147	**0,256**	37,142
29	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	38,899
30	RF	IMB	WN	0,940	0,814	0,960	0,774	**0,793**	4,860

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
30	RF	IMB	MM	0,940	0,818	0,961	0,776	**0,797**	8,000
30	RF	IMB	ZS	0,940	0,815	0,961	0,776	**0,795**	4,843
30	RF	SUB	WN	0,662	0,963	0,612	0,292	**0,448**	1,482
30	RF	SUB	MM	0,656	0,963	0,604	0,290	**0,446**	1,414
30	RF	SUB	ZS	0,657	0,964	0,606	0,289	**0,445**	1,438
30	RF	SMOTE	WN	0,160	0,992	0,022	0,144	**0,252**	12,165
30	RF	SMOTE	MM	0,160	0,992	0,022	0,144	**0,252**	9,988
30	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	10,004
31	RF	IMB	WN	0,920	0,596	0,974	0,795	**0,681**	10,364
31	RF	IMB	MM	0,914	0,516	0,980	0,812	**0,631**	11,394
31	RF	IMB	ZS	0,927	0,660	0,971	0,793	**0,720**	7,771
31	RF	SUB	WN	0,521	0,893	0,459	0,215	**0,347**	1,102
31	RF	SUB	MM	0,489	0,946	0,412	0,213	**0,348**	1,281
31	RF	SUB	ZS	0,579	0,960	0,515	0,247	**0,393**	1,868
31	RF	SMOTE	WN	0,191	0,973	0,061	0,147	**0,255**	20,066
31	RF	SMOTE	MM	0,191	0,973	0,061	0,147	**0,255**	19,989
31	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	17,739
32	RF	IMB	WN	0,834	0,882	0,826	0,457	**0,602**	9,968
32	RF	IMB	MM	0,831	0,886	0,821	0,453	**0,600**	16,000
32	RF	IMB	ZS	0,830	0,885	0,821	0,451	**0,598**	10,130
32	RF	SUB	WN	0,379	0,988	0,278	0,185	**0,312**	2,947
32	RF	SUB	MM	0,382	0,990	0,279	0,188	**0,315**	2,934
32	RF	SUB	ZS	0,381	0,990	0,280	0,186	**0,313**	2,987
32	RF	SMOTE	WN	0,291	0,879	0,193	0,153	**0,261**	22,567
32	RF	SMOTE	MM	0,173	0,977	0,039	0,145	**0,252**	22,879
32	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	22,875
33	RF	IMB	WN	0,878	0,766	0,897	0,552	**0,642**	13,230
33	RF	IMB	MM	0,870	0,752	0,890	0,532	**0,623**	14,212
33	RF	IMB	ZS	0,867	0,811	0,876	0,522	**0,635**	12,982
33	RF	SUB	WN	0,316	0,985	0,205	0,171	**0,291**	2,438

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
33	RF	SUB	MM	0,353	0,988	0,246	0,181	**0,305**	2,574
33	RF	SUB	ZS	0,355	0,986	0,250	0,179	**0,303**	2,848
33	RF	SMOTE	WN	0,413	0,745	0,358	0,162	**0,266**	39,454
33	RF	SMOTE	MM	0,263	0,891	0,158	0,150	**0,256**	32,771
33	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	29,051
34	RF	IMB	WN	0,893	0,830	0,903	0,588	**0,688**	8,767
34	RF	IMB	MM	0,889	0,840	0,898	0,578	**0,685**	8,998
34	RF	IMB	ZS	0,889	0,838	0,898	0,577	**0,684**	9,055
34	RF	SUB	WN	0,488	0,980	0,406	0,215	**0,353**	2,721
34	RF	SUB	MM	0,489	0,980	0,407	0,217	**0,356**	2,705
34	RF	SUB	ZS	0,490	0,980	0,409	0,216	**0,354**	2,693
34	RF	SMOTE	WN	0,420	0,707	0,372	0,158	**0,258**	22,765
34	RF	SMOTE	MM	0,413	0,713	0,363	0,157	**0,257**	15,326
34	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	15,273
35	RF	IMB	WN	0,909	0,685	0,946	0,680	**0,683**	12,017
35	RF	IMB	MM	0,906	0,654	0,947	0,675	**0,665**	10,569
35	RF	IMB	ZS	0,907	0,742	0,935	0,655	**0,695**	11,845
35	RF	SUB	WN	0,411	0,963	0,320	0,191	**0,318**	2,070
35	RF	SUB	MM	0,433	0,978	0,342	0,200	**0,332**	2,341
35	RF	SUB	ZS	0,427	0,974	0,336	0,196	**0,326**	2,589
35	RF	SMOTE	WN	0,549	0,549	0,549	0,168	**0,258**	26,593
35	RF	SMOTE	MM	0,544	0,557	0,542	0,168	**0,258**	26,144
35	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	24,320
36	RF	IMB	WN	0,834	0,923	0,820	0,460	**0,614**	10,833
36	RF	IMB	MM	0,832	0,925	0,817	0,457	**0,612**	11,063
36	RF	IMB	ZS	0,836	0,924	0,821	0,463	**0,617**	11,118
36	RF	SUB	WN	0,313	0,995	0,199	0,171	**0,292**	3,317
36	RF	SUB	MM	0,319	0,995	0,205	0,174	**0,296**	3,313
36	RF	SUB	ZS	0,317	0,995	0,204	0,172	**0,293**	3,270
36	RF	SMOTE	WN	0,148	0,998	0,007	0,143	**0,251**	26,601

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
36	RF	SMOTE	MM	0,144	1,000	0,001	0,143	**0,250**	26,331
36	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	26,267
37	RF	IMB	WN	0,864	0,896	0,858	0,512	**0,652**	16,077
37	RF	IMB	MM	0,857	0,893	0,851	0,500	**0,641**	14,176
37	RF	IMB	ZS	0,859	0,907	0,851	0,503	**0,647**	15,550
37	RF	SUB	WN	0,285	0,993	0,168	0,165	**0,284**	2,994
37	RF	SUB	MM	0,354	0,994	0,247	0,182	**0,307**	3,462
37	RF	SUB	ZS	0,351	0,994	0,244	0,179	**0,304**	3,879
37	RF	SMOTE	WN	0,157	0,996	0,018	0,144	**0,252**	39,522
37	RF	SMOTE	MM	0,149	0,998	0,008	0,143	**0,251**	39,029
37	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	34,956
38	RF	IMB	WN	0,905	0,846	0,914	0,621	**0,716**	9,670
38	RF	IMB	MM	0,903	0,854	0,911	0,615	**0,715**	9,493
38	RF	IMB	ZS	0,901	0,852	0,909	0,610	**0,711**	9,570
38	RF	SUB	WN	0,527	0,978	0,452	0,229	**0,371**	2,759
38	RF	SUB	MM	0,533	0,979	0,458	0,233	**0,376**	2,750
38	RF	SUB	ZS	0,530	0,978	0,456	0,230	**0,372**	2,762
38	RF	SMOTE	WN	0,179	0,981	0,045	0,146	**0,254**	16,664
38	RF	SMOTE	MM	0,179	0,981	0,046	0,146	**0,254**	16,842
38	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	16,685
39	RF	IMB	WN	0,917	0,758	0,944	0,691	**0,723**	13,286
39	RF	IMB	MM	0,915	0,734	0,945	0,691	**0,712**	11,705
39	RF	IMB	ZS	0,915	0,797	0,935	0,670	**0,728**	12,988
39	RF	SUB	WN	0,463	0,962	0,380	0,205	**0,338**	2,260
39	RF	SUB	MM	0,483	0,976	0,401	0,215	**0,352**	2,522
39	RF	SUB	ZS	0,539	0,975	0,467	0,233	**0,376**	3,151
39	RF	SMOTE	WN	0,209	0,963	0,084	0,149	**0,258**	28,555
39	RF	SMOTE	MM	0,211	0,961	0,087	0,149	**0,258**	28,769
39	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	26,085
40	RF	IMB	WN	0,672	0,642	0,677	0,248	**0,358**	8,984

Table 47 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	Time
40	RF	IMB	MM	0,666	0,655	0,668	0,248	**0,359**	8,767
40	RF	IMB	ZS	0,666	0,649	0,669	0,246	**0,357**	8,658
40	RF	SUB	WN	0,194	0,986	0,062	0,149	**0,259**	3,189
40	RF	SUB	MM	0,196	0,987	0,063	0,151	**0,261**	3,188
40	RF	SUB	ZS	0,194	0,987	0,063	0,149	**0,259**	3,200
40	RF	SMOTE	WN	0,176	0,970	0,044	0,144	**0,251**	23,395
40	RF	SMOTE	MM	0,164	0,981	0,028	0,144	**0,251**	23,145
40	RF	SMOTE	ZS	0,143	1,000	0,000	0,143	**0,250**	23,009

B.2- Neural Networks

B.2.1 Conventional Threshold

Table 48 – Results of Neural Networks Workflow Tests - Conventional

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
18	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	3,033
18	NN	IMB	MM	0,902	0,355	0,993	0,899	0,509	1,842
18	NN	IMB	ZS	0,859	0,013	1,000	0,863	0,026	1,353
18	NN	SUB	WN	0,785	0,631	0,811	0,356	0,456	51,144
18	NN	SUB	MM	0,755	0,672	0,769	0,328	0,441	40,813
18	NN	SUB	ZS	0,758	0,674	0,772	0,329	0,443	51,410
18	NN	SMOTE	WN	0,141	0,987	0,000	0,141	0,247	3,631
18	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	6,479
18	NN	SMOTE	ZS	0,143	0,999	0,001	0,143	0,249	1,377
19	NN	IMB	WN	0,859	0,013	1,000	0,880	0,026	1,430
19	NN	IMB	MM	0,902	0,354	0,993	0,896	0,508	4,203

Table 48 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
19	NN	IMB	ZS	0,903	0,371	0,992	0,889	0,523	3,862
19	NN	SUB	WN	0,804	0,610	0,836	0,382	0,470	1,049
19	NN	SUB	MM	0,754	0,673	0,767	0,327	0,440	44,465
19	NN	SUB	ZS	0,760	0,672	0,775	0,331	0,444	57,384
19	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	5,959
19	NN	SMOTE	MM	0,141	0,987	0,000	0,141	0,247	6,999
19	NN	SMOTE	ZS	0,319	0,865	0,228	0,157	0,266	1,889
20	NN	IMB	WN	0,909	0,423	0,990	0,872	0,570	3,088
20	NN	IMB	MM	0,859	0,013	1,000	0,899	0,026	1,925
20	NN	IMB	ZS	0,859	0,013	1,000	0,869	0,027	2,657
20	NN	SUB	WN	0,784	0,634	0,808	0,355	0,455	1,348
20	NN	SUB	MM	0,754	0,675	0,767	0,328	0,442	1,577
20	NN	SUB	ZS	0,855	0,019	0,994	0,347	0,035	42,433
20	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	13,350
20	NN	SMOTE	MM	0,347	0,831	0,266	0,158	0,266	3,581
20	NN	SMOTE	ZS	0,337	0,851	0,251	0,159	0,268	3,762
21	NN	IMB	WN	0,909	0,422	0,990	0,876	0,570	5,219
21	NN	IMB	MM	0,901	0,350	0,993	0,894	0,503	3,964
21	NN	IMB	ZS	0,903	0,369	0,992	0,889	0,522	4,187
21	NN	SUB	WN	0,784	0,634	0,809	0,356	0,456	49,842
21	NN	SUB	MM	0,755	0,673	0,769	0,328	0,441	1,108
21	NN	SUB	ZS	0,759	0,675	0,773	0,330	0,443	1,245
21	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	6,981
21	NN	SMOTE	MM	0,143	0,999	0,001	0,143	0,249	1,346
21	NN	SMOTE	ZS	0,311	0,873	0,218	0,156	0,265	2,403
22	NN	IMB	WN	0,857	0,000	1,000	0,000	0,000	2,199
22	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	6,115
22	NN	IMB	ZS	0,859	0,013	1,000	0,869	0,026	2,809
22	NN	SUB	WN	0,791	0,630	0,817	0,365	0,462	57,399
22	NN	SUB	MM	0,598	0,223	0,661	0,100	0,138	32,969

Table 48 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
22	NN	SUB	ZS	0,858	0,015	0,998	0,516	0,028	26,592
22	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	10,795
22	NN	SMOTE	MM	0,353	0,819	0,276	0,158	0,265	3,634
22	NN	SMOTE	ZS	0,311	0,875	0,217	0,157	0,266	3,052
23	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	2,951
23	NN	IMB	MM	0,902	0,355	0,993	0,899	0,509	1,787
23	NN	IMB	ZS	0,859	0,013	1,000	0,863	0,026	1,404
23	NN	SUB	WN	0,785	0,631	0,811	0,356	0,456	53,624
23	NN	SUB	MM	0,755	0,672	0,769	0,328	0,441	40,734
23	NN	SUB	ZS	0,758	0,674	0,772	0,329	0,443	51,137
23	NN	SMOTE	WN	0,141	0,987	0,000	0,141	0,247	3,718
23	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	6,248
23	NN	SMOTE	ZS	0,143	0,999	0,001	0,143	0,249	1,399
24	NN	IMB	WN	0,857	0,000	1,000	0,000	0,000	50,456
24	NN	IMB	MM	0,859	0,014	1,000	0,903	0,028	1,060
24	NN	IMB	ZS	0,896	0,375	0,983	0,786	0,507	6,430
24	NN	SUB	WN	0,785	0,635	0,810	0,357	0,457	1,305
24	NN	SUB	MM	0,845	0,022	0,983	0,175	0,038	32,337
24	NN	SUB	ZS	0,352	0,836	0,272	0,160	0,269	30,016
24	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	1,084
24	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	12,076
24	NN	SMOTE	ZS	0,378	0,777	0,312	0,158	0,263	5,128
25	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	2,992
25	NN	IMB	MM	0,902	0,355	0,993	0,899	0,509	1,836
25	NN	IMB	ZS	0,859	0,013	1,000	0,863	0,026	1,377
25	NN	SUB	WN	0,785	0,631	0,811	0,356	0,456	53,093
25	NN	SUB	MM	0,755	0,672	0,769	0,328	0,441	42,876
25	NN	SUB	ZS	0,758	0,674	0,772	0,329	0,443	48,161
25	NN	SMOTE	WN	0,141	0,987	0,000	0,141	0,247	3,570
25	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	6,277

Table 48 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
25	NN	SMOTE	ZS	0,143	0,999	0,001	0,143	0,249	1,270
27	NN	IMB	WN	0,857	0,000	1,000	0,000	0,000	2,280
27	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	6,047
27	NN	IMB	ZS	0,859	0,013	1,000	0,869	0,026	2,820
27	NN	SUB	WN	0,791	0,630	0,817	0,365	0,462	58,617
27	NN	SUB	MM	0,598	0,223	0,661	0,100	0,138	32,769
27	NN	SUB	ZS	0,858	0,015	0,998	0,516	0,028	27,531
27	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	10,761
27	NN	SMOTE	MM	0,353	0,819	0,276	0,158	0,265	3,442
27	NN	SMOTE	ZS	0,311	0,875	0,217	0,157	0,266	2,871
28	NN	IMB	WN	0,857	0,000	1,000	0,000	0,000	2,243
28	NN	IMB	MM	0,903	0,360	0,993	0,896	0,514	5,918
28	NN	IMB	ZS	0,859	0,013	1,000	0,869	0,026	2,743
28	NN	SUB	WN	0,791	0,630	0,817	0,365	0,462	52,133
28	NN	SUB	MM	0,598	0,223	0,661	0,100	0,138	29,838
28	NN	SUB	ZS	0,858	0,015	0,998	0,516	0,028	25,002
28	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	10,836
28	NN	SMOTE	MM	0,353	0,819	0,276	0,158	0,265	3,329
28	NN	SMOTE	ZS	0,311	0,875	0,217	0,157	0,266	2,826
29	NN	IMB	WN	0,909	0,422	0,990	0,873	0,569	4,468
29	NN	IMB	MM	0,903	0,362	0,993	0,900	0,516	5,478
29	NN	IMB	ZS	0,902	0,357	0,993	0,890	0,510	5,514
29	NN	SUB	WN	0,788	0,635	0,813	0,361	0,460	1,364
29	NN	SUB	MM	0,756	0,677	0,769	0,330	0,444	49,953
29	NN	SUB	ZS	0,854	0,017	0,993	0,298	0,032	36,639
29	NN	SMOTE	WN	0,144	0,999	0,001	0,143	0,250	13,616
29	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	18,068
29	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	11,678
31	NN	IMB	WN	0,872	0,118	0,998	0,899	0,208	4,809
31	NN	IMB	MM	0,868	0,083	0,999	0,948	0,152	4,721

Table 48 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
31	NN	IMB	ZS	0,869	0,088	0,999	0,935	0,161	5,651
31	NN	SUB	WN	0,630	0,545	0,644	0,203	0,296	1,192
31	NN	SUB	MM	0,639	0,567	0,651	0,215	0,312	49,773
31	NN	SUB	ZS	0,639	0,569	0,651	0,213	0,310	1,245
31	NN	SMOTE	WN	0,299	0,906	0,198	0,158	0,269	10,256
31	NN	SMOTE	MM	0,143	0,999	0,001	0,143	0,249	1,337
31	NN	SMOTE	ZS	0,315	0,870	0,222	0,157	0,266	2,939
33	NN	IMB	WN	0,857	0,000	1,000	0,000	0,000	1,689
33	NN	IMB	MM	0,857	0,000	1,000	0,000	0,000	2,355
33	NN	IMB	ZS	0,858	0,005	1,000	0,789	0,009	2,502
33	NN	SUB	WN	0,430	0,777	0,372	0,171	0,280	53,056
33	NN	SUB	MM	0,470	0,719	0,428	0,175	0,281	46,964
33	NN	SUB	ZS	0,474	0,729	0,432	0,176	0,283	59,082
33	NN	SMOTE	WN	0,697	0,300	0,763	0,174	0,220	17,126
33	NN	SMOTE	MM	0,330	0,862	0,242	0,159	0,268	5,214
33	NN	SMOTE	ZS	0,303	0,840	0,214	0,151	0,256	7,732
35	NN	IMB	WN	0,858	0,002	1,000	0,653	0,003	3,831
35	NN	IMB	MM	0,857	0,000	1,000	0,533	0,001	2,829
35	NN	IMB	ZS	0,858	0,005	1,000	0,751	0,009	2,084
35	NN	SUB	WN	0,851	0,020	0,989	0,244	0,038	22,958
35	NN	SUB	MM	0,842	0,023	0,980	0,160	0,041	23,141
35	NN	SUB	ZS	0,846	0,021	0,983	0,174	0,038	33,412
35	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	1,357
35	NN	SMOTE	MM	0,143	0,999	0,001	0,143	0,249	1,731
35	NN	SMOTE	ZS	0,337	0,818	0,257	0,155	0,260	11,923
37	NN	IMB	WN	0,907	0,409	0,990	0,876	0,557	6,359
37	NN	IMB	MM	0,859	0,014	1,000	0,902	0,027	2,032
37	NN	IMB	ZS	0,904	0,372	0,992	0,888	0,524	7,679
37	NN	SUB	WN	0,767	0,648	0,786	0,335	0,442	1,983
37	NN	SUB	MM	0,846	0,021	0,985	0,192	0,038	26,862

Table 48 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
37	NN	SUB	ZS	0,859	0,014	0,999	0,659	0,027	36,127
37	NN	SMOTE	WN	0,164	0,969	0,031	0,143	0,248	17,281
37	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	2,226
37	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	16,474
39	NN	IMB	WN	0,873	0,119	0,998	0,899	0,210	6,511
39	NN	IMB	MM	0,868	0,084	0,999	0,939	0,154	4,818
39	NN	IMB	ZS	0,867	0,081	0,999	0,907	0,149	4,012
39	NN	SUB	WN	0,854	0,019	0,993	0,315	0,036	42,847
39	NN	SUB	MM	0,643	0,571	0,655	0,218	0,316	52,736
39	NN	SUB	ZS	0,642	0,589	0,650	0,218	0,319	1,685
39	NN	SMOTE	WN	0,297	0,909	0,195	0,158	0,269	14,547
39	NN	SMOTE	MM	0,252	0,953	0,136	0,155	0,267	13,075
39	NN	SMOTE	ZS	0,317	0,892	0,221	0,160	0,271	7,971

B.2.2 Majority Threshold

Table 49 – Results of Neural Networks Workflow Tests - Majority

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
18	NN	IMB	WN	0,859	0,537	0,913	0,505	0,521	3,033
18	NN	IMB	MM	0,760	0,655	0,778	0,330	0,439	1,842
18	NN	IMB	ZS	0,677	0,327	0,735	0,171	0,224	1,353
18	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	51,144
18	NN	SUB	MM	0,211	0,991	0,080	0,153	0,266	40,813
18	NN	SUB	ZS	0,218	0,990	0,090	0,153	0,265	51,410
18	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	3,631
18	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	6,479
18	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	1,377

Table 49 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
19	NN	IMB	WN	0,859	0,013	1,000	0,878	0,026	1,430
19	NN	IMB	MM	0,772	0,636	0,795	0,341	0,444	4,203
19	NN	IMB	ZS	0,780	0,639	0,803	0,351	0,453	3,862
19	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	1,049
19	NN	SUB	MM	0,209	0,991	0,078	0,153	0,265	44,465
19	NN	SUB	ZS	0,205	0,992	0,074	0,151	0,262	57,384
19	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	5,959
19	NN	SMOTE	MM	0,141	0,988	0,000	0,141	0,247	6,999
19	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	1,889
20	NN	IMB	WN	0,857	0,541	0,910	0,500	0,520	3,088
20	NN	IMB	MM	0,839	0,029	0,974	0,155	0,048	1,925
20	NN	IMB	ZS	0,857	0,016	0,998	0,539	0,031	2,657
20	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	1,348
20	NN	SUB	MM	0,217	0,990	0,087	0,154	0,267	1,577
20	NN	SUB	ZS	0,143	1,000	0,000	0,142	0,249	42,433
20	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	13,350
20	NN	SMOTE	MM	0,177	0,980	0,043	0,146	0,253	3,581
20	NN	SMOTE	ZS	0,167	0,982	0,032	0,144	0,252	3,762
21	NN	IMB	WN	0,854	0,537	0,906	0,488	0,512	5,219
21	NN	IMB	MM	0,771	0,630	0,794	0,338	0,440	3,964
21	NN	IMB	ZS	0,779	0,635	0,803	0,350	0,451	4,187
21	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	49,842
21	NN	SUB	MM	0,214	0,991	0,083	0,154	0,266	1,108
21	NN	SUB	ZS	0,222	0,989	0,095	0,153	0,266	1,245
21	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	6,981
21	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	1,346
21	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	2,403
22	NN	IMB	WN	0,305	0,913	0,204	0,160	0,272	2,199
22	NN	IMB	MM	0,764	0,660	0,782	0,336	0,445	6,115
22	NN	IMB	ZS	0,854	0,022	0,993	0,348	0,041	2,809

Table 49 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
22	NN	SUB	WN	0,174	0,997	0,038	0,147	0,256	57,399
22	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	32,969
22	NN	SUB	ZS	0,142	1,000	0,000	0,142	0,249	26,592
22	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	10,795
22	NN	SMOTE	MM	0,161	0,983	0,024	0,144	0,250	3,634
22	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	3,052
23	NN	IMB	WN	0,859	0,537	0,913	0,505	0,521	2,951
23	NN	IMB	MM	0,760	0,655	0,778	0,330	0,439	1,787
23	NN	IMB	ZS	0,677	0,327	0,735	0,171	0,224	1,404
23	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	53,624
23	NN	SUB	MM	0,211	0,991	0,080	0,153	0,266	40,734
23	NN	SUB	ZS	0,218	0,990	0,090	0,153	0,265	51,137
23	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	3,718
23	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	6,248
23	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	1,399
24	NN	IMB	WN	0,857	0,000	1,000	0,000	0,000	50,456
24	NN	IMB	MM	0,849	0,022	0,987	0,213	0,039	1,060
24	NN	IMB	ZS	0,773	0,640	0,795	0,343	0,446	6,430
24	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	1,305
24	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	32,337
24	NN	SUB	ZS	0,142	1,000	0,000	0,142	0,249	30,016
24	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	1,084
24	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	12,076
24	NN	SMOTE	ZS	0,158	0,988	0,020	0,144	0,251	5,128
25	NN	IMB	WN	0,859	0,537	0,913	0,505	0,521	2,992
25	NN	IMB	MM	0,760	0,655	0,778	0,330	0,439	1,836
25	NN	IMB	ZS	0,677	0,327	0,735	0,171	0,224	1,377
25	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	53,093
25	NN	SUB	MM	0,211	0,991	0,080	0,153	0,266	42,876
25	NN	SUB	ZS	0,218	0,990	0,090	0,153	0,265	48,161

Table 49 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
25	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	3,570
25	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	6,277
25	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	1,270
27	NN	IMB	WN	0,305	0,913	0,204	0,160	0,272	2,280
27	NN	IMB	MM	0,764	0,660	0,782	0,336	0,445	6,047
27	NN	IMB	ZS	0,854	0,022	0,993	0,348	0,041	2,820
27	NN	SUB	WN	0,174	0,997	0,038	0,147	0,256	58,617
27	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	32,769
27	NN	SUB	ZS	0,142	1,000	0,000	0,142	0,249	27,531
27	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	10,761
27	NN	SMOTE	MM	0,161	0,983	0,024	0,144	0,250	3,442
27	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	2,871
28	NN	IMB	WN	0,305	0,913	0,204	0,160	0,272	2,243
28	NN	IMB	MM	0,764	0,660	0,782	0,336	0,445	5,918
28	NN	IMB	ZS	0,854	0,022	0,993	0,348	0,041	2,743
28	NN	SUB	WN	0,174	0,997	0,038	0,147	0,256	52,133
28	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	29,838
28	NN	SUB	ZS	0,142	1,000	0,000	0,142	0,249	25,002
28	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	10,836
28	NN	SMOTE	MM	0,161	0,983	0,024	0,144	0,250	3,329
28	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	2,826
29	NN	IMB	WN	0,855	0,541	0,907	0,491	0,514	4,468
29	NN	IMB	MM	0,765	0,657	0,783	0,336	0,445	5,478
29	NN	IMB	ZS	0,778	0,626	0,803	0,346	0,446	5,514
29	NN	SUB	WN	0,145	1,000	0,003	0,143	0,250	1,364
29	NN	SUB	MM	0,222	0,990	0,092	0,155	0,268	49,953
29	NN	SUB	ZS	0,143	1,000	0,001	0,142	0,249	36,639
29	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	13,616
29	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	18,068
29	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	11,678

Table 49 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
31	NN	IMB	WN	0,659	0,541	0,678	0,218	0,311	4,809
31	NN	IMB	MM	0,665	0,509	0,691	0,215	0,303	4,721
31	NN	IMB	ZS	0,652	0,521	0,674	0,211	0,300	5,651
31	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	1,192
31	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	49,773
31	NN	SUB	ZS	0,143	1,000	0,000	0,142	0,249	1,245
31	NN	SMOTE	WN	0,151	0,994	0,011	0,143	0,250	10,256
31	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	1,337
31	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	2,939
33	NN	IMB	WN	0,853	0,018	0,992	0,280	0,034	1,689
33	NN	IMB	MM	0,542	0,610	0,530	0,178	0,276	2,355
33	NN	IMB	ZS	0,849	0,023	0,987	0,220	0,041	2,502
33	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	53,056
33	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	46,964
33	NN	SUB	ZS	0,142	1,000	0,000	0,142	0,249	59,082
33	NN	SMOTE	WN	0,161	0,974	0,026	0,143	0,249	17,126
33	NN	SMOTE	MM	0,146	0,998	0,004	0,143	0,250	5,214
33	NN	SMOTE	ZS	0,145	0,998	0,003	0,143	0,250	7,732
35	NN	IMB	WN	0,540	0,624	0,526	0,180	0,279	3,831
35	NN	IMB	MM	0,523	0,649	0,501	0,179	0,280	2,829
35	NN	IMB	ZS	0,848	0,023	0,986	0,213	0,042	2,084
35	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	22,958
35	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	23,141
35	NN	SUB	ZS	0,142	1,000	0,000	0,142	0,249	33,412
35	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	1,357
35	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	1,731
35	NN	SMOTE	ZS	0,143	0,999	0,001	0,143	0,250	11,923
37	NN	IMB	WN	0,852	0,528	0,906	0,482	0,504	6,359
37	NN	IMB	MM	0,843	0,025	0,980	0,175	0,044	2,032
37	NN	IMB	ZS	0,777	0,642	0,799	0,348	0,451	7,679

Table 49 continued from the previous page

WF	ML	BL	NM	AC	SS/R	SP	P	F1-Score	T.EL
37	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	1,983
37	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	26,862
37	NN	SUB	ZS	0,143	1,000	0,000	0,142	0,249	36,127
37	NN	SMOTE	WN	0,143	1,000	0,000	0,143	0,250	17,281
37	NN	SMOTE	MM	0,143	1,000	0,000	0,143	0,250	2,226
37	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	16,474
39	NN	IMB	WN	0,687	0,507	0,717	0,229	0,316	6,511
39	NN	IMB	MM	0,665	0,513	0,691	0,217	0,305	4,818
39	NN	IMB	ZS	0,718	0,385	0,773	0,221	0,281	4,012
39	NN	SUB	WN	0,143	1,000	0,000	0,143	0,250	42,847
39	NN	SUB	MM	0,144	1,000	0,000	0,144	0,252	52,736
39	NN	SUB	ZS	0,144	1,000	0,002	0,143	0,249	1,685
39	NN	SMOTE	WN	0,153	0,993	0,013	0,143	0,250	14,547
39	NN	SMOTE	MM	0,151	0,995	0,010	0,143	0,250	13,075
39	NN	SMOTE	ZS	0,143	1,000	0,000	0,143	0,250	7,971