UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO MESTRADO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

Mirele Marques Borges

MACHINE LEARNING COMO FERRAMENTA GERENCIAL PARA PREDIÇÃO DE INDICADORES E DETECÇÃO DE ANOMALIAS

Porto Alegre
2020

Mirele Marques Borges

Machine Learning como ferramenta gerencial para predição de indicadores e detecção de anomalias

Dissertação submetida ao Programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Profissional, na área de concentração em Sistemas de Produção.

Orientador: Professor Cláudio José Müller

Porto Alegre

2020

Mirele Marques Borges

Machine Learning como ferramenta gerencial para predição de indicadores e detecção de anomalias

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Profissional e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Cláudio José Müller
Orientador PMPEP/UFRGS

_____

Profa. Christine Tessele Nodari
Coordenadora PMPEP/UFRGS

Banca Examinadora:

Professor Néstor Fabián Ayala, Dr. (UFRGS)

Professor Ricardo Augusto Cassel, Ph.D. (PMPEP/UFGRS)

Professor Vinicius Andrade Brei, Dr. (PPGA/UFRGS)

AGRADECIMENTOS

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho. Ao orientador, Prof. Dr. Cláudio José Müller, pelo inestimável apoio na orientação deste trabalho. Aos professores e colegas, que contribuíram singularmente para a construção deste trabalho com o compartilhamento de conhecimento e experiências em sala de aula.

Agradeço também a minha Família e ao meu Marido, pelo apoio, paciência e por sempre estarem à disposição para discutir novas ideias.

RESUMO

A presente dissertação tem o objetivo de identificar as técnicas de *Machine Learning* utilizadas nas áreas de Engenharia e Medicina e proporcionar um conhecimento da aplicação de modelos de *Machine Learning* e métodos de detecção de anomalias para problemas gerenciais, pois muitas vezes veem-se estas técnicas sendo incorporadas a problemas complexos e de larga escala, sem muitos exemplos dentro dos ambientes gerenciais. O trabalho está dividido em dois artigos: o primeiro artigo é uma revisão de literatura focada em apresentar os algoritmos de *Machine Learning*, os tipos de problemas aos quais são aplicados e métodos de validação utilizados nas áreas de Engenharia e Medicina. O segundo artigo apresenta o processo de criação de um modelo de *Machine Learning* capaz de predizer um indicador gerencial bem como propor uma métrica para detecção de anomalias. Todos os artigos utilizaram ferramentas *open source*, como o *software* estatístico R. As contribuições dos artigos foram: (1) Identificação dos algoritmos mais utilizados nas áreas de Engenharias e Medicinas, os métodos de validação mais utilizados e as contribuições dos autores sobre desempenho dos algoritmos; (2) Demonstração do processo de criação de um modelo de *Machine Learning* (coleta e preparação dos dados, seleção das variáveis, escolha do algoritmo, seleção dos hiperparâmetros, treino do modelo e avaliação dos resultados), além do método de detecção de anomalia através da análise da distribuição da diferença entre predito e observado.

Palavras-chave: *Machine Learning*, Predição, Detecção de Anomalias, Revisão de Literatura, Algoritmos.

ABSTRACT

The present dissertation aims to identify the Machine learning techniques used in the fields of engineering and medicine and provides knowledge of the application of models of machine learning and anomaly detection methods to management problems, since often these techniques are incorporated into complex and large-scale problems without many examples within the management environments. The work is divided into two articles: the first article is a literature review focused on presenting Machine Learning algorithms, the types of problems to which they are applied and validation methods used in the areas of Engineering and Medicine. The second article presents the process of creating a Machine Learning model capable of predicting a managerial indicator as well as proposing a metric for detecting anomalies. All articles used open-source tools, such as the statistical software R. The contributions of the articles were: (1) Identification of the most used algorithms in the areas of Engineering and Medicine, the most used validation methods and the contributions of the authors about algorithms performance; (2) Demonstration of the process of creating a Machine Learning model (collection and preparation of data, variables selection, choice of algorithm, selection of hyperparameters, training of the model and evaluation of results), in addition to the method of detection of anomaly by analyzing the distribution of the difference between predicted and observed.

Key words: Machine Learning, Prediction, Anomaly Detection, Literature Review, Algorithm.

LISTA DE FIGURAS

Capítulo 2

Capítulo 3

LISTA DE TABELAS

Capítulo 2

Capítulo 3

LISTA DE QUADROS

SUMÁRIO

1 **INTRODUÇÃO**

Os algoritmos de *Machine Learning* podem ser aplicados para resolução de diversos tipos de problemas, ou seja, são usados tanto para predizer um indicador num setor industrial quanto dentro da área da saúde, contribuindo assim para a eficácia dos processos, diagnósticos, tomadas de decisão entre outros. Sob a ótica do uso de algoritmos de *Machine Learning*, as áreas do conhecimento Engenharia e Medicina tornam-se similares, pois ambas podem fazer uso de técnicas semelhantes para a solução de problemas distintos. Por exemplo, o mesmo algoritmo utilizado para predizer a probabilidade de infecção por uma determinada doença poderá ser o mesmo utilizado na detecção de defeito de uma peça em uma cadeia de produção.

Para continuar este tema, se faz necessário incluir algumas definições dos termos mais utilizados nesta área. Segundo Mohri, Rostamizadeh e Talwalkar (2018), *Machine Learning* pode ser definida como um método computacional para predição de dados, ou seja, a partir de uma base de dados, como por exemplo, uma série temporal que apresenta os dados ao longo do tempo, a "máquina" aprende sobre os padrões de comportamento destes dados através de algoritmos com o objetivo de predizer um resultado futuro. Um algoritmo, segundo Moschovakis (2001), é comumente definido como uma máquina abstrata ou modelos matemáticos computadorizados que através de um *input* (base de dados) inicial consegue realizar inúmeras operações a fim de devolver um resultado ou a solução de um problema (*output*).

Mohri, Rostamizadeh e Talwalkar (2018), definem que os tipos de problemas nos quais são aplicados algoritmos de *Machine Learning* estão em constante expansão. Alguns exemplos de casos nos quais algoritmos de *Machine Learning* têm sido aplicados amplamente são: classificação de texto ou documento, processamento de linguagem natural (PLN), aplicativos de processamento de fala, aplicativos de visão computacional, aplicativos de biologia computacional, detecção de fraude, diagnóstico médico, sistemas de extração de informações, pontuação de crédito e muito mais.

Um ponto importante para aplicação de modelos de *Machine Learning* é qualidade e tamanho da base de dados que será utilizada, pois as bases de dados precisam contemplar variáveis que expliquem a variável resposta de maneira adequada. Para evitar alguns problemas quanto ao tamanho e multi-colinearidade, se faz necessária a aplicação de algumas técnicas, tais como *Feature Engineering* e *Variable Selection.* De acordo com Murdoch

(2019), *Feature Engineering* é uma técnica de criação de variável a partir de uma ou de um grupo de variáveis já existentes, esta técnica visa aprimorar a precisão do modelo agregando novas variáveis ao algoritmo. Por exemplo, uma variável de "data" pode dar origem a inúmeras variáveis, tais quais: dias da semana, mês, ano, ser ou não um dia útil entre outros. Assim, com a criação de novas variáveis através do *Feature Engineering*, temos a possibilidade de explicar melhor a variável resposta. Porém, também se faz necessário verificar a presença de multi-colienaridade, ou seja, quando variáveis independentes possuem relações lineares exatas ou aproximadamente exatas, causando redundância e possivelmente distorção no resultado esperado, por isso, para evitar estes tipos de problemas, são aplicadas técnicas de seleção de variáveis tais como: *Stepwise backward*, *Stepwise forward*, *Lasso*, entre outras.

Os principais objetivos dos modelos de *Machine Learning* consistem em gerar predições acuradas para todos os tipos de problemas, desde problemas de pequena escala até maiores escalas, através de um modelo eficiente. Segundo Mohri, Rostamizadeh e Talwalkar (2018), *Machine Learning* está altamente relacionado à análise de dados e estatística, normalmente combinando as técnicas de aprendizado com conceitos da ciência da computação, estatística, probabilidade e otimização.

A Detecção de anomalias, de acordo com Chandola, Banerjee, Kumar (2009) e Khairi (2018), é um problema importante de ampla aplicação e que tem sido utilizado em diversas áreas do conhecimento. O termo "detecção de anomalias" pode ser associado com "detecção de *outliers*", que consiste na identificação de itens, eventos ou observações que não estão em conformidade com um padrão esperado ou outros itens em um conjunto de dados. Normalmente, os itens anômalos se traduzem em algum tipo de problema, como fraude bancária, defeito estrutural, problemas médicos ou erros em um texto.

Normalmente, quando o tema *Machine Learning* é abordado na literatura, vê-se a aplicação destas técnicas em problemas de larga escala e com alta complexidade. Porém, é importante ressaltar que estes métodos podem ser aplicados para problemas de pequena escala principalmente na área gerencial, pois estas técnicas garantem uma maior velocidade no processamento de dados e na identificação de informações relevantes trazendo agilidade na tomada de decisão, redução de custos e competitividade.

Esta dissertação é composta por dois artigos relacionados ao tema *Machine Learning*, que objetivam tornar mais abrangente o conhecimento sobre a aplicabilidade desses

algoritmos através da exploração dos artigos já publicados sobre o tema, além de proporcionar uma compreensão do processo de criação de um modelo de *Machine Learning* através do desenvolvimento e aplicação em um problema real.

O primeiro artigo traz ao conhecimento dois pontos importantes sobre a aplicação dos algoritmos de *Machine Learning*, sendo o primeiro, os tipos de algoritmos mais presentes nas áreas de engenharia e medicina, e o segundo, os métodos de validação mais utilizados. Esses métodos têm como objetivo identificar a acuracia de um algoritmo ao resolver determinado problema e assim, realizar a comparação entre modelos a fim de escolher o de melhor desempenho.

O segundo artigo traz ao foco todas as etapas relacionadas ao processo de criação e aplicação de um modelo de *Machine Learning* para a predição de um indicador real na área de saúde suplementar brasileira. Além disso, o estudo também propõe um método de detecção de anomalias no indicador de números de consultas médicas realizadas por dia, embasado no entendimento da distribuição e comportamento dos dados.

Objetiva-se com o primeiro artigo demonstrar todas as similaridades entre os algoritmos e métodos de validação utilizados nas áreas de Engenharia e Medicina e assim expor as contribuições de cada autor na aplicação destas técnicas. Assim como, objetiva-se com o segundo artigo demonstrar como algoritmos de *Machine Learning* podem unir conhecimentos de várias áreas e construir um modelo eficiente para predizer e detectar anomalias em um indicador real, e assim colaborar para a eficácia de processos de uma organização.

1.1 TEMA DO TRABALHO E DELIMITAÇÃO

Machine Learning, como já definido por Mohri, Rostamizadeh e Talwalkar (2018), é um método computacional para predição de dados. Segundo Domingos (2017), o termo *Machine Learning* começou a se tornar conhecido na década de 1980, quando foi usado no setor de finanças para predição da flutuação de ações, passando a crescer na década de 1990, quando expandiu a sua aplicação para a área bancária para predição de fraudes e *credit scorecard*. Em seguida, estas técnicas foram aplicadas no *e-commerce* ganhando força na publicidade automatizada e divulgação de produtos e serviços personalizados. Atualmente, as possibilidades de aplicação das técnicas de *Machine Learning* são muito amplas, pois cada vez mais as empresas buscam respostas rápidas, automatizadas, com menos recursos humanos

e mais recursos computacionais, trazendo agilidade para o dia-a-dia organizacional. Ainda segundo Domingos (2017), os algoritmos de *Machine Learning* surgiram para aperfeiçoar tarefas e recursos: enquanto anteriormente era necessário construir um modelo computacional para realizar somente uma tarefa específica, ou possuir um grupo de pessoas para analisar determinada informação e encontrar uma resposta, hoje os algoritmos podem realizar tarefas simultâneas, podem ser aplicados a diferentes problemas e bases de dados, além de depender de menos recursos humanos.

Como visto o tema *Machine Learning* é bastante amplo, podendo ser estudado por diversas vertentes, tais como: aplicação, desenvolvimento de novas técnicas, otimização, explicabilidade, desempenho computacional entre outros. Para este trabalho, propõem-se somente identificar os algoritmos mais utilizados e aplicar os mesmos a um problema real de predição de um indicador, bem como entender o processo de criação como um todo, abstraindo a complexidade matemática e computacional das técnicas apresentadas.

1.1.1 PROBLEMA E QUESTÃO DE PESQUISA

Diariamente, as organizações geram e armazenam uma infinidade de dados, que quando analisados tornam-se informação e podem trazer ganhos, tais como: eficiência operacional, redução de custos, projeção de resultados entre outros. Porém, é comum que muitos destes dados não sejam propriamente utilizados. Isso acontece mais frequentemente em áreas gerenciais e menos estratégicas para as organizações. Assim, existe ainda um grande potencial de utilização de técnicas de *Machine Learning* em áreas gerenciais.

Com o problema exposto, surge a questão de pesquisa: Como integrar um modelo de *Machine Learning* a um indicador gerencial utilizando os algoritmos mais aplicados nas áreas de engenharia e medicina?

A questão de pesquisa demonstra como a escolha do algoritmo é uma etapa importante do processo e faz parte da criação do modelo de *Machine Learning*. Assim, ao identificar os algoritmos mais utilizados e os tipos de problemas aos quais estes são aplicados, pode-se embasar os critérios de seleção do mesmo.

1.2 OBJETIVOS

Os objetivos deste trabalho estão divididos em geral e específicos.

1.2.1 OBJETIVO GERAL

O objetivo geral do trabalho é identificar os tipos de algoritmos de *Machine Learning* mais utilizados nas áreas de Engenharia e Medicina, e realizar a aplicação destes algoritmos a um problema real de predição.

1.2.2 OBJETIVOS ESPECÍFICOS

Com este trabalho pretende-se também:

- Identificar, além dos algoritmos, os tipos de problemas aos quais são aplicados e os métodos de validação com os quais são medidos.
- Descrever o processo de criação e aplicação de um modelo de *Machine Learning*, propondo um método de detecção de anomalias para um indicador.

1.3 JUSTIFCATIVA

Tendo em vista que as técnicas de *Machine Learning*, usualmente, são aplicadas a problemas complexos e de grande escala, o tema *Machine Learning* como ferramenta gerencial para predição de indicadores e detecção de anomalias trás ao foco a necessidade de difundir estes conhecimentos para áreas gerenciais contribuindo na otimização de processos e agilidade na tomada de decisão, trazendo como principais benefícios da utilização de algoritmos de *Machine Learning* a simplificação dos modelos de previsão e a redução nos tempos de execução computacional (GEYER et al. 2018).

De acordo com Domingos (2017), *Machine Learning* está mais presente em nosso dia-a-dia do que se pode perceber. A quantidade de informações geradas todos os dias trás a oportunidade de entender melhor os padrões. Assim, para Elangovan et al. (2015) e Puranik et al. (2016), apesar do avanço tecnológico das máquinas industriais, ainda há como aprimorar a qualidade da fabricação de produtos através da aplicação de técnicas de *Machine Learning* para predição de falhas em produtos, falha de máquinas entre outros. Para Dos Santos (2019), as aplicações de algoritmos de *Machine Learning* na área de saúde ainda são poucas se comparadas a outras áreas do conhecimento, porém Battineni et al. (2019) e Leighton et al. (2019) demonstram que a aplicação destes algoritmos na área médica está evoluindo rapidamente, trazendo diversos estudos quanto a diagnóstico de doenças, transmissão de vírus, resultados de tratamentos entre outros.

O setor de saúde brasileiro é composto pelo sistema de saúde público e o sistema privado, chamado de sistema suplementar de saúde. Segundo De Araujo et al. (2015) e a ANS (Agência Nacional de Saúde Suplementar) hoje o Brasil possui mais de 1500 operadoras de planos de saúde e, para manter este serviço, é necessário um grande estudo dos custos assistenciais, sendo que muitas destas empresas têm dificuldade de prever o custo assistencial gerando instabilidade financeira para a organização e aumentando assim o custo para o cliente final.

Visto que a área de Medicina ainda possui poucos estudos com o tema *Machine Learning*, sendo estes mais voltados para a área de diagnósticos de doenças, é importante a incorporação de outra área do conhecimento, como a Engenharia, para trazer o contraponto das técnicas aplicadas e demonstrar o uso dos algoritmos em problemas do tipo predição de indicadores, predição de eventos, predição de falhas entre outros.

Segundo Da Silva (2017), nos últimos vinte anos, a maioria dos artigos publicados sobre *Machine Learning* estão localizados nas áreas de Engenharia e Ciências da Computação, por este motivo a inclusão de uma destas áreas mostra-se essencial para a ampliação do entendimento do tema. De acordo com Mitchell (2006), as contribuições da área da Ciência da Computação ligadas a *Machine Learning* têm uma forte relação no desenvolvimento computacional, sendo assim, optou-se pela inclusão da área de Engenharia para complemento do estudo, pois o foco pretendido está na aplicação das técnicas e não no desenvolvimento computacional.

1.4 MÉTODO

Apresentam-se nesta seção os métodos de pesquisa e métodos de trabalho aplicados.

1.4.1 MÉTODO DE PESQUISA

Os métodos de pesquisa utilizados para desenvolvimento deste trabalho foram: Revisão de literatura e Estudo de caso aplicado.

Segundo Bento (2012), a revisão de literatura é uma parte muito importante para o processo de investigação, pois consiste em identificar através de critérios previamente definidos, um conjunto de conteúdo (artigos, livros entre outros) que discursam sobre o objeto de estudo, para que se consiga identificar as contribuições, discordâncias e concordâncias ou novos conceitos evidenciados pela literatura. Já o estudo de caso descritivo é um método de

investigação qualitativa. Segundo Yin (2001), este método é utilizado para explicar as ações tomadas referente ao objeto de estudo, como estas foram implementadas e quais os resultados obtidos.

Conforme Quadro 1, os procedimentos metodológicos aplicados nos artigos consistem em Revisão de literatura e estudo de caso aplicado. A revisão de literatura realizada como procedimento metodológico no primeiro artigo buscou definir e filtrar um grupo de artigos disponíveis nas bases eletrônicas públicas no período de Janeiro de 2015 e Setembro de 2019, que possuíam como tema *Machine Learning* e predição nas áreas de engenharia e medicina. O estudo de caso aplicado no segundo artigo buscou demonstrar todas as etapas do processo de criação de um modelo de *Machine Learning* aplicado a uma base dados e um problema real.

Quadro 1: Questão de pesquisa, capítulos, métodos, artigos e publicação

Questão de pesquisa	**Capítulo**	**Método aplicado**	**Artigo**	**Atual status referente a publicação**
Quais são os algoritmos mais utilizados para predição nas áreas de engenharia e medicina?	2	Revisão de literatura de artigos publicados em base eletrônicas: *science direct* e *scielo.*	"*Machine learning algorithms and their applications: a literature review in the fields of engineering and medicine*"	Submetido para a revista Independent Journal Of Management & Production em 02/10/2020.
Como integrar um modelo de machine learning a um indicador gerencial?	3	Estudo aplicado com a utilização dos algoritmos de *Machine Learning*: *Random Forest, Linear Regression, Extreme Gradient Boosting* e *Neural Network*. Validação de algoritmos através de MAE. Detecção de anomalia.	"*Prediction of indicators through machine learning and anomaly detection: a case study in the supplementary health system in Brazil*"	Submetido para a revista Independent Journal Of Management & Production em 02/10/2020.

Fonte: Elaborado pelo autor

1.4.2 MÉTODO DE TRABALHO

Para alcançar os objetivos propostos, o trabalho seguiu determinadas etapas:

- Definição da problemática a ser estudada;
- Levantamento da base teórica através da revisão de literatura;
- Compreensão do conteúdo e contribuições encontradas na revisão de literatura;
- Identificação de um problema a ser estudado;

- Compreensão do problema e construção de uma solução para o estudo de caso;
- Aplicação da solução;
- Análise dos resultados obtidos;
- Conclusão do trabalho.

1.5 DELIMITAÇÃO DO TRABALHO

Não se pretende discutir todo o universo relacionado a *Machine Learning*, principalmente, não se pretende analisar o tema através da ótica de operações matemáticas e recursos computacionais.

Não se pretende explorar todos os indicadores utilizados na área da Saúde Suplementar brasileira.

Pretende-se apenas discutir o tema abordado e demonstrar a possibilidade de implementação das técnicas no meio gerencial.

Pretende-se, ao estudar o tema *Machine Learning* dentro da área de Engenharia e Medicina, estender a aplicabilidade de suas técnicas bem como demonstrar que tais técnicas podem ajudar setores gerenciais a obter agilidade na tomada de decisão.

Não serão abordadas outras técnicas de Machine Learning, a não ser as pré-selecionadas.

Os algoritmos pré-selecionados serão comparados entre si, somente quanto ao desempenho ao predizer o indicador proposto.

Não será abordado *Machine Learning* associado com outras técnicas que visam a explicabilidade da predição, ou seja, como as variáveis estão influenciando na predição.

A aplicação é específica, não permitindo extrapolação de resultados.

1.6 ESTRUTURA DO TRABALHO

A dissertação está dividida em quatro capítulos: (i) introdução, (ii) capítulo 2 (Artigo 1), (iii) capítulo 3 (Artigo 2) e (iv) conclusão.

A introdução apresenta uma visão geral do trabalho, definições de termos e etapas realizadas, seguindo para o capítulo 2, o qual apresenta o artigo "*Machine learning algorithms and their applications: a literature review in the fields of engineering and medicine*", continuando com o capítulo 3 que apresenta o artigo "*Prediction of indicators through*

machine learning and anomaly detection: a case study in the supplementary health system in Brazil" e finalizando com uma conclusão geral sobre as contribuições dos dois artigos.

2 MACHINE LEARNING ALGORITHMS AND THEIR APPLICATIONS: A LITERATURE REVIEW IN THE FIELDS OF ENGINEERING AND MEDICINE

Abstract

The research aimed to review the literature on the subject of data prediction through Machine Learning algorithms and their applications in areas of knowledge such as engineering and medicine, to identify which Machine Learning algorithms are being used for data prediction, what kind of problems, and what are the methods of validation for each type of models. It was used the literature review method to select a group of articles to be analyzed. The selection was performed in the electronic data source Science Direct and Scielo, using the keywords Machine Learning and Prediction and filtering just the research papers published from January 2015 to September 2019 in the English language in the areas of Engineering and Medicine. It was identified that the most used algorithms to predict results in both areas were Neural Network, Linear Regression, and Random Forest, followed by Support Vector Machine in the Medicine area and Extreme Learning Machine in the Engineering area, and in the side of validation methods were identified as the most used methods R^2 (R-squared), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). In addition, this study also provides an overview of the different techniques used by the authors and the analyzed factors to choose a Machine Learning model, such as the performance and required computational resources. With the results analyzed, it was possible to conclude that the Machine Learning algorithms have high applicability in diverse areas, and it is possible to use the same method to predict behavior data, failures, risks, or group classification.

Keywords: Machine Learning, Prediction, Algorithms Methods, Validation Methods, Literature Review.

2.1 INTRODUCTION

The term Machine Learning can be defined as a set of mathematical and statistical techniques, which through computational tools and historical data are able to predict data, patterns, and behaviors or even classify groups (MARSLAND, 2014 and BISHOP, 2006). Every day computers from all over the world store data of the most varied types, such as: from stores, banks, hospitals, scientific laboratories, and even personal data from users of social networks, it means, there are banks registering spending behavior of clients, hospitals recorded histories of illnesses and treatments of patients, systems recording the performance of engines and more. According to Marsland (2014), Machine Learning algorithms are a set of instructions used to predict data that can be applicable in different areas of knowledge and for the most diverse types of problems, such as: How to detect bank fraud? How to detect the best treatments for certain diseases? Among others.

According to Da Silva (2017), in the last twenty years, most of the articles published about Machine Learning are located in the areas of Engineering and Computer Sciences and Mitchell (2006) says that the contributions of the Computer Science area linked to Machine Learning have a strong relationship in computational development. So it's important to bring to the discussion other areas that have less articles exploring the Machine Learning theme, areas like healthcare, that has studies with distinct objectives. Thus, the question becomes apparent: "What algorithms are most used in predicting data in the field of engineering and medicine?" Engineering and Medicine are two distinct areas of knowledge, however, when analyzed from the perspective of Machine Learning with the predictive objective, both probably use the same techniques and algorithms, making it possible to compare and identify the most common techniques in each area and complement with others techniques that are used to solve the same type of problem.

Also according to Marsland (2014), some steps must be followed to use Machine Learning to predict data: (i) collecting and preparing the database, (ii) selecting the variables, (iii) choosing the algorithm, (iv) selection of parameters, (v) training of the model and (vi) evaluation of the results.

Therefore, the objective of this article will focus on the steps (iii) choice of the algorithm and (vi) evaluation of the results according to steps defined by Marsland (2014), that means, it proposes to identify which Machine Learning algorithms are being used for prediction of data and what types of problems are studied in the areas of engineering and

medicine. In addition, the methods used to validate the results obtained by the algorithms in each application will also be exposed.

The article is divided into five main sections: (i) introduction, (ii) methodology, (iii) literature review, (iv) results and discussion and (v) conclusions. The methodology, presented in the second section, defines the search and selection methods of the articles to be analyzed in the next chapters. Then, in the third section, the literature review aims to identify the Machine Learning algorithms used by the authors. The fourth section, in which the results are presented, exposes the phase of analysis of the identified contributions and discussion. Finally, the last section presents the conclusions obtained from the literature review.

2.2 METHODOLOGICAL PROCEDURES

The literature review was made in a systematic process, which consists of analyzing the existing content in a structured way in relation to the proposed research question, understanding the topics and to identifying conflicting information or necessary complements. The selection of the studied articles was defined prior, in the beginning of the study, to reduce sample bias, in other words, without removing or adding articles manually, accepting only the selected articles through the filters defined in the research.

The systematic literature review was performed according to articles available in the electronic databases Science Direct and Scielo. First, all articles that presented the keywords Machine Learning and Prediction in the titles were selected. The research was fulfilled by filtering the articles published in Journals in the fields of Engineering and Medicine with applied research, published in the English language and open access. It was selected the articles published between January 2015 and September 2019, according to the search results after filtering by the parameters defined for this research the articles volume starts to increase in 2015, and this range of period represents 80% of the articles published since 2000.

A total of 209 articles related to the research question were found. First, the possibility of duplicate articles was verified and for this, it was created an R script capable of standardizing the selected titles and checking possible duplications of it automatically. After standardize, 5 articles were identified and removed for duplicity, resulting in the amount of 204. Subsequent to the selection of the articles, the titles were read in order to select the ones most closely related to the topic, resulting in a total of 100 selected articles. At the end it was read the abstracts to refine the selection, and it was only considered those articles that used

existing algorithms or that contained the comparison between algorithms, resulting in 20 final articles with greater relevance, as shown in Figure 1.

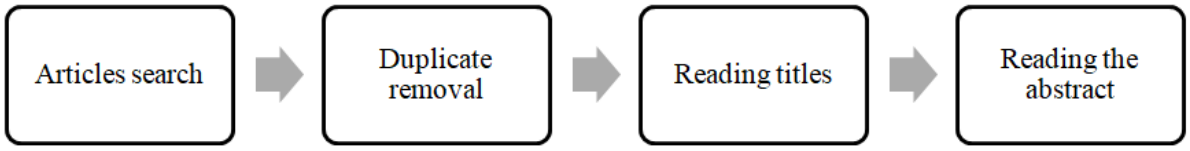

Figure 1: Article selection process

2.3 LITERATURE REVIEW

In this section will be presented the articles identified as relevant in the area of data prediction through Machine Learning algorithms, as well as their applications and the methods used to evaluate the results obtained in each case study.

Marsland (2014) and James et al. (2013) classify Machine Learning algorithms as: supervised learning, unsupervised learning or reinforced learning. According to James et al. (2013), supervised learning algorithms are used when the problem to be answered by the developed model can be qualified as a classification or regression problem and also when the response variable is already known. Some examples of supervised learning algorithms are: (i) k-nearest neighbors: which consists of allocating observations to existing groups, using similarity measures; (ii) linear regression: whose objective, in short, is to adjust a linear equation between the explanatory variables and the response variable; (iii) logistic regression: it was described as the probability of the occurrence of a given event in relation to a set of exploratory variables; (iv) support vector machines: it was defined as a set of methods used to analyze and recognize patterns; (v) decision trees: can be described as the representation of a decision table in the form of a tree, but the decision trees in Machine Learning consist of creating trees in a non-parametric way, that means, the parameters of the trees created, such as the depth and number of branches, does not need to be defined; (vi) random forests: it was described as the creation of a lot of random decision trees to generate an average forecast of individual trees and (vii) neural networks: can be defined as computational models used to recognize complex patterns.

Also according to James et al. (2013), unsupervised learning algorithms are a little more challenging because there is no response variable to be predicted, and it is usually used for models that aim to identify relationships between variables or observations. Some examples of algorithms in unsupervised learning are: (i) k-means: which consists of a clustering method that aims to divide the observations into several groups; (ii) hierarchical cluster analysis: defined as an analysis method that aims to build a hierarchy of clusters; (iii) main component analysis: a mathematical method that uses an orthogonal transformation, in other words, it creates linearly independent components, to convert a set of observations of variables. Machine learning algorithms also can be classified as reinforced learning. According to Sutton and Barto (2018), reinforced learning is learning by trial and error and delayed rewards, which means, the algorithm has not a setting path to follow to get the results, it will learning by trying many actions and discovering which one gives more rewards.

Another point of interest in this research is how the different algorithms are evaluated and compared, as well as the most used performance metrics to define what technique is better to achieve the final objectives of each article. Some of the validation methods used in Machine Learning models are: (i) the AUC index derived from the ROC curve (Receiver Operating Characteristic), where higher the value of the AUC curve, better will be the performance; (ii) Confusion Matrix or error matrix, which consists of generating a table where it is possible to evaluate the numbers of false positives and false negatives (this method compares the results obtained by a forecast with the actual results observed); (iii) F-measure, used to measure results of binary classifications; (iv) MAE (Mean Absolute Error), used to summarize and verify the quality of a model according to the measurement of the difference between two continuous variables; (v) MAPE (Mean Absolute Percentage Error), used to estimate the loss trend or function in regression problems; (vi) MSE (Mean Squared Error), used to calculate the mean of the squared error in a regression problem; (vii) R and R^2, which are measures that indicate the degree of adjustment of the model, that is, the portion of variation in the response variable explained by the explanatory variables; (Viii) RMS or RMSE (Root Mean Squared Error), which is a measure used to verify the difference between the predicted values and actual values.

Also, it's important to expose some of the main kind of problems that Machine Learning algorithms are applied, such as: text or document classification, natural language processing (NLP), speech processing applications, computer vision applications,

computational biology applications, fraud detection, medical diagnosis, information extraction systems, credit score and more. According to Mohri, Rostamizadeh and Talwalkar (2018), those are just a few examples of prediction problems using Machine Learning methods, in practical applications, the use of machine learning keeps expanding.

In terms of application, in the Engineering area, the use of Machine Learning algorithms is applicable for the prediction of events, behaviors, results, failure or risk, among others. According to Cheng and Xiong (2017), the use of the Extreme Learning Machine algorithm proved to be efficient when combined with two other prediction models previously applied in the database, integrating the results of the first two models as an input vector for the new model, developing an efficient model for the prediction of the risk of dam detachment, validated by comparing the values of MSE and MAE. Lei, Zhao and Cai (2015) also used Extreme Learning Machine, applied to predict a result and not an event: the proposed objective was to forecast the duration of the day. Using a database from the International Service of Reference Systems and Earth Rotation (IERS), the proposed new model was compared with other Machine Learning models such as Neural Network, contracting similar results, but with better computational performance.

Mozaffari, Mozaffari and Azad (2015) used Least Machine Learning algorithms to predict behavior, aiming to predict the speed of a vehicle in urban areas using time series in order to improve the performance of the vehicle in relation to fuel economy and greenhouse gas emissions. The proposed model was compared with other Machine Learning models such as the Extreme Learning Machine, Regression and Neural Network. The comparison was based on the calculation of MSE, RMSE, MAPE and R^2, considering the model that uses Least Learning Machine suitable for measuring the performance of vehicles.

According to Ding, Lam and Feng (2017), also in the engineering area, Machine Learning algorithms can be used to propose a new index to assess the potential of natural ventilation in the internal environment of buildings, to be used in urban planning. For this problem, the use of the Gradient Boosting algorithm was more effective, as it presented higher R^2 and lower MAPE compared to the Linear Regression algorithm.

According to Ahmad, Chen and Huang (2019), French et al. (2017), Kreutz et al. (2019) and Mozaffari, Mozaffari and Azad (2015), Neural Networks can be applied to predict events. Kreutz et al. (2019) used Neural Network algorithms to create a Machine Learning model through a database with historical data in order to predict the risk of freezing wind

turbines. The result proved to be effective, when more meteorological data such as air humidity and liquid water content were included.

In the paper by Ahmad, Chen and Huang (2019), it was used Neural Network algorithms to predict the short-term power requirements to the district level and for measure the accuracy of the models it was performed the CV index (coefficient of variation) and MAPE. FRENCH et al. (2017) also used Neural Network for short-term predictions of extreme water levels in estuarine ports, using the city of Immingham as a test for the Neural Network model, applying information on tide, waves, wind and atmospheric pressure as a vector input. The model achieved accuracy comparable to the UK's national tidal rise model with a lower computational cost. According to Geyer and Singaravel (2018), the main benefits of using Machine Learning algorithms are simplifying predictive modeling and reduction in computational runtime. In the area of civil construction, there are limitations regarding the use of Machine Learning models for data prediction and may be more accurate depending on the range of data that make up the training base.

According to Gouarir et al. (2018), Neural Networks can be used to predict behavior, such as performing tool wear prediction in process. This production process uses a force sensor to monitor the progression of tool edge wear. Thus, the Neural Network algorithm was applied to the data captured by the sensor. In addition to the Machine Learning algorithm, an adaptive control called self-adaptive was used to calculate the force that should be used and thus optimize the equipment's life cycle, showing this methodology an estimated accuracy of 90%. Wang and Kim (2018) also used the Neural Network algorithms to predict behavior, besides to comparing the Gated Recurrent Unit and Random Forest algorithms, to predict the available short-term number of bicycles in docking stations.

For Elangovan et al. (2015) and Puranik, Deshpande and Chandrasekaran (2016), despite the technological advancement of industrial machines, there is still a way to improve the quality of product manufacturing. In their paper, Elangovan et al. (2015) expose the use of regression for the prediction of failures, that is, perform the forecast of the roughness of a surface by multiple regression analysis and obtain greater predictability and low computational effort. Puranik, Deshpande and Chandrasekaran (2016) use the regression algorithms to predict the occurrence of errors and failures in new software. The techniques already known were analyzed, such as Linear Regression and the R^2 metric, as a basis to

propose and compare a new algorithm to predict the propensity index of the occurrence of an error or failure in new software.

As in engineering, Neural Network algorithms are also commonly applied to the health care area. In the medical area, besides using Machine Learning methods to predict events, results, and behaviors, the algorithms are also used to classify groups. For Battineni, Chintalapudi and Amenta (2019) and Leighton et al. (2019), the use of Machine Learning algorithms in the medical field is rapidly evolving, carrying out studies on previous disease diagnosis and disease transmission. Battini et al. (2019) explored the use of the Support Vector Machine algorithm in classifying patients with dementia and validating their performance through statistical analysis. Leighton et al. (2019), in turn, used Machine Learning algorithms to build models for classifying psychosis cases. The Logistic Regression algorithm was used, as well as the cross-validation technique, to evaluate the model's performance, and thus classify possible cases of psychosis based on the results of remission and recovery collected during a year of people with first episode psychosis. It was identified that the prediction models can reliably and prospectively identify the poor results of remission and recovery in one year for patients with first episode psychosis, using basic clinical variables in the first clinical contact.

According to Kourou et al. (2015), Machine Learning algorithms are increasingly used for disease prediction, making the comparison of several Machine Learning techniques, such as Neural Network, Bayesian Network, Support Vector Machine and Decision Trees, important and widely applicable in research of cancer for the development of predictive models, resulting in effective and accurate decision making.

For data prediction, Campos et al. (2019) compared variable selection techniques in order to the model obtained better results, although the Machine Learning models trained with subsets of data performed better than a random selection. Mathur, Glesk and Buis (2016) used adaptive inference strategy methods in Neural Network to predict the residual temperature of prosthetic sockets by monitoring the temperature between the socket and the liner, in order to alleviate complaints about increased temperature and perspiration in prosthetic sockets. The result of this new model was compared to the previous one, finding efficient performance, but similar to the model already used.

Funkner, Kovalchuk and Bochenina (2016), Raj and Nandhini (2018) and Birjali, Beni-Hssane, Erritali (2017) brought an approach by merging the supervised and

unsupervised learning algorithms to achieve a better result in their research. For Funkner, Kovalchuk and Bochenina (2016), in addition to the definition of the Machine Learning algorithm to be used, it is necessary to perform the pre-classification of the data using the K-means method. In this study, the goal was to predict the preoperative time for patients with acute coronary syndrome and the pre-classification of the data improved the regression result by up to two times.

Raj and Nandhini (2018) proposed the application of a set of algorithms for data prediction, a technique known as an ensemble model. The objective of the study was to predict the patterns of human movement sequences in an internal environment, to improve the trajectory of nurses in a hospital, to optimize the movements of elderly or disabled people to minimize their routine efforts, among other patterns or anomalies. For Birjali, Beni-Hssane, Erritali (2017), the analysis of feelings is a very present theme, and its use is increasingly widespread given the amount of data generated daily by social networks. The objective of the study was to create a vocabulary associated with suicide, using a Text Mining tool and comparing Machine Learning algorithms, such as Support Vector Machine and Naive Bayes, used to classify potential Tweets related to this topic.

It was seen that Machine Learning algorithms can be used to solve several problems in order to bring operational efficiency, both for complex problems such as those found in the health area, like as disease diagnoses, and for operational problems, like as the reduction of waste in engineering.

2.4 RESULTS AND DISCUSSION

Data prediction through Machine Learning is extensively discussed in different areas of knowledge. The methodology used for the articles' selection and the literature review, made it possible to identify 20 articles, 60% of it was classified with subjects related to the areas of Engineering and 40% with the areas of Medicine. The origin of the authors of the articles predominate in countries China, India and the UK, and 50% of the articles selected for the literature review were published in the last two years.

From Figure 2, it is possible to observe the main Machine Learning algorithms used by the authors in the areas of knowledge Engineering and Medicine in order to predict data through regression or classification.

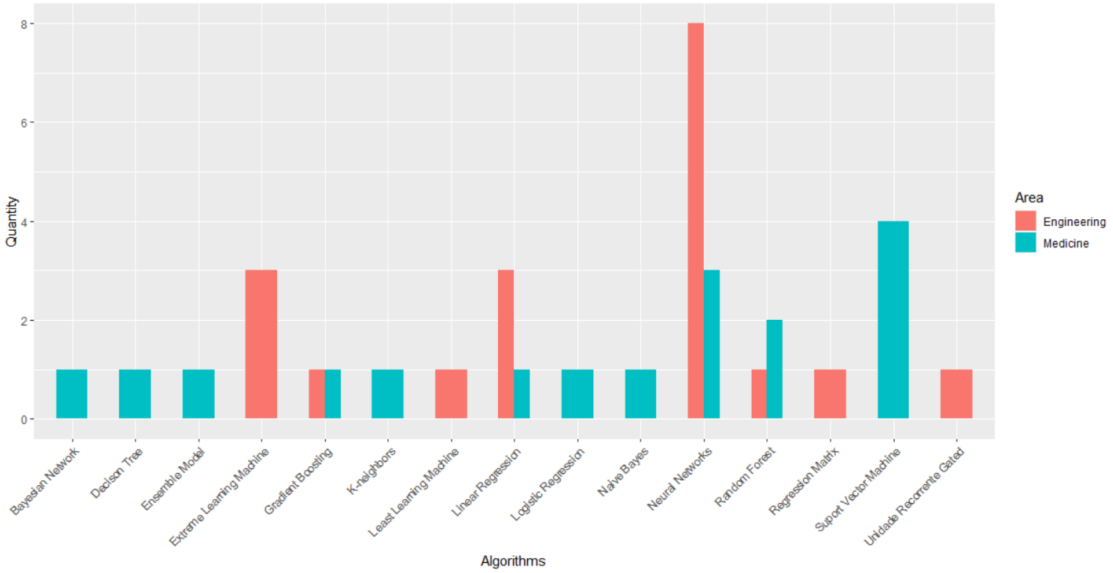

Figure 2: Most cited algorithms

In the engineering area, there was a greater use of Neural Network algorithms, followed by Linear Regression and Extreme Learning Machine algorithms. Kreutz et al. (2019), Ahmad, Chen and Huang (2019), Wang and Kim (2018), Gouarir et al. (2018), French et al. (2017) and Mozaffari, Mozaffari and Azad (2015) described the use of Neural Network algorithms to solve a classification problem through a time series, while Geyer and Singaravel (2018) and Lei, Zhao and Cai (2015) used these same algorithms for regression problems, bringing contributions such as the importance of Machine Learning to optimize computational resources.

The Extreme Learning Machine algorithms used in the Engineering area by the authors Cheng and Xiong (2017), Lei, Zhao and Cai (2015) and Mozaffari, Mozaffari and Azad (2015) also used a time series to predict new data, that is, predicting a future data based on the past. However, Mozaffari, Mozaffari and Azad (2015) and Cheng and Xiong (2017), presented a new modeling technique that consists of including the result of Machine Learning models previously applied as a vector of the new model developed.

In the medicine area, there was no predominance of a single algorithm, with the Neural Network and Support Vector Machine algorithms being used by more than one author. Kourou et al. (2015) compared several of these algorithms in order to identify the best model for predicting health data. The Neural Network algorithm was also used by Kourou et al. (2015) compared to other models for groups classification assisting decision making.

The Machine Learning algorithms for Sentiment Analysis were exposed by the authors Birjali, Beni-Hssane, Erritali (2017), in order to classify, through Support Vector Machine, Naive Bayes and Text Mining algorithms, potential Tweets related to the theme of suicide. Only the authors Funkner, Kovalchuk and Bochenina (2016), Raj and Nandhini (2018) and Birjali, Beni-Hssane, Erritali (2017) showed applications of unsupervised learning algorithms, and Funkner, Kovalchuk and Bochenina (2016) have applied unsupervised learning algorithms, such as k-means, such as a pre-step for rated supervised learning model, such as Linear Regression and Random Forest.

The authors Campos et al. (2019), Ding, Lam and Feng (2017), Kourou et al. (2015) and Funkner, Kovalchuk and Bochenina (2016), conducted studies comparing different algorithms in order to select the best in performance, even though these authors using similar algorithms, the choice of the final model was different for each study.

There were different contributions by the authors when comparing the objectives of each study. According to Table 1, both in the medical and engineering fields, the authors presented studies demonstrating the comparison of algorithms with inclusion of indicators to improve the model's performance; the creation of regression or classification models based on time series, foreseeing a future data based on the past; the creation of models to assist decision making; some authors have also presented the combination of unsupervised models and supervised models to achieve a better result; use of social networks and analysis of feelings for classification problems; simplification of models through Machine Learning algorithms optimizing computational resources; and the inclusion of results from other models as a vector for a new model.

Table 1: Authors' Contributions

Contributions	Author
Comparison of algorithms	Campos et al. (2019)
	Ding, Lam and Feng (2017)
	Funkner, Kovalchuk and Bochenina (2016)
	Kourou et al. (2015)
Comparison of algorithms with inclusion of indicators	Gouarir et al. (2018)
	Mathur, Glesk and Buis (2016)
	Wang and Kim (2018)
Creation of classification models based on time	Battineni, Chintalapudi and Amenta (2019)

series	Leighton et al. (2019)
Creation of models to assist decision making	Ahmad, Chen and Huang (2019)
	Ding, Lam and Feng (2017)
	French et al. (2017)
	Kourou et al. (2015)
	Kreutz et al. (2019)
	Mozaffari, Mozaffari and Azad (2015)
Unsupervised models using social media	Berjali, Beni-Hssane and Erritali (2017)
Creation of regression models based on time series	Cheng and Xiong (2017)
	Lei, Zhao and Cai (2015)
	Mozaffari, Mozaffari and Azad (2015)
Simplification of models through ML, bringing optimization of computational resources	Ahmad, Chen and Huang (2019)
	Elangovan et al. (2015)
	Geyer and Singaravel (2018)
	Lei, Zhao and Cai (2015)
	Puranik, Deshpande and Chandrasekaran (2016)
	Wang and Kim (2018)
Use of unsupervised and supervised algorithms together	Berjali, Beni-Hssane and Erritali (2017)
	Funkner, Kovalchuk and Bochenina (2016)
	Raj and Nandhini (2018)
Inclusion of results from other models as a vector for the new model	Cheng and Xiong (2017)
	Mozaffari, Mozaffari and Azad (2015)

The validation methods also proved to be applicable for more than one type of model, as shown in Table 2.

Table 2: Algorithm Validation Methods

Validation methods	Problem type	Algorithms
Accuracy	Classification	Neural Networks
	Regression	K-neighbors, Random Forest, Linear Regression
AUC	Classification	Neural Networks, Gradient Boosting, Support Vector Machine, Random Forest, Bayesian Network, Decision Tree, Logistic Regression
Confusion Matrix	Classification	Support Vector Machine, Neural Networks
CV	Classification	Neural Networks
F-measure	-	Ensemble Model
F-score	Classification	Support Vector Machine, Naive Bayes

MAE	Classification	Extreme Learning Machine, Neural Networks, Random Forest, Gated Recurrent Unit
	Regression	Gradient Boosting, Linear Regression, K-neighbors, Random Forest, Extreme Learning Machine, Neural Networks
MAPE	Classification	Neural Networks, Least Learning Machine, Extreme Learning Machine, Regression Matrix, Random Forest, Gated Recurrent Unit
	Regression	Gradient Boosting, Linear Regression
MSE	Classification	Extreme Learning Machine, Least Learning Machine, Neural Networks, Regression Matrix, Random Forest, Gated Recurrent Unit
R^2	Classification	Logistic Regression, Least Learning Machine, Extreme Learning Machine, Neural Networks, Regression Matrix
	Regression	Gradient Boosting, Linear Regression, Neural Networks
RMS or RMSE	Classification	Linear Regression, Neural Networks
	Regression	Gradient Boosting, Linear Regression, Neural Networks, Extreme Learning Machine

According to the literature review, it was possible to notice that the authors used the algorithm comparison technique to define, through validation methods such as R^2, MSE, MAE, RMSE, MAPE, among others, the best algorithm to be applied in the prediction. It was also understood that the Neural Network algorithms have greater applicability in different areas of knowledge and in different types of data prediction due to their generality, and it was possible to solve problems of both classification and regression. The Linear Regression algorithms also proved to be applicable for data prediction in the areas of knowledge classified in this study, reaching significant results when considering its low complexity and high interpretability of results.

Table 3 presents an overview of all the information collected by the study, indicating the algorithms, validation methods, types of problems and contributions made by each author.

Table 3: Overview by area of knowledge and authors

Area	Problem type	Authors	Year	Country	Algorithms	Validation Methods	Contribution
Engineering	Classification	Ahmad, Chen and Huang	2019	China	NN	MAPE CV	Creation of models to assist decision making. Simplification of models through ML, bringing optimization of computational resources.
Engineering	Classification	Cheng and Xiong	2017	China	ELM	MAE MSE	Creation of regression models based on time series. Inclusion of results from other models as a vector for the new model.
Engineering	Classification	Elangovan et al.	2018	India	LR	R^2 RMSE	Simplification of models through ML, bringing optimization of computational resources.
Engineering	Classification	French et al.	2017	United Kingdom	NN	RMSE	Creation of models to assist decision making.
Engineering	Classification	Gouarir et al.	2018	United Kingdom	NN	Confusion Matrix	Comparison of algorithms with inclusion of indicators.
Engineering	Classification	Kreutz et al.	2019	Germany	NN	Accuracy	Creation of models to assist decision making.
Engineering	Classification	Mozaffari, Mozaffari and Azad	2015	Iran / Canada	LLM ELM NN RM	MSE MAPE R^2	Creation of models to assist decision making. Creation of regression models based on time series. Inclusion of results from other models as a vector for the new model.
Engineering	Classification	Puranik, Deshpande and Chandrasekaran	2016	India	LR	R^2	Simplification of models through ML, bringing optimization of computational resources.
Engineering	Classification	Wang and Kim	2018	Australia	NN RF GRU	MSE MAPE MAE	Comparison of algorithms with inclusion of indicators. Simplification of models through ML, bringing optimization of computational resources.
Engineering	Regression	Ding, Lam and Feng	2017	China	GB LN	R^2 MAE RMSE MAPE	Comparison of algorithms. Creation of models to assist decision making.
Engineering	Regression	Geyer and Singaravel	2018	Belgium	NN	R^2	Simplification of models through ML, bringing optimization of computational resources.
Engineering	Regression	Lei, Zhao and Cai	2015	China	ELM NN	MAE RMS	Creation of regression models based on time series. Simplification of models through ML, bringing optimization of computational resources.

Medicine	-	Raj and Nandhini	2018	India	EM	F-measure	Use of unsupervised and supervised algorithms together.
Medicine	Classification	Battineni, Chintalapudi and Amenta	2019	India / Italy	SVM	Confusion Matrix	Creation of classification models based on time series.
Medicine	Classification	Berjali, Beni-Hssane and Erritali	2017	Morocco	SVM NV	F-score	Unsupervised models using social networks. Using unsupervised and supervised algorithms together.
Medicine	Classification	Campos et al.	2019	Australia / Brazil	NN GB SVM RF	AUC	Comparison of algorithms.
Medicine	Classification	Kourou et al.	2015	Greece	BN DT NN SVM	AUC	Comparison of algorithms. Creation of models to assist decision making
Medicine	Classification	Leighton et al.	2019	United Kingdom	LOG	AUC	Creation of classification models based on time series.
Medicine	Regression	Funkner, Kovalchuk and Bochenina	2016	Russia	K-n RF LR	MAE Accuracy	Comparison of algorithms. Use of unsupervised and supervised algorithms together.
Medicine	Regression	Mathur, Glesk and Buis	2016	United Kingdom	NN	MAE RMSE R^2	Comparison of algorithms with inclusion of indicators.

2.5 CONCLUSION

The study provided a great understanding of the Machine Learning algorithms used to predict data in the areas of knowledge Engineering and Medicine. It was possible to identify different techniques used by the authors both in the modeling and comparison of algorithms and in the validation methods, as well as the creation of regression and classification models based on historical data, in addition to the combination of unsupervised Machine Learning and supervised algorithms for achieve a better result in data prediction.

It was possible to identify that the same algorithm can be applied to problems of different types, such as: prediction of failure or risk, behavior prediction, event prediction and even for classification of groups, and also these algorithms were compared by the same validation methods. It was identified that the most used algorithms were Neural Network, Linear Regression, Random Forest, Extreme Learning Machine and Support Vector Machine and the most used validation methods were R^2, MAE and MAPE. There were algorithms that were applied in both areas, Engineering and Medicine, those were: Neural Network, Linear Regression, Random Forest and Gradient Boosting, and there were algorithms that were

identified just in one of the studied areas, those were: Support Vector Machine, Naive Bayes, Extreme Learning Machine, K-neighbors, Bayesian Network, Decision Tree, Logistic Regression, Least Learning Machine, Regression Matrix, Ensemble Model and Gated Recurrent Unit.

Another important contribution made by authors was about the analyzed factor to compare and select a model to be applied for a prediction, which was the importance to be aware of the performance and the computational resource require because it can influence the decision making.

The methodology applied in this study can be replicated to a new literature review on the topic as well as to update this review, since this topic is increasingly discussed both in the professional and academic fields, thus receiving new applications and discoveries to be made each new study.

2.6 REFERENCES

AHMAD, Tanveer; CHEN, Huanxin; HUANG, Yao. Short-Term Energy Prediction for District-Level Load Management Using Machine Learning Based Approaches. **Energy Procedia**, v. 158, p. 3331-3338, 2019.

BATTINENI, Gopi; CHINTALAPUDI, Nalini; AMENTA, Francesco. Machine Learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). **Informatics in Medicine Unlocked**, v. 16, p. 100200, 2019.

BIRJALI, Marouane; BENI-HSSANE, Abderrahim; ERRITALI, Mohammed. Machine Learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. **Procedia Computer Science**, v. 113, p. 65-72, 2017.

CAMPOS, Tulio L. et al. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. **Computational and Structural Biotechnology Journal**, v. 17, p. 785-796, 2019.

CHENG, Jiatang; XIONG, Yan. Application of extreme learning machine combination model for dam displacement prediction. **Procedia Computer Science**, v. 107, p. 373-378, 2017.

DA SILVA, Emerson Correia; LENGLER, Fernando Ramos. A PRODUÇÃO CIENTÍFICA SOBRE MACHINE LEARNING NA ÁREA EDUCACIONAL NO BRASIL (1999-2017). **CADERNOS DE INICIAÇÃO CIENTÍFICA**, v. 2, n. 1, 2017.

DING, Chao; LAM, Khee Poh; FENG, Wei. An evaluation index for cross ventilation based on CFD simulations and ventilation prediction model using Machine Learning algorithms. **Procedia engineering**, v. 205, p. 2948-2955, 2017.

ELANGOVAN, M. et al. Machine Learning approach to the prediction of surface roughness using statistical features of vibration signal acquired in turning. **Procedia Computer Science**, v. 50, p. 282-288, 2015.

FRENCH, Jon et al. Combining Machine Learning with computational hydrodynamics for prediction of tidal surge inundation at estuarine ports. **Procedia IUTAM**, v. 25, p. 28-35, 2017.

FUNKNER, Anastasia; KOVALCHUK, Sergey; BOCHENINA, Klavdiya. Preoperational Time Prediction for Percutaneous Coronary Intervention Using Machine Learning Techniques. **Procedia Computer Science**, v. 101, p. 172-176, 2016.

GEYER, Philipp; SINGARAVEL, Sundaravelpandian. Component-based Machine Learning for performance prediction in building design. **Applied energy**, v. 228, p. 1439-1453, 2018.

GOUARIR, A. et al. In-process tool wear prediction system based on Machine Learning techniques and force analysis. **Procedia CIRP**, v. 77, p. 501-504, 2018.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. **An introduction to statistical learning**. New York, 2013.

KOUROU, Konstantina et al. Machine Learning applications in cancer prognosis and prediction. **Computational and structural biotechnology journal**, v. 13, p. 8-17, 2015.

KREUTZ, Markus et al. Machine Learning-based icing prediction on wind turbines. **Procedia CIRP**, v. 81, p. 423-428, 2019.

LEI, Yu; ZHAO, Danning; CAI, Hongbing. Prediction of length-of-day using extreme learning machine. **Geodesy and Geodynamics**, v. 6, n. 2, p. 151-159, 2015.

LEIGHTON, Samuel P. et al. Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a Machine Learning approach. **The Lancet Digital Health**, v. 1, n. 6, p. e261-e270, 2019.

MARSLAND, Stephen. **Machine Learning: an algorithmic perspective**. Chapman and Hall/CRC, 2014.

MATHUR, Neha; GLESK, Ivan; BUIS, Arjan. Comparison of adaptive neuro-fuzzy inference system (ANFIS) and Gaussian processes for Machine Learning (GPML) algorithms for the prediction of skin temperature in lower limb prostheses. **Medical engineering & physics**, v. 38, n. 10, p. 1083-1089, 2016.

MITCHELL, Tom Michael. **The discipline of machine learning**. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. Foundations of machine learning. **MIT press**, 2018.

MOZAFFARI, Ladan; MOZAFFARI, Ahmad; AZAD, Nasser L. Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on San Francisco urban roads. **Engineering science and technology, an international journal**, v. 18, n. 2, p. 150-162, 2015.

MURDOCH, W. James et al. Definitions, methods, and applications in interpretable machine learning. **Proceedings of the National Academy of Sciences**, v. 116, n. 44, p. 22071 - 22080, 2019.

PURANIK, Shruthi; DESHPANDE, Pranav; CHANDRASEKARAN, K. A novel Machine Learning approach for bug prediction. **Procedia Computer Science**, v. 93, p. 924-930, 2016.

RAJ, S. Sridhar; NANDHINI, M. Ensemble human movement sequence prediction model with Apriori based Probability Tree Classifier (APTC) and Bagged J48 on Machine Learning. **Journal of King Saud University-Computer and Information Sciences**, v. 31, p. 55-62, 2018.

SUTTON, Richard S.; BARTO, Andrew G. Reinforcement learning: An introduction. **MIT press**, 2018.

WANG, Bo; KIM, Inhi. Short-term prediction for bike-sharing service using Machine Learning. **Transportation research procedia**, v. 34, p. 171-178, 2018.

2.7 APPENDIX A – ACRONYMS

Acronyms	Algorithms
NN	Neural Networks
SVM	Support Vector Machine
NB	Naive Bayes
GB	Gradient Boosting
RF	Random Forest
ELM	Extreme Learning Machine
LR	Linear Regression
K-n	K-neighbors
BN	Bayesian Network
DT	Decision Tree
LOG	Logistic Regression
LLM	Least Learning Machine
RM	Regression Matrix
EM	Ensemble Model
GRU	Gated Recurrent Unit

3 PREDICTION OF INDICATORS THROUGH MACHINE LEARNING AND ANOMALY DETECTION: A CASE STUDY IN THE SUPPLEMENTARY HEALTH SYSTEM IN BRAZIL

Abstract

The research aimed to investigate the stages of a Machine Learning model process creation in order to predict the indicator over the number of medical appointments per day done in the area of supplementary health in the region of Porto Alegre / RS - Brazil and to propose a metric for anomalies detection. Literature review and descriptive case study was used as a methodology in this paper, besides was used the statistical software called R, in order to prepare the data and create the model. The stages of the case study were: understanding the problem to be studied, database extraction, division of the base in training and testing, creation of functions and feature engineering, variables selection and correlation analysis, choice of the algorithms with cross-validation and tuning, training of models, application of the models in the test data, selection of the best model and proposal of the metric for anomalies detection. At the end of these stages, it was possible to select the best model in terms of MAE (Mean Absolute Error), the Random Forest, which was the algorithm with better performance when compared to Extreme Gradient Boosting, Linear Regression and Neural Network. It also makes possible to identified nine anomaly points and thirty-eight warning points using the standard deviation metric. It was concluded, through the proposed methodology and the results obtained, that the steps of feature engineering and variables selection were essential for the creation and selection of the model, in addition, the proposed metric achieved the objective of generates alerts in the indicator, showing cases with possible problems or opportunities.

Keywords: Machine Learning, Indicators, Anomaly Detection, Feature Engineering e Supplementary Health System.

3.1 INTRODUCTION

Every day computers from all over the world are generating and storing data of the most varied types. According to James et al. (2013), with the emergence of Big Data, modeling and understanding complex databases through statistical models has become a subject in evidence in different areas of science. For Burrel (2016) and Lee (2019), Machine Learning algorithms are good tools when the goals is to make predictions in large databases, because, by combining statistical techniques and artificial intelligence, Machine Learning algorithms are able to solve problems, such as: fraud detection, prediction of indicators, prediction of behaviors and even early diagnosis of diseases.

In the Brazilian health sector, in addition to the public health system, there is also the private system called supplementary health. According to De Araujo et al. (2015), ANS (National Supplementary Health Agency) regulates more than 1500 health plan operators existing in Brazil today. Many of these operators are in an unstable financial situation, due to the difficulty in forecasting assistance costs. For this reason, these companies have been implementing technology to detect unnecessary exams, costly procedures without justification and medical fraud, thus guaranteeing a better service. To predict in advance a cost behavior, number of medical appointments, number of patients and more, improves the strategic and financial sector of health plan operators that subsidize the cost of assistance. With an accurate prediction of the indicators, a fairer value for the population can be guaranteed.

Therefore, this article aims to demonstrate the process of creating a Machine Learning model to predict the number of medical appointments in the area of supplementary health in the region of Porto Alegre / RS, in addition to proposing metrics for detecting anomalies for this indicator. The necessary steps to create a predictive model will be presented, such as: collecting and preparing the database, selecting the variables, choosing the algorithm, selecting the parameters, training the model and evaluating the results. All analyzes and codes presented in this article were developed using the statistical software R, which is an open-source programming language that makes it possible to share the knowledge developed in this paper with the whole R community.

3.2 MACHINE LEARNING

Marsland (2014) and Bishop (2006) define the term Machine Learning, as a set of techniques that aims to learn from historical data, that is, using computational strength to

better predict patterns, behaviors, or even perform the classification and creation of groups. The numerous algorithm techniques, according to Marsland (2014) and James et al. (2013) can be classified as: supervised learning and unsupervised learning. Supervised learning algorithms are used for classification or regression problems, when the response variable is also known a priori. Still according to James et al. (2013), the unsupervised learning algorithms are a little more challenging because there is no response variable to be predicted, it is normally used for models that aim to identify relationships between the variables or observations.

According to Marsland (2014), the process of creating a supervised Machine Learning model for predicting continuous data (regression problem) must follow some steps: (i) collecting and preparing the data, (ii) selecting the variables, (iii) choice of algorithm, (iv) selection of hyperparameters, (v) training of the model and (vi) evaluation of results.

The collection and preparation of data can be done by selecting a group of variables potentially important for the proposed objective. According to Marsland (2014), this set can be tested in order to choose the best set of variables present in the original base. It is at this stage that the division of the base in training and testing is also carried out. The training data used for all the steps related to the discovery of knowledge and the test data having its use restricted only to the stage of validation of the results.

According to Garcia et al. (2007), it is in the stage of preparing the database that descriptive analyzes are carried out in order to know, clean, group, transform and enrich the data, the latter has known technically as Feature Engineering. According to Garla et al. (2012), Feature Engineering is the process of creating new variables from those already existing in a database, either by combining two or more variables, or by creating a new variable extracted from an existing one.

As Marsland (2014), the selection of variables is an important step, at this stage, are selected the most useful variables to explain the problem. According to James et al. (2013), there are several types of variable selection a method that allow a better interpretability of the model and reduce the risk of overfitting, such as: Subset Selection and Stepwise Selection. The Subset selection method is used for any addition or removal procedure variables in a pre-existing set, or selects exhaustively the subset of variables which maximize the desired result. As the number of combinations depends on the number of variables in base, it becomes costly to obtain metrics for all possible subsets of variables. To solve this problem, it is common to

use the Stepwise Selection Forward or Stepwise Selection Backward Method, which are Subset Selection techniques that respectively add and remove variables sequentially.

According to Burrel (2016), the choice of the Machine Learning algorithm must consider the computational capacity available and the type of problem to be solved. In addition to these items, it is very common to perform a combination of models or to compare models in order to identify the best result. Also according to Burrel (2016), the most popular examples of Machine Learning are Neural Network, Decision Tree, Logistic Regression, in addition to Linear Regression and Random Forest.

According to Marsland (2014) and Kraska et al. (2013), many algorithms require tuning hyperparameters, which can be selected manually or through tests in order to identify the most appropriate hyperparameter. This selection is commonly made in an exhaustive way, that is, an interval is tested for each hyperparameter and all possible combinations of these intervals are performed. In order to guarantee the robustness of the results generated by such algorithms, the Cross-validation technique can be used.

Refaeilzadeh et al. (2009) defines Cross-validation as a statistical method used to validate Machine Learning algorithms, by crossing the training base with a validation base, so that all data sets are validated. The most common method is the k-fold, which divides k sub-samples of the training base, if the number chosen for k is equal to 10, the training base will be divided into 10 sub-samples that will be crossed and validated with each other.

According to Marsland (2014) and Vinyals et al. (2019), after the steps of data preparation, selection of variables, selection of algorithms and selection of parameters, it is possible to perform the training stage of the models through some computational resource. In this step, the model will receive inputs and generate outputs, that is, it will receive explanatory variables and estimate the response variable.

In the results assessment stage, the models are compared with each other to determine which technique has better performance for solving the target problem (RODRIGUEZ-GALIANO et al, 2015; SUNDAYS, 2012). Here the performance metrics obtained from the application of the trained models in the test base are generated. The metrics commonly used for the evaluation of regression models are: Rsquared, RMSE and MAE. Rsquared indicates the correlation between observed result values and the values predicted by the model, in this case, the higher the Rsquared the better the model. The RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are two metrics used to measure the prediction error of each

model. In this case, the metric is evaluated considering that the lower the RMSE and the MAE, the better the model.

3.3 METHODOLOGICAL PROCEDURES

Seeking to analyze the proposed subject, the method used in this research consists of a descriptive case study. The data was extracted from January 2018 to July 2019 in order to predict the number of medical appointments per day in the health sector in the region of Porto Alegre / RS - Brazil and propose metrics to identify potential anomalies in this indicator, according to the steps described in the Figure 1.

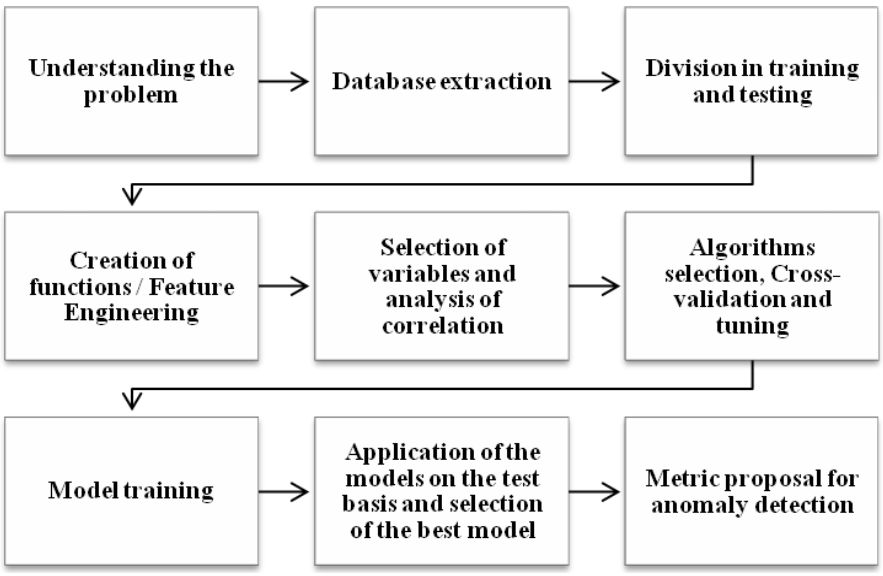

Figure 1: Case study steps

As shown in Figure 1, this research begins with the understanding the problem to be studied, described in item 3.3.1, followed by the data extraction, described in item 3.3.2. and proceeds with the division of the base in training and testing, also detailed in item 3.3.2 .; continuing with the creation of functions and feature engineering, described in item 3.3.3.; moving on to the variable selection and analysis of the correlation, also detailed in item 3.3.3 .; after this stage, the algorithms selection, cross-validation and tuning, described in item 3.3.4; and then the models were trained and the models were applied to the test base to select the best model, also described in item 3.3.4.; and finally, the proposed metric for anomaly detection was elaborated, according to item 3.3.5.

The software used to prepare the database and create the Machine Learning model whole process was the software R. The databases were saved in a repository using the extension .csv.

3.3.1 UNDERSTANDING THE PROBLEM

The objective of this study consists in the creation of a Machine Learning model capable of predicting the response variable: number of medical appointments per day in a company of the supplementary health system in the region of Porto Alegre / RS, and then create anomaly detection metric. The number of medical appointments per day indicator, along with other indicators was created in order to help the responsible department for the company's operational cost be more efficient to identify the events that can contribute to the actual monthly cost exceeds the budget. The development of a Machine Learning model for this indicator will bring agility and precision to the team responsible for the cost. Identifying anomalous behaviors in the next day of the event, it can be understood as fraud, or unforeseen demand that will cause impact in the monthly cost.

3.3.2 BASE EXTRACTION AND DIVISION

A database was used with 563 observations provided by a supplementary health company that operates in the region of Porto Alegre / RS, considering all dates with registered medical appointments referring to the period of January 2018 to July 2019, any identifying information was disregarded in order to ensure data anonymity. This set was selected according to the objective of the study, which is the prediction of the indicator number of medical appointments per day.

The division of the original database in training and testing aims to reserve a portion of the observations to simulate the real conditions for the implementation of the trained models. The training data was composed for the period from January 1, 2018 to December 31, 2018, representing 63.6% of the original base. The test data was created from the remaining information, which is, the period from January 1, 2019 to July 31, 2019, representing the remaining 36.4%.

3.3.3 CREATION OF FUNCTIONS AND FEATURE ENGINEERING

Feature Engineering aimed to create new variables from the date variable. In order to make the process of creating variables reproducible, two functions written in R language were created: “var_date” and “var_lag_diff”, which are exposed in Appendix A and Appendix B.

The "var_date" function returns a data frame with the addition of six new variables, which are: "bus_day", "week", "is_bus", "dist_holiday" and "week_month". The variable “bus_day” is a categorical variable to classify dates as the first working day of the month, second working day, for example, the date 7/1/2018 was a Sunday, so the first working day of the month will be 7/2/2018; “Week” is a categorical variable with seven categories representing the names of the seven days of the week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday); “Is_bus” which is a binary variable, that is, 1 for dates corresponding to business days and 0 for dates corresponding to non-business days; “Dist_holiday” is a continuous variable that measures the distance in days to the nearest holiday, with zero for the date that represents a holiday; “Week_month” is a categorical variable that indicates which week of the month a specific date belongs to. The base of holidays used in the function was obtained from the “bizdays” package (v1.0.6) function “holidaysANBIMA” available on CRAN for software R.

The "var_lag_diff" function returns a data frame with the addition of thirteen new variables, which are: "lag1", "lag2", "lag3", "lag4", "lag5", "lag6", "lag7", "lag14 ”,“ Lag30 ”,“ diff_lag7lag14 ”,“ diff_lag1lag2 ”,“ diff_lag1lag30 ”and“ diff_lag1lag7 ”. The “lagX” variables represent the value of the response variable with X days gaps, for example, “lag2” is the value of the response variable two days ago, and “lag30” is the value of the response variable 30 days ago. The “diff_lagXlagY” variables represent the difference between the “lagX” variable and the “lagY” variable.

In order to identify linear relationships between the lag variables and the response variable, Pearson's linear correlation calculation was performed two by two, as shown in Figure 2. All variables whose Pearson correlation coefficient module was greater or equal to 0.5 were considered satisfactory. After selecting variables by analyzing the Pearson correlation, it was still necessary to verify whether this composition is in fact the best possible, for this reason the Stepwise Backward technique was applied, using the lowest MAE (Mean Absolute Error) value as the selection metric.

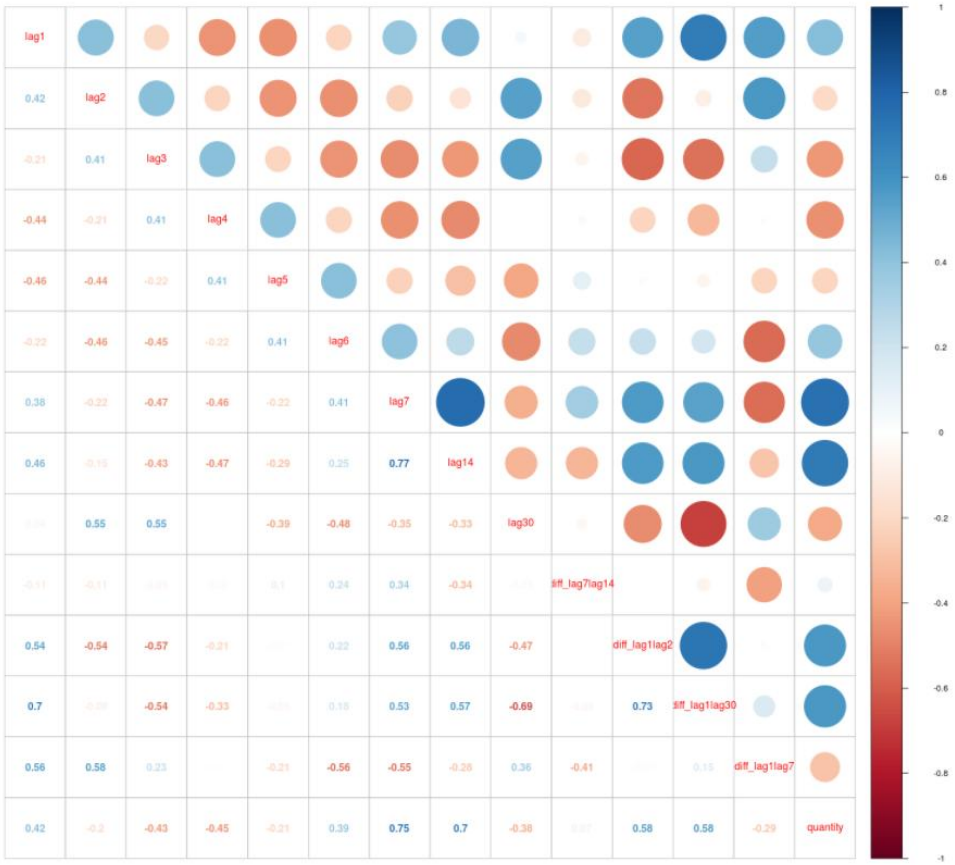

Figure 2: Pearson correlation between variables

3.3.4 ALGORITHMS SELECTION, CROSS-VALIDATION AND TUNING

The Random Forest, Extreme Gradient Boosting, Linear Regression and Neural Network algorithms were selected due to the nature of the response variable, which is continuous, limits the choice of algorithms to the subgroup of techniques defined in the set of supervised learning regression problems, these were the least complex algorithms for implementation and widely used in the literature to solve similar problems. As a cross-validation method, 5-fold was used, and the tuning grids for the Random Forest algorithm were 500 and 1000 for the ntree parameter, which is the number of trees, and 2, 4, 8 and 10 for the parameter mtry representing the number of branches. For the Neural Network algorithm, the tuning grids used were size and decay, size is the number of units in hidden layer and decay is the regularization parameter used to avoid overfitting. It was used as hyperparameters for the Extreme Gradient Boosting algorithm, the eta, max_depth, gamma, and nrounds.

The Random Forest, Extreme Gradient Boosting, Linear Regression and Neural Network algorithms were trained on the training data so that they could later be applied on the test data to select the best model according to the MAE. In total, 1 Linear Regression model, 10 Random Forest models were trained, one for each hyperparameter combination, and 50 Neural Network models, also considering the hyperparameter combination. At the end of the training stage, the Random Forest, Extreme Gradient Boosting, Linear Regression and Neural Network model was selected, whose hyperparameters represented a better MAE, to be applied to the test base.

In order to evaluate and select the best algorithm, among Random Forest, Extreme Gradient Boosting, Linear Regression and Neural Network, the selected models were applied to the test data. The methodologies were compared according to their MAE values and an algorithm was chosen that obtained superior performance.

3.3.5 METRIC PROPOSAL FOR THE DETECTION OF ANOMALIES

After the implementation of the higher performance algorithm, it was possible to compare the real value with the predicted value. Therefore, a metric to detect anomalies was proposed, based on the occasion when the real value is one or two standard deviations higher than the predicted value, indicating anomalous behavior, serving as a warning of possible fraud or irregular behavior.

The definition of alert and anomaly was based on the understanding of the distribution of the difference between the observed and predicted values of the training base, as shown in Figure 3. The calculation of the standard deviation as a metric was used to measure the dispersion of the curve, and the definition of the warning and anomaly points. The standard deviation of the differences between observed and predicted was equal to 244.55.

When analyzing the density of the differences, it was noted that 93% of the observations are less than one standard deviation, that is, it is expected that only 7% of the cases have a difference between observed and predicted greater than 244.55 medical appointments, represented to the right of the yellow line in Figure 3. Likewise, it is noted that 98.5% of the data are less than twice the standard deviation, that is, 1.5% of the observations have a difference between observed and predicted greater than 489.10 medical appointments, represented to the right of the red line in Figure 3.

Thus, it was defined that the detection of suspicious behavior will be through two situations: points above one deviation will be considered "Alerts" and will serve as a warning about possible irregularities, and points above two deviations will be considered "Anomalies", and should have immediate attention.

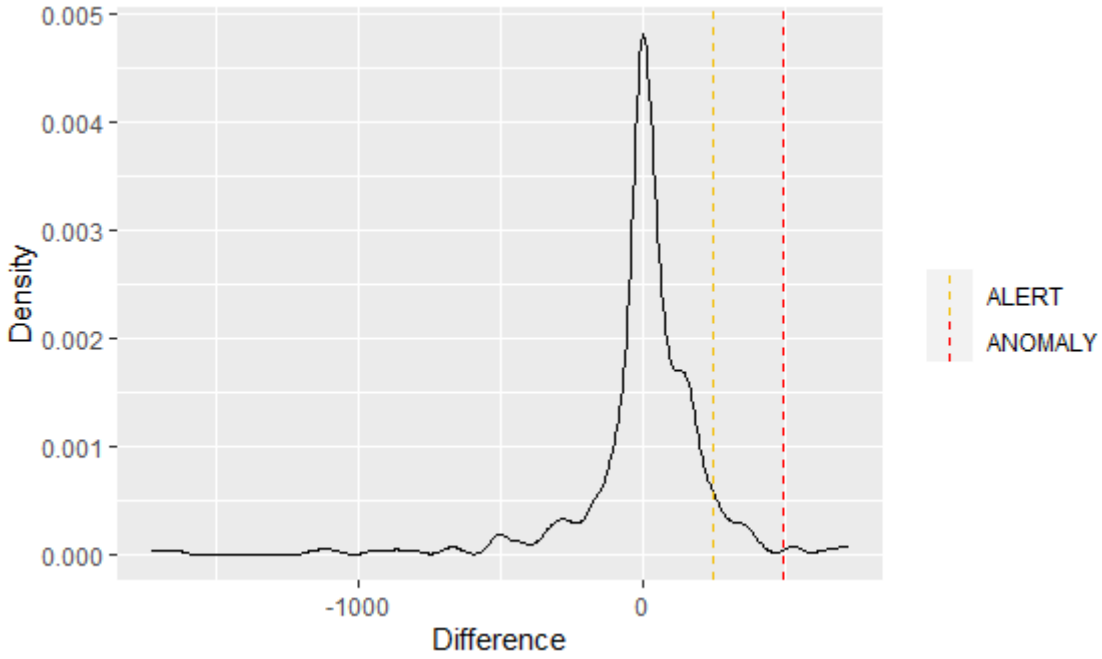

Figure 3: Distribution of the difference between the observed value and the predicted value of the training data

3.4 RESULTS AND DISCUSSION

From the application of the presented methodology, it was possible to create a Machine Learning model, with sufficient accuracy to predict the number of medical appointments per day in the area of supplementary health in the region of Porto Alegre / RS. According to the methodology, the applied steps were: database extraction, division of the base in training and testing, creation of functions and feature engineering, variable selection and analysis of correlation, choice of algorithms with cross-validation and tuning, training of models, application of the models on the test data, selection of the best model and, finally, the metric for anomaly detection was proposed.

The initial database contained two variables, namely the number of medical appointments grouped by date of completion. After dividing the base into training and testing,

the “var_date” function was applied to create new variables. Table 1 shows a sample of the database after the creation of the new variables with the application of the “var_date” function.

Table 1: Variables after applying the “var_date” function

Variable	Format	Example
date_register	Date, format	"2018-01-01" "2018-01-02" "2018-01-03" "2018-01-04" . . .
Quantity	int	1 1168 2104 2380 1478 68 3 2506 2825 2765 . . .
bus_day	num	0 1 2 3 4 5 6 7 8 9 . . .
Week	chr	"Monday Tuesday Wednesday Thursday Friday Saturday" . . .
is_bus	num	0 1 1 1 1 1 0 1 1 1 . . .
dist_holiday	num	0 1 2 3 4 5 6 7 8 9 . . .
week_month	int	6 1 1 1 1 1 1 2 2 2 . . .

A second function was created in order to further increase the number of explanatory variables and thereby also generate a greater understanding of the response variable. Table 2 shows a sample of the database after the creation of the new variables with the application of the “var_lag_diff” function.

Table 2: Variables after applying the “var_lag_diff” function

Variable	Format	Example
date_register	Date, format:	"2018-01-31" "2018-02-01" "2018-02-02" "2018-02-03" . . .
Quantity	int	2539 2215 123 25 2152 2478 2359 2375 1180 30 . . .
bus_day	num	30 0 1 2 4 5 6 7 8 9 . . .
Week	chr	"Monday Tuesday Wednesday Thursday Friday Saturday" . . .
is_bus	num	1 1 1 1 1 1 1 1 1 1 . . .
dist_holiday	num	12 11 10 9 7 6 5 4 3 2 . . .
week_month	int	5 5 1 1 2 2 2 2 2 2 . . .
lag1	int	NA 2539 2215 123 25 2152 2478 2359 2375 1180 . . .
lag2	int	NA NA 2539 2215 123 25 2152 2478 2359 2375 . . .
lag3	int	NA NA NA 2539 2215 123 25 2152 2478 2359 . . .
lag4	int	NA NA NA NA 2539 2215 123 25 2152 2478 . . .
lag5	int	NA NA NA NA NA 2539 2215 123 25 2152 . . .
lag6	int	NA NA NA NA NA 2539 2215 123 25 2152 . . .
lag7	int	NA NA NA NA NA NA NA 2539 2215 123 . . .
lag14	int	NA NA NA NA NA NA NA NA NA NA . . .
lag30	int	NA NA NA NA NA NA NA NA NA NA . . .
diff_lag1lag2	int	NA NA -324 -2092 -98 2127 326 -119 16 -1195 . . .
diff_lag1lag7	int	NA NA NA NA NA NA NA -180 160 1057 . . .

diff_lag7lag14	int	NA NA NA NA NA NA NA NA NA NA . . .
diff_lag1lag30	int	NA NA NA NA NA NA NA NA NA NA . . .

Making use of variables "date_register" and "quantity," it was possible to create 18 new variables, shown in table 2. These variables were subjected to selection techniques, the selection were performed primarily by the Pearson correlation coefficient. As shown in Figure 2, from the criterion, correlation module greater than or equal to 0.5, the selected variables were: “lag7”, “lag14”, “diff_lag1lag2” and “diff_lag1lag30”. It is worth noting that the variables “lag7” and “lag14” have a high correlation with each other (0.77), as well as the variables “diff_lag1lag2” and “diff_lag1lag30” (0.73). In order to avoid possible multicollinearity problems, only one “lag” and “diff” variable was considered, the “lag7” and “diff_lag1lag2”. Both presented the bigger correlation in relation to the response variable, 0.75 and 0.58 respectively.

From this stage, the explanatory variables "bus_day", "week", "is_bus", "dist_holiday", "week_month", "lag7", "diff_lag1lag2", compose the equation to be used to predict the indicator of the number of medical appointments per day. However, to improve the variable selection, the Stepwise Backward technique was applied, using the lowest MAE as the selection metric. The set of variables that obtained the lowest MAE value (237.79) using the Stepwise Backward technique were: “week”, “is_bus”, “week_month” and “diff_lag1lag2”. Therefore, in Equation 1, the final equation to be applied to all algorithms is presented.

$$quantity = week + is_{bus} + week_{month} + diff_{lag1lag2} \quad [1]$$

With the predictor selection phase completed, training began on the two algorithms presented in the methodology, which were compared according to the MAE. In the training phase, the Random Forest algorithm with hyperparameters mtry equal to 6 and ntree equal to 1000 showed a subtle gain in relation to the other models, as can be seen in Table 3. However, the Neural Network algorithm performed much lower when compared to the Extreme Gradient Boosting, Linear Regression and Random Forest algorithms, for this reason the Neural Network algorithm was not applied to the test data.

Table 3: MAE comparison of models applied to the training base

Algorithm	ntree	mtry	size	decay	eta	max_ depth	gamma	n rounds	RMSE	R^2	MAE	RMSE SD	R^2 SD	MAE SD
Linear Regression	NA	NA	NA	NA	NA	NA	NA	NA	414.75	0.89	242.45	71.83	0.04	16.78
Random Forest	500	2	NA	NA	NA	NA	NA	NA	487.63	0.88	360.54	41.67	0.03	31.85
Random Forest	500	4	NA	NA	NA	NA	NA	NA	415.52	0.89	235.36	52.32	0.03	24.55
Random Forest	500	6	NA	NA	NA	NA	NA	NA	418.14	0.89	228.06	57.47	0.03	22.46
Random Forest	500	8	NA	NA	NA	NA	NA	NA	421.76	0.89	228.31	59.25	0.03	21.02
Random Forest	500	10	NA	NA	NA	NA	NA	NA	433.98	0.88	232.74	61.55	0.03	21.84
Random Forest	1000	2	NA	NA	NA	NA	NA	NA	486.21	0.88	359.16	41.72	0.03	31.71
Random Forest	1000	4	NA	NA	NA	NA	NA	NA	416.39	0.89	235.41	53.21	0.03	25.63
Random Forest	**1000**	**6**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**416.41**	**0.89**	**227.19**	**57.15**	**0.03**	**21.00**
Random Forest	1000	8	NA	NA	NA	NA	NA	NA	422.39	0.89	228.24	59.49	0.03	21.08
Random Forest	1000	10	NA	NA	NA	NA	NA	NA	434.56	0.88	232.44	62.22	0.03	22.89
Neural Network	NA	NA	1	0.1	NA	NA	NA	NA	2219.32	0.23	1819.36	18.98	0.16	34.60
Neural Network	NA	NA	1	0.2	NA	NA	NA	NA	2219.32	0.13	1819.36	18.98	0.06	34.60
Neural Network	NA	NA	1	0.3	NA	NA	NA	NA	2219.32	0.17	1819.36	18.98	0.02	34.60
Neural Network	NA	NA	1	0.4	NA	NA	NA	NA	2219.32	0.16	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	1	0.5	NA	NA	NA	NA	2219.32	0.13	1819.36	18.98	0.06	34.60
Neural Network	NA	NA	2	0.1	NA	NA	NA	NA	2219.32	0.10	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	2	0.2	NA	NA	NA	NA	2219.32	0.13	1819.36	18.98	0.06	34.60
Neural Network	NA	NA	2	0.3	NA	NA	NA	NA	2219.32	0.12	1819.36	18.98	0.09	34.60
Neural Network	NA	NA	2	0.4	NA	NA	NA	NA	2219.32	0.14	1819.36	18.98	0.06	34.60
Neural Network	NA	NA	2	0.5	NA	NA	NA	NA	2219.32	0.14	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	3	0.1	NA	NA	NA	NA	2219.32	0.27	1819.36	18.98	0.25	34.60
Neural Network	NA	NA	3	0.2	NA	NA	NA	NA	2219.32	0.15	1819.36	18.98	0.05	34.60
Neural Network	NA	NA	3	0.3	NA	NA	NA	NA	2219.32	0.18	1819.36	18.98	0.12	34.60
Neural Network	NA	NA	3	0.4	NA	NA	NA	NA	2219.32	0.22	1819.36	18.98	0.20	34.60
Neural Network	NA	NA	3	0.5	NA	NA	NA	NA	2219.32	0.16	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	4	0.1	NA	NA	NA	NA	2219.32	0.13	1819.36	18.98	0.06	34.60
Neural Network	NA	NA	4	0.2	NA	NA	NA	NA	2219.32	0.17	1819.36	18.98	0.17	34.60
Neural Network	NA	NA	4	0.3	NA	NA	NA	NA	2219.32	0.32	1819.36	18.98	0.30	34.60
Neural Network	NA	NA	4	0.4	NA	NA	NA	NA	2219.32	0.10	1819.36	18.98	0.07	34.60
Neural Network	NA	NA	4	0.5	NA	NA	NA	NA	2219.32	0.27	1819.36	18.98	0.23	34.60
Neural Network	NA	NA	5	0.1	NA	NA	NA	NA	2219.32	0.12	1819.36	18.98	0.09	34.60
Neural Network	NA	NA	5	0.2	NA	NA	NA	NA	2219.32	0.18	1819.36	18.98	0.12	34.60
Neural Network	NA	NA	5	0.3	NA	NA	NA	NA	2219.32	0.07	1819.36	18.98	0.07	34.60
Neural Network	NA	NA	5	0.4	NA	NA	NA	NA	2219.32	0.11	1819.36	18.98	0.07	34.60
Neural Network	NA	NA	5	0.5	NA	NA	NA	NA	2219.32	0.16	1819.36	18.98	0.11	34.60
Neural Network	NA	NA	6	0.1	NA	NA	NA	NA	2219.32	0.14	1819.36	18.98	0.04	34.60

Neural Network	NA	NA	6	0.2	NA	NA	NA	NA	2219.32	0.17	1819.36	18.98	0.24	34.60
Neural Network	NA	NA	6	0.3	NA	NA	NA	NA	2219.32	0.16	1819.36	18.98	0.09	34.60
Neural Network	NA	NA	6	0.4	NA	NA	NA	NA	2219.32	0.23	1819.36	18.98	0.20	34.60
Neural Network	NA	NA	6	0.5	NA	NA	NA	NA	2219.32	0.18	1819.36	18.98	0.16	34.60
Neural Network	NA	NA	7	0.1	NA	NA	NA	NA	2219.32	0.10	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	7	0.2	NA	NA	NA	NA	2219.32	0.13	1819.36	18.98	0.09	34.60
Neural Network	NA	NA	7	0.3	NA	NA	NA	NA	2219.32	0.11	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	7	0.4	NA	NA	NA	NA	2219.32	0.33	1819.36	18.98	0.26	34.60
Neural Network	NA	NA	7	0.5	NA	NA	NA	NA	2219.32	0.21	1819.36	18.98	0.10	34.60
Neural Network	NA	NA	8	0.1	NA	NA	NA	NA	2219.32	0.25	1819.36	18.98	0.22	34.60
Neural Network	NA	NA	8	0.2	NA	NA	NA	NA	2219.32	0.14	1819.36	18.98	0.07	34.60
Neural Network	NA	NA	8	0.3	NA	NA	NA	NA	2219.32	0.09	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	8	0.4	NA	NA	NA	NA	2219.32	0.26	1819.36	18.98	0.21	34.60
Neural Network	NA	NA	8	0.5	NA	NA	NA	NA	2219.32	0.08	1819.36	18.98	0.06	34.60
Neural Network	NA	NA	9	0.1	NA	NA	NA	NA	2219.32	0.10	1819.36	18.98	0.08	34.60
Neural Network	NA	NA	9	0.2	NA	NA	NA	NA	2219.32	0.12	1819.36	18.98	0.11	34.60
Neural Network	NA	NA	9	0.3	NA	NA	NA	NA	2219.32	0.12	1819.36	18.98	0.11	34.60
Neural Network	NA	NA	9	0.4	NA	NA	NA	NA	2219.32	0.09	1819.36	18.98	0.09	34.60
Neural Network	NA	NA	9	0.5	NA	NA	NA	NA	2219.32	0.29	1819.36	18.98	0.27	34.60
Neural Network	NA	NA	10	0.1	NA	NA	NA	NA	2219.32	0.13	1819.36	18.98	0.07	34.60
Neural Network	NA	NA	10	0.2	NA	NA	NA	NA	2219.32	0.11	1819.36	18.98	0.05	34.60
Neural Network	NA	NA	10	0.3	NA	NA	NA	NA	2219.32	0.20	1819.36	18.98	0.17	34.60
Neural Network	NA	NA	10	0.4	NA	NA	NA	NA	2219.32	0.16	1819.36	18.98	0.17	34.60
Neural Network	NA	NA	10	0.5	NA	NA	NA	NA	2219.32	0.08	1819.36	18.98	0.08	34.60
Extreme Gradient Boosting	NA	NA	NA	NA	0.05	5	0.01	50	476.52	0.88	327.63	57.42	0.03	31.75
Extreme Gradient Boosting	NA	NA	NA	NA	0.05	5	0.01	100	424.77	0.89	234.78	55.26	0.03	21.19
Extreme Gradient Boosting	NA	NA	NA	NA	0.05	5	0.01	200	432.30	0.88	232.29	55.22	0.03	19.50
Extreme Gradient Boosting	NA	NA	NA	NA	0.05	5	0.01	300	443.89	0.88	237.94	58.38	0.03	24.20

After defining the best combination of hyperparameters for the Random Forest algorithm and Extreme Gradient Boosting, these, together with the Linear Regression, were applied to the test data, in order to identify which algorithm would present a better performance when simulated real implementation conditions. The Random Forest model performed better than the Linear Regression and Extreme Gradient Boosting model when applied to the test base, as shown in Table 4.

Table 4: MAE comparison of the models applied to the test base

Algorithm	ntree	mtry	eta	max_depth	gamma	nrounds	MAE
Random Forest	**1000**	**6**	NA	NA	NA	NA	**182.89**
Linear Regression	NA	NA	NA	NA	NA	NA	216.44
Extreme Gradient Boosting	NA	NA	0.05	5	0.01	200	196.53

It can also be seen from Figures 3 and 4, which compare the real values with the predicted values, that the selected model was able to predict the number of medical appointment in a satisfactory way both in the training data and on the test data.

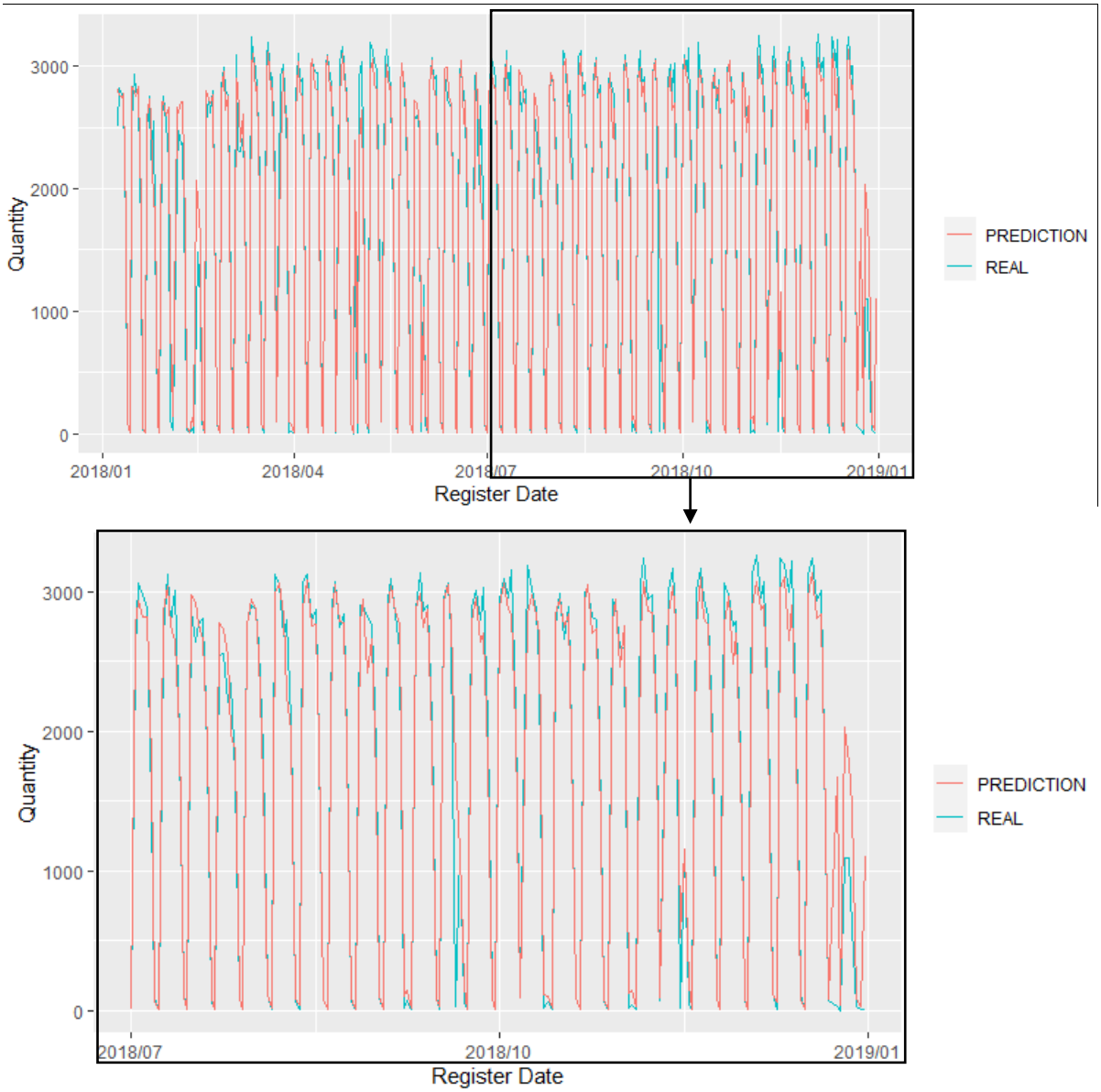

Figure 4: Comparison of the prediction of the selected model with the real value in the training base

After the prediction of the variable number of medical appointments, the anomalies and alerts identification metric that is defined by the rule was applied: If the real value is greater than the predicted value plus a standard deviation, the occurrence will be defined as an alert, if the real value is greater than the predicted value plus two standard deviations, the occurrence will be defined as anomalous. According to Figures 5 and 6, nine anomaly situations can be identified in the test data, on March 25, 2019, March 27, 2019, March 29, 2019, April 26, 2019, May 2, 2019, May 9, 2019, May 24, 2019, June 24, 2019 and July 29, 2019, in addition to thirty-eight alert points during the period from January to July. The prior identification of these points, as well as the investigation of the reason for such behaviors, can bring gains both in the identification of frauds, as well as in the opportunity to improve the provision of the service through a better distribution of resources.

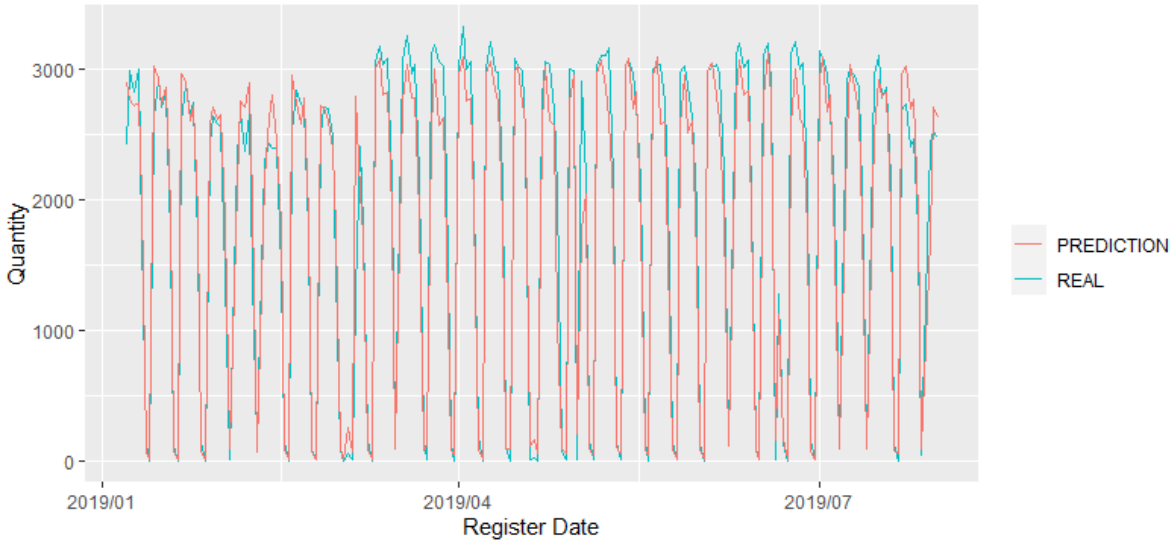

Figure 5: Comparison of the prediction of the selected model with the real value in the test base

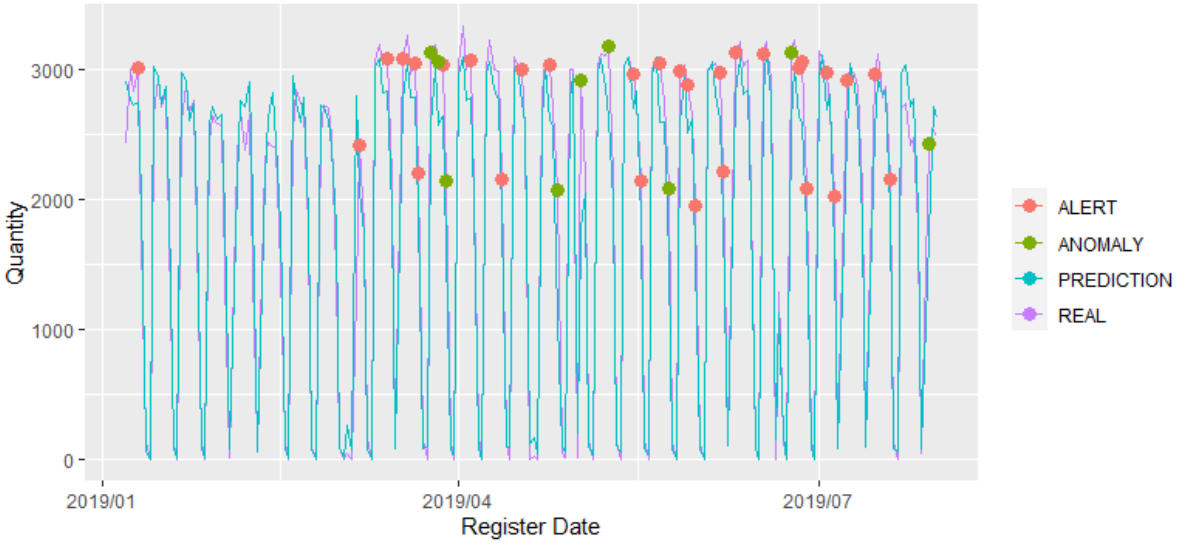

Figure 6: Identification of anomalous occurrence through standard deviation

3.5 CONCLUSION

The study provided a broad theoretical and applied understanding of important steps for the creation of a Machine Learning model (collecting and preparing the data, selecting the variables, choice of algorithm, selection of hyperparameters, training of the model and evaluation of results). The creation of Feature Engineering was an essential step to understanding the behavior of the data, and combined with the development of auxiliary functions made it possible to reproduce in a more efficient and fast way this very important part of this Machine Learning process. Another essential point for the creation of a high-performance model was the step of selecting the variables, because using the techniques of analysis of the correlations followed by the Stepwise Backward selection, it was possible to identify the predictors with the greatest impact on the response variable preventing overfitting.

With the functions created, the implementation, maintenance, or replication of the methodology as a whole becomes more simplified, this is easily replicable for other similar studies whose goals are the prediction of a continuous variable using as features any kind of time frames (daily, monthly, annually).

The indicator of the number of medical appointments per day is an important indicator for the area of supplementary health because this kind of KPI is strongly correlated to the operational costs and to predict it improves the decision making. With the Random Forest

model selected and the proposed metric for detecting anomalies through the standard deviation, it was possible not only to predict the KPI results with almost 90% of accuracy, but also to compare the prediction with the real value producing alerts of anomalous occurrences. This can enable the investigation of possible problems, or new demands, in order to improve health services.

3.6 REFERENCES

BISHOP, Christopher M. Pattern recognition and machine learning. Springer, 2006.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms**. Big Data & Society**, v. 3, n. 1, p. 2053951715622512, 2016.

DE ARAÚJO, Flávio Henrique Duarte; SANTANA, André Macedo; DOS SANTOS NETO, Pedro de Alcântara. Uma Abordagem Influenciada por Pré-processamento para Aprendizagem do Processo de Regulação Médica. **Journal of Health Informatics**, v. 7, n. 1, 2015.

DOMINGOS, Pedro M. A few useful things to know about machine learning. **Commun. acm**, v. 55, n. 10, p. 78 - 87, 2012.

GARCÍA, Enrique et al. Drawbacks and solutions of applying association rule mining in learning management systems. **Proceedings of the International Workshop on Applying Data Mining in e-Learning,** p. 13 - 22, 2007.

GARLA, Vijay N.; BRANDT, Cynthia. Ontology-guided feature engineering for clinical text classification. **Journal of biomedical informatics**, v. 45, n. 5, p. 992 - 998, 2012.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2013). **An introduction to statistical learning (Vol. 112, p. 18)**. New York, Springer, 2013

KRASKA, Tim et al. MLbase: A Distributed Machine-learning System. **CIDR**. p. 2.1, 2013

LEE, Polly Po Yee et al. **Interactive interfaces for machine learning model evaluations**. U.S. Patent n. 10, 452, 992, 22 out. 2019.

MARSLAND, Stephen. Machine learning: an algorithmic perspective. Chapman and Hall/CRC, 2014.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. Foundations of machine learning. **MIT press**, 2018.

MURDOCH, W. James et al. Definitions, methods, and applications in interpretable machine learning. **Proceedings of the National Academy of Sciences**, v. 116, n. 44, p. 22071 - 22080, 2019.

REFAEILZADEH, Payam; TANG, Lei; LIU, Huan. Cross-validation. **Encyclopedia of database systems**, p. 532-538, 2009.

RODRIGUEZ-GALIANO, V. et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804 - 818, 2015.

VINYALS, Oriol; DEAN, Jeffrey A.; HINTON, Geoffrey E. **Training distilled machine learning models**. U.S. Patent n. 10, 289, 962, 14 maio 2019.

3.7 APPENDIX A - 'VAR_DATE' FUNCTION (R CODE)

```
# Required packges
require(dplyr)
require(bizdays)
require(purrr)
require(lubridate)

#--- FUNCTION ---#
var_date <- function(db, col_name_date, format_date = "%d/%m/%Y", feriados = bizdays::holidaysANBIMA)
{

 # auxiliar function for calculate the distance to a holiday
 dist_holiday<-function(x, y = feriados)n
 {
  # distance between two dates
  menor_dist <- purrr::map2(x, y, difftime)
  db_min <- min(abs(unlist(menor_dist)))
  return(db_min)
 }

 db <- db %>%
  mutate(
   dt_ok = as.Date(get(col_name_date), format = format_date),
   first_day = ymd(format(dt_ok, "%Y-%m-01")),
   week = weekdays(dt_ok),
   is_bus = case_when(
    week == "sábado" | week == "domingo" | dt_ok %in% feriados ~ 0,
    TRUE ~ 1
   ),
   dist_holiday = unlist(purrr::map(.x = dt_ok, .f = dist_holiday)),
   week_month = stringi::stri_datetime_fields(get(col_name_date), tz = 'Etc/GMT-3')$WeekOfMonth
  )

 # creating the business day of the month
 temp <- db %>%
  filter(is_bus == 1) %>%
  mutate(
   var_temp = 1
  ) %>%
  group_by(first_day) %>%
  mutate(
   bus_day = cumsum(var_temp)) %>%
  ungroup() %>%
  select(dt_ok, bus_day)

 db <- db %>%
  left_join(temp, by = c("dt_ok" = "dt_ok")) %>%
  mutate(bus_day = if_else(is.na(bus_day),0,bus_day))
```

```
 return(db)

}
```

3.8 APPENDIX B - 'VAR_LAG_DIFF' FUNCTION (CODE IN R)

```
# Required packges
require(dplyr)
require(lubridate)
require(corrplot)

#--- FUNCTION ---#
var_lag_diff <- function(db, col_name_date, target_var, reference_value = 0.2)
{
 db <- db %>%
   arrange(get(col_name_date)) %>%
   mutate(
     lag1 = lag(get(target_var), 1),
     lag2 = lag(get(target_var), 2),
     lag3 = lag(get(target_var), 3),
     lag4 = lag(get(target_var), 4),
     lag5 = lag(get(target_var), 5),
     lag6 = lag(get(target_var), 6),
     lag7 = lag(get(target_var), 7),
     lag14 = lag(get(target_var), 14),
     lag30 = lag(get(target_var), 30),
     diff_lag7lag14 = lag7 - lag14,
     diff_lag1lag2 = lag1 - lag2,
     diff_lag1lag30 = lag1 - lag30,
     diff_lag1lag7 = lag1 - lag7
     )
 aux <- db %>%
   filter(!is.na(lag30)) %>%
   select(
     lag1,
     lag2,
     lag3,
     lag4,
     lag5,
     lag6,
     lag7,
     lag14,
     lag30,
     diff_lag7lag14,
     diff_lag1lag2,
     diff_lag1lag30,
     diff_lag1lag7,
     target_var)

 # Function return list
 list_return <- as.list(NULL)

 # return 1
```

```
  ## plot das correlacoes dos lags e diffs e a variavel target
  correl <- cor(aux)
  list_return[[1]] <- correl

  # return 2
  ## data frame with lag or diff variables that were bigger then predetermined amount

  correl <- as.data.frame(correl)
  nome_linha <- row.names(correl)

  correl <- correl %>%
    mutate(linha = nome_linha) %>%
    filter(abs(get(target_var)) >= reference_value)

  selection <- c(correl$linha, 'linha')
  tirar <- colnames(correl)[!colnames(correl) %in% selection]

  list_return[[2]] <- db

  db <- db %>%
    select(-tirar)

  list_return[[3]] <- db

  return(list_return)
}
```

CONCLUSÃO

Com os resultados obtidos em ambos os artigos é possível destacar que o objetivo geral pretendido com este trabalho foi alcançado, ou seja, foi possível identificar os tipos de algoritmos de *Machine Learning* mais utilizados nas áreas de Engenharia e Medicina, além de realizar a aplicação destes algoritmos a um problema real de predição de um indicador, e assim propor um método para detecção de anomalias. Os objetivos específicos também foram alcançados, visto que foi possível identificar os tipos de problemas, métodos de validação, além de descrever o processo de criação e aplicação de um modelo de *Machine Learning*.

Os dois artigos demonstraram a ampla aplicabilidade dos algoritmos de *Machine Learning*, e como é possível utilizar os mesmos métodos para diversos tipos de problemas. Os artigos apresentados se complementam, pois através da revisão de literatura, presente no primeiro artigo, foi possível trazer ao conhecimento contribuições importantes dos autores para a criação e implementação de um Modelo de *Machine Learning*. Sendo assim, o primeiro artigo trás o embasamento teórico que justifica a aplicação dos algoritmos *Random Forest*, *Linear Regression, Extreme Gradient Boosting* e *Neural Network* aplicados para a solução do problema apresentado no segundo artigo.

Além de apresentar os algoritmos mais presentes nas áreas de Engenharia e Medicina, o primeiro artigo também expos os métodos de validação mais utilizados, R^2 (*R-squared*), MAE (*Mean Absolute Error*) e MAPE (*Mean Absolute Percentage Error*). Com a revisão de literatura foi possível identificar algumas contribuições importantes apresentadas pelos autores, tais como: criação de modelos para auxiliar na tomada de decisão, criação de modelos de classificação com base em uma serie histórica, modelos não supervisionados utilizando as redes sociais, comparação de algoritmos, simplificação dos modelos através de *Machine Learning*, otimização dos recursos computacionais, criação de modelos para auxiliar na tomada de decisão, comparação de algoritmos com inclusão de indicadores, utilização de algoritmos não supervisionados e supervisionados em conjunto e a inclusão de resultados de outros modelos como vetor para o novo modelo. Outro ponto importante foi a realização de comparação entre algoritmos, pois dependendo dos recursos computacionais disponíveis, tipo de dados, tipo de variável resposta, um modelo pode apresentar um desempenho superior aos demais.

O segundo artigo, embasado nas contribuições trazidas pelo primeiro artigo, descreveu todos os passos realizado para a criação de um modelo, desde a coleta dos dados, até a escolha e aplicação do modelo de melhor performance. O artigo também ressaltou a importância dos processos de seleção de variáveis evitando assim problemas de multi-colinearidade, além de demonstrar o processo de *Feature Engeneering* com a criação das funções 'VAR_LAG_DIFF', 'VAR_DATE', onde foi possível criar através de uma única variável (Data_register) dezoito novas variáveis utilizadas na construção de um modelo mais performático. Foi também apresentado um método de detecção de anomalias para o indicador gerencial "quantidade de consultas médicas por dia" através do estudo da distribuição da diferença entre predito e observado. Identificar previamente as ocorrências anômalas proporciona ao setor gerencial analisar e identificar de maneira mais ágil as possíveis causas, aprimorando o serviço prestado, prevenindo fraude médica ou até mesmo identificando oportunidades de melhoria.

A metodologia aplicada nos dois artigos pode ser facilmente replicada, e todos os códigos utilizados na construção deste estudo estão disponíveis no repositório pessoal da autora no Github (https://github.com/mirelemborges/dissertation_codes).

Em termos de trabalhos futuros, sugere-se a implementação deste modelo para outros indicadores da área da saúde e assim confrontar os resultados obtidos, além de realizar um aprofundamento nas demais técnicas de *Machine Learning*, principalmente um estudo com foco em *Feature Engineering*, pois este se mostrou um importante recurso quando se obtém uma base de dados com poucas variáveis explicativas. Sugere-se também realizar a comparação de desempenho de outros algoritmos, trazer ao foco a análise da explicabilidade da predição incluindo um entendimento da matemática e dos recursos computacionais. Sugere-se ainda, reproduzir a revisão de literatura, incorporando outras áreas do conhecimento, e assim alcançar um entendimento maior sobre o tema.

REFERÊNCIAS

AHMAD, Tanveer; CHEN, Huanxin; HUANG, Yao. Short-Term Energy Prediction for District-Level Load Management Using Machine Learning Based Approaches. **Energy Procedia**, v. 158, p. 3331-3338, 2019.

BATTINENI, Gopi; CHINTALAPUDI, Nalini; AMENTA, Francesco. Machine Learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). **Informatics in Medicine Unlocked**, v. 16, p. 100200, 2019.

BENTO, António. Como fazer uma revisão da literatura: Considerações teóricas e práticas. **Revista JA (Associação Académica da Universidade da Madeira)**, v. 7, n. 65, p. 42-44, 2012.

BIRJALI, Marouane; BENI-HSSANE, Abderrahim; ERRITALI, Mohammed. Machine Learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. **Procedia Computer Science**, v. 113, p. 65-72, 2017.

BISHOP, Christopher M. Pattern recognition and machine learning. Springer, 2006.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms**. Big Data & Society**, v. 3, n. 1, p. 2053951715622512, 2016.

CAMPOS, Tulio L. et al. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. **Computational and Structural Biotechnology Journal**, v. 17, p. 785-796, 2019.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, v. 41, n. 3, p. 1-58, 2009.

CHENG, Jiatang; XIONG, Yan. Application of extreme learning machine combination model for dam displacement prediction. **Procedia Computer Science**, v. 107, p. 373-378, 2017.

DA SILVA, Emerson Correia; LENGLER, Fernando Ramos. A PRODUÇÃO CIENTÍFICA SOBRE MACHINE LEARNING NA ÁREA EDUCACIONAL NO BRASIL (1999-2017). **CADERNOS DE INICIAÇÃO CIENTÍFICA**, v. 2, n. 1, 2017.

DE ARAÚJO, Flávio Henrique Duarte; SANTANA, André Macedo; DOS SANTOS NETO, Pedro de Alcântara. Uma Abordagem Influenciada por Pré-processamento para Aprendizagem do Processo de Regulação Médica. **Journal of Health Informatics**, v. 7, n. 1, 2015.

DING, Chao; LAM, Khee Poh; FENG, Wei. An evaluation index for cross ventilation based on CFD simulations and ventilation prediction model using Machine Learning algorithms. **Procedia engineering**, v. 205, p. 2948-2955, 2017.

DOMINGOS, Pedro M. A few useful things to know about machine learning. **Commun. acm**, v. 55, n. 10, p. 78 - 87, 2012.

DOMINGOS, Pedro. **O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo**. Novatec Editora, 2017.

DOS SANTOS, Hellen Geremias et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cad. Saúde Pública**, v. 35, n. 7, p. e00050818, 2019.

ELANGOVAN, M. et al. Machine Learning approach to the prediction of surface roughness using statistical features of vibration signal acquired in turning. **Procedia Computer Science**, v. 50, p. 282-288, 2015.

FRENCH, Jon et al. Combining Machine Learning with computational hydrodynamics for prediction of tidal surge inundation at estuarine ports. **Procedia IUTAM**, v. 25, p. 28-35, 2017.

FUNKNER, Anastasia; KOVALCHUK, Sergey; BOCHENINA, Klavdiya. Preoperational Time Prediction for Percutaneous Coronary Intervention Using Machine Learning Techniques. **Procedia Computer Science**, v. 101, p. 172-176, 2016.

GARCÍA, Enrique et al. Drawbacks and solutions of applying association rule mining in learning management systems. **Proceedings of the International Workshop on Applying Data Mining in e-Learning,** p. 13 - 22, 2007.

GARLA, Vijay N.; BRANDT, Cynthia. Ontology-guided feature engineering for clinical text classification. **Journal of biomedical informatics**, v. 45, n. 5, p. 992 - 998, 2012.

GEYER, Philipp; SINGARAVEL, Sundaravelpandian. Component-based Machine Learning for performance prediction in building design. **Applied energy**, v. 228, p. 1439-1453, 2018.

GOUARIR, A. et al. In-process tool wear prediction system based on Machine Learning techniques and force analysis. **Procedia CIRP**, v. 77, p. 501-504, 2018.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2013). **An introduction to statistical learning**.v. 112, New York, Springer, 2013.

KHAIRI, Mutaz HH et al. A review of anomaly detection techniques and distributed denial of service (DDoS) on software defined network (SDN). **Engineering, Technology & Applied Science Research**, v. 8, n. 2, p. 2724-2730, 2018.

KOUROU, Konstantina et al. Machine Learning applications in cancer prognosis and prediction. **Computational and structural biotechnology journal**, v. 13, p. 8-17, 2015.

KRASKA, Tim et al. MLbase: A Distributed Machine-learning System. **CIDR**. p. 2.1, 2013.

KREUTZ, Markus et al. Machine Learning-based icing prediction on wind turbines. **Procedia CIRP**, v. 81, p. 423-428, 2019.

LEE, Polly Po Yee et al. **Interactive interfaces for machine learning model evaluations**. U.S. Patent n. 10, 452, 992, 22 out. 2019.

LEI, Yu; ZHAO, Danning; CAI, Hongbing. Prediction of length-of-day using extreme learning machine. **Geodesy and Geodynamics**, v. 6, n. 2, p. 151-159, 2015.

LEIGHTON, Samuel P. et al. Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a Machine Learning approach. **The Lancet Digital Health**, v. 1, n. 6, p. e261-e270, 2019.

MARSLAND, Stephen. Machine learning: an algorithmic perspective. Chapman and Hall/CRC, 2014.

MATHUR, Neha; GLESK, Ivan; BUIS, Arjan. Comparison of adaptive neuro-fuzzy inference system (ANFIS) and Gaussian processes for Machine Learning (GPML) algorithms for the

prediction of skin temperature in lower limb prostheses. **Medical engineering & physics**, v. 38, n. 10, p. 1083-1089, 2016.

MITCHELL, Tom Michael. **The discipline of machine learning**. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. **Foundations of machine learning**. MIT press, 2018.

MOSCHOVAKIS, Yiannis N. What is an algorithm? In: **Mathematics unlimited—2001 and beyond**. Springer, Berlin, Heidelberg, 2001. p. 919-936.

MOZAFFARI, Ladan; MOZAFFARI, Ahmad; AZAD, Nasser L. Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on San Francisco urban roads. **Engineering science and technology, an international journal**, v. 18, n. 2, p. 150-162, 2015.

MURDOCH, W. James et al. Definitions, methods, and applications in interpretable machine learning. **Proceedings of the National Academy of Sciences**, v. 116, n. 44, p. 22071-22080, 2019.

PURANIK, Shruthi; DESHPANDE, Pranav; CHANDRASEKARAN, K. A novel Machine Learning approach for bug prediction. **Procedia Computer Science**, v. 93, p. 924-930, 2016.

RAJ, S. Sridhar; NANDHINI, M. Ensemble human movement sequence prediction model with Apriori based Probability Tree Classifier (APTC) and Bagged J48 on Machine Learning. **Journal of King Saud University-Computer and Information Sciences**, v. 31, p. 55-62, 2018.

REFAEILZADEH, Payam; TANG, Lei; LIU, Huan. Cross-validation. **Encyclopedia of database systems**, p. 532-538, 2009.

RODRIGUEZ-GALIANO, V. et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804 - 818, 2015.

SUTTON, Richard S.; BARTO, Andrew G. Reinforcement learning: An introduction. **MIT press**, 2018.

VINYALS, Oriol; DEAN, Jeffrey A.; HINTON, Geoffrey E. **Training distilled machine learning models**. U.S. Patent n. 10, 289, 962, 14 maio 2019.

WANG, Bo; KIM, Inhi. Short-term prediction for bike-sharing service using Machine Learning. **Transportation research procedia**, v. 34, p. 171-178, 2018.

YIN, R. K. Estudo de Caso _ Planejamento e Método. 2. ed. São Paulo: **Bookman**, 2001.