UNIVERSIDADE FEDERAL DE GOIÁS
UNIDADE ACADÊMICA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
MAÍZA BIAZON DE OLIVEIRA

PREVISÃO DO LEAD TIME DE PROCESSOS USANDO MINERAÇÃO DE DADOS

Catalão
2021

UNIVERSIDADE FEDERAL DE GOIÁS
UNIDADE ACADÊMICA ESPECIAL DE ENGENHARIA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese

2. Nome completo do autor

Maíza Biazon de Oliveira

3. Título do trabalho

PREVISÃO DO LEAD TIME DE PROCESSOS USANDO MINERAÇÃO DE DADOS

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Nubia Rosa da Silva Guimarães**, **Coordenadora de**

Pós-Graduação, em 06/04/2021, às 13:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **MAÍZA BIAZON DE OLIVEIRA**, **Discente**, em 06/04/2021, às 22:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1985589** e o código CRC **E58F04A9**.

Referência: Processo nº 23070.012643/2021-19 SEI nº 1985589

MAÍZA BIAZON DE OLIVEIRA

PREVISÃO DO LEAD TIME DE PROCESSOS USANDO MINERAÇÃO DE DADOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Goiás, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção.

Área de Concentração: Engenharia de Operações e Processos Industriais

Orientadora: Profa. Dra. Núbia Rosa da Silva

Co-Orientador: Prof. Dr. Douglas Farias Cordeiro

Catalão
2021

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Oliveira, Maíza Biazon de
Previsão do Lead Time de Processos usando Mineração de Dados [manuscrito] / Maíza Biazon de Oliveira. - 2021.
81 f.

Orientador: Profa. Dra. Núbia Rosa da Silva Guimarães; co orientador Dr. Douglas Farias Cordeiro.
Dissertação (Mestrado) - Universidade Federal de Goiás, Unidade Acadêmica Especial de Engenharia e Administração, Programa de Pós graduação em Engenharia de Produção, Catalão, 2021.

1. Lead time. 2. Mineração de dados. 3. Inteligência artificial. I. Guimarães, Núbia Rosa da Silva, orient. II. Título.

CDU 658.5

UNIVERSIDADE FEDERAL DE GOIÁS

UNIDADE ACADÊMICA ESPECIAL DE ENGENHARIA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **21** da sessão de Defesa de Dissertação de **MAÍZA BIAZON DE OLIVEIRA**, que confere o título de Mestra em **Engenharia de Produção**, na área de concentração em **Engenharia de Operações e Processos Industriais.**

"Banca Examinadora de Qualificação/Defesa Pública de Dissertação/Tese realizada em conformidade com a Portaria da CAPES n. 36, de 19 de março de 2020, de acordo com seu segundo artigo:

Art. 2o A suspensão de que trata esta Portaria não afasta a possibilidade de defesas de tese utilizando tecnologias de comunicação à distância, quando admissíveis pelo programa de pós-graduação stricto sensu, nos termos da regulamentação do Ministério da Educação."

Aos **nove dias do mês de março do ano de dois mil e vinte e um**, a partir das **nove horas**, na Sala Virtual (https://meet.google.com/hao-zuhz-qqv?hs=122&authuser=1), realizou-se a sessão pública de Defesa de Dissertação intitulada **"PREVISÃO DO LEAD TIME DE PROCESSOS USANDO MINERAÇÃO DE DADOS"**. Os trabalhos foram instalados pela Orientadora, Professora Doutora **NÚBIA ROSA DA SILVA (PPGEP/ UFG)** com a participação dos demais membros da Banca Examinadora: Professor Doutor **CARLOS ANTONIO RIBEIRO DUARTE (PPGEP/ UFG)**, membro titular interno, cuja participação ocorreu via videoconferência; Professor Doutor **SERGIO FRANCISCO DA SILVA (PPGCOM/UFG)**, cuja participação ocorreu via videoconferência, membro titular externo e do coorientador **DOUGLAS FARIAS CORDEIRO (PPGEP/ UFG)**, cuja participação ocorreu via videoconferência. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido a candidata **aprovada** pelos seus membros. Proclamados os resultados pela Professora Doutora **NÚBIA ROSA DA SILVA**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos **nove dias do mês de março do ano de dois mil e vinte e um.**

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Nubia Rosa da Silva Guimarães**, **Professora do Magistério Superior**, em 17/03/2021, às 14:36, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carlos Antonio Ribeiro Duarte**, **Professor do Magistério Superior**, em 17/03/2021, às 15:57, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Sérgio Francisco Da Silva**, **Professor do Magistério Superior**, em 22/03/2021, às 20:03, conforme horário oficial de Brasília, com fundamento no art. 6°, § 1°, do Decreto n° 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Douglas Farias Cordeiro**, **Professor do Magistério Superior**, em 23/03/2021, às 21:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1927265** e o código CRC **B82C8FE1**.

Referência: Processo nº 23070.012643/2021-19 SEI nº 1927265

A Deus, aos meus pais e minha irmã, pelo incentivo e apoio.

AGRADECIMENTOS

Agradecimentos

A Deus por sempre me ajudar e tornar meus sonhos em realidade. A meus pais, Iolita e Ademir pelo apoio, incentivo e suporte antes e durante o mestrado.

A minha irmã Monique por me apoiar, incentivar e ajudar durante minha jornada acadêmica e pós acadêmica.

A minha professora orientadora Núbia Rosa, pela excelente orientação, direcionamento, suporte e amizade que foram decisivos para tornar este trabalho e o intercâmbio realidade e conhecimento de técnicas de Ciência da Computação para aplicação na Engenharia.

A meu professor co-orientador Douglas Farias pelo excelente suporte, orientação e amizade que foram decisivos para que eu pudesse aprender novas ferramentas de análise de dados e tornar esse trabalho realidade.

Aos professores Manuel Iori e Marco Lippi que me receberam e excelentemente co-orientaram durante o período de intercâmbio. Aos integrantes da banca de qualificação Carlos Ribeiro, Núbia Rosa e Douglas Cordeiro pelas as importantes considerações que contribuíram para a realização deste trabalho.

Aos professores do mestrado, Núbia Rosa, André Alves, José Waldo, Carlos Ribeiro, Marco Paulo, Stella Jacyszyn que ministraram as aulas do mestrado e proporcionaram encorajamento para todos os alunos.

A minha amiga Débora Paula Cechin, que me ofereceu apoio e suporte durante o mestrado.

Agradecimento individual às agências de fomento, que financiaram essa pesquisa.

A Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), Conselho Nacional das Fundações Estaduais de Amparo à Pesquisa (CONFAP-ITALLY) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

RESUMO

BIAZON, M.. **Previsão do *Lead Time* de processos usando Mineração de Dados**. 2021. 83 f. Dissertação de Mestrado (Mestrado em Comunicação) – Universidade Federal de Goiás (UFG), Goiânia – GO.

A era da Indústria 4.0 leva a constantes adaptações de processos produtivos e gera uma quantidade significativa de informações. Dessa forma, a gestão das informações torna-se um fator crucial para garantir a estratégia competitiva nas indústrias. Uma das informações a ser gerenciada trata-se do *lead time*, tempo entre o cliente solicitar um pedido e este estar disponível. Usualmente, ele pode ser estimado por meio de mensurações dispendiosas ou métodos tradicionais que normalmente não refletem o comportamento real dos dados ou não suportam a quantidade significativa de informação gerada na Indústria 4.0. Além disso, existem lacunas na literatura sobre previsão de *lead time*, como o uso de métodos inteligentes para prever o *lead time* em toda cadeia de suprimentos. Nesse contexto, o objetivo dessa pesquisa é utilizar mineração de dados com uso de algoritmos de aprendizado de máquina para previsão do *lead time* em processos reais. A metodologia proposta fez o uso do ciclo de *Knowledge Discovery in Databases* (KDD) estruturado nas fases seleção, pré-processamento, transformação, mineração de dados e descoberta de conhecimento. Foram testados os algoritmos de aprendizado para predição *Linear Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e *Multilayers Perceptron* (MLP). Para validação dos experimentos foram utilizadas três bases de dados advindas do Sistema Eletrônico de Informações (SEI), cadeia de suprimentos de um setor de logística farmacêutica e do setor de automação industrial para o setor cerâmico. Os resultados mostraram que a mineração de dados é uma ferramenta eficaz para análise de dados gerados na quarta revolução industrial para previsão *Lead time* e a tomada de decisão sobre o planejamento e controle da produção.

Palavras-chave: *Lead time*, mineração de dados, inteligência artificial, *Knowledge discovery in databases*, aprendizado de máquina..

ABSTRACT

BIAZON, M.. **Previsão do *Lead Time* de processos usando Mineração de Dados**. 2021. 83 f. Dissertação de Mestrado (Mestrado em Comunicação) – Universidade Federal de Goiás (UFG), Goiânia – GO.

The era of Industry 4.0 leads to constant adaptations of production processes and generates a significant amount of information. In this way, information management becomes a crucial factor to guarantee the competitive strategy in the industries. One of the information to be managed is textit lead time, time between the customer requesting an order and it being available. Usually, it can be estimated using expensive measurements or traditional methods that do not normally reflect the actual behavior of the data or do not support the significant amount of information generated in Industry 4.0. In addition, there are gaps in the literature on textit lead time forecasting, such as the use of smart methods to predict textit lead time across the supply chain. In this context, the objective of this research is to use data mining using machine learning algorithms to predict the textit lead time in real processes. The proposed methodology made use of the textit Knowledge Discovery in Databases (KDD) cycle structured in the selection, pre-processing, transformation, data mining and knowledge discovery phases. The learning algorithms for textit Linear Regression (LR), textit Random Forest (RF), textit Support Vector Machine (SVM), textit K-Nearest Neighbors (KNN) were tested and textit Multilayers Perceptron (MLP). To validate the experiments, three databases from the Electronic Information System (SEI), a supply chain from a pharmaceutical logistics sector and from the industrial automation sector for the ceramic sector, were used. The results showed that data mining is an effective tool for analyzing data generated in the fourth industrial revolution for forecasting textit Lead time and decision making on production planning and control.

Key-words: Lead time, datamining, artificial intelligence,Knowledge discovery in databases, machine learning..

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Aspectos gerais

O crescimento e as conexões das tecnologias da Indústria 4.0 incidiram no aumento das informações geradas nos processos industriais (FRANK; DALENOGARE; AYALA, 2019a; DALENOGARE *et al.*, 2018; BUMBLAUSKAS *et al.*, 2017; SCHUH *et al.*, 2017; SHENG; AMANKWAH-AMOAH; WANG, 2017). Com volumes significativos de informações e processos autônomos, a tomada de decisão sobre os principais parâmetros e estimativas do processo torna-se ainda mais complexa, demandando o uso de técnicas inteligentes (SCHUH *et al.*, 2017). Neste sentido, uma das tomadas de decisões fundamentais no planejamento e controle da produção trata-se da estimativa do *lead time*, o qual se refere ao tempo despendido entre a solicitação do pedido de um item e/ou lote pelo cliente até o momento em que o mesmo esteja disponível, isto é, seja efetivada a entrega (WERKEMA, 2011).

O *lead time* é um dos parâmetros essenciais para o gerenciamento de qualquer processo industrial (GYULAI *et al.*, 2018; LINGITZ *et al.*, 2018). Decisões relacionadas ao *Just in Time*, sistema de gerenciamento da produção que define o momento e quantidade exata de quanto, quando e como produzir e comprar são dependentes da estimativa do *lead time* (CHUNG; TALLURI; KOVÁCS, 2018; KONG *et al.*, 2018). De maneira similar, a definição de custos de produção, otimização do processo, estimativas de estoque e problemas de otimização utilizam estimativas do *lead time* para tomada de decisão (CHUNG; TALLURI; KOVÁCS, 2018). Além disso, metodologias como “seis sigma” e “manufatura enxuta” também utilizam o *lead time* como indicador de melhorias de processos (COSTA *et al.*, 2020). Ademais, o *lead time* é um dos fatores decisivos para a maioria dos clientes na aquisição de um produto ou serviço de determinada empresa (NOORI-DARYAN; TALEIZADEH; JOLAI, 2019).

1.2 Descrição da problemática

Visto que o *lead time* é um fator essencial na gestão industrial, faz-se necessário estabelecer um *lead time* ótimo como um meio de aumentar a participação no mercado, fidelizar clientes e permitir um planejamento e controle da produção mais eficiente (NOORI-DARYAN; TALEIZADEH; JOLAI, 2019). No entanto, obter o *lead time* de forma assertiva não é uma atividade simples. Sendo de fundamental importância destacar que estimativas não assertivas do *lead time* provocam inconsistências em diversas análises e estimativas da gestão da produção, comprometendo a qualidade da gestão, produto e serviço gerado. Além disso, a estimação do *lead time* torna-se mais complexa se forem usadas estimativas incompletas ou incertas nos processos de avaliação (ÇEBI; OTAY, 2016). Logo, um *lead time* assertivo, torna-se um fator fundamental no diferencial competitivo da empresa e satisfação do cliente (ALBANA; FREIN; HAMMAMI, 2018; NOORI-DARYAN; TALEIZADEH; JOLAI, 2019).

A estimativa do *lead time* tem sido arduamente discutida na literatura. Diversos mecanismos para obtenção do *lead time* vem sendo propostos por diferentes autores, de forma a obter estimativas próximas ao real (SIEVERS *et al.*, 2017; PFEIFFER *et al.*, 2016a; BERLING; FARVID, 2014; IOANNOU; DIMITRIOU, 2012). A obtenção dos valores de *lead time* era comumente realizada através da média de mensurações obtidas da crono-análise de processos, porém, estudos mais orientados propuseram o cálculo da estimativa do *lead time* através do uso de simulação e testes como: ANOVA, cadeias de Markov, técnicas modulares, métodos estatísticos, heurísticas, métodos matemáticos e outros (SIEVERS *et al.*, 2017; PFEIFFER *et al.*, 2016a; BERLING; FARVID, 2014; IOANNOU; DIMITRIOU, 2012). No entanto, a principal desvantagem dos métodos convencionais abordados ocorre por estes suporem que as tendências do passado se manterão no futuro, sendo esta uma das principais desvantagens destes métodos (IOANNOU; DIMITRIOU, 2012).

Pesquisas recentes mostram que apenas métodos usuais e formulações matemáticas baseadas em tendências passadas não são suficientes para uma previsão assertiva do *lead time* (LINGITZ *et al.*, 2018). Logo, o uso de soluções inteligentes, como mineração de dados, inteligência artificial, aprendizado de máquina e metodologias baseadas em KDD, devem ser investigadas na pesquisa científica. Recentemente, técnicas inteligentes vêm sendo empregadas na previsão do *lead time* como medida para preencher as lacunas dos métodos não inteligentes (GYULAI *et al.*, 2018). A maioria das pesquisas propostas que se baseiam em técnicas inteligentes não fazem uso de dados reais, mas utilizam bancos de dados obtidos por meio de simulações computacionais, considerando um sistema ideal e linear de produção (GYULAI *et al.*, 2018). Neste sentido, é possível notar que ainda existem lacunas a serem investigadas nas pesquisas já propostas nos métodos inteligentes, como o uso de técnicas inteli-

gentes na previsão do *lead time* de processo considerando fatores internos da cadeia de suprimentos, ou ainda, o uso de métodos inteligentes na previsão utilizando dados reais e não simulados.

Uma técnica que pode ser aplicada para a investigação das lacunas existentes, considerando um sistema real e fatores externos, trata-se da aplicação do processo denominado *knowledge discovery in databases* (descoberta de conhecimento em bases de dados - KDD) (FRANK; DALENOGARE; AYALA, 2019a). O KDD, proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996), é utilizado para descoberta de conhecimento, previsões e classificações em banco de dados Frank, Dalenogare e Ayala (2019a), tendo como etapa principal a aplicação de soluções baseadas em técnicas de mineração de dados. A mineração de dados pode ser descrita como uma área de estudo que promove o uso de técnicas advindas da inteligência artificial, estatística, álgebra linear, ciência da computação e *big data analytics*, podendo ser utilizada como um potencial estratégico na Indústria 4.0 (FRANK; DALENOGARE; AYALA, 2019b).

1.3 Objetivos

Diante das lacunas e obsolescência dos métodos de predição convencionais e inteligentes, da necessidade de estimar o *lead time* corretamente como vantagem competitiva, esse trabalho propõe o uso do KDD, mineração de dados e aprendizado de máquina como método para obter o *lead time* de processos reais. De forma específica, é proposto investigar o uso do KDD para predição do *lead time* da cadeia de suprimentos, usando como amostra bancos de dados provenientes do setor farmacêutico e do setor de automação industrial para o setor cerâmico, assim como de processos internos, ou seja, trâmites de processos administrativos processados na plataforma, do banco de dados do Sistema Eletrônico de Informações (SEI) do Governo Federal.

Nesse sentindo, essa pesquisa fez o uso da técnica de mineração de dados para predição do *Lead time*. Os algortimos escolhidos para predição foram os com menor erro quadrático médio (mean squared error-MSE) identificados na etapa de validação cruzada. Sendo esses algoritmos: *Linear Regression* (Regressão Linear), *Random Forest* (Florestas Aleatórias), *Support Vector Machine* (Máquinas de Vetores de Suporte), *K-Nearest Neighbors* (K-vizinhos mais Próximos - KNN) e *Multilayers Perceptron* (Redes Neurais Perceptron Multicamadas). De forma particular, para a base de dados do SEI além dos algoritmos supracitados foi utilizado um método hibrido contendo os algotimos KNN, K-means e RL. Predições utilizando dados numéricos, nominais (dados do tipo *string*) e binários (0 ou 1) foram investigados.

1.4 Justificativa

Diante do exposto, esta pesquisa se justifica devido à necessidade da estimação do *lead time* do processo produtivo, com métricas de tratamento de grandes volumes de dados, tendo em vista as abordagens convencionais que não refletem os parâmetros do processo com avaliação de todo um banco de dados. Já que, segundo Noori-Daryan, Taleizadeh e Jolai (2019), estimar corretamente o *lead time* trata-se de um processo crítico, pois pode ser considerado um dos parâmetros mais importantes, assim como, fundamental para manter a fidelidade do cliente, agregar em competitividade e produtividade. Além disso, por ser considerado um dos parâmetros essenciais, fornece subsídios necessários para a adoção das tecnologias da chamada quarta revolução industrial (NAGAHARA; NONAKA, 2018; FRANK; DALENOGARE; AYALA, 2019b; BUMBLAUSKAS *et al.*, 2017).

A partir disso, a pesquisa apresenta contribuições importantes no campo de atuação da engenharia e desenvolvimento de produtos e processos. Sendo que contribui como uma nova abordagem para definição de parâmetros de processo dentro das novas perspectivas da Indústria 4.0. Posto que, de acordo com Dalenogare *et al.* (2018), tem-se como tendência a adequação do mercado como estratégia competitiva a novas técnicas da quarta revolução industrial.

Ademais, nota-se que a adoção de técnicas de gerenciamento de dados tem perspectivas significantes no âmbito de pesquisas acadêmicas, conforme descrito por Esmaeilian, Behdad e Wang (2016), que afirma que o estudo de técnicas de análise e monitoramento de processos de menor custo e maior eficiência por meio de dados se emerge como de uma linha de pesquisa promissora para os novos processos de manufatura. Logo, esta pesquisa vai de encontro a essa nova perspectiva do mercado, contribuindo enquanto uma nova abordagem, advinda da mineração de dados, para ser utilizada na tomada de decisão da estimativa do *lead time*, assim como no auxílio à identificação prévia do *lead time* de grandes bancos de dados.

1.5 Estrutura da dissertação

Essa pesquisa está dividida em oito capítulos, sendo eles: Introdução, Referencial teórico, Metodologia, Experimentos caso 1: setor farmacêutico, Experimentos caso 2: setor de automação industrial para o setor cerâmico, Experimentos caso 3: setor de serviços eletrônicos, Considerações finais.

O Capítulo 1 apresenta a contextualização da temática abordada, problemática de pesquisa, métodos de predição do *lead time* usuais e recentes na literatura assim

a proposta da pesquisa que visa contornar as lacunas na literatura. Além disso, o Capítulo 1 discorre a respeito dos objetivos gerais e específicos e contribuições.

O Capítulo 2 traz o embasamento teórico a respeito da relevância do *lead time* e uma exposição dos métodos já propostos para a previsão do *lead time*. Além de apresentar a importância do *lead time* no contexto da Industria 4.0, apresentando os métodos inteligentes que foram empregados na pesquisa.

O Capítulo 3 apresenta o enquadramento metodológico da pesquisa e etapas de cada fase do ciclo KDD.

Os Capítulos 4, 5 e 6 apresentam a aplicação da mineração de dados para previsão do *lead time* em 3 bancos de dados reais, sendo estes do setor farmacêutico, de automação cerâmico e serviço eletrônicos, respectivamente.

O Capítulo 7 aborda as considerações finais e sugestões de pesquisas futuras.

CAPÍTULO

2

REFERENCIAL TEÓRICO

2.1 *Lead time*

O *lead time* compreende o tempo despendido entre o pedido de um item e este estar disponível ao cliente final, sendo um dos indicadores de desempenho essenciais e importantes para o gerenciamento de processos produtivos de manufatura ou serviço (GYULAI *et al.*, 2018; LINGITZ *et al.*, 2018; KIM; KIM; LEE, 2014). Este também é considerado um dos fatores determinantes para fidelidade de um cliente na empresa (GYULAI *et al.*, 2018). Além disso, o planejamento e programação da produção é dependente das previsões do *lead time*, ou seja, os tempos de fabricação influenciam a eficiência e qualidade do planejamento e gerenciamento da produção (LINGITZ *et al.*, 2018). Logo, a eficácia dos métodos de planejamento e controle da produção é dependente da precisão da previsão do *lead time*, pois, o *lead time* assertivo auxilia na disponibilização de forma mais rápida e na quantidade certa de produtos e serviços (GYULAI *et al.*, 2018).

Diversos fatores estratégicos de gestão e aprimoramento de processo são dependentes do *lead time* (WERKEMA, 2011). Em relação a esses fatores, pode-se citar como exemplos a gestão e desenvolvimento do processo e projeto, o *Just in Time* (JIT), o PCP (Planejamento e controle da produção), *lean manufacturing* (manufatura enxuta), *lean* seis sigma e gestão de estoques (WERKEMA, 2011). Além disso, o *lead time* auxilia na tomada de decisão sobre qual será o tempo de resposta ao cliente na fabricação industrial, a taxa de saída do processo, a quantidade de itens em processos iniciados e não finalizados (Work in Progres - WIP), a eficiência do processo (Process Cycle Efficiency- PCE), variáveis diretas para elaboração do plano mestre de produção (PMP), planejamento da capacidade dos materiais (Material Requirements Planning - MRP I e II), definição de estoques de segurança, lotes econômicos, pontos

de ressuprimento, seleção de fornecedores e outras estimativas (WERKEMA, 2011; NOORI-DARYAN; TALEIZADEH; JOLAI, 2019).

No entanto, prever o *lead time* de forma assertiva não é uma tarefa simples (GYULAI *et al.*, 2018; KIM; KIM; LEE, 2014; LEE; BAGHERI; KAO, 2015), principalmente devido à variedade e complexidade de produtos nos sistemas produtivos (GYULAI *et al.*, 2018). Métodos tradicionais de obtenção do *lead time* afetam de forma negativa diversos critérios do sistema produtivo, uma vez que, calculam principalmente valores médios baseados em dados históricos resultando em deficiências no Planejamento e controle da Produção (LINGITZ *et al.*, 2018).

2.2 Métodos da literatura para previsão do *lead time*

O problema de estimação e gerenciamento do *lead time* tem sido abordado na literatura desde a década de 60 (IOANNOU; DIMITRIOU, 2012). Por meados de 1980 estudos sobre tempos de operação e estimativas do *lead time* por meio de formulações matemáticas e métodos estatísticos com análise de variância ANOVA foram propostos (CHANG, 1997; TATSIOPOULOS; KINGSMAN, 1983). A previsão aproximada do *lead time* por meio de formulações matemáticas foi proposto para um sistema sob encomenda desconsiderando a carga de trabalho atual do sistema (VANDAELE; BOECK; CALLEWIER, 2002). A comparação entre as redes de circuito elétricos e o fluxograma de atividades de um processo de produtos complexos (com alta diversidade e variabilidade) foi realizada com o intuito de encontrar o valor total do *lead time* na rede (JUN; PARK; SUH, 2006). Algumas pesquisas propuseram o uso de técnicas de avaliação e revisão de programas (*Program Evaluation and Review Technique - PERT*), método do caminho crítico (*Critical Path Method - CPM*) e cadeias de Markov para estimar o *lead time* (JUN; PARK; SUH, 2006).

Em Ioannou e Dimitriou (2012), foi proposto o uso de formulações matemáticas para obtenção do *lead time* dinâmico de um sistema de encomenda. Outras pesquisas propuseram o uso de abordagem de teoria das filas para análise e predição do *lead time* (IOANNOU; DIMITRIOU, 2012). Berling e Farvid (2014) propuseram o uso de simulação de eventos discretos por meio de expressões matemáticas para fazer um estudo sobre o *lead time* assumindo uma demanda contínua verificando o valor da variância do *lead time*. Em Mourtzis *et al.* (2014), propôs-se a análise e predição do *lead time* um sistema de produção sob encomenda baseado em um fluxograma gráfico de atividades que envolvem a validação da estimativa baseada na experiência de engenheiros. Pfeiffer *et al.* (2016b) fez o uso de métodos estatísticos de regressão multivariada utilizando dados simulados para obter o *lead time* de um sistema *flow-shop*.

Formulações matemáticas e estatísticas foram propostas para estimativas do *lead time* em plantas modulares de produção do setor químico (SIEVERS *et al.*, 2017).

2.3 Indústria 4.0

A quarta revolução Industrial, conhecida também como Indústria 4.0, possui como característica as empresas alcançarem um desempenho industrial mais alto por meio, entre outras coisas, dos chamados Sistemas Ciber-Físicos, que se refere a integração de tecnologia da informação com processos fabris (FRANK; DALENOGARE; AYALA, 2019a; JESCHKE *et al.*, 2017). Uma vez que, a Indústria 4.0, promove o uso e integração destas diversas tecnologias de processamento de informação que impactam em mudanças na realização das operações do ambiente industrial e de serviços (TRSTENJAK; COSIC, 2017; NETO *et al.*, 2018), promovendo um incremento na qualidade do processo e produtos assim como uma redução do *lead time* e consequentemente maior produtividade (TRSTENJAK; COSIC, 2017; NETO *et al.*, 2018). Alguns países adotaram programas para fornecer incentivo e subsídios para auxiliar as empresas a adotarem tecnologias da Indústria 4.0, esses programas podem ser vistos na Alemanha, onde o conceito de Indústria 4.0 surgiu, Brasil e China (DALENOGARE *et al.*, 2018). No entanto, a adoção de tecnologias e sua integração pelas empresas pode ser considerado um desafio em diversas empresas, principalmente em países emergentes (DALENOGARE *et al.*, 2018).

Dentre as tecnologias da Indústria 4.0, tem-se a integração do real com o virtual com o emprego de sistemas ciber-físicos, internet das coisas (*Internet of things - IOT*), big data e data analytics (DALENOGARE *et al.*, 2018; WANG; TÖRNGREN; ONORI, 2015). Estas tecnologias permitiram que as operações automatizadas, algumas tomadas de decisão e o tráfego de informações sejam realizadas em tempo real (BAUERNHANSL; HOMPEL; VOGEL-HEUSER, 2014). Com emprego dessas tecnologias, as fábricas assumiram o perfil de fábricas inteligentes com seus diversos mecanismos ligados por meio da Inteligência Artificial, Internet das coisas e o fluxo da informação (BAUERNHANSL; HOMPEL; VOGEL-HEUSER, 2014). Em relação ao *big data analysis*, análise de grande quantidade de dados, Oliff e Liu (2017) afirma que este tornou-se uma ferramenta fundamental no desenvolvimento da Indústria 4.0, uma vez que, o gerenciamento e interpretação dos bancos de dados tornou possível o surgimento da Indústria 4.0. Dentre as técnicas do *big data* e *data analytics* tem-se a mineração de dados, ferramenta advinda do processo de aprendizado de máquina (OLIFF; LIU, 2017).

2.4 Inteligência artificial: mineração de dados

Diariamente são gerados volumes significativos de dados que em sua maioria não podem ser ignorados, uma vez que, estes dados tornam-se imprescindíveis, entre outras coisas, para consulta, análise e tomadas de decisão (WITTEN; FRANK, 2002). Ainda para Witten e Frank (2002), visando atender esta necessidade de reconhecimento, análise e tratamento dessas informações, tem-se a metodologia mineração de dados.

A mineração de dados trata-se de uma metodologia com significativo potencial para oferecer subsídios a tomada de decisão às empresas por meio da extração de conhecimento de seus bancos de dados. Sob outra perspectiva, a mineração de dados é compreendida como uma técnica para extração de conhecimento preditivo oculto de um conjunto relativamente grande de dados (ARMANO; FARMANI, 2016; BHOWMIK; CHATTOPADHYAY; CHATTERJEE, 2016). Dessa forma, visa gerar informação estratégica que auxilie na redução de custos e incremento na receita (DEEPASHRI; KAMATH, 2017). Esta, trata-se da análise e conversão de uma quantidade significativa de dados em conhecimento relevantes por meio da identificação de padrões, comportamento dos dados, associações, correlações e entre outras de forma a refletir medidas estratégicas (FRANK; HALL, 2011; DEEPASHRI; KAMATH, 2017).

A mineração de dados também é uma das etapas do ciclo *Knowledge Discovery in Databases* (KDD), descoberta de conhecimento em bancos de dados, que permite a extração automatizada de informações implícitas em bancos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O ciclo KDD trata-se de uma metodologia de Inteligência Artificial (BHOWMIK; CHATTOPADHYAY; CHATTERJEE, 2016). Em geral, os passos iniciais de sua aplicação tratam da definição do problema, banco de dados, a coleta, limpeza e pré-processamento dos dados (BHOWMIK; CHATTOPADHYAY; CHATTERJEE, 2016). Ainda para (BHOWMIK; CHATTOPADHYAY; CHATTERJEE, 2016), posteriormente, define-se técnicas de modelagem dos dados e algoritmos inteligentes de mineração de dados.

A informação extraída por meio da mineração de dados, ocorre por meio de cinco técnicas usadas para determinar padrões, sendo estas a análise de regras de associação, análise de padrões sequenciais, classificação e predição, análise de agrupamentos e análise e remoção de *outliers* (SAGAERT *et al.*, 2019; BORAH; NATH, 2018).

De forma sumarizada, a técnica de associação visa descobrir relacionamentos entre itens e conjuntos de dados (DOGAN; BIRANT, 2020). No que se refere a técnica de predição, tem-se a classificação e a regressão, entende-se que se trata da predição de classes nominais e valores números respectivamente (SHAO *et al.*, 2018). Em outras palavras, tem a finalidade de predizer determinada classe ainda não categori-

zada (SHAO *et al.*, 2018). Outro tipo de método que integra a classificação com regras de associação é chamado de classificação associativa (AZMI; RUNGER; BERRADO, 2019). De acordo com Sagaert *et al.* (2019), quanto aos agrupamentos, visa-se identificar grupos com itens com padrões similares. Os algoritmos de agrupamento podem ser categorizados em três tipos, os baseados em aprendizagem, em restrição e em rótulos (SAGAERT *et al.*, 2019).

A mineração de dados tem muitas aplicabilidades em diferentes campos, uma vez que a mineração de dados e o aprendizado de máquina são considerados impulsionadores tecnológicos essenciais na industria. (MOONAM; QIN; ZHANG, 2019; AHUETT-GARZA; KURFESS, 2018; SUN; MEDAGLIA, 2019; ALLAHVERDI, 2015). Entre outros exemplos, a mineração de dados também pode ser aplicada em diversos setores, como o setor de saúde, de eficiência energética, área de otimização, mecânica, químico, mineração, em botânica, finanças, manufatura aditiva, segurança, qualidade de materiais e outros (ALDEN *et al.*, 2019; RAMEZANIAN; PEYMANFAR; EBRAHIMI, 2019; ZHANG *et al.*, 2019; GHOLAMI *et al.*, 2019; LEE *et al.*, 2017; ZHAO; ROSEN, 2017; BHOWMIK; CHATTOPADHYAY; CHATTERJEE, 2016; ARMANO; FARMANI, 2016; RAMEZANIAN; PEYMANFAR; EBRAHIMI, 2019).

2.4.1 Predição numérica com mineração de dados

A análise preditiva, tanto numérica (regressão) quanto nominal (classificação) trata-se de uma técnica que agrega valor e conhecimento para empresas por meio da previsão numérica de informações de grandes bancos de dados (LEPENIOTI *et al.*, 2020; ŠIKŠNYS *et al.*, 2016). Essa previsão, gera suporte para as tomadas de decisões e ações implementadas, podendo, conforme os parâmetros de entrada detectar padrões que indicam um problema ou oportunidade futura para o negócio (LEPENIOTI *et al.*, 2020). Antes de iniciar o processo de predição é preciso realizar processamento dos dados, organizando e removendo *outliers*, valores fora do comportamento normal esperado do processo, valores extremos e atípicos (ZHANG *et al.*, 2019).

Esses *outliers* ou ruídos podem ser definidos como pontos fora do comportamento padrão da maioria dos dados, erros grosseiros, que tornam os dados de baixa qualidade e as saídas das análises de mineração de dados com esses dados precárias (MANIKANDAN; ABIRAMI, 2018; NURUNNABI; WEST; BELTON, 2015). Faz-se importante tratar os *outliers* e valores extremos, pois a qualidade da predição também depende da qualidade dos dados utilizados (SAGAERT *et al.*, 2019). A análise de pontos fora do padrão pode ser feita por meio de *box plots*, usando o conceito de interquartil, onde os dados fora das faixas de valores dos quartis são considerados *outliers* (SAGAERT *et al.*, 2019; HU *et al.*, 2018). Os dados são distribuídos na faixa dos quartis que são divididos em quatro valores de 25 quartis. Os valores extremos

são os valores que excedem o 75 quartil ou que estão abaixo do 25 quartil, sendo considerados *outliers* (WITTEN; FRANK, 2002).

Além disso, técnicas de aprendizado de máquina, assim como outras, podem ser aplicadas em todas as etapas do ciclo KDD, utilizado por exemplo nas fases de pré-processamento, limpeza e separação dos dados em arquivos de treinamento, teste, classificação e predição das amostras (SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019; ZHANG *et al.*, 2019). O subconjuntode treinamento normalmente representa de 60% a 70% do banco de dados utilizado para treinar o algoritmo e o restante dos dados são utilizados como amostras de teste (SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019; ZHANG *et al.*, 2019). O pré-processamento dos dados, pode ser feito por diversas técnicas e ferramentas (H'NG; LOH, 2019).

2.4.2 Algoritmos de mineração de dados

Lepenioti *et al.* (2020) categoriza os métodos de análise de predição em três categorias, sendo estes modelos probabilísticos, aprendizado de máquina ou mineração de dados e análise estatística. Em geral, as técnicas preditivas utilizando modelos probabilísticos estão segmentadas em métodos de redes bayesianas, cadeias e modelos de camadas de Markov simulações de monte Carlo (LEPENIOTI *et al.*, 2020). As técnicas de predição, aprendizado de máquina e mineração de dados concentram-se no uso de técnicas inteligentes. Entre esses pode-se citar reconhecimento de padrões, procura aleatória, heurísticas baseadas em clusters e métodos kernels (LEPENIOTI *et al.*, 2020). Além do mais, os algoritmos de aprendizado de máquina: *Artificial Neural Network* (Redes neurais artificiais), *Random Forest* (Floresta aleatória), *Support Vector Machine* (Máquina de vetores de suporte), *k-nearest neighbors* algoritmo (k-vizinhos mais próximos) são alguns dos mais importantes e utilizados na literatura (QIN *et al.*, 2020; MERCADIER; LARDY, 2019; CHEN; GUO, 2015). De forma sumarizada, os métodos de análise estatística podem ser segmentados em regressão linear, regressão linear múltipla, regressão logística, regressão de vetores de suporte e estimação de densidade (LEPENIOTI *et al.*, 2020). Devido a importância dos métodos de aprendizado de máquina supracitados, estes serão tratados nos tópicos subsequentes (GOU *et al.*, 2019b; H'NG; LOH, 2019; NING; MA; ZHAO, 2019; SETTOUTI; BECHAR; CHIKH, 2016).

2.4.2.1 K-Nearest Neighbor

O algoritmo *k-nearest neighbors* (KNN) trata-se de um classificador e preditor simples e eficaz, usualmente utilizado em reconhecimento de padrões, aprendizado de máquina e inteligência artificial. Este algoritmo possui como parâmetro a medida de distância dos dados de uma amostra de teste (valores que se deseja prever) às

amostras de treinamento (padrões de comportamento existente) (GOU *et al.*, 2019b; NING; MA; ZHAO, 2019; ZHAO *et al.*, 2020; DOGAN; OZTAYSI, 2019).

Este algoritmo considera que os itens que tendem a concentrar-se em uma mesma região de uma vizinhança possuem o mesmo padrão de comportamento. Diante disso, considera a distância entre o item que se deseja classificar ou predizer em relação aos k-vizinhos mais próximos de classes ou padrões de comportamentos já existentes (NING; MA; ZHAO, 2019). Posteriormente, as distâncias são ordenadas em ordem crescente e a quantidade de distâncias consideradas são delimitadas por um parâmetro k (JIMÉNEZ *et al.*, 2020; NING; MA; ZHAO, 2019). Que por sua vez, este parâmetro trata-se do número de vizinhos mais próximos de uma amostra (ZHANG *et al.*, 2017). Em relação ao valor ótimo para o parâmetro K_{Limite}, (BECKER; THRÄN, 2017) afirmam que esse valor pode ser definido arbitrariamente. Uma alternativa viável para definição do número ótimo de k seria comparar diferentes subconjuntos de dados de treinamento (BECKER; THRÄN, 2017). A predição ou classe é definida conforme maior número de ocorrência, ou seja, a predição considerada refere-se à maioria dos votos dos vizinhos mais próximos (JIMÉNEZ *et al.*, 2020; NING; MA; ZHAO, 2019; GOU *et al.*, 2019a; ZHANG *et al.*, 2019). Em conformidade com (JIMÉNEZ *et al.*, 2020) a decisão pela maioria de vizinhos mais próximos pode ser sumarizada pela Equação 2.1.

$$Y_0^{KNN} = argmax \sum_{i_{NN}=1}^{n} \left(x_i^{KNN}, y_i^{NN}\right) \in T' \delta \left(y_c = y_i^{NN}\right) \tag{2.1}$$

Onde, y_c simboliza o rótulo da classe, e o y_i^{NN} é o rótulo do da classe para o i-ésimo vizinho mais próximo. Onde x_i^{KNN} representam variáveis de entrada dos k-vizinhos mais próximos y_i^{NN}, para $i_{NN} = 1, 2, ..., n$ dentre os K_{Limite} vizinhos mais próximos (JIMÉNEZ *et al.*, 2020). O objeto matemático , equivale a função delta de Dirac, onde as seguintes condições são expostas $\delta\{1\text{se } y_c = y_i KNN), \ 0$ caso contrário $\}$ (JIMÉNEZ *et al.*, 2020).

Em relação ao parâmetro distância, o KNN pode adotar o uso de diversas distancias, a Euclidiana, de Manhattan e outras (ZHANG et al., 2019). No entanto, é necessário verificar a que melhor se adéqua aos bancos de dados utilizados (JIMÉNEZ *et al.*, 2020; CHEN; GUO, 2015; CHENG; CHAN; SHEU, 2019; GOU *et al.*, 2019b; ZHANG *et al.*, 2019).

2.4.2.2 Random forest

Random forest trata-se do algoritmo de florestas aleatórias, considerado um dos mais eficazes e robustos métodos de predição (regressão e classificação) (CAO *et al.*, 2019; GONG *et al.*, 2018; HU *et al.*, 2018; KANG *et al.*, 2020; PAUL; MUKHERJEE, 2019; VERIKAS; GELZINIS; BACAUSKIENE, 2011). Este algoritmo refere-se a uma

técnica de aprendizado de máquina com aplicações em diversos setores, entre eles o de medicina, análise de imagens e outros (NI *et al.*, 2017). Considerando vetores de conjuntos de entrada, em conformidade com (HALIM; REHAN, 2020) e (KANG *et al.*, 2020), de forma sumarizada, esse algoritmo trata-se da combinação e avaliação de múltiplas árvores de decisão (*decision trees*). A avaliação é feita por meio de um sistema de votação, onde cada árvore de decisão gera uma resposta, uma predição. A predição é definida pela média das respostas de cada árvore (HALIM; REHAN, 2020; WANG *et al.*, 2016). O processo de florestas aleatórias seleciona como resposta a classificação ou predição mais votada (HALIM; REHAN, 2020). Essa pode ser considerada uma vantagem em relação a árvores de decisão, pois quando se faz avaliação de uma árvore de decisão pode haver um ajuste excessivo dos dados a uma determinada condição (HALIM; REHAN, 2020).

O conceito do algoritmo florestas aleatórias, de acordo com (LI *et al.*, 2018a; NI *et al.*, 2017), pode ser sumarizado como a média dos resultados da combinação aleatória de árvores de decisão, conforme expresso na Equação 2.2.

$$Y_{RF} = \frac{1}{i_{RF}} \left(\sum_{t=1}^{i} {}_{RF} y_{i_{RF}}^{RF} \right) = \frac{1}{i_{RF}} \left(\sum_{t=1}^{i} {}_{RF} \hat{h} \left(x_i^{RF}, S_{a_{RF}}^{\Theta l} \right) \right) \tag{2.2}$$

Onde $y_{i_{RF}}^{RF}$ trata-se dos valores preditos, ou seja *outputs* gerados em cada árvore, $RF_{prediction}$ o valor majoritário predito pelo algoritmo florestas aleatórias (LI *et al.*, 2018b). Além disso, a variável $\hat{h}$ refere-se à função de previsão tendo como parâmetros as variáveis de entrada, $x^{RF} = \{x_1^{RF}, x_2^{RF}, ..., x_n^{RF}\}$, de cada árvore com a_{RF} características e o vetor de dados de treinamento $S_{i_{RF}}^{\Theta l}$, que variam de $l = 1, 2, ..., n$ (MERCADIER; LARDY, 2019). Os conjuntos de treinamento partem do conjunto de entradas que são posteriormente separados em cada nó (MERCADIER; LARDY, 2019; LI *et al.*, 2018b). É importante ressaltar que o processo de geração e combinação das árvores são aleatórios, conhecido como processo *bootstrap*, que gera aleatoriamente uma coleção de a_{RF} árvores, considerando que cada uma tem a probabilidade de $1/a_{RF}$ de ser escolhida (LI *et al.*, 2018b).

Faz-se fundamental o entendimento das árvores de decisão para compreensão do funcionamento do algoritmo florestas aleatórias (GONG *et al.*, 2018). A árvore de decisão, conhecida também como árvore de regressão ou classificação, usa regras para dividir os dados iniciando pelo nó raiz, em grupos de forma recursiva que serão posteriormente divididos em cada nó de decisão até que se chegue nas folhas (GONG *et al.*, 2018; TSAGKRASOULIS; MONTANA, 2018). De forma geral, quando se trabalha com predição numérica, (regressão), as variáveis com valores de respostas similares são repartidas na mesma região. Onde cada divisão é representada por árvores, o número de variáveis incorporadas em cada árvore pode ser definida com o método

de validação cruzada (GONG *et al.*, 2018). Os pontos e variáveis de divisão são selecionados de forma que se minimize a função de perda, para tanto, faz-se o uso do erro quadrático médio (*Mean squared error - MSE*), à medida que se minimiza a função de perda um par de divisão pode ser escolhido. Em conformidade com (AL-WAELI *et al.*, 2019), MSE pode ser obtido por meio da Equação 2.3.

$$MSE = \frac{1}{i_{MSE}} \sum_{t=1}^{i} MSE(y_{real} - y_{predito})^2 \tag{2.3}$$

Onde n trata-se do número de observações de entrada, y_{real} o valor real que se deseja prever, e $y_{predito}$ o valor predito (AL-WAELI *et al.*, 2019). Em conformidade com (GONG *et al.*, 2018) e (JIMÉNEZ *et al.*, 2020).

O processo da árvore de decisão pode ser sumarizado em quatro etapas, considerando que Y_{RF} é a variável a predita e x_i^{RF}, $x_i^{RF}[x_1^{arv}, x_2^{arv}, ..., x_n^{arv}]$ é um vetor de variáveis preditoras, ou seja, *inputs*. Esse processo é apresentado a seguir:

1. Inicie com todas as alternativas em uma região, ou seja, toda população em um único nó, que é o nó raiz;
2. Aplique um teste para uma das variáveis preditora em cada nó interno da árvore;
3. As observações são divididas em sub-regiões, chamadas de sub-nós, a cada resultado do teste a direita ou esquerda de cada árvore;
4. A etapa antecedente é repetida até que uma folha seja alcançada. Cada folha representa uma predição e os nós são tomadas de decisão que são divididas em ramificações.

Dessa forma, no algorítimo de Florestas aleatórias, a floresta desenha sub-conjuntos aleatórios de a_{RF} árvores de decisão e toma uma decisão final a partir do conjunto de predições geradas.

2.4.2.3 Regressão linear

A análise de regressão linear é uma análise simples e eficaz para muitas aplicações, empregada tanto em abordagens estatísticas quanto em análise inteligentes, como mineração de dados (LEPENIOTI *et al.*, 2020; GOU *et al.*, 2019b).

A análise de regressão linear fornece a relação entre variáveis de entrada e pesos aleatórios e uma variável a ser predita (MAIRIZAL *et al.*, 2020). Em conformidade

com (AL-WAELI *et al.*, 2019; JIMÉNEZ *et al.*, 2020), a expressão matemática que descreve a equação de regressão linear pode ser dada pela Equação 2.4.

$$Y_{LR} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i^{LR} + \varepsilon \tag{2.4}$$

Onde β_0 é o valor de interceptação da reta, Y_{LR}, saída ou resposta da predição, x_i^{LR} variáveis de entrada, e ε refere-se ao erro presente na saída gerada (AL-WAELI *et al.*, 2019; GOU *et al.*, 2019b).

2.4.2.4 Support vector machine

A máquina de vetores de suporte (*Support Vector Machine*) busca encontrar os hiperplanos ótimos que separam de forma assertiva as observações em diferentes classes (JIMÉNEZ *et al.*, 2020; SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019). Os hiperplanos de $p^{SVM} - 1$ observações separam as variáveis x_i^{SVM} em d^{SVM} dimenções no espaço. O melhor hiperplano é aquele que maximiza a distância entre os hiperplanos existentes (LIU, 2017). Podem ocorrer três casos diferentes na análise do SVM (LIU, 2017). Sendo estes os que tem apenas duas classes, mais de duas classes e resolução de problemas linearmente não separáveis (LIU, 2017).

Para o caso de problemas com duas classes o SVM separa os dados em dois hiperplanos, um positivo e outro negativo (LIU, 2017). Nesses planos bidimensionais tem-se que w_{SVM} representa um vetor dimensional interceptado por b_{SVM} (LIU, 2017). Os pontos mais próximos aos planos positivo e negativo são convencionalmente chamados de vetores de suporte e o plano paralelo aos planos negativos e positivos trata-se do plano de decisão (LIU, 2017). Como supracitado, o melhor hiperplano é aquele que maximiza a distância, a margem, entre o ponto de dados mais próximo do lado positivo e entre a distância entre o ponto de dados mais próximo do lado negativo (LIU, 2017). Ou seja, quanto mais longe os hiperplanos estiverem, mais longe as classes estarão uma da outra, aumentando dessa forma a confiabilidade de que a amostra pertence a determinada classe.

De forma geral, em conformidade com (LIU, 2017), tem-se um problema de otimização, para maximizar a distância entre os hiperplanos, ou seja, minimizar a norma Euclidiana do vetor g^{SVM}. Esse conceito pode ser sumarizado na Equações 2.5 e 2.6.

$$Minimizar \|g^{SVM}\| \tag{2.5}$$

$$u^{(i)}(g^{SVM} j^{(i)} + b^{SVM}) \geq 1 \tag{2.6}$$

Para dados de treinamento $\left(x_{SVM}^{(1)}, y_{SVM}^{(1)}\right), \left(x_{SVM}^{(2)}, y_{SVM}^{(2)}\right), ..., \left(x_{SVM}^{(n)}, y_{SVM}^{(n)}\right)$.

Em casos de problemas com mais de duas classes, trabalha-se com dois conceitos *one-vs-all* (um contra todos) e *one-vs-one* (um contra a um) para um problema com $C_{classes}^{SVM}$ classes (LIU, 2017). O conceito um contra um, analisa de duas em duas classes, por pares, até que todas tenham sido confrontadas umas com as outras (LIU, 2017). Já no conceito um contra tudos, as classes são divididas em duas classes, a classe é separada em um hiperplano positivo e as demais de forma agrupada no hiperplano negativo (LIU, 2017). Como no caso de duas classes, a melhor classificação trata-se daquela que apresenta o maior valor, $argmax$, de distância entre os hiperplanos $\left(g^{SVM}x_{SVM}^{(1)} + b^{SVM}\right)$ (LIU, 2017), conforme expresso na Equação 2.7.

$$SVM_{pred}^{duasclasses} = argmax\left(g^{SVM}x_{SVM}^{(1)} + b^{SVM}\right) \tag{2.7}$$

Em problemas linearmente não separáveis, não é possível determinar um hiperplano linear de forma que separe os dados em duas classes (HALIM; REHAN, 2020; LIU, 2017). Partindo do pressuposto que dados de uma determinada classe estão mais próximas da origem que outras nesses casos, faz-se necessário usar uma função. *Kernel* (HALIM; REHAN, 2020; LIU, 2017).

A função *Kernel* mais utilizada trata-se da função Gaussiana, conhecida também com função de base radial (RBF) (HALIM; REHAN, 2020; LIU, 2017; SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019). O sistema Gaussiano pode ser compreendido como a generalização da distribuição de probabilidade, onde faz-se o uso de um vetor de entradas, a média e variância escalares para se calcular um vetor de média e covariância para obter a probabilidade (HALIM; REHAN, 2020; SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019). De forma sumarizada, trata-se da exponencial do produto do coeficiente Kernel (Y_{kernel}^{SVM}) e o quadrado da diferença entre o ponto de análise $x_{svm}^{(i)}$ e a origem $o_{svm}^{(j)}$, conforme exposto na Equação 2.8.

$$SVM_{pred}^{Kernel}\left(x_{svm}^{(i)}, o_{svm}^{(j)}\right) = exp\left(-Y_{Kernel}^{SVM}\|x_{svm}^{(i)} - o_{svm}^{(j)}\|^2\right) \tag{2.8}$$

Onde o coeficiente Kernel $Y_{kernel}^{SVM} = \frac{1}{w\sigma^{(2)}}$ representa o quanto a função se distancia das observações, quando maior o coeficiente de Kernel menor é a variação e quanto menor, maior a variação (LIU, 2017). Torna-se importante ressaltar que se pode utilizar outras funções Kernel dependendo de cada problema estudado (LIU, 2017).

2.4.2.5 Artificial neural network: multilayer perceptron

A Artificial Neural Network (Rede Neural Artificial) trata-se de uma rede similar a rede neural humana, considerando o processamento de informações do cérebro humano (HAYKIN, 2007). A rede neural trata-se de u processador, que por meio de um processo de aprendizagem capaz de armazenar conhecimento e disponibiliza-lo

(HAYKIN, 2007). Essa rede é composta de camadas, de forma geral, tem-se a camada de entrada, as camadas ocultas, que são as camadas intermediárias e camada de saída que se trata da última camada (HALIM; REHAN, 2020; NOGUEIRA *et al.*, 2018; YADAV; CHANDEL, 2017). A primeira camada, camada de entrada, é composta por um número de vetores de entradas $x_i^{MLP} = \left\{x_1^{MLP}, x_2^{MLP}, ..., x_n^{MLP}\right\}$. Essas entradas são ponderadas por pesos aleatórios $w_i^{MLP} = \left\{w_1^{MLP}, w_2^{MLP}, ..., w_n^{MLP}\right\}$, acrescidos por uma constante de tendência de camada b_i^{MLP} (NOGUEIRA *et al.*, 2018). A soma do produto entre os pesos aleatórios iniciais e cada variável de entrada forma um neurônio, z_i^{MLP}, da próxima camada, camada oculta o número de neurônios é igual ao número de saídas entre a soma do produto gerada pelas n entradas e m pesos aleatórios (NOGUEIRA *et al.*, 2018). Ainda para (NOGUEIRA *et al.*, 2018), de forma matemática, o neurônio para cada camada pode ser obtido conforme a Equação 2.9.

$$z_i^{MLP} = \sum_{i=1}^{N} x_i^{MLP} w_i^{MLP} + b_i^{MLP} \tag{2.9}$$

A partir da primeira camada oculta são geradas saídas parciais, Y_{MLP} preditas em cada neurônio, essas predições são calculadas usando uma função de ativação, $\sigma(z_i^{MLP})$, que tem como finalidade gerar uma não linearização nos componentes neurais (YADAV; CHANDEL, 2017; FRANK; HALL, 2011; FRANK; DALENOGARE; AYALA, 2019b). Quando se faz predições, utilizando valores numéricos os valores são convertidos em unidades lineares sem restrições (FRANK; HALL, 2011). Essas predições parciais serão multiplicadas pelos pesos aleatórios (YADAV; CHANDEL, 2017). Logo, os *outputs* parciais, $Y_{MLP}^{parcial}$ saídas de cada neurônio se comportaram como as novas variáveis de entrada, $x_i^{MLP} = Y_{MLP}^{parcial}$, que formaram as próximas camadas ocultas gerando novos neurônios, até chegar na camada de saída onde será gerada a predição da camada de saída Neurônio $Y_{MLP(z_n,w_n)}^{final}$ (CARDONA; NEDJAH; MOURELLE, 2017). De acordo com (CARDONA; NEDJAH; MOURELLE, 2017), a predição gerada na última camada pode ser expressa pela Equação 2.10.

$$Y_{MLP(z_n,w_n)}^{final} = \sigma_{(z_i^{MLP}=)} = \sigma\left(\sum_{i=1}^{N} x_i^{MLP} w_i^{MLP} + b_i^{MLP}\right) \tag{2.10}$$

Onde z_i^{MLP} é apresentado na Equação 2.9 e a função de ativação normalmente refere-se a uma função sigmoide das variáveis de entrada (YADAV; CHANDEL, 2017). Ainda para (YADAV; CHANDEL, 2017), a função de ativação pode ser obtida pela Equação 2.11.

$$\sigma\left(x^{MLP}\right) = \frac{1}{1 + e^{\left(-x_i^{MLP}\right)}} \tag{2.11}$$

A quantidade de neurônios, H_n, das camadas ocultas, em conformidade com (YADAV; CHANDEL, 2017), pode ser calculada utilizando a Equação 2.12.

$$H_n = \frac{i_n + s_n^{MLP}}{2} + \sqrt{S_n} \tag{2.12}$$

Onde i_n e s_n^{MLP} se referem respectivamente ao número de entradas e saídas, e s_n é o número de amostras usadas no modelo de predição (YADAV; CHANDEL, 2017).

As redes neurais possuem diferentes modelos, o *feedforward* (alimentação para frente) e com uso de *backpropagation* (realimentação) (NOGUEIRA *et al.*, 2018). O modelo mais comum trata-se da *feedforward*, como o próprio nome sugere o processo acontece do início para frente e não há *feedback* entre as camadas neurais (NOGUEIRA *et al.*, 2018). Esse sistema, embora não tenha retroalimentação pode ser aplicado em problemas de maior complexidade, como reconhecimento de padrões (NOGUEIRA *et al.*, 2018). Entretanto, o modelo *feedforward* não é adequado para sistemas dinâmicos (NOGUEIRA *et al.*, 2018; YADAV; CHANDEL, 2017; CARDONA; NEDJAH; MOURELLE, 2017). Nesse caso faz necessário usar um modelo mais avançado com feedback, onde a saída da camada final é ligada a entrada de forma retroativa visando minimizar o erro entre cada camada (GHOLAMI *et al.*, 2019; NOGUEIRA *et al.*, 2018).

Em um modelo com *feedback*, faz-se o uso *backpropagation* (realimentação) da saída para as entradas recalculando o vetor peso com o conjunto de treinamento de forma minimizar o erro das saídas intermediarias e final de cada camada (GHOLAMI *et al.*, 2019; SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019; YADAV; CHANDEL, 2017). O processo se repete até que o erro entre a saída desejada e saída predita seja minimizado, essa diferença, erro entre o real e o predito (NOGUEIRA *et al.*, 2018). Ou seja, processo de realimentação tem como objetivo de recalcular os pesos de forma que reduza o erro entre o valor real e predito. De forma sumarizada, o processo de realimentação pode ser entendido como o ajustes dos pesos w_i^{MLP} iniciando o peso precedente, w_{i-1}^{MLP}, subtraído pelo produto entre o coeficiente de aprendizado, MLP, e a derivada parcial do MSE por cada vetor de pesos, $\frac{\delta MSE}{\delta w^{MLP}}$ (SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019; AVUÇLU; BAŞÇIFTÇI, 2018; FRANK; HALL, 2011). Em conformidade com (SHARIFZADEH; SIKINIOTI-LOCK; SHAH, 2019; AVUÇLU; BAŞÇIFTÇI, 2018; FRANK; HALL, 2011) o cálculo do *backpropagation* pode ser compreendido como o gradiente decrescente, pois parte-se da saída para a entrada do sistema de camadas. Este encontra-se na Equação 2.13.

$$w_i^{MLP} =\leftarrow w_i^{MLP} - \alpha^{MLP} \frac{\delta MSE}{\delta w^{MLP}} \tag{2.13}$$

Onde o coeficiente de aprendizagem, α^{MLP}, pode ser compreendido a um coeficiente proporcional ao ajuste realizado em cada peso (AVUÇLU; BAŞÇIFTÇI, 2018).

2.4.2.6 K-means Clustering

O agrupamento de dados na mineração é considerado um método não supervisionado que agrupa em amostras com mesmo padrão de comportamento e características similares (AFSHOON; MIRI; MOUSAVI, ; RAHMAN; ISLAM, 2018). A avaliação da qualidade do agrupamento é obtido por meio da Equação 2.14.

$$\min \sum_{k_c=1}^{K} \sum_{i=1}^{n} y_{ik} D(x_i^{K-means}, K_c). \tag{2.14}$$

Onde $x_i^{K-means}$ são entradas de dados, n o número de amostras, $D(x_i^{K-means}, K_c)$ é a função de distância, K_c os centroides dos grupos e y_{ik} é uma variável que se atribui o valor de 1 se a entidade i pertencer ao *clustering* K_c.

O método de K-means pode ser sumarizado no seguintes passos (AFSHOON; MIRI; MOUSAVI, ; AHMAD; KHAN, 2020):

1. Seleção do número de *clusterings* K_c;
2. São selecionados aleatoriamente os centroides K_c;
3. Compara-se a distância entre os dados e cada centroide dos grupos, agrupa-se os dados no grupo do centroide mais próximo;
4. Os centros dos grupos são atualizados a cada iteração de acordo com a média das amostras atribuidas a cada grupo;
5. A partir dos novos centros os dados são agrupados novamente;
6. As etapas 3, 4 e 5 são repetidas até que os centroides sejam fixados.

2.4.2.7 Cross-validation

A *Cross-Validation* (Validação cruzada) trata-se de um método científico para avaliar a confiabilidade de um modelo de previsão, reduzindo dessa forma a lacuna entre o conjunto de treinamento e de teste e consequentemente no valor real e predito (ZHANG *et al.*, 2019). Entre outras técnicas, o banco de dados após a etapa de pré-processamento é dividido em K_i *cross* subconjuntos (ZHANG *et al.*, 2019). Em K_j^{cross} iterações um subconjunto é utilizado como dados de teste e todos os outros K_{i-1}^{cross} subconjuntos são utilizados como treinamento (ZHANG *et al.*, 2019). Na próxima

iteração o subconjuntoK_{i+1}^{cross} será o próximo subconjuntode teste a ser treinado com cada um dos outros K_{i-1}^{cross} subconjuntos de treinamento e assim sucessivamente até a última iteração $K_{i=n}^{cross}$ (ZHANG *et al.*, 2019).

Em seguida, são gerados valores preditos para cada amostra de teste (ZHANG *et al.*, 2019). O modelo de previsão adotado refere-se ao que apresenta o menor erro em relação aos dados preditos e originais (GONG *et al.*, 2018). Segundo (GONG *et al.*, 2018; AL-WAELI *et al.*, 2019), a métrica mais utilizada como critério de escolha do melhor algoritmo para problemas de regressão é o MSE.

Em alguns casos, os resultados obtidos não conseguem convergir para uma predição assertiva, ou seja, apresentam em sua maioria um MSE muito grande ou uma taxa de acerto baixa ou negativa (SAIDI *et al.*, 2018; DJELLOULI *et al.*, 2018). Um dos principais fatores para a ineficácia da previsão com métodos de regressão a ausência de correlação das variáveis uma vez que, a mineração de dados investiga e compreende as correlações entre as variáveis (SAIDI *et al.*, 2018; DJELLOULI *et al.*, 2018). A correlação pode ser obtida pela Equação 2.15.

$$R^2 = 1 - \frac{\sum(y_{real - y_{predito}})^2}{\sum(y_{real} - \overline{y}_{predito})^2} \tag{2.15}$$

Onde o y_real trata-se do valor predito e y_{real} o valor real.

CAPÍTULO

3

METODOLOGIA

3.1 Classificação metodológica da pesquisa

Em conformidade com Prodanov e Freitas (2013), esse estudo caracteriza-se enquanto uma abordagem de análise quantitativa, visto que, as variáveis da pesquisa serão tratadas de forma estatística. Ainda para Prodanov e Freitas (2013), no que diz respeito aos procedimentos técnicos, esta pesquisa enquadra-se em experimental, uma vez que visa a análise da influência de variáveis conhecidas e controladas de dados de literatura. Ademais, devido a necessidade de determinar o *lead time* do processo por meio de uma abordagem de avaliações de padrões, a pesquisa classifica-se em relação a seus objetivos como explicativa, uma vez que investiga, mensura, registra e analisa fatores que influenciam uma ocorrência de um determinado fenômeno (PRODANOV; FREITAS, 2013). Ainda em conformidade com os autores, este estudo enquadra-se enquanto a sua natureza em básica, por propor o uso de uma nova metodologia para análise e estimativas do *lead time* de processos.

3.2 KDD: mineração de dados

Para o desenvolvimento do presente estudo, foi usada a metodologia estruturada no processo *knowledge discovery in database* (KDD) propostas por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). As etapas presentes no ciclo KDD encontram-se ilustradas na Figura 1.

A qualidade do conhecimento e do comportamento descoberto está relacionada às ferramentas e técnicas empregadas em cada uma das etapas da KDD (RISTOSKI; PAULHEIM, 2016). Essas etapas serão apresentadas nas secções subsequentes.

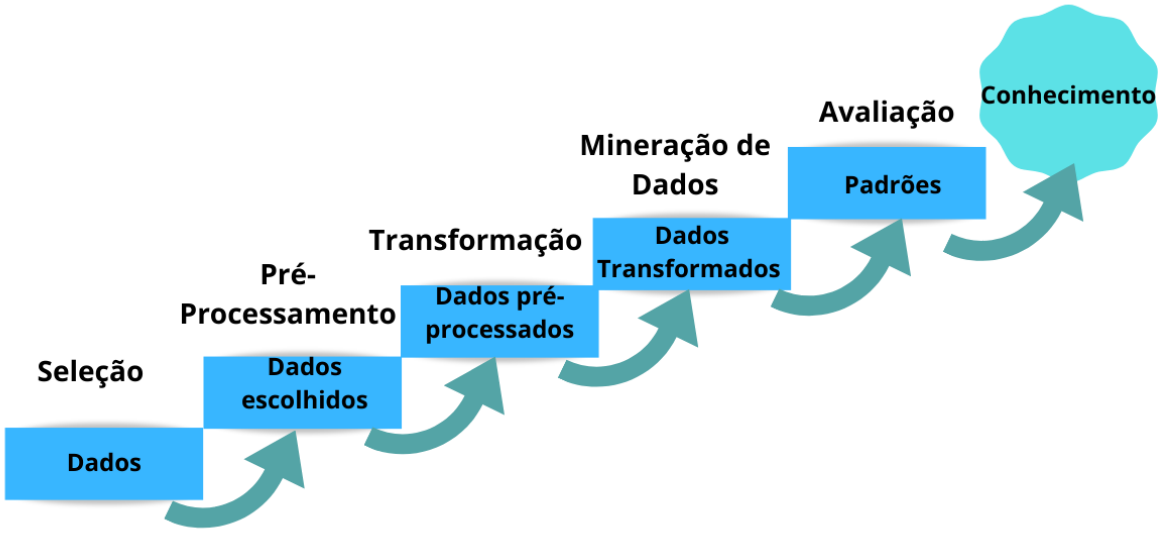


Figura 1 – Etapas do ciclo KDD.

3.2.1 Etapa 1: Seleção

A eficácia do desempenho da mineração de dados e do conhecimento extraído depende de uma análise preliminar dos dados e sua compreensão (RISTOSKI; PAULHEIM, 2016). Diante disso, a fase de seleção trata-se da fase preliminar de análise do banco de dados (RISTOSKI; PAULHEIM, 2016). Nessa fase, os objetivos da mineração são revisados (RISTOSKI; PAULHEIM, 2016). Nessa etapa análises gráficas de todo o banco de dados com o intuito de identificar inicialmente o comportamento das amostras no banco de dados, suportando dessa forma a decisão de quais *ranges* de amostras e atributos do banco de dados utilizar nas etapas subsequentes.

3.2.2 Etapa 2: Pré-processamento

Na etapa de pré-processamento é realizada uma limpeza dos ruídos e anomalias do banco de dados (RISTOSKI; PAULHEIM, 2016). Os dados duplicados e corrompidos, valores extremos, valores ausentes, não verídicos e caracteres especiais são identificados e removidos da base de dados (RISTOSKI; PAULHEIM, 2016). Além disso, a fusão e correlação de dados de diferentes bancos de dados são construídas e os possíveis conflitos e erros dessa fusão são corrigidos (RISTOSKI; PAULHEIM, 2016). Diversas ferramentas podem ser aplicadas nesta etapa, como algoritmos, softwares e métodos estatísticos.

Sobre os métodos estatísticos, a análise de box plot pode ser aplicada para identificar os valores extremos e *outliers* (SAGAERT *et al.*, 2019; HU *et al.*, 2018). As

amostras fora dos interquartis, extremidades do primeiro e terceiro quartil, do gráfico box plot são considerados *outliers*.

3.2.3 Etapa 3: Transformação

Na etapa Transformação, os dados do banco de dados são transformados para um formato que possam ser utilizados na mineração de dados (RISTOSKI; PAULHEIM, 2016). Esta etapa pode incluir diversas atividades de transformação como, como geração, seleção de e transformação de atributos, agregação, discretização de dados, amostragem de instância de teste e treinamento, mudança do tipo do caráter das amostras e transformação por meio de funções (RISTOSKI; PAULHEIM, 2016). Além disso, o método de validação cruzada pode ser utilizado para definir o melhor algoritmo de machine learning antes da predição final. Normalmente os algoritmos mais usuais na literatura são testados na base de dados. Além disso, o melhor algoritmo para base de dados pode ser escolhido com base no o menor valor erro quadrado médio (MSE) da predição entre todos os algoritmos testados.

3.2.3.1 Validação cruzada: algoritmos de machine learning propostos

Na transformação, para melhor entendimento e aprimoramento da análise, os dados podem ser divididos em K_i folds, ou seja, K_i amostras de dados que serviram como arquivo de teste e treino. A etapa de validação cruzada não é obrigatória, no entanto, todas as análises preliminares, segmentações, transformações e a validação cruzada podem ser contribuir para a melhora decisão da previsão. Se necessário, os valores dos atributos podem ser transformados em diferentes tipos de dados, como numérico, nominal e binário. Os resultados da previsão podem ser melhores usando determinados tipos de dados, portanto, todas as possibilidades podem ser testadas. O processo de validação cruzada é a fase para testar essas possibilidades, como tipos de algoritmo e tipo de dados. Em outras palavras, a validação cruzada é o processo de teste e treinamento do banco de dados com o objetivo de definir o algoritmo de aprendizado de máquina apropriado. Durante a validação cruzada, o banco de dados é dividido em K_i número de folds e cada folds é utilizada como arquivo de teste em uma interação e todos os demais como arquivos de treinamento durante K_j iterações.

Os métodos mais comuns de aprendizado de máquina aplicados na validação cruzada são os *k-nearest neighbors*, florestas aleatórias, regressão linear e *multi-layer perceptron*. Estes algoritmos foram abordados de forma detalhada na seção do referencial teórico.

3.2.4 Etapa 4: Mineração de dados

Na etapa de mineração de dados, a tarefa de mineração, classificação, predição, associação, já foi definida com base nos objetivos anteriores, nos resultados de validações cruzadas (RISTOSKI; PAULHEIM, 2016). As categorias da tarefa podem ser separadas em dois objetivos de decisão: descrição e previsão (RISTOSKI; PAULHEIM, 2016). A descrição é frequentemente associada a algoritmos de mineração de dados não supervisionados, que pretendem descobrir comportamentos e padrões interpretáveis no banco de dados (RISTOSKI; PAULHEIM, 2016). Em contrapartida, a previsão está associada a algoritmos supervisionados, que pretendem prever a variável desconhecida de um banco de dados (RISTOSKI; PAULHEIM, 2016). Normalmente, nessa etapa, o banco de dados é dividido em 20% das amostras para o conjunto de testes e cerca de 80% das amostras para o conjunto de treinamento. Além disso, o desempenho dos resultados da mineração de dados pode ser medido com base em diversos indicadores, como taxa de acertos e erros.

Consequentemente, esta pesquisa escolherá o menor erro MSE, para decisão do melhor algoritmo e tipo de dado obtido durante os experimentos realizados na validação cruzada. Além disso, esse estudo fara uso da taxa de acerto e erro para verificar o desempenho de previsão em relação ao valor real.

3.2.5 Etapa 5: Interpretação do conhecimento

Por fim, nesta seção, são examinados o comportamento e padrões fornecidos durante as análises do ciclo KDD (RISTOSKI; PAULHEIM, 2016). Se necessário, esta etapa pode usar os painéis de visualização, gráficos, valores numéricos extraídos pelo processo de mineração de dados e outros (RISTOSKI; PAULHEIM, 2016).

A interpretação do conhecimento, dos dados gerados pelo método de previsão são analisadas e pode auxiliar na consolidação da tomada de decisão, podendo ser capaz de identificar comportamentos de padrões já explícitos ou novos.

CAPÍTULO

4

EXPERIMENTOS E RESULTADOS CASO 1: SETOR FARMACÊUTICO

Como já descrito, essa pesquisa fará o uso de três bancos de dados. Os bancos de dados do setor farmacêutico, setor de automação para o segmento cerâmico e de trâmites de serviço eletrônico os quais são chamados de Caso 1, Caso 2 e Caso 3 respectivamente. Esta seções posteriores apresenta a descrições da empresa, seleção de dados, pré-processamento, transformação de dados, mineração de dados e interpretação para cada caso.

4.1 Banco de dados do setor farmacêutico

O banco de dados analisado foi disponibilizado por uma empresa de serviços integrados, o Grupo Coopservice, fundado em 1992, que presta serviços especializados a empresas do âmbito público e privado. A empresa com sede na Itália, atua mundialmente, contando com cerca de 22.000 funcionários. A Coopservice oferece todos os serviços da instalação, especialmente aqueles que não fazem parte do negócio principal dos clientes, incluindo limpeza industrial, comercial e de saúde; gestão e manutenção de edifícios e sistemas; gestão de suprimentos de energia; segurança e vigilância; transporte e manuseio de mercadorias; movimentação industrial e comercial; coleta e transporte de resíduos especiais. Além disso, a empresa possui 18 armazéns logísticos com uma área de armazenamento de mais de 150000 metros quadrados e é líder em logística de assistência médica e farmacêutica na Itália e fornecedora principal de serviços de gerenciamento e distribuição de produtos farmacêuticos, dispositivos médico-cirúrgicos e consumíveis não médicos. A empresa busca aprimoramentos contínuos para melhorar e manter a qualidade e desempenho consistente, além de usar ferramentas, equipamentos e veículos de alta tecnologia.

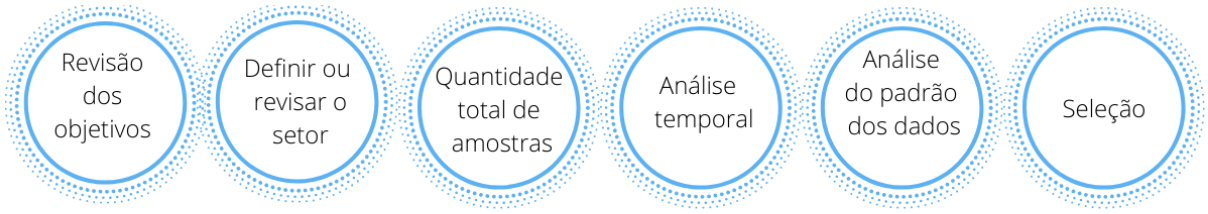
A previsão do *lead time* na Coopservice é crucial, porque com uma previsão precisa é possível otimizar e gerenciar o agendamento de cargas, bem como prever o processo de descarregamento para a área de entrada. Ademais, é possível realizar de forma mais eficaz o planejamento da produção, reorganizando de forma mais eficaz os turnos dos funcionários no armazém. A previsão do *lead time* permite que a empresa controle os fornecedores e avalie o desempenho. Finalmente, com uma previsão precisa, o gerenciamento do estoque no armazém pode ser mais seguro, evitando fenômenos como excesso ou falta de estoque.

4.1.1 Seleção

Na etapa de seleção, foi realizada a primeira análise gráfica do comportamento das amostras de dados. A análise gráfica foi dividida em seis análises, são elas: a reafirmação dos objetivos e e do setor do banco de dados, a quantidade total de amostras, uma análise temporal das amostras por mês, o comportamento dos dados em relação a distribuição do *lead time* e seleção de amostras e variáveis relevantes.

As análises gráficas foram resumidas na Figura 2.

Figura 2 – Etapas da fase seleção.



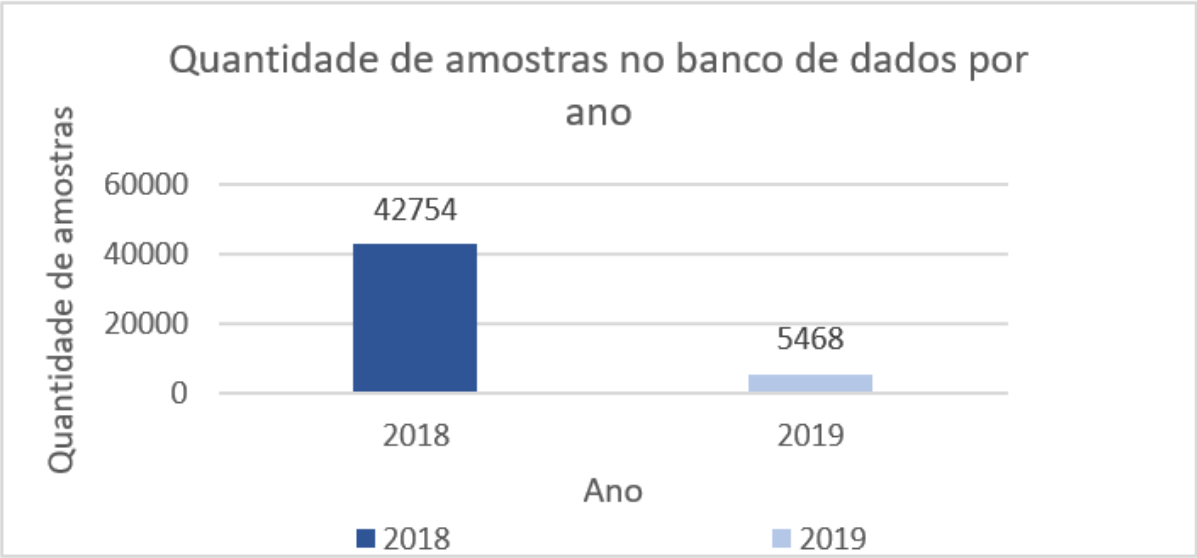
Fonte: Os autores

O objetivo da análise continua com foco em técnicas de predição e o uso do banco de dados da cadeia de suprimentos do setor farmacêutico. Com o intuito de identificar a quantidade de amostras, foi feito uma análise gráfica da quantidade de amostras por ano e por categoria de produto. O gráfico com a quantidade de amostras por ano encontra-se na Figura 3.

O banco de dados totalizou 48.222 amostras, sendo que, 42.754 amostras estão relacionadas ao ano de 2018 e 5.468 ao ano de 2019. Isso significa que 88,6% das amostras estão relacionadas ao ano de 2018 e 11,4% ao ano de 2019. Um dos motivos para essa discrepância entre os dados de cada ano pode estar relacionado ao fato de que o ano de 2018 possui amostras de dados de todos os meses do ano, já o ano de 2019 possui apenas as amostras do mês de janeiro e fevereiro.

Todos os produtos farmacêuticos no banco de dados analisados estão associados a categorias de produtos. As categorias relacionadas aos produtos são tumores,

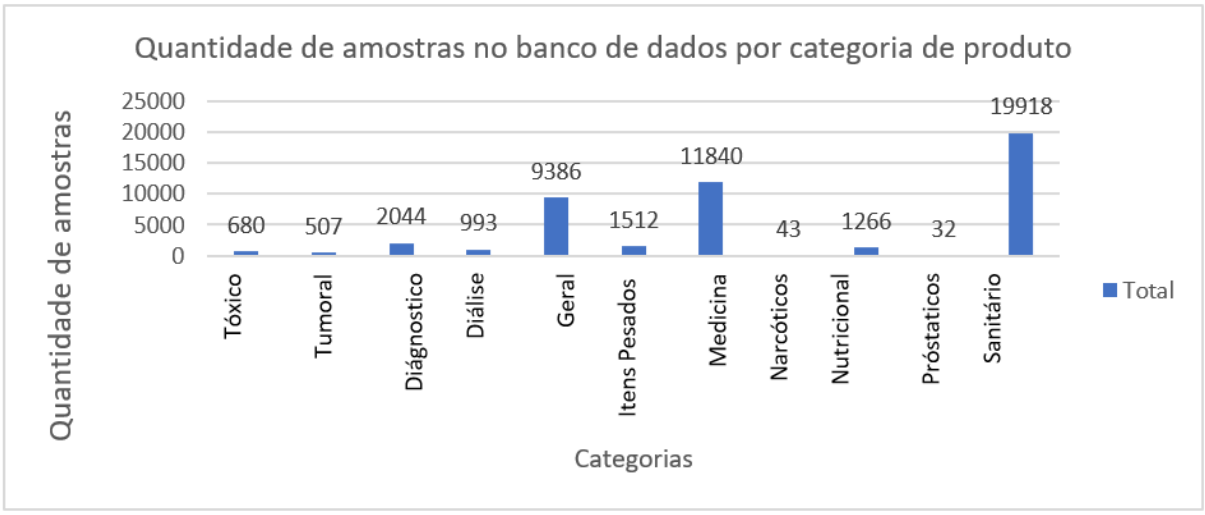
Figura 3 – Número de amostras do banco de dados do Setor Farmacêutico de 2018 a 2019 por ano.



Fonte: Os autores

diagnóstico geral, medicamentos, nutrição, prostáticos, sanitário, diálise, produtos pesados, tóxicos e narcóticos. A Figura 4 mostra a relação entre o número de amostras por categoria de produto.

Figura 4 – Número de amostras do banco de dados do setor farmacêutico de 2018 e 2019 por categoria.



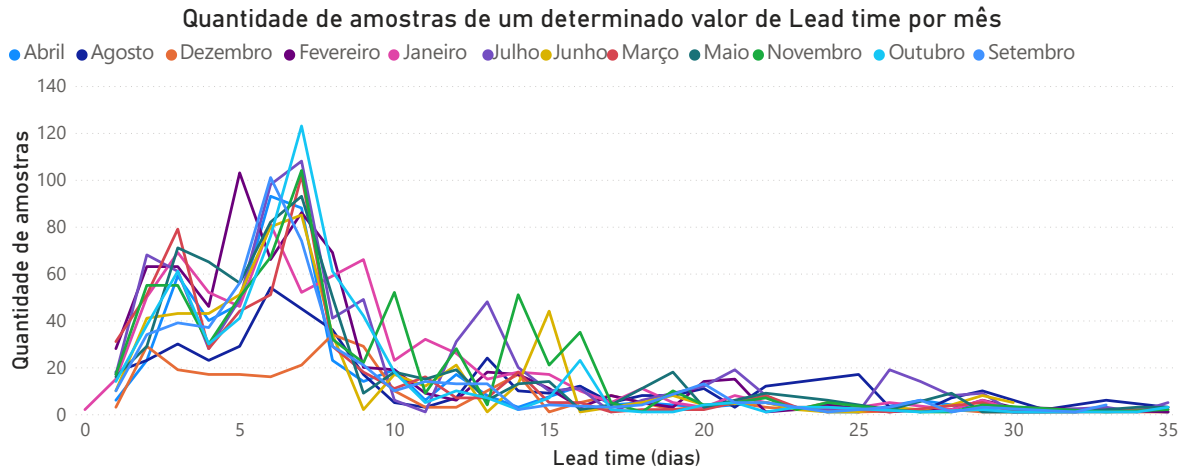
Fonte: Os autores

As amostras de dados estão concentradas nas categorias sanitária, medicina e geral. Esta pesquisa investigou o *lead time* considerando os dados do ano de 2018, pois há dados para todos os meses do ano. Em relação às categorias, embora existissem categorias com poucos dados, devido ao grau de relevância das categorias na previsão do *lead time* todas as categorias foram mantidas.

Visando compreender quais os valores de *lead time* são mais frequentes e atípicos foi avaliado o comportamento do em cada mês na cadeia de suprimentos

do *lead time* em relação setor farmacêutico foi obtido um gráfico da quantidade de amostras por valor de *lead time* por mês presente no banco de dados. Esses valores podem ser observados na Figura 5.

Figura 5 – Quantidade de amostras associadas a cada valor de *lead time*.



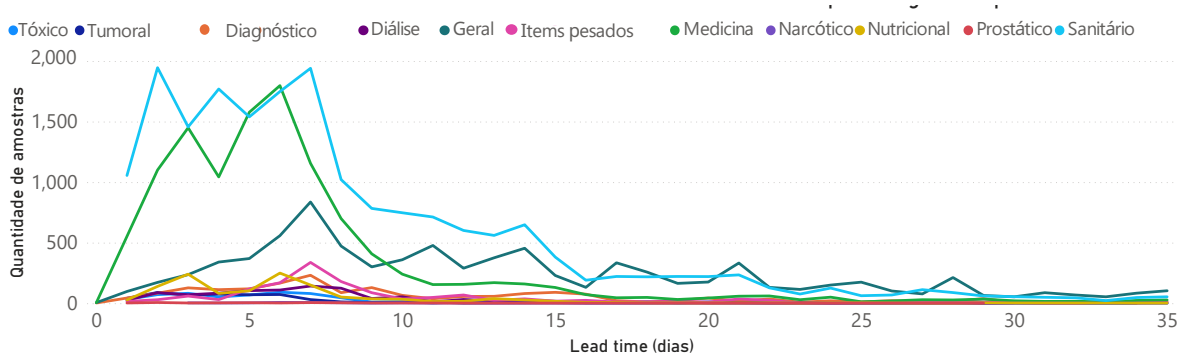
Fonte: Os autores.

Na Figura 5 pode ser visto que os valores de *lead time* mais frequentes concentram-se nos valores entre 1 a 22 dias. Após 32 dias as concentrações de amostras de *lead time* diminuem significativamente, podendo indicar que valores de *lead time* acima de 32 dias poderiam ser considerados valores atípicos e portanto *outliers*.

Nota-se que existiram *lead times* com valores inferiores a 1 dia, nesses casos, normalmente, o pedido foi realizado e entregue no mesmo dia que foi solicitado. Esse padrão de comportamento pode ocorrer em empresas que trabalham com sistemas de estoque e ressuprimento *make-to-stock* (feito para estocar - MTS). Além disso, o comportamento gráfico do *lead time* foi similar para todos os meses o que pode indicar que indiferença na distinção na detecção de padrões de dados com a presença do atributo mês. A relação entre distribuição das amostras de *lead time* por categoria de produto, está representada no gráfico da Figura 6. As categorias presentes são Tóxicas, Tumoral, Diagnóstico, Diálise, Itens Pesados, Narcótico Nutricional e Prostático.

Analisando essas categorias, a Figura 6 mostra que o número de amostras permanece concentrado em valores do *lead time* entre 0 e 20.

Figura 6 – Quantidade amostras de *lead time* por categorias de produto Tóxico, Tumoral, Diagnóstico, Diálise, Itens Pesados, Narcótico, Nutricional, Prostático, Geral, Sanitário e Medicina.



Fonte: Os autores.

No geral, tanto a quantidade de amostras de cada valor de *lead time* por mês quanto por por categorias possui mais amostras concentradas em valores de dias entre 0 a 32 dias. Existem poucas amostras de dados concentradas em *lead time* mais altos, acima de 32 dias por exemplo. Diante do exposto, para a continuidade da análise, o range de valores de lead time utilizados na predição do *lead time* consideram *lead time* de até 32 dias. Na seleção de variáveis relevantes, foram removidos aqueles com informação duplicadas, porém formato diferente e aqueles cuja ausência não aumentasse o erro de previsão. Os variáveis originais que permaneceram no banco de dados são:

- Dia, de um a vinte e oito ou trinta, do pedido do cliente (Numérico);
- Dia da semana, de segunda a sexta-feira (Numérico);
- Mês, de um a doze, (Numérico);
- Código do Fornecedor da cadeia de suprimentos (Numérico e String);
- Nome do produto (String);
- Categoria de tipo de produto farmacêutico (String);
- Quantidade solicitada (Numérica);
- Distância geográfica entre cada fornecedor e o armazém da farmácia (km);
- *Lead time* de fornecimento para cada fornecedor por pedido na cadeia de suprimentos (Numérico).

Em resumo, na seleção, foram definidos que para a previsão se utilizará dados de 2018, considerando um prazo de entrega de até 32 dias e usando 8 variáveis,

dia, dia da semana, mês, fornecedor, produto, categoria de produto e 1 variável a ser predita, *lead time* de fornecimento.

4.2 Pré-processamento

Nesta etapa, os ruídos das amostras do banco de dados foram identificados e eliminados. Para a identificação dos *outliers* e valores extremos a análise de *box plot* foi feita para apontar possíveis discrepâncias restantes e eliminá-las do banco de dados. Além disso, valores duplicados, dados desnecessários e corrompidos foram removidos. A Tabela 1 mostra a quantidade de dados com e sem ruídos, *outliers* e valores extremos, identificados na análise *box plot* para o total de amostras.

Tabela 1 – Quantidade de amostras com a presença de *outliers* e valores extremos do caso 1

Total	Porcentagem com ruídos(%)
Outliers	3,9%
Valores extremos	5,9%

Fonte: Os autores

Este banco de dados contém poucos valores extremos, não mais que 6%. No banco de dados, em 96,1% das amostras não foram detectados *outliers*, sendo 3,9% das amostras foram identificadas como *outliers*. Além disso, 94,1% das amostras não apresentaram valores extremos e 5,9% dos dados amostrais foram detectadas como valores extremos. Esses valores extremos e *outliers* foram removidos do banco de dados. Além disso, todos os dados duplicados e corrompidos também foram removidos.

4.3 Transformação

Inicialmente, os dados foram segmentados em dois bancos de dados: por mês e por categoria de produto. Para cada banco de dados, por mês e por categoria, foram gerados três bancos de dados com as mesmas informações e diferentes tipos de formato de dados. Os formatos utilizados foram nominais, binário e numérico.

No tipo numérico, todas as variáveis não numéricas no banco de dados foram convertidas em tipo numérico, como o nome do produto e o nome do fornecedor. No tipo nominal, todos os tipos de variáveis foram transformados em string, por exemplo, o código do produto, o código do fornecedor, o dia e o mês. No tipo binário, foi atribuído o valor 1 se identificou a presença de uma determinada variável ou 0 se estava ausente. Em casos de variáveis com valores constates, como por exemplo distância geográfica, foi atribuído 1 para o valor real da variável e zero para todos os demais valores presentes no banco de dados. Além disso, foram formados um total de 36 bancos de dados para

a categoria mês e 33 bancos de dados para a segmentação por categoria. Todos os 69 bancos de dados foram testados na validação cruzada com os cinco algoritmos mais comuns na literatura sendo esses: regressão linear (Equação 2.4), máquinas vetoriais de suporte (Equação 2.7), k-vizinhos mais próximos (Equação 2.1), perceptron multicamada (Equação 2.9, 2.10, 2.11, 2.12 e 2.13) e floresta aleatória (Equação 2.2). Na validação cruzada, os bancos de dados foram divididos em K_i folds, a cada iteração cada fold foi usado como um conjunto de teste e todas os outros foram usados como um conjunto de treinamento até que todos K_i folds foram usados em K_j iterações. Nesta pesquisa foi definido o uso de $K_j = K_i = 10$ número de folds. O melhor algoritmo para um banco de dados foi o que apresenta o menor erro quadrático médio.

Considerando um sistema com o $K_j = K_i = 10$ número de folds, foi obtido na validação cruzada o MSE para os dados segmentados por categoria de produto. Foram considerados os dados no formato nominal, binário e numérico para cada algoritmo testado na pesquisa encontram-se na Tabela 2.

Tabela 2 – Média do MSE dos dados segmentados por categoria de produtos de cada algoritmo com dados do tipo de dados nominais, numérico e binário do setor farmacêutico.

Formato	Média MLP	Média LR	Média KNN	Média RF	Média SVM
Binário	4,88	3,59	2,98	2,32	2,45
Nominal	4,09	3,44	3,08	2,84	2,50
Numérico	4,67	3,76	2,92	2,63	2,36

Fonte: Os autores

Nota-se que o melhor algoritmo se trata do RF com os dados no formato binário com valor de MSE de 2,32. O calculo do MSE na etapa de validação cruzada foi realizada também para os dados segmentados por mês. Os resultados do MSE na etapa de validação cruzada obtidos por cada mês com os dados no formato binário, numérico e nominal encontram-se encontra-se na Tabela 3.

Tabela 3 – Média do MSE dos dados segmentados por por mês de produtos de cada algoritmo com dados do tipo de dados nominais, numérico e binário do setor farmacêutico.

Formato	Média MLP	Média LR	Média KNN	Média RF	Média SVM
Binário	7,84	4,94	3,72	2,80	1,89
Nominal	7,84	4,97	3,17	3,42	1,88
Numérico	7,71	3,06	3,18	3,80	2,23

Fonte: Os autores

Embora os valores obtidos por mês e por categoria de produto tivesse valores próximos, o algoritmo SVM apresentou o melhor desempenho em todos as experimentações realizadas na validação cruzada. Os valores binários foram melhores que os

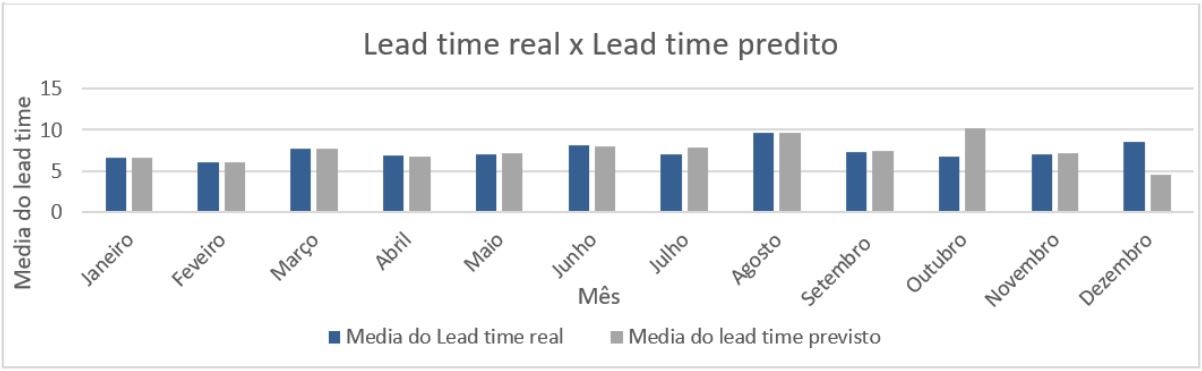
demais ou iguais ao nominal. Portanto, para a previsão do *lead time* para esse banco de dados foi realizada com os dados no formato binários usando o algoritmo SVM,com os dados segmentados por mês.

4.4 Mineração de dados

Conforme exposto na seção anterior, a mineração de dados foi realizada com dados particionados por mês, o uso dos dados no formato binário, usando o algoritmo SVM. Os parâmetros usados no algoritmo SVM, Equação 2.7, para obter um resultado com o menor erro possível, foram o kernel polinomial RegSMO-aprimorado epislon: $(1.0)^{-}12$

Além disso, para a predição final cada um dos 12 partições de dados por mês no formato binário foram divididos em um conjunto de teste com 33,3% do total de amostras e um conjunto de treinamento com 66,7% do total de amostras para cada mês. Foram obtidos individualmente para cada amostra pertencente aos 12 bancos de dados a taxa de erro e de acerto para cada uma das 12 partições de forma individual. Diante da quantidade significativa de informações a Figura 7 apresenta a média dos resultados obtidos para cada amostra por mês.

Figura 7 – Comparação entre o *lead time* real e o *lead time* previsto.

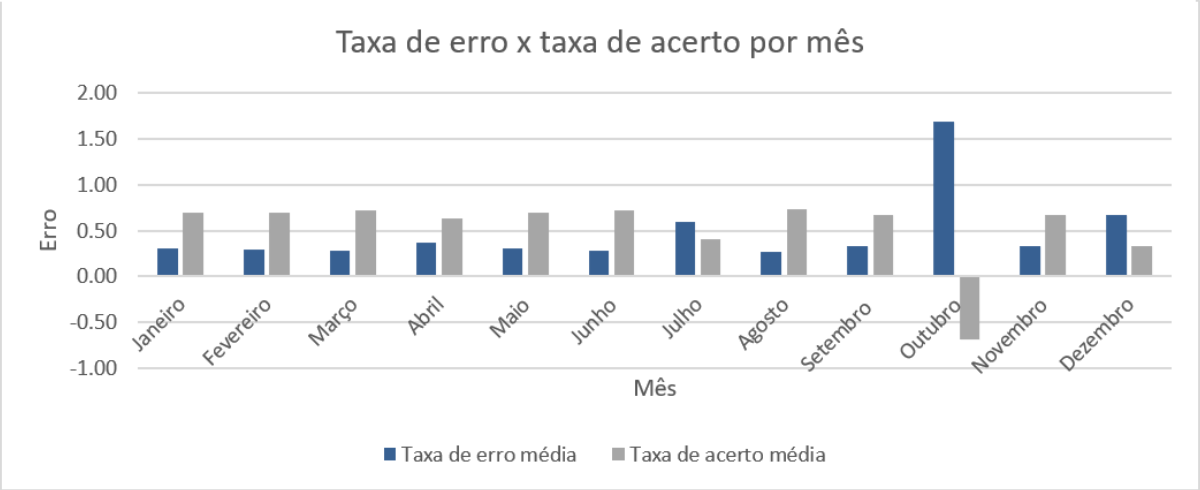


Fonte: Os autores

No geral, observa-se que os valores previstos foram próximos aos reais. Em média o *lead time* previsto é igual ao *lead time* real, exceto em outubro e dezembro, cuja diferença foi de 3 dias. O gráfico da taxa de acertos e erros é apresentado na Figura 8.

Note-se que a taxa de acerto foi superior à taxa de erro em todos os meses, exceto julho, outubro e dezembro. Entre as possíveis razões esse comportamento nos meses de julho, outubro e dezembro pode estar associado ao fato de o período de festividades aumentar a demanda por um determinado tipo de produto. Portanto, para pesquisas futuras, sugere-se investigar esse comportamento atípico. Além disso, em

Figura 8 – Comparação entre taxa de erro x taxa de acerto.



Fonte: Os autores

média, as amostras atingiram entre 63% e 73% de taxa de acerto, exceto nos meses de julho, outubro e dezembro.

4.4.1 Interpretação

A maioria dos pedidos está relacionada às categorias de produtos como sanitários, gerais e medicamentos, Figura 4. Embora a categoria geral possua uma das maiores quantidades de amostras, o tempo do *lead time* está concentrado entre 0 e 22 dias, Figura 6. Além disso, as categorias nutricionais, itens pesados, diagnóstico, saúde e medicina têm uma quantidade de pedidos alta concentrada em pequenos valores de *lead time*, com um *lead time* de 10 dias, Figura 6. No entanto, as categorias de diagnóstico, itens pesados e nutricionais apresentam entre 250 e 300 amostras para o intervalo de *lead time* mencionado. A categoria medicamento tem cerca de 1500 a 2000 amostras com *lead time* concentrado em até 8 dias, Figura 6. A distribuição das amostras de *lead time* por mês, Figura 5, segue um padrão de comportamento semelhante para todos os meses. Estes possuem maior quantidades de amostras concentradas em valores baixos de *lead time*. Observa-se na Figura 5 os meses de agosto, dezembro e abril têm o menor número de pedidos, com quantidades de amostra variando de 100 a 200 para prazos de entrega de até 24 dias. Os meses de janeiro e fevereiro, Figura 5, têm maiores quantidades de pedidos que varia entre 300 a 600 amostras com prazo de entrega de até 23 dias.

Em resumo, é possível observar que algoritmos inteligentes têm valores de previsão mais significativos, ou seja, mais próximos dos valores reais conforme Figura 7. Isso ocorre porque os métodos inteligentes não apenas trabalham com tendências passadas, mas buscam aprender o comportamento dos dados. Na validação cruzada, o melhor método para o banco de dados farmacêutico foi o SVM, Tabela 3. Observou-se

que o *lead time* previsto por cada amostra foi em sua maioria satisfatório, uma vez que cerca de 74,85% das predições apresentaram um valor de taxa de acerto superior a 60%. Isso demostra a eficácia da previsão do *lead time* utilizando métodos inteligentes. Além disso, em média as amostras previstas, Figura 8, alcançaram entre 63% e 73% de taxa de acerto.

CAPÍTULO

5

EXPERIMENTOS E RESULTADOS CASO 2: SETOR DE AUTOMAÇÃO INDUSTRIAL PARA O SETOR CERÂMICO

5.1 Base de dados do setor de automação industrial para o setor cerâmico

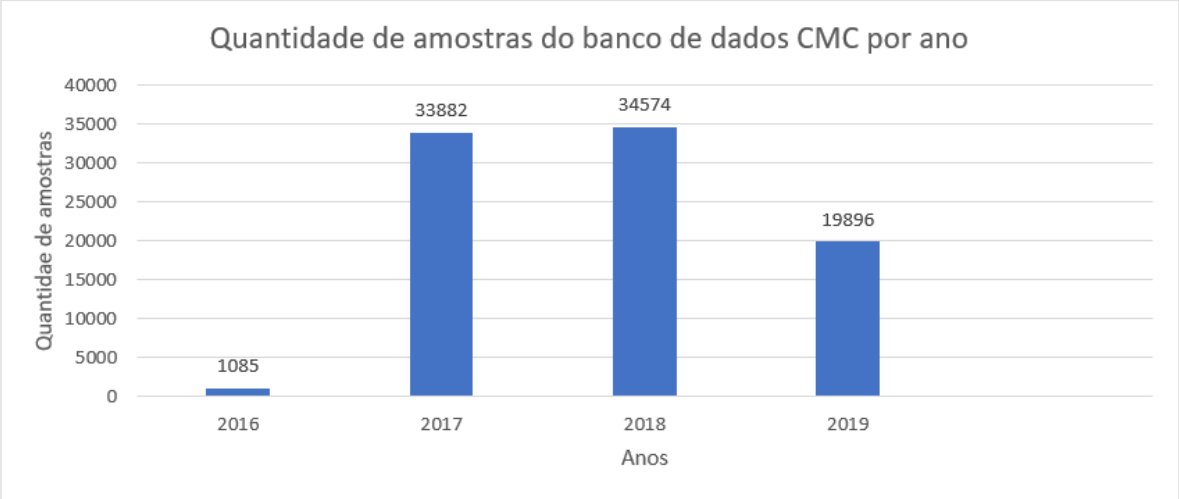
O banco de dados analisado foi disponibilizado pela oficina CMC SACMI líder mundial de fornecimento de produtos para diversos segmentos. A oficina CMC SACMI, com sede em Salvaterra di Casalgrande Itália,foi fundada em 1981, fabricando automações para empresas cerâmicas. Posteriormente, a empresa ampliou seu portifólio, fornecendo produtos de automação para modelagem plástica, embalagens, engarrafamento, processos alimentares, sistemas de transporte que se conectam com fornos, prensas, secadoras, vidraças e outros.

A oficina CMC SACMI continua crescendo e ampliando a fabricação de seus produtos de automação industrial para setores como o de componentes metálicos e laminados, componentes sinterizados, prensa e processamento de mármore. O Grupo SACMI opera globalmente nos cinco continentes, está presente em cerca de 30 países com mais de 80 empresas de produção, distribuição e serviços (SACMI GROUP, 2019). Além disso, o Grupo SACMI desenvolve constantemente quase 85% de seus negócios no exterior. Nos últimos 5 anos, o Grupo SACMI investiu mais de € 220 milhões em atividades de pesquisa e desenvolvimento (P&D) (SACMI GROUP, 2019). Além disso, o SACMI Group possui as certificações de qualidade ISO 14001, ISO 9001.

5.2 Seleção

Assim como realizado no Caso 1, nesta fase foram revisadas as metas e o setor de interesse do banco de dados, foram feitas algumas análises gráficas descritivas, como quantidade total de amostras, análise temporal do banco de dados e sua relação com o *lead time*, assim como a seleção das amostras e atributos relevantes. Na revisão das metas e objetivos do setor foi validada a análise da previsão do *lead time* usando a mineração no banco de dados da cadeia de suprimentos da indústria de automação cerâmica. Sobre a análise descritiva gráfica, o gráfico sobre o número de amostras no banco de dados por ano é mostrado na Figura 9.

Figura 9 – Número de amostras do banco de dados do setor de automação cerâmica de 2016 a 2019 por ano.



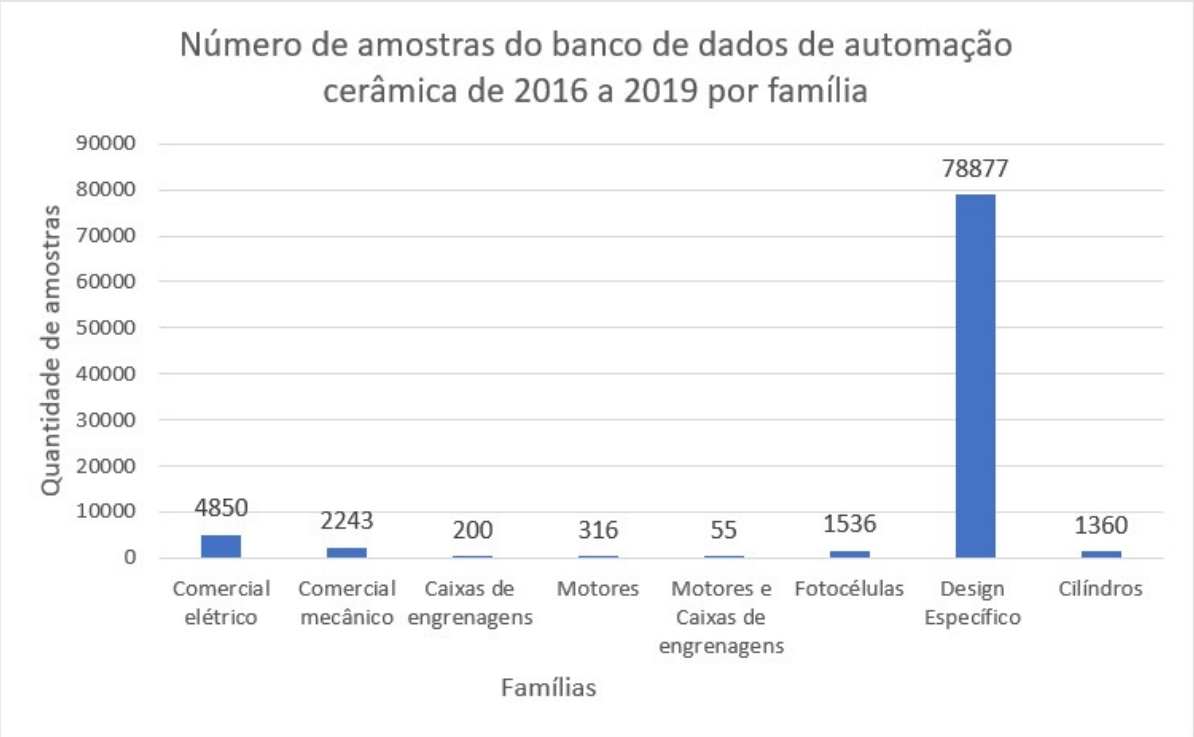
Fonte: Os autores

No banco de dados do setor de automação cerâmica, o número total de amostras por ano foi de 89.437 mil, onde 1.085 amostras estão relacionadas com o ano de 2016, 33.882 com o ano de 2017, 34.574 com o ano de 2018 e 19.896 com o ano de 2019. Isso significa que 1.2% das amostras dos dados estão relacionados ao ano de 2016, 37.8% ao ano de 2017, 38.6% ao ano de 2018 e 22% ao ano de 2019. A diferença de quantidade de amostras entre os anos ocorre porque os anos de 2017 e 2018 tem amostras de todos os meses, mas para o ano de 2016 e 2019 tem apenas amostras relacionadas a alguns meses.

Além disso, todos os produtos no banco de dados do setor de automação cerâmica estão relacionadas as famílias de produtos. Foi realizada a análise gráfica das categorias por família, que são grupos de produtos similares agrupados. O gráfico com o número de amostras segmentado por família pode ser observado na Figura 10.

Na Figura 10 observa-se que as amostras estão concentradas na família *Design específico*, seguida da família *comercial elétrica*, *comercial mecânica* e *fotocélulas*.

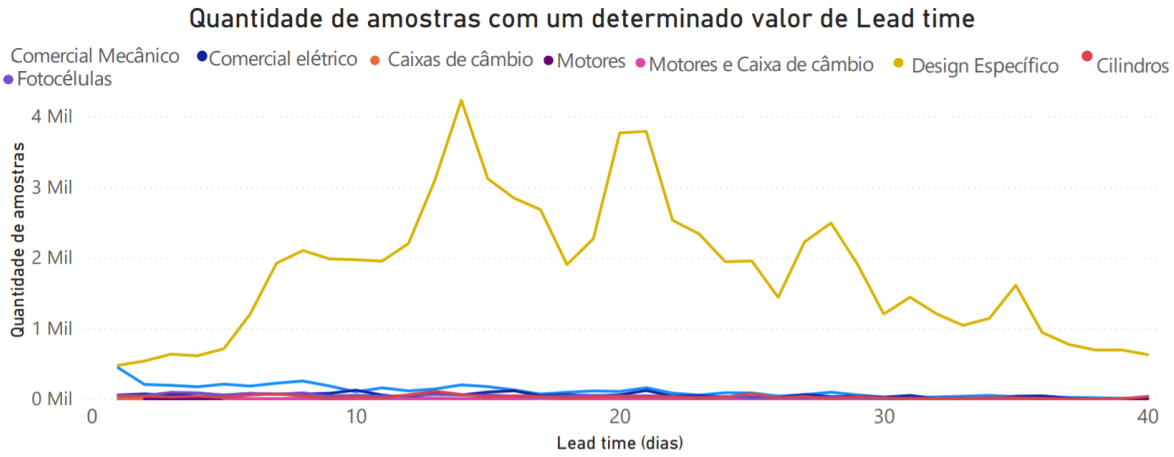
Figura 10 – Número de amostras do banco de dados de automação cerâmica de 2016 a 2019 por família.



Fonte: Os autores

Com o intuito de compreender quais os valores de *lead time* são mais frequentes e atípicos em cada família de produtos no banco de dados do setor cerâmico foi gerado um gráfico com a quantidade de amostras por valor de *lead time*. A Figura 11 mostra a relação da quantidade de amostras por valor de *lead time* por família de produto.

Figura 11 – Quantidade de amostras com valores de *lead time* por família de produtos.



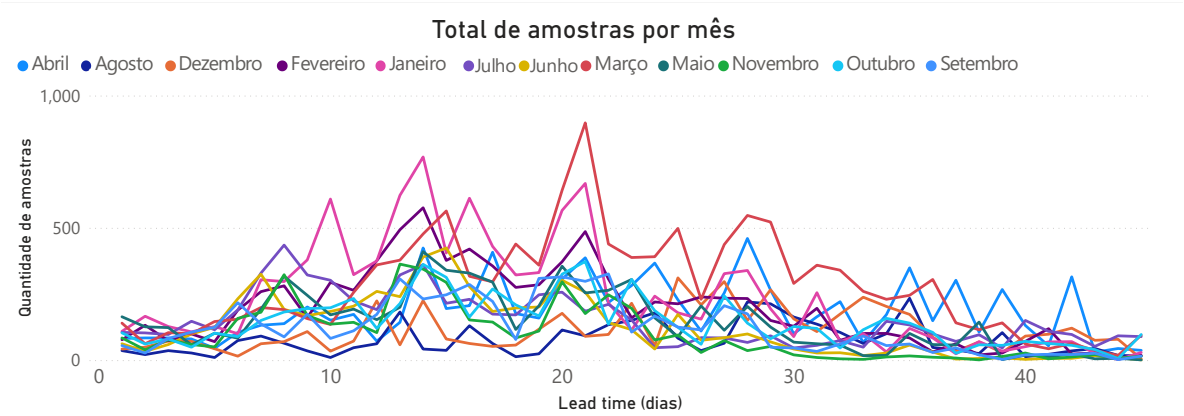
Fonte: Os autores

Analisando a Figura 11 observa-se há uma significativa quantidade de dados relacionadas a família *Design específico*. Além disso, a família *comercial elétrico* possui

uma quantidade significativa de amostras de *lead time* em relação as famílias *comercial elétrico*, *fotocélula* e *cilindros*.

Visando compreender quais os valores de *lead time* são mais frequentes e atípicos em cada mês, foi gerado um gráfico, Figura 12, com a quantidade de amostras por valor de *lead time* em cada mês. Observa-se que a quantidade de amostras de *lead time* mais recorrentes encontram-se mais concentradas em valores de *lead time* de até 32 dias para os dados segmentados por mês.

Figura 12 – Quantidade de dados com um determinado valor de *lead time* por mês.



Fonte: Os autores

Nota-se que um comportamento de *Lead time* similar entre os meses. Ademais as amostras se concentram em até 32 dias.

Em geral, as amostras de dados por família estão concentradas na família *design específico*, relacionada a produtos fabricados pela cadeia de suprimentos exatamente conforme as especificações dos clientes do setor de automação cerâmica. Em resumo, o *Design específico* representa 88% do total das amostras no banco de dados. Todas as outras categorias de famílias de produtos representam juntas 22% das amostras no banco de dados para todos os anos. Esta pesquisa investigou o *lead time* sem considerar o ano de 2016. Em relação às famílias, embora existam famílias com diferentes quantidades de amostras, todas foram mantidas para investigar a interferência do tipo de família na previsão.

Esse banco de dados possui 18 atributos utilizados na análise de mineração de dados:

- Dia (numérico);
- Mês (numérico);
- Ano (numérico);
- Dia da semana (numérico);

- Código do fornecedor;
- Fornecedor trabalha com sistema make-to-order (0) ou make-to-stock (1) (binário);
- Número de empregados (numérico);
- Quantidade de anos do fornecedor no mercado (numérico)
- Quantidade de empresas que o fornecedor atende (numérico);
- Atraso médio (numérico);
- Médio de entrega antecipada (numérico);
- Percentual de entrega antecipada por fornecedor (Porcentagem);
- Atraso percentual na entrega pelo fornecedor (Porcentagem);
- Família de produto
- Família total de produtos por fornecedor (Numérico);
- Distância euclidiana entre cada fornecedor e o armazém da empresa do setor cerâmico (Km);
- *Lead time* de fornecimento por pedido (Numérico).

Conforme as análises gráficas supracitadas, foi definido o uso de dados de 2017 a 2018, considerando um prazo de entrega de até 34 dias. Além disso, a análise foi realizada considerando os 100 produtos mais solicitados para todos os fornecedores do banco de dados. Além disso, usando esses filtros e limites, o banco de dados ficou sem amostras associadas a família *motores e caixas de engrenagens*.

5.3 Pré-processamento

Nesta etapa, os dados foram reduzidos de acordo com as avaliações realizadas na fase anterior, eliminando os valores discrepantes. Além disso, a análise de box plot foi utilizada para identificar o número de amostras com e sem ruídos. Além disso, variáveis com a mesma informação em diferentes formatos, dados corrompidos com informações ausentes, valores duplicados, com erros de qualquer tipo ou amostras que não aumentem a precisão da previsão foram removidos. A quantidade de amostras detectadas com e sem presença de *outliers* e valores extremos encontra-se na Tabela 4.

De acordo com a Tabela 4 para o banco de dados do setor de automação cerâmico, existem 16.997 amostras com *outliers*, o que significa que 19% dos dados de amostras têm *outliers* e 72.440 amostras não têm *outliers*. Além disso, 33.164 amostras

Tabela 4 – Quantidade de amostras com a presença de *outliers* e valores extremos do caso 2

Tipo	Porcentagem com ruídos (%)
Outliers	19%
Valores extremos	37%

Fonte: Os autores

de dados, 37%, têm valores extremos. Todas as amostras com os valores extremos e *outliers* foram removidas do banco de dados.

5.4 Transformação

Nesta etapa, assim como no Caso 1 os dados foram segmentados por mês e por família com tipos de dados numéricos, nominais e binários, totalizando 36 bancos de dados por mês e 12 bancos de dados por família de produtos. Os dados foram testados na validação cruzada com os cinco algoritmos mais usados na literatura, por meio da Equação 2.4 para regressão linear, Equação 2.7 para máquinas vetoriais de suporte, Equação 2.1 para k-vizinhos mais próximos, Equação 2.9, 2.10, 2.11, 2.12 e 2.13 para perceptron multicamada e Equação 2.2 para floresta aleatória. Além disso, a análise do erro quadrático médio, Equação 2.3, foi usada como critério para escolha do melhor algoritmo. O valor do MSE na etapa de validação cruzada por família encontra-se na Tabela 5.

Tabela 5 – Média do MSE dos dados segmentados por família de produtos de cada algoritmo com dados do tipo nominal, numérico e binário do setor automação cerâmica.

Formato	Média RF	Média KNN	Média LR	Média SVM	Média MLP
Binário	6,15	6,25	6,73	7,21	10,47
Nominal	6,88	7,30	6,89	6,96	10,93
Numérico	6,58	7,85	6,86	7,08	9,31

Fonte: Os autores

Nota-se que os melhores resultados foram no formato binário e RF, com MSE de 6,15, e KNN, com MSE 6,25.

Na etapa de validação cruzada foi obtido o MSE para os dados segmentados por família de produto conforme Tabela 6.

Nota-se que em relação à média dos valores do MSE, Tabela 6, que o algoritmo de menor MSE trata-se do KNN para os dados no formato binário, sendo 5,92. Ademais, a segmentação por mês teve menores valores de erro do que por familia de produto. Considerando a base de dados que apresentou menor MSE na validação cruzada, foi utilizado na previsão do *lead time* do setor de automação, dados segmentados por mês do tipo binário usando o algoritmo KNN.

Tabela 6 – Média da validação cruzada por mês do MSE obtido para cada algoritmo nos tipo nominal, numérico e binário dos dados do setor automação cerâmica.

Formato	Média KNN	Média LR	Média RF	Média SVM	Média MLP
Binário	5,92	6,18	6,44	6,46	7,69
Nominal	5,93	6,29	6,48	6,44	7,67
Numérico	6,00	6,38	6,49	6,52	7.88

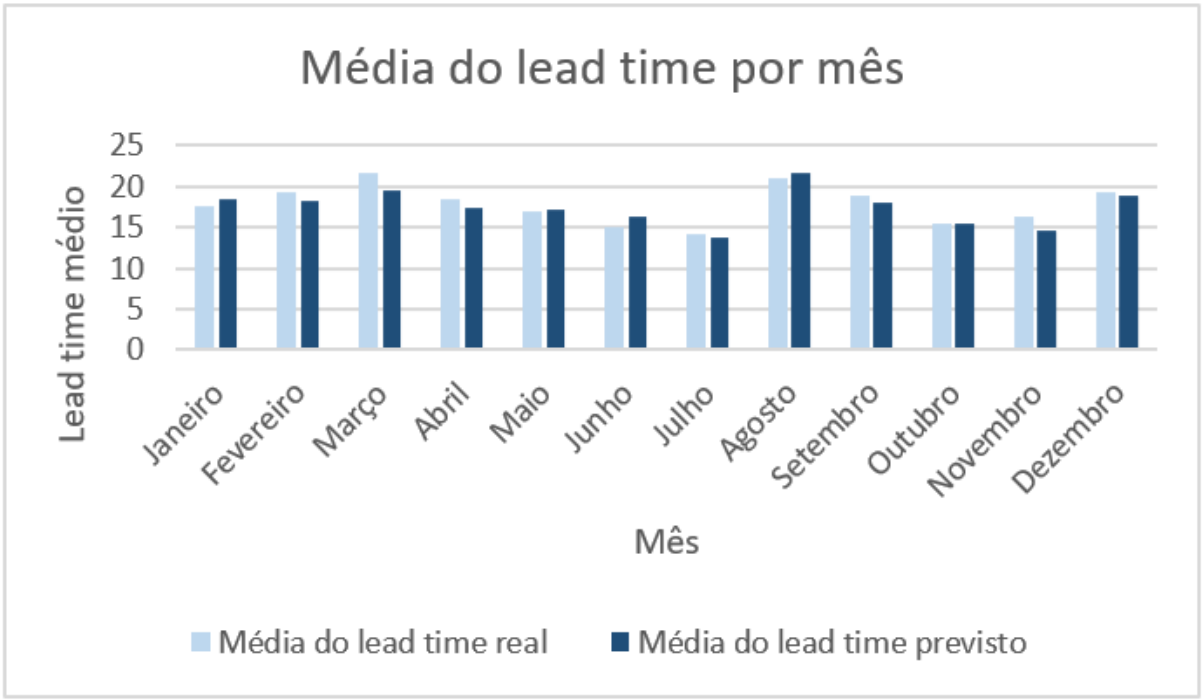
Fonte: Os autores

5.5 Mineração de dados

A mineração de dados foi realizada com dados segmentados por mês, com dados do tipo binário, usando o algoritmo KNN. Os parâmetros utilizados no algoritmo KNN para obter um resultado com o menor erro possível o melhor valor para o parâmetro K foi 9 e usando a distância de Manhattan.

Além disso, o banco de dados foi dividido em 33,3% no conjunto de teste e 66,7% no conjunto de treinamento. A comparação entre os valores reais e preditivos e a taxa de erro e a taxa de acerto foi calculada por mês. O gráfico com a comparação entre o *lead time* real e o *lead time* previsto encontra-se ilustrado na Figura 13.

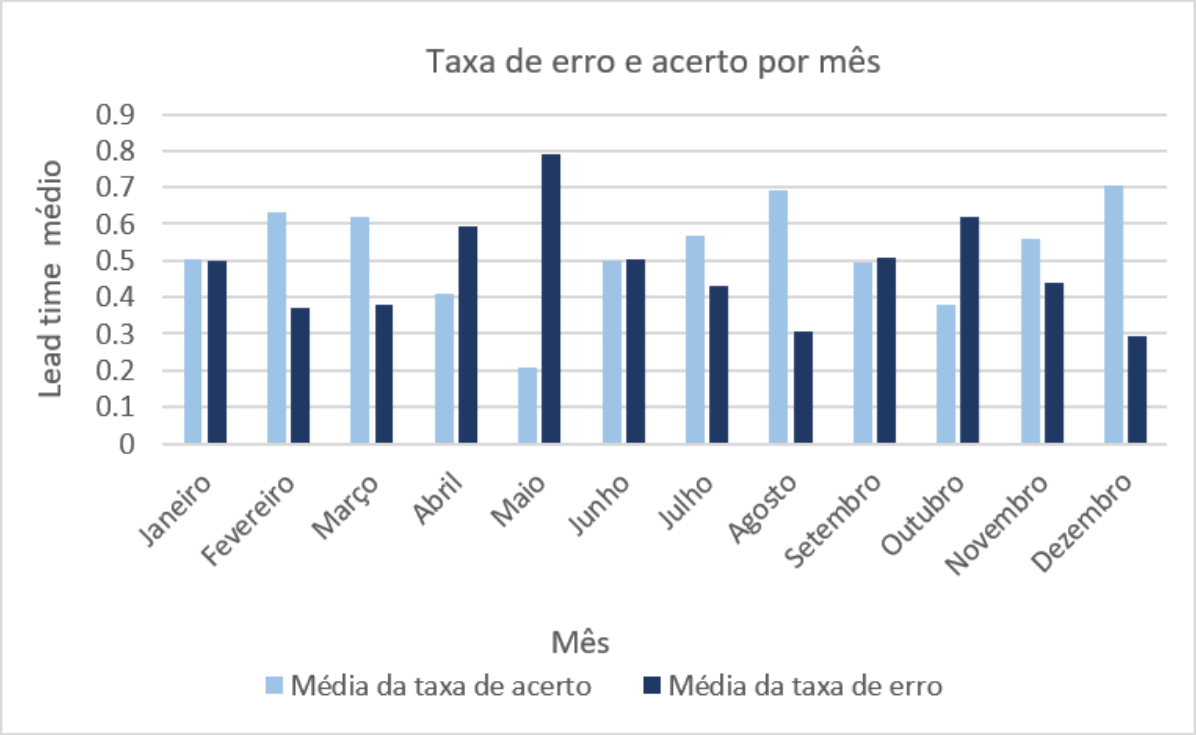
Figura 13 – Comparação entre o *lead time* médio previsto e o real



Fonte: Os autores

A Figura 13 mostra que o *lead time* previsto estava próximo do real. De forma geral, os valores reais e previstos do *lead time* concentram-se entre 15 a 20 dias para todos os meses, e a diferença entre os valores reais e previstos foi menor que 5 dias. A comparação entre a taxa de erro e taxa de acerto encontra-se presente na Figura 14.

Figura 14 – Comparação entre a taxa de acerto e erro na previsão do *lead time* por mês



Fonte: Os autores

Em média, na maioria dos meses, fevereiro, março, julho, agosto, novembro e dezembro, a taxa de acerto foi maior que a taxa de erro. Além disso, há alguns meses com a taxa média de acertos igual a de erros, como janeiro e junho. Além disso, os meses de maio, abril e outubro apresentaram uma taxa de erro maior que a taxa de acertos.

5.5.1 Interpretação

Em resumo, na fase de seleção a maioria das amostras de pedidos de clientes está associada aos anos de 2017 e 2018, 37% e 38% das amostras, respectivamente. De acordo com a Figura 5 a maioria das amostras estão associadas à família de design específicos. Além disso, há mais amostras de *lead times* concentrados nos meses de janeiro, fevereiro e março, Figura 10. Essas amostras de *lead times*, Figura 10, estão concentradas em valores de *lead time* de até 32 dias. A Figura 11 mostra que as amostras de *lead time* segmentadas por família possuem mais amostras acumuladas na família design específico, com mais de 4.000 amostras por valor de *lead time*. Em segundo lugar, as famílias fotocélulas, comercial elétrica e comercial mecânica possuem dados entre 300 e 400 amostras por valor do *lead time*, Figura 11. No mais, poucas amostras, cerca de 20, para cada valor do *lead time*, Figura 11, estão associadas às famílias Motores e caixas de câmbio, caixas e cilindros. Ainda na Figura 11, para todas as famílias, as amostras com maior tempo de processamento têm no máximo 32 dias.

Na fase de pré-processamento, Tabela 4, foram encontrados 19% de *outliers* e 37% de valores extremos no banco de dados do setor de automação. Na fase de transformação, os 19 atributos foram transformados em binário, nominal e numérico. Na validação cruzada, os resultados obtidos no algoritmo KNN, Tabela 6, com dados binários apresentaram o menor erro quadrático médio com dados segmentados por mês. O erro quadrático médio por família para KNN teve valor entre 5–6 dias e por mês entre 4–5 dias.

Na fase de mineração de dados, observou-se que o *lead time* previsto foi satisfatório, a diferença entre o *lead time* real e o *lead time* previsto foi em média 2 dias, Figura 14. Para todos os meses, na Figura 14, o tempo médio de entrega foi de 10 a 20 dias.

CAPÍTULO

6

EXPERIMENTOS E RESULTADOS CASO 3: SETOR DE SERVIÇOS ELETRÔNICOS

6.1 Descrição do setor da base de dados

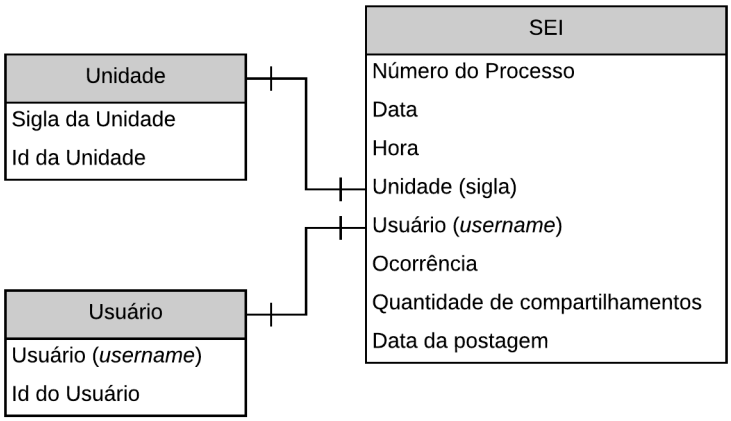
Os dados utilizados neste capítulo tratam-se de informações dos trâmites de documentos do Sistema Eletrônico de informações (SEI), considerando, especificamente, sua utilização no âmbito da Universidade Federal de Goiás. O SEI refere-se a um sistema de gestão de informações desenvolvido pelo Tribunal Regional Federal da 4° Regional, que permite o processamento de documentos eletrônicos, processos administrativos, oferecendo serviços de produção envio, edição, tramites, assinaturas e armazenamento de documentos, e sendo utilizado em larga escala em todo o território nacional, no âmbito de todas as autarquias, federal, estadual e municipal (SEI, 2017). Esse sistema possibilita a edição de documentos sem a necessidade da digitalização ou de impressão de documentos. Ademais, o SEI é suportado pela esfera administrativa pública no contexto do Processo Eletrônico Nacional (PEN), que tem como objetivo construir e manter uma estrutura pública de documentos e processos eletrônicos administrativos (SEI, 2017).

6.2 Seleção

Conforme realizado nos casos anteriores, nessa fase as análises preliminares do banco de dados são realizadas. No banco de dados do SEI encontram-se informações do trâmites de documentos da UFG, as quais incluem uma série de atributos que permitem desde a identificação do tipo de atividade realizada, data e hora da

efetivação da ação e usuário responsável. A Figura 15 apresenta o diagrama entidade-relacionamento para a base de dados do SEI, onde é importante destacar que já no momento da extração de dados foi considerada a aplicação de máscaras de proteção à identificação dos dados de nome de usuário (*username*) e sigla da unidade, com o propósito de proteção de informações de identificação individual.

Figura 15 – Diagrama Entidade-Relacionamento da base de dados do SEI.



Fonte: Os autores.

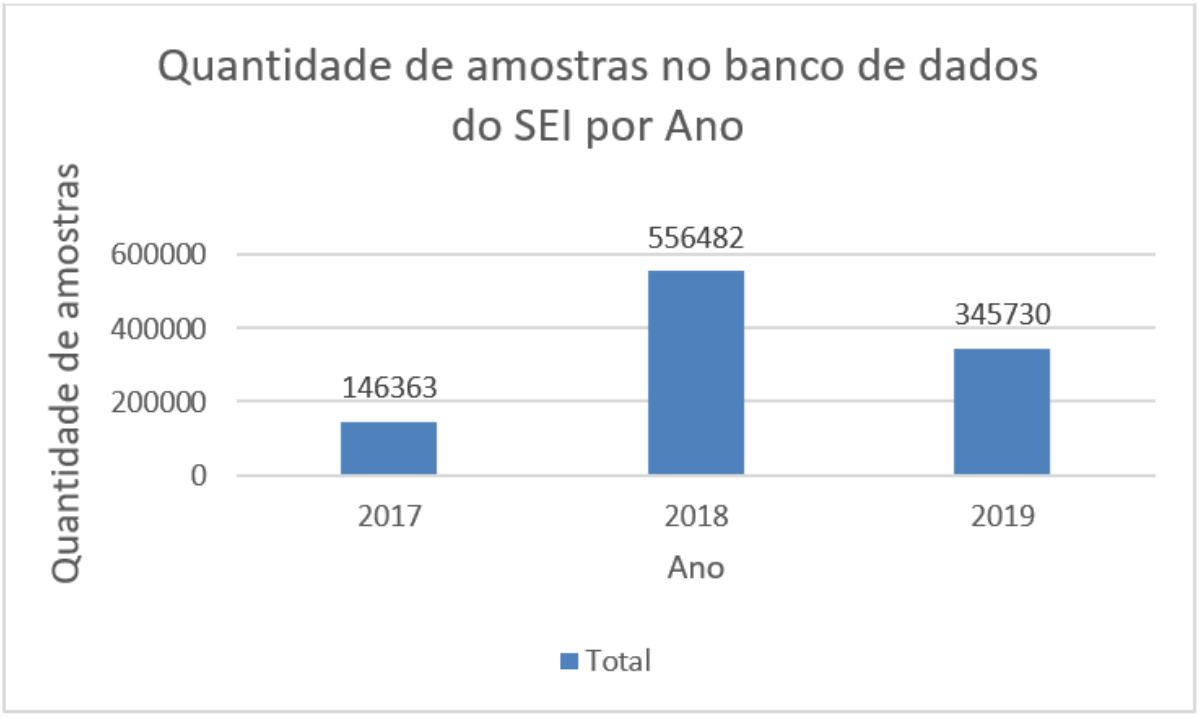
Na fase de seleção os objetivos da mineração nesse banco de dados são validados, e realizadas análises gráficas preliminares para avaliação da qualidade dos dados, sendo investigado a quantidade total de amostras relacionada a cada ano, a quantidade de dados para os dez primeiros processos com mais amostras análise temporal do banco de dados e sua relação com o *lead time*, assim como a definição atributos significativos. Nessa sentido, a Figura 16 apresenta uma distribuição do quantitativo de dados por cada ano.

O número total de amostras foi de 1.048.575 milhões de dados, onde 146.363 estão relacionados a amostras do ano de 2017, 556.482 estão relacionados ao ano de 2018 e 345.730 com ano de 2019. Em termos percentuais, observa-se que, 13,95% das amostras dos dados estão relacionados ao ano de 2017, 53,07% ao ano de 2018 e 32,97% ao ano de 2019. A diferença entre a quantidade de amostras entre anos ocorre porque o ano de 2018 possui amostras de todos os meses, já o ano de 2017 apenas amostras de outubro a dezembro e de 2019 entre janeiro a março.

Outra análise gráfica relevante refere-se à quantidade de dados em face das unidades e processos. O banco de dados analisado possui um total de 269 tipos de processos, um total de unidades de 275 diferentes entre si, número de processos 35.427, ocorrências 71. Na Figura 17 são destacados os 17 processos processos mais frequentes.

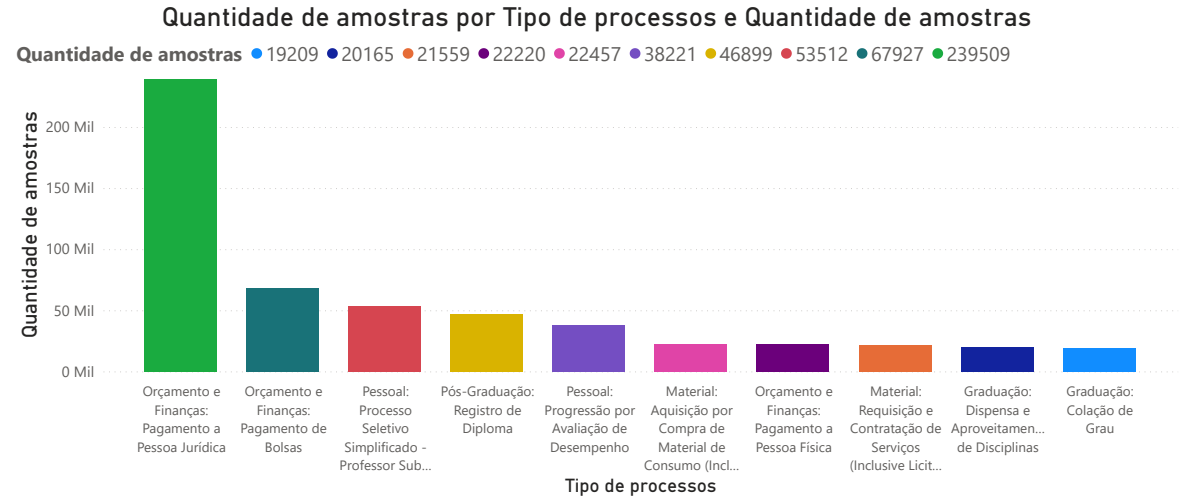
Em relação as amostras de dados por processo da Figura 17 (a) as amostras se concentram em ordem decrescentes nos processos Orçamento e Finanças: Pagamento a Pessoa Jurídica, 47%, Orçamento e Finanças: Pagamento de Bolsas, 13%, Pessoal:

Figura 16 – Número de amostras do banco de dados do serviço eletrônico de informações dos anos de 2017 a 2019 por ano.



Fonte: Os autores

Figura 17 – Número de amostras por processo do banco de dados SEI das 10 primeiras categorias com mais amostras.

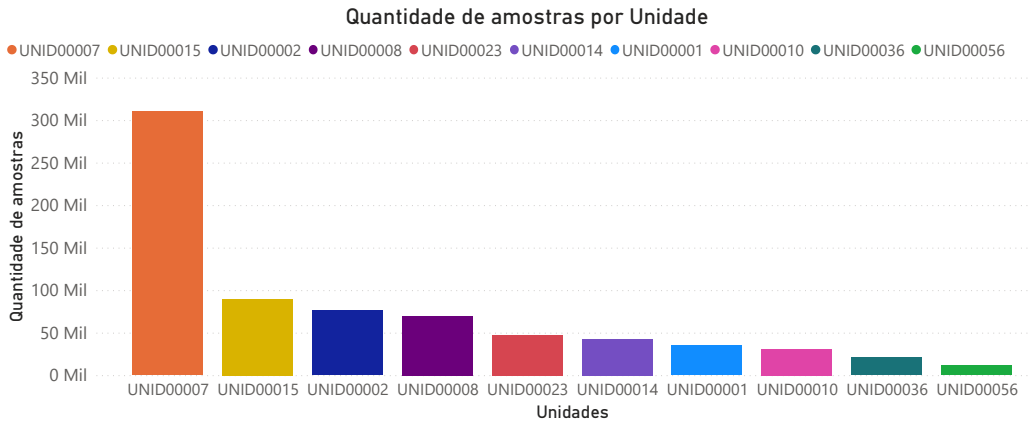


Fonte: Os autores

Processo Seletivo Simplificado - Professor Substituto, 10%, Material: Aquisição por Compra de Material de Consumo (Inclusive Licitação), 4%, Orçamento e Finanças:

Pagamento a Pessoa Física, 4%, Material: Requisição e Contratação de Serviços (Inclusive Licitações), 4%, Pessoal: Afastamento para Estudo ou Missão no Exterior, 3%, Administração Geral: Contratos, 3%, Material: Aquisição por Compra de Material, Permanente (Inclusive Licitação), 2%, Pessoal: Pagamento de servidor, 2%, Pessoal: Processo Seletivo Simplificado, 2%, Orçamento e Finanças: Pagamento de Diárias, 2%, Orçamento e Finanças: Suprimento de Fundos, 1%, Graduação: Auxílio para Eventos (Discentes), 1%, Orçamento e Finanças: Empenho Estimativo , 1%, Orçamento e Finanças: Pagamento Multas e Juros, 0.11% e Orçamento e Finanças: Recolhimento de PASEP 0.08%. Na Figura 18 são destacadas com as dez unidades com mais amostras.

Figura 18 – Número de amostras por processo do banco de dados SEI das 10 primeiras unidades com mais amostras.



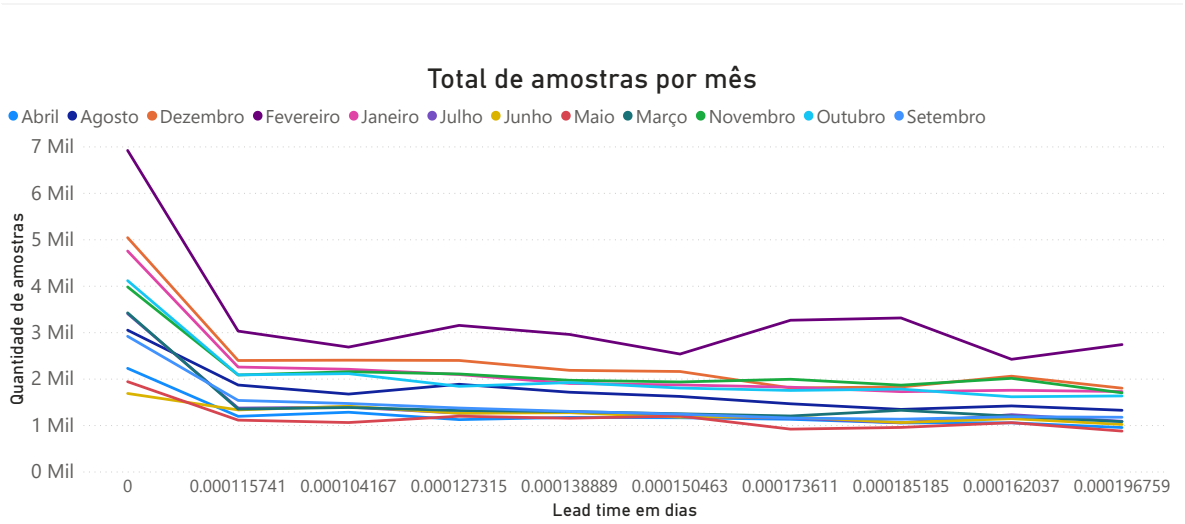
Fonte: Os autores

A Figura 18 mostra que as unidades que mais possuem amostras tratam-se das unidades UNID00007, UNID00015, UNID00002, UNID00008, UNID00023, UNID00014, UNID00001, UNID00010, UNID00036, UNID0005 6, representando respectivamente 42%, 12%, 10%, 9%, 6%, 6%, 6%, 4%, 3% e 2%.

Com relação a análise temporal dos dados, foi analisado o número de amostras do *lead time* por mês, essa análise encontra-se na Figura 19.

Observa-se que em todos os meses existem mais amostras com valores de de *lead time* concentrados em valores menores de 0.000104 dias, o que indica que

Figura 19 – Quantidade de dados com um determinado valor de *lead time* por mês.



Fonte: Os autores

possivelmente, em sua maioria, os processos na plataforma do SEI são processados em menos de um dia. concentram em valores de *lead time* menores de um dia.

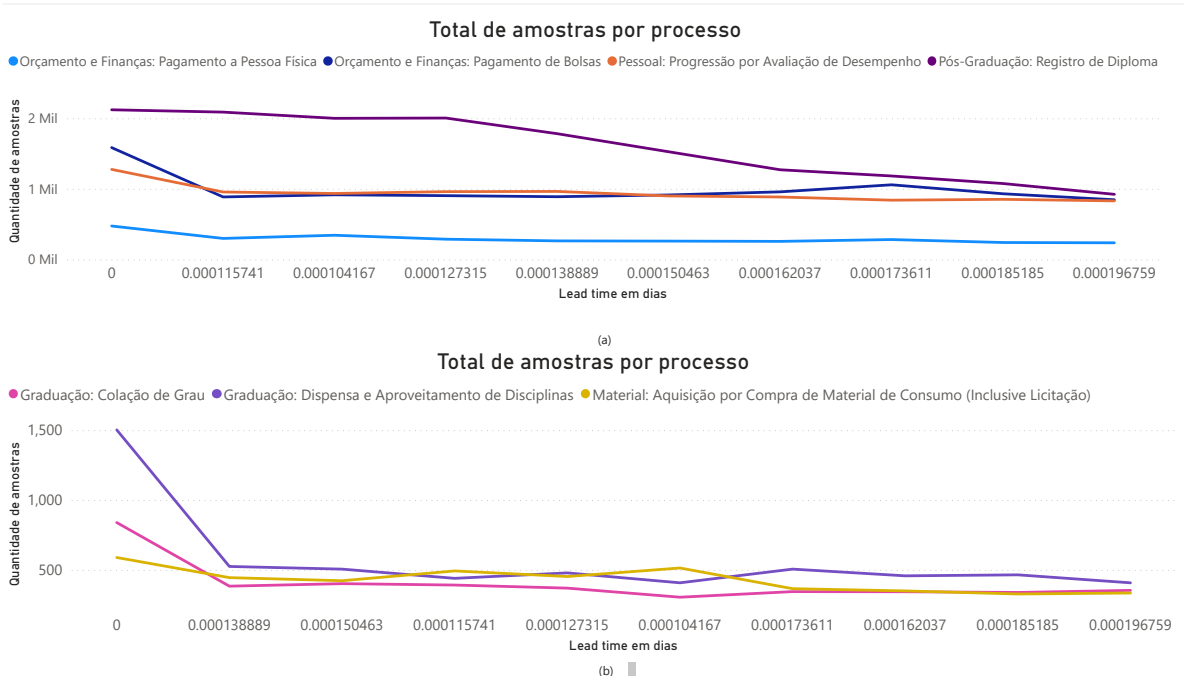
Foi investigada a quantidade de amostras para cada valor de *lead time* para os dados segmentados por categoria por processo conforme a Figura 20 (a) para os processos orçamento e finanças: pagamento a pessoa jurídica, orçamento e finanças: pagamento de bolsas , pessoal: progressão por avaliação de desempenho e pós-graduação: registro de diploma e (b) para os processo graduação: colação de grau, graduação: dispensa e aproveitamento de disciplinas, material: aquisição por compra de material de consumo (inclusive licitação).

Nota-se que na Figura 20 (a) e (b) segmentação por tipo de processo apresenta valores de *lead time* concentrados em menos de 1 dia.

Além disso, Foi investigada a quantidade de amostras para cada valor de *lead time* para os dados segmentados por categoria por processo (Orçamento e finanças: pagamento a pessoa jurídica, pessoal:avaliação de desempenho, pessoal: processo seletivo simplificado - professor substituto) conforme a Figura 21 (a) e unidades 21 (b).

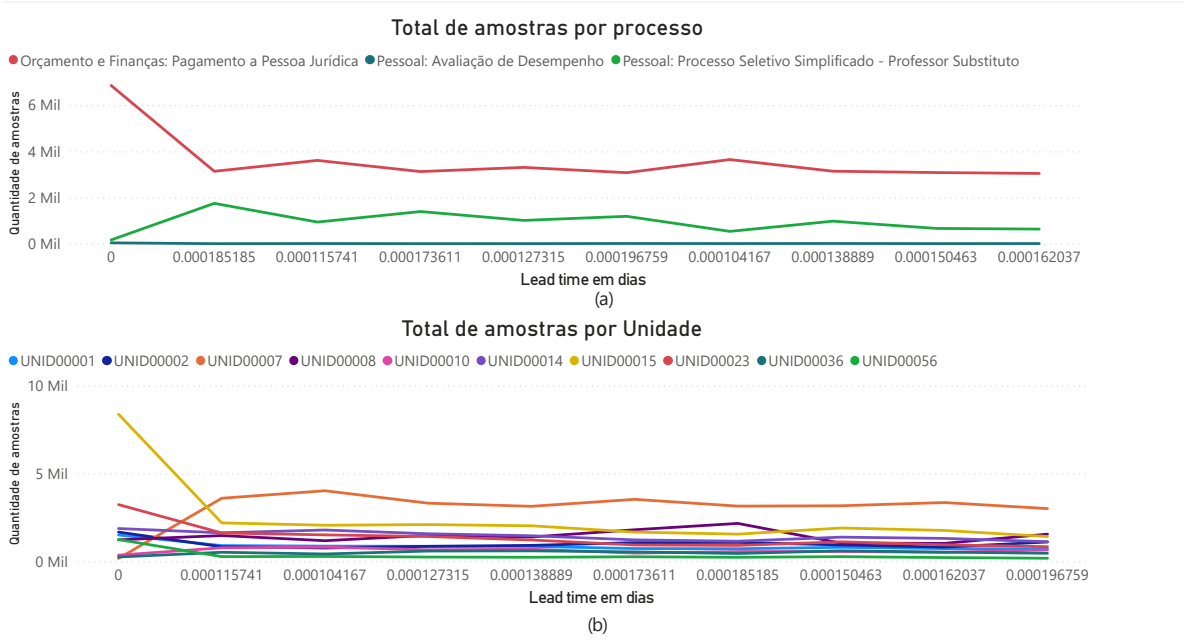
Nota-se na Figura 21 (a) e (b) que a segmentação por tipo de processo e unidade apresenta valores de *lead time* concentrados em menos de 1 dia. Em resumo, através do apresentado nas Figuras 16, 17 e 19, observa-se que as amostras do banco de dados concentram-se no ano de 2018. Além disso, como banco de dados possui ao total 35.427 processos diferentes e um total 269 tipos de processos, 275 unidades

Figura 20 – Quantidade de dados com um determinado valor de *lead time* por processo (a) e (b).



Fonte: Os autores

Figura 21 – Quantidade de dados com um determinado valor de *lead time* por processo (a) e unidade (b).



Fonte: Os autores

diversas e um total de 6696 usuários, essa pesquisa optou por utilizar os 100 processos mais utilizados. O uso dos 100 processos mais frequentes gerou um total de 17 tipos de processos, sendo eles: Orçamento e finanças: pagamento a pessoa jurídica, Orçamento

e finanças: pagamento de bolsas, Pessoal: processo seletivo simplificado - professor substituto, Pós-graduação: registro de diploma, Pessoal: Progressão por avaliação de desempenho, Material aquisição por compra de material de consumo (inclusive licitação), Orçamento e finanças: pagamento a pessoa física, Material: requisição e contratação de serviços (inclusive licitações), Graduação: dispensa e aproveitamento de disciplinas, Graduação: colação de grau.

Orçamento e Finanças: Pagamento a Pessoa Jurídica, Orçamento e Finanças: Pagamento de Bolsas, Pessoal: Processo Seletivo Simplificado - Professor Substituto, Material: Aquisição por Compra de Material de Consumo (Inclusive Licitação), Orçamento e Finanças: Pagamento a Pessoa Física, Material: Requisição e Contratação de Serviços (Inclusive Licitações), Pessoal: Afastamento para Estudo ou Missão no Exterior, Administração Geral: Contratos, Material: Aquisição por Compra de Material, Permanente (Inclusive Licitação), Pessoal: Pagamento de servidor, Pessoal: Processo Seletivo Simplificado, Orçamento e Finanças: Pagamento de Diárias, Orçamento e Finanças: Suprimento de Fundos, Graduação: Auxílio para Eventos (Discentes), Orçamento e Finanças: Empenho Estimativo, Orçamento e Finanças: Pagamento Multas e Juros e Orçamento e Finanças: Recolhimento de PASEP

A partir das análises gráficas supracitadas, foi definido o uso de dados de 2018. Além disso, a análise foi realizada considerando os cem processos mais solicitados para todas as unidades e ocorrências do banco de dados.

Além disso, o banco de dados do SEI possui 9 variáveis atributos utilizadas na análise de mineração de dados foram:

- Dia, um a trinta, do pedido do cliente (Numérico);
- Ano, de 2018, do pedido do cliente (Numérico);
- Mês, de um a trinta, a partir do pedido do cliente (Numérico);
- Código da Unidade (String);
- Ocorrência (String);
- Processos (Númerico);
- Tipo de processos (String)
- Usuários (String)
- Lead time em dias (numérico).

6.3 Pré-processamento

No contexto dos experimentos realizados junto aos dados obtidos do SEI, o pré-processamento consiste na fase de identificação de *outliers* e valores discrepantes. Para tanto, a análise gráfica através de *box plot* foi utilizada com o intuito de identificar o número de amostras discrepantes em termos de valores. No mais, assim como nos casos anteriores, informações duplicadas, e corrompidos foram identificados e removidos do banco de dados. A quantidade de amostras com e sem *outliers* e valores extremos encontra-se na Tabela 7. É possível notar que não foi constatado *outliers* e valores extremos no banco de dados SEI. O que pressupõe que os dados possuem um comportamento de dados parecido entre as amostras.

Tabela 7 – Quantidade de amostras com a presença de *outliers* e valores extremos do caso 3

Tipo	Porcentagem com ruídos(%)
Outliers	0%
Valores extremos	0%

Fonte: Os autores

6.4 Transformação

Nesta etapa, os dados foram segmentados por mês e por tipo de processo. Como já supracitado o uso dos 100 processos mais frequentes gerou um total de 17 tipos de processos. Dessa forma foram obtidos um total de 87 bancos de dados, 1 banco de dados para cada tipo de processo e mês.

Os dados foram testados na validação cruzada com os cinco algoritmos mais usados na literatura, por meio da Equação 2.4 para regressão linear, Equação 2.7 para suporte a máquinas vetoriais, Equação 2.1 para k-vizinhos mais próximos, Equação 2.9, 2.10, 2.11, 2.12 e 2.13 para perceptron multicamada e Equação 2.2 para floresta aleatória. Além disso, a análise do erro quadrático médio foi usada como critério para escolha do melhor algoritmo.). Embora tenham sido testados os cinco algoritmos(LR, MLP, RF,SVM e KNN), não foi gerado resultados para os algoritmos LR e MLP. Um dos possíveis motivos para que não se gerasse resultados com o uso dos algoritmos LR e MLP trata-se da falta de correlação entre os atributos observados na Figura 22.

A Figura 22 apresenta o valor da correlação de cada novo atributo, atributo derivado da base de dados atual, e tambem a correlação entre o método K-means com os atributos. A Figura 22 mostra que a correlação entre os atributos não foi significativa já que nenhuma correlação entre os atributos foi maior ou igual a um, e em alguns casos como a relação entre o numero de usuário e tipo de categoria de processo a correlação foi negativa. Em contrapartida é possível observar que existe uma correlação

Figura 22 – Correlação entre os atributos utilizados na predição

Fonte: Os autores

positiva e significativa próxima de 1 entre o método de classificação K-means e o *lead time* de processos.

A Tabela 8 mostra o valor do MSE obtido na validação cruzada realizada para cada por tipo de processo, com os dados no formato binário, numérico e nominal.

Tabela 8 – Média do MSE dos dados segmentados por processo para cada algoritmo nos tipo de dados nominal, numérico e binário.

Formato	Soma de KNN	Soma de SVM	Soma de RF
Binário	2.06	2.16	2.03
Nominal	2.13	2.17	1.96
numérico	2.24	2.16	1.93

Fonte: Os autores

Conforme a Tabela 8 é possível observar que o algoritmo com menor MSE trata-se do RF com os tipos de dados Numérico no valor de 1.93.

A validação cruzada com os dados segmentados por mês também foi avaliada, Tabela 9.

Tabela 9 – Média do MSE obtido dos dados segmentados por mês para cada algoritmo nos tipos de dados nominais, numérico e binário.

Formato	Soma de RF	Soma de SVM	Soma de KNN
Nominal	1.91	1.75	2.00
Numérico	1.96	1.79	1.93
Binário	1.51	1.75	1.73

Fonte: Os autores

A Tabela 9 mostra que o algoritmo RF com dados do tipo binário apresentou os menores de MSE 1,50 valores foram obtidos para o algoritmo.

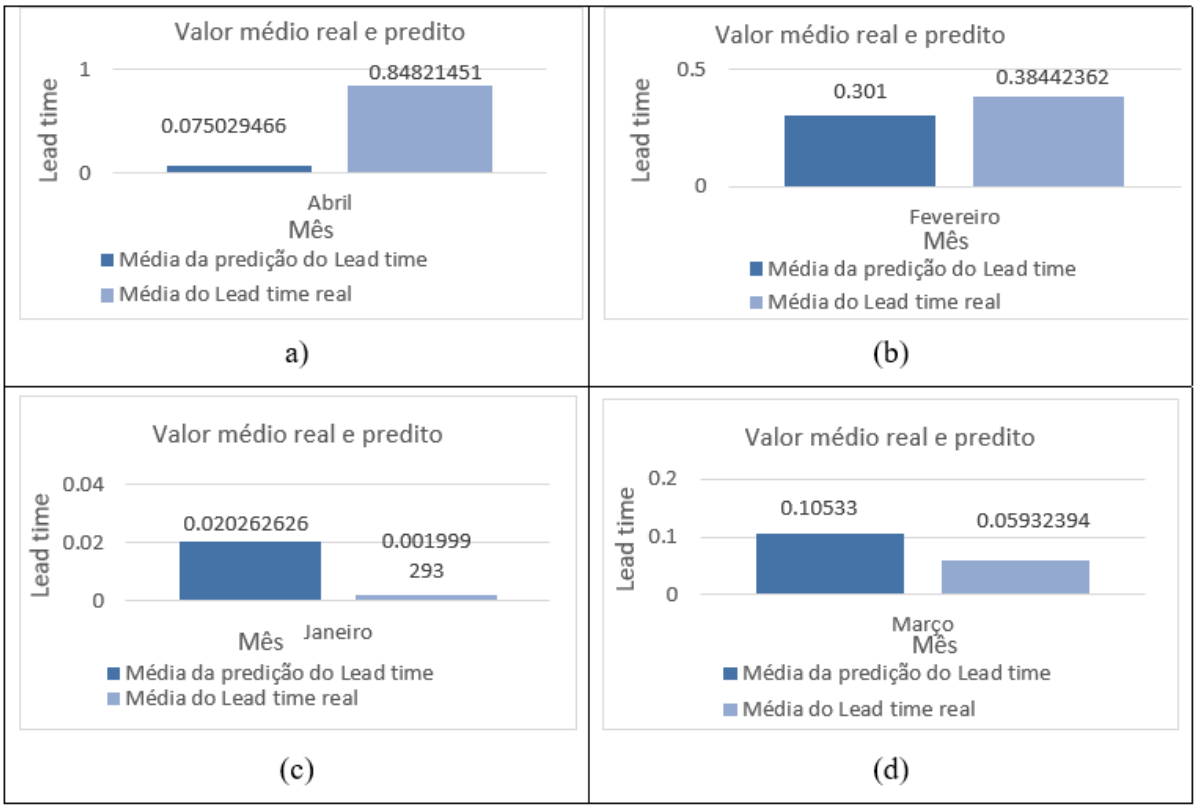
Comparando as duas validações cruzadas, Tabela 8 e Tabela 9, nota-se que os melhores resultados de MSE foi obtido por meio do algoritmo RF para ambas análises e no entanto para os dados segmentados por mês o valor médio do MSE foi mais baixo 1.51 em relação ao o valor médio do MSE segmentados por processos, 1.93. Além disso, a segmentação por mês mostrou-se com resultados melhores que a validação por processos, logo a presença do variável tipo de processo aumenta a acurácia da predição.

6.5 Mineração de dados

De acordo com a seção acima, para a mineração de dados foram foi gerada utilizando os dados segmentados por mês com os dados no formato binário fazendo uso do algoritmo RF. Os parâmetros utilizados no algoritmo RF para obter um resultado com o menor MSE foram número de arvores 100, profundidade da árvore ilimitada, semente para gerador de número aleatórios 1. Além disso, o banco de dados foi dividido em 33,3% no arquivo de teste e 66,7% no arquivo de treinamento. No momento da geração da predição os resultados não convergiram para todos os meses, levando cerca de 25 a 20 dias para gerar um resultado. Logo foram obtidas predições apenas para os meses de Abril, Fevereiro, Janeiro e Março Figura 14 (a), (b), (c) e (d) respectivamente. A comparação entre os valores reais e preditivos e a taxa de erro e a taxa de acerto foi calculada por mês.

O gráfico com a comparação entre o *lead time* real e o *lead time* previsto para os meses de abril, fevereiro, janeiro e março encontra-se ilustrado na Figura 23. A Figura 23 mostra que o valor médio do *lead time* previsto apresentou-se próximos com variação média de 0,23. De forma geral os valores reais e previstos do *lead time* concentraram-se em menos de 1 dia. A comparação entre a taxa de erro e taxa de acerto encontra-se presente na Figura 24.

Figura 23 – Comparação entre o *lead time* médio previsto e o real



Fonte: Os autores

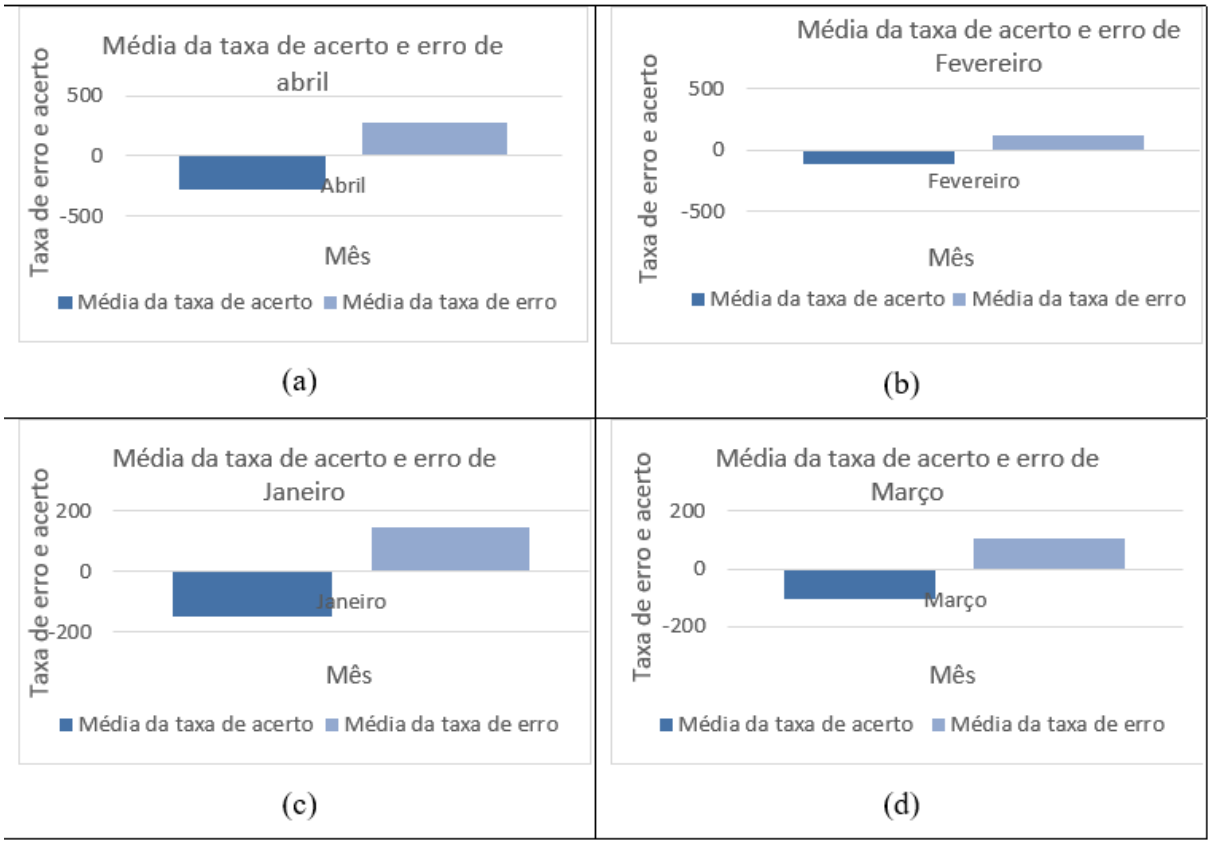
Em média, na maioria dos meses analisados, abril, fevereiro, janeiro, março, a taxa de erro foi maior que a taxa de erro. Observa-se, portanto, que além de nem todos os meses conseguirem convergir para uma predição os meses que obtiveram resultados da predição apresentaram taxa de acerto negativa e taxa de erro significativa.

6.6 Interpretação

Na fase seleção, as quantidades de amostras de dados concentram-se nos anos de 2018, 53,07% Figura 16. Além disso, os dados concentram-se no tipo de processo Orçamento e Finanças: Pagamento a Pessoa Jurídica com 43% do total de dados em relação aos outros dez primeiros processos com mais dados amostrais. As unidades UNID007 trata-se da unidade com mais amostras de dados, 42% em relação as 10 primeiras unidades com mais dados amostrais.

Na fase de pré-processamento, Tabela 7, não foram encontrados *outliers* e valores extremos no banco de dados do SEI. Na fase de transformação, os dados segmentados por mês e por processo foram utilizados no formato binário, nominal e numérico onde foram testados na validação cruzada para os cinco algoritmos RF, KNN, SVM, LR e MLP. Desses em ambos os casos o algoritmo RF apresentou um menor MSE, no entanto os dados segmentados por mês com os dados no formato binário

Figura 24 – Comparação entre a taxa de acerto e erro na previsão do *lead time* por mês



Fonte: Os autores

apresentaram o melhor valor de MSE de 1.51, Tabela 8, em relação ao menor valor de MSE obtido para os dados segmentados por processo 1.93 com dados numéricos Tabela 9.

Na fase de mineração de dados, observou-se que o *lead time* previsto não foi satisfatório, uma vez que a taxa de acerto foi negativa e taxa de erro superior a 100%, Figura 24. Além disso, embora a diferença entre o *lead time* real e o *lead time* previsto foi em média menor que 1 dia, Figura 24, não se pode afirmar com exatidão que esse foi valor significativo já que os *lead time* para os dados do SEI tinham valores em horas, minutos e segundos, podendo ter uma grande variação dentre o período de 1 dia.

Nota-se que uma possível justificativa para uma previsão não satisfatória para os dados do SEI deve-se a falta de correlação entre as variáveis do banco de dados do SEI, conforme presente na Figura 22. Diante da inviabilidade da previsão do *lead time* por um método de regressão decorrente da falta de correlação entre as variáveis, essa pesquisa sugeriu como uma solução viável a criação de um método híbrido, que mescla mais de um método. Nesse caso utilizaremos um método de agrupamento KNN, um de classificação e *clustering* K-means e um de regressão, regressão linear, como alternativa para predição dos dados. A proposta apresentada na pesquisa encontra-se no tópico a seguir.

6.7 Método híbrido proposto

Essa pesquisa propôs o uso de um método híbrido, para a base de dados do SEI, que não apresentaram correlação significativa entre os atributos. Foi observado, conforme exposto na Figura 22 que os dados do banco de dados do SEI não apresentavam correlação significativa entre si, logo a predição nesse caso utilizando apenas métodos de regressão tornavam a predição inviável. Diante disso, verificando a correlação dos dados com um algoritmo de classificação (K-means), foi possível verificar que a correlação era significativa, positiva e forte de 0,85, Figura 22.

Diante do exposto, essa pesquisa propôs utilizar o algoritmo KNN para definir grupos nos quais os valores de *lead time* poderiam ser segregados conforme sua similaridades. Uma vez por que, o KNN trabalha com o conceito de vizinhanças, ou seja, os amostras próximas a sua vizinhança, possem comportamentos similares e consequentemente pertencem aquele grupo. Com o KNN os dados foram agrupados e rotulados. Os próximos dados foram classificados como pertencente a determinado rótulo de grupo por meio do K-means. Posteriormente o valor do *lead time* presente nos dados rotulados foram preditos por meio da regressão linear.

O processo de aplicação do método híbrido foi aplicado com base na seguinte lógica:

Foram utilizados os atributos secundários para definição dos grupos,sendo eles:

1. Tempo total;
2. Número de usuários;
3. Número de unidades;
4. Número de ocorrências;

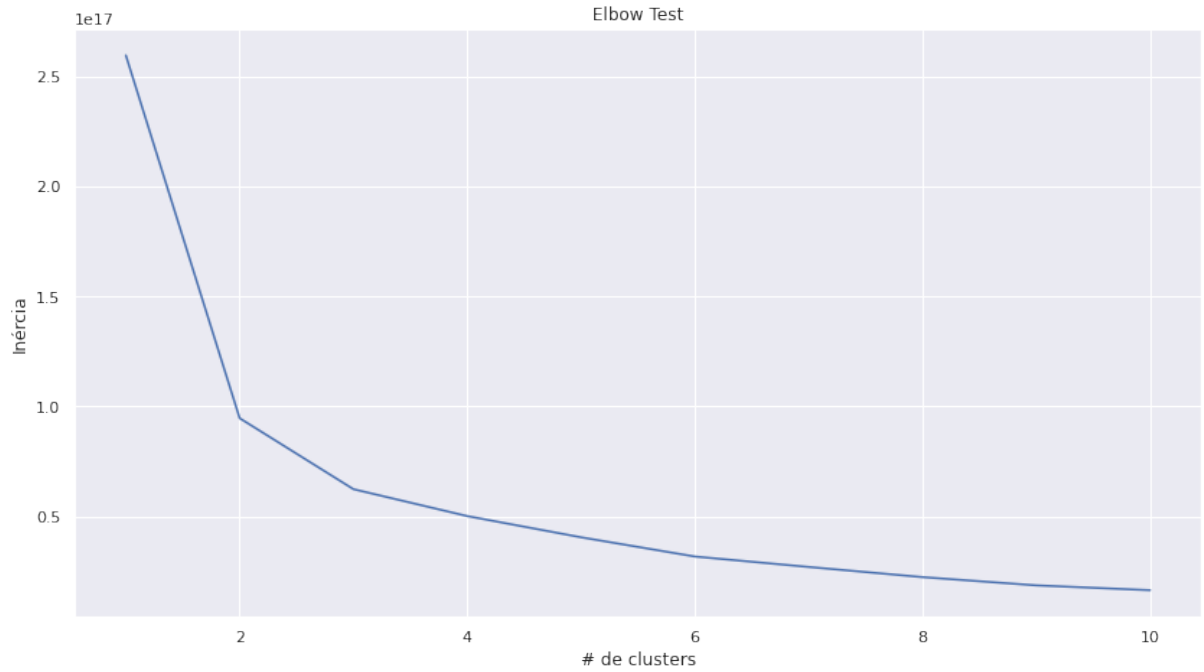
A primeira etapa da utilização desse método foi definir o número de grupos Kc. Nessa pesquisa utilizamos o método *Elbow Test* e *Silhouette Method*. O número ótimo de grupos Kc para o *Elbow Test* encontra-se presente na Figura 25.

Na Figura 25 é possível observar que o menor ponto de inflexão, onde o número ótimo Kc foi definido, é 2.

Utilizando o *Silhouette Method* observamos o valor ótimo para Kc na Figura 26

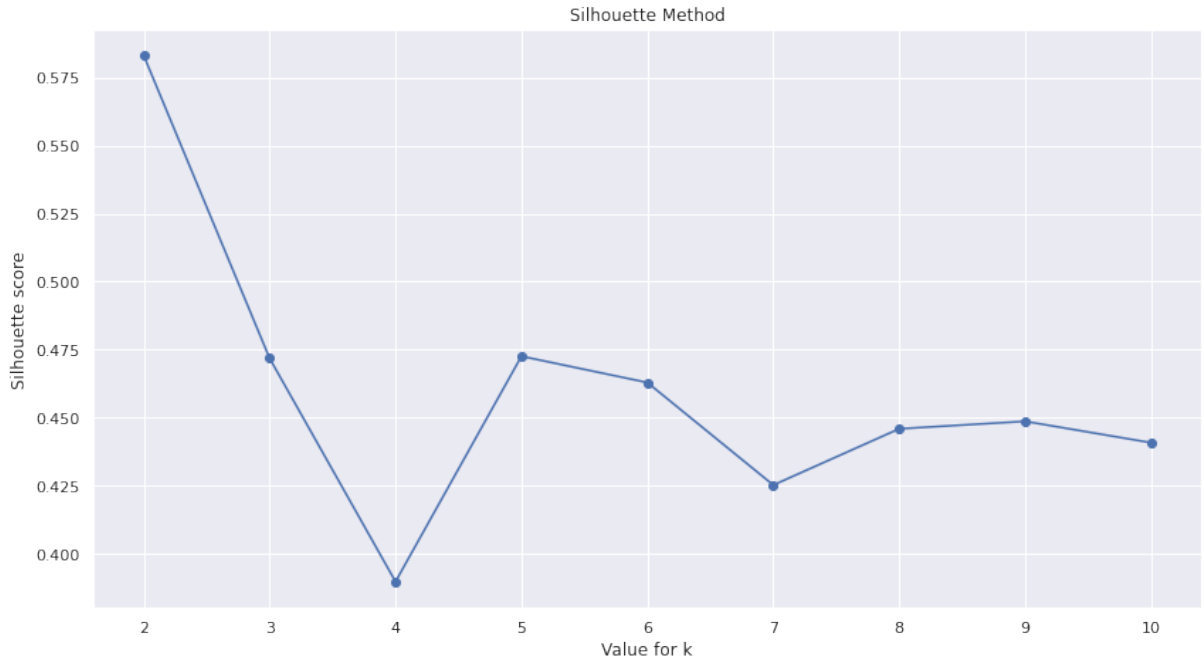
No caso do método da silhueta o melhor Kc é o ponto da curva com menor inflexão, notamos que para esse método os melhores valores poderiam assumir Kc igual a 2 ou 3 onde o angulo é de 180. Embora no teste Silhouette Method os número

Figura 25 – Número de Kc ótimos para com o método de *Elbow Test*



Fonte: Os autores

Figura 26 – Número de Kc ótimos para com o método de *Silhouette Method*



Fonte: Os autores

de Kc possam parecer ambíguo o método *Elbow Test* confirma que o valor ótimo de Kc é de 2.

Com valor ótimo de Kc igual a dois, os dados foram agrupados em dois grupos e rotulados como 0 e 1. Nota-se que os dados pertencentes ao grupo 0 são os que apresentam a soma valores de *lead time* menores, inferiores a 115 dias e outro, grupo

1, cujo a soma dos valores *lead time* é de aproximadamente de 347. O que indica que o método consegue agrupar e rotular os dados do *lead time* em grupos onde o *lead time* pode ser mais curto ou longo.

Com base nesses agrupamentos e rótulos é possível definir por meio do K-means se as novas variáveis pertencem a um determinado grupo ou não. Além disso, foi obtido um modelo de regressão para qual o valor do *lead time* dos dados rotulados e agrupados.

Em resumo os resultados obtidos estão listados a seguir:

- Aplicação do KNN (usado para realizar agrupamentos e gerar um atributo de rótulo de grupo - como se fosse parte do pré-processamento):
  1. Verificação através do *Elbow Test* e *Silhouette Method*, indicação de dois grupos principais;
  2. Agrupamento para dois grupos, aqui foi gerado um novo atributo contendo o rótulo do grupo (grupo 0 representa os processos que apresentam um *lead time* menor, inferior a 10^7 segundos (aproximadamente 115 dias), enquanto o grupo 1 apresenta *lead time* superior a $3x10^7$ segundos (aproximadamente 347 dias).
- Verificação de correlação. Aqui foi constatado que a categorização obtida pelo K-means com o *lead time* apresenta 0.86 de correlação;
- Foi gerado um modelo de regressão linear usando estes parâmetros para indicação do *lead time* com base nos atributos Número de usuários, Tipo, Número de ocorrências, Número de Unidades, Grupo (o grupo é obtido através da proximidade de vizinhança (K-means));
- O score obtido (R^2) foi de 0.77.

CAPÍTULO

7

CONSIDERAÇÕES FINAIS

Nota-se que essa pesquisa não só propôs o uso de métodos inteligentes para previsão do *lead time*. A pesquisa propôs um método de predição que além da predição, fornecer informações sobre o processo, fornecedores e produto. Dentre os principais informações, a pesquisa apresenta um método que identifica e análise dos gargalos nos processos, produtos e fornecedores, quais são os fornecedores que fornecem determinado produto em tempo hábil, quais os produtos com maior demanda, maior *lead time* que possa requerer otimizações. Além disso, quais os processo e usuários mais lentos, com maior demanda, necessidade de otimizações de tempo de número de funcionários, qual produto deve ser atendido primeiro conforme as dinâmicas da cadeia de suprimentos. A pesquisa fez o uso de três bancos de dados, dois relacionados a cadeia de suprimentos logística farmacêutica e de automação industrial e do processo e outro relacionado ao processamento interno de tramites administrativos de um sistema eletrônico. A metodologia empregada para previsão de *lead time* compreendeu o uso de técnicas inteligentes por meio processo de descoberta de conhecimento em banco de dados (KDD), mineração de dados.

Para todos os bancos de dados foram feitas análises gráficas para selecionar a quantidade de dados e atributos importantes. Posteriormente, os ruídos foram identificados e removidos por meio da ferramenta *blox plot*. Em terceiro lugar, os bancos de dados de cada setor foram segmentados por mês e por categoria de produto, para o banco de dados farmacêutico, por família para o banco de dados de automação industrial e processo para o banco de dados do setor de sistema de informações eletrônicos. Além disso, os dados de cada banco de dados foram transformados no formato dados binários, nominais e numéricos. A fase de validação cruzada utilizou 66,7% das amostras de cada banco de dados para o conjunto de treinamento, 33,3% das amostras para conjuntos de testes. Na validação cruzada foram testados para todos os bancos de dados os algoritmos mais utilizados na literatura, k-vizinhos mais próximos (KNN),

floresta aleatória (RF), regressão linear (LR), multicamadas perceptron (MLP). O erro quadrado médio (MSE) foi calculado para todos os algoritmos com o intuito de definir o algoritmo mais adequado, de menor erro, para cada banco de dados e tipos de dados.

Em relação ao Caso 1, banco de dados do setor farmacêutico, algoritmo SVM com o formato dos dados no tipo binário segmentados por mês alcançou o menor MSE. O erro MSE do SVM foi menor que 2. Além disso, os valores de previsão foram próximos do real e na maioria dos meses a taxa de acerto foi maior que a taxa de erro. Em geral, a média do *lead time* previsto foi igual ao *lead time* real, exceto nos meses de outubro e dezembro, cuja diferença foi de 3 dias, Figura 8. Além disso, em média, as amostras mais previsíveis atingiram entre 63% e 73% de taxa de acerto. Consequentemente, isso significa que a previsão teve maior precisão para a maioria das amostras na análise do *lead time*.

Em relação ao Caso 2, banco de dados do setor de automação cerâmico, na fase de validação cruzada, o algoritmo KNN alcançou o menor erro quadrado médio e foi usado para prever o *lead time* com dados do tipo binário segmentado por mês. Além disso, a previsão para esse banco de dados foi significativa. Em média, Figura 14, a diferença entre *lead time* real e o *lead time* previsto foi inferior a 3 dias. Além disso, na maioria dos meses, a Figura 14 apresentou a maior taxa de acerto do que erro.

No que diz respeito ao Caso 3, banco de dados do sistema eletrônico de informações, os algoritmos RF com os dados segmentados por mês no formato binário apresentaram o menor MSE. Dessa forma o RF com os dados segmentados por mês e com formato binário foi usado para predição do *lead time* do SEI. No momento da predição os valores a predição que rodou de forma eficiente, ou seja, a predição foi obtida em poucos minutos, foram os dos meses de abril, fevereiro, janeiro, março, os demais meses o algoritmo ficou rodando por mais de uma semana e não foram obtidos resultados desses dentro de um tempo hábil. Ademais os valores obtidos dos meses abril, fevereiro, janeiro, março apresentaram um valor predito bastante diverso do valor real. Uma das justificativas avaliadas trata-se de o fato das variáveis não terem correlação entre si conforme exposto na Figura 22. Diante disso foi investigado a correlação entre aos atributos. Notou-se uma correlação fraca e negativa entre os atributos, o que nos indica que métodos de regressão sozinhos não seriam tão eficientes na predição dos dados desse banco de dados. Uma nova correlação foi realizada comparando os atributos e o método K-means no qual apresentou um valor de correlação significativo e positivo, 0.85. Foi obtidos resultados significativos utilizando os métodos híbridos a correlação entre as variáveis foi de 0,85.

O estudo apresentado demonstrou que as técnicas de mineração de dados podem ser aplicadas a problemas prático de engenharia de produção, pois essa pesquisa usou métodos inteligentes para prever o tempo de espera na cadeia de

suprimentos farmacêutica. Além disso, esta pesquisa propôs uma pesquisa científica que visa preencher algumas lacunas na literatura em relação à previsão de *lead time* . Primeiro porque, esta trata-se de mais uma das poucas pesquisas que usam técnicas inteligentes para alcançar de forma mais assertiva o *lead time*. Essa pesquisa contribui como mais uma pesquisa com aplicação de tecnologias da Industria 4.0, big data, machine learning, mineração de dados, KDD para auxiliar na tomada de decisões nas mais diversas áreas, como logística farmacêutica, automação, sistemas eletrônicos públicos, controle do planejamento da produção.

Outra lacuna que essa pesquisa contribui para o preenchimento é o uso de técnicas inteligentes e mineração de dados, em oposição às técnicas usuais na literatura na previsão do *lead time* . Além disso, este estudo é o primeiro a prever o *lead time* de toda a cadeia de suprimentos. Consequentemente, contribuiu como um novo método para prever o *lead time* da cadeia de suprimentos usando técnicas inteligentes, mineração de dados, diante da quantidade de dados da quarta revolução industrial. Além disso, esta pesquisa fornece uma técnica para analisar e reduzir a espera do recebimento de medicamentos no setor farmacêutico, como hospitais e farmácias. Bem como uma técnica de seleção de fornecedores baseado no menor do *lead time* de toda uma cadeia de suprimentos. Uma vez, que se prevê antecipadamente o *lead time* do setor industrial pode-se escolher o fornecedor que entrega o produto em menor tempo possível. Este estudo, também demonstrou que as análises preliminares, como as análises gráficas na etapa de seleção, fornecem informações efetivas sobre o comportamento dos dados.

No entanto, ainda existem novos desafios na pesquisa acadêmica a serem explorados. Especialmente o uso de técnicas da Indústria 4.0, como mineração de dados, big data e inteligência artificial para previsão do *lead time* em uma análise de forma online em tempo real.

Finalmente, tópicos de pesquisas futuras devem incorporar a previsão do *lead time* interno de processos reais de serviço. Além disso, estudos científicos futuros necessitam abordar a previsão do *lead time* com uso da mineração de dados em bancos de dados de processos computacionalmente simulados, considerando sistemas não-lineares para investigar como o tempo de inatividade de uma máquina e as interrupções para manutenção podem influenciar no valor do *lead time*.

REFERÊNCIAS

AFSHOON, I.; MIRI, M.; MOUSAVI, S. R. Combining kriging meta models with u-function and k-means clustering for prediction of fracture energy of concrete. **Journal of Building Engineering**, Elsevier, v. 35, p. 102050. Citado na página 32.

AHMAD, A.; KHAN, S. S. initkmix-a novel initial partition generation algorithm for clustering mixed data using k-means-based clustering. **Expert Systems with Applications**, Elsevier, p. 114149, 2020. Citado na página 32.

AHUETT-GARZA, H.; KURFESS, T. A brief discussion on the trends of habilitating technologies for industry 4.0 and smart manufacturing. **Manufacturing Letters**, Elsevier, v. 15, p. 60–63, 2018. Citado na página 23.

AL-WAELI, A. H.; SOPIAN, K.; YOUSIF, J. H.; KAZEM, H. A.; BOLAND, J.; CHAICHAN, M. T. Artificial neural network modeling and analysis of photovoltaic/thermal system based on the experimental study. **Energy conversion and management**, Elsevier, v. 186, p. 368–379, 2019. Citado 3 vezes nas páginas 27, 28 e 33.

ALBANA, A. S.; FREIN, Y.; HAMMAMI, R. Effect of a lead time-dependent cost on lead time quotation, pricing, and capacity decisions in a stochastic make-to-order system with endogenous demand. **International Journal of Production Economics**, Elsevier, v. 203, p. 83–95, 2018. Citado na página 15.

ALDEN, K. M.; OMID, M.; RAJABIPOUR, A.; TAJEDDIN, B.; FIROUZ, M. S. Quality and shelf-life prediction of cauliflower under modified atmosphere packaging by using artificial neural networks and image processing. **Computers and Electronics in Agriculture**, Elsevier, v. 163, p. 104861, 2019. Citado na página 23.

ALLAHVERDI, A. The third comprehensive survey on scheduling problems with setup times/costs. **European Journal of Operational Research**, Elsevier, v. 246, n. 2, p. 345–378, 2015. Citado na página 23.

ARMANO, G.; FARMANI, M. R. Multiobjective clustering analysis using particle swarm optimization. **Expert Systems with Applications**, Elsevier, v. 55, p. 184–193, 2016. Citado 2 vezes nas páginas 22 e 23.

AVUÇLU, E.; BAŞÇIFTÇI, F. New approaches to determine age and gender in image processing techniques using multilayer perceptron neural network. **Applied Soft Computing**, Elsevier, v. 70, p. 157–168, 2018. Citado 2 vezes nas páginas 31 e 32.

AZMI, M.; RUNGER, G. C.; BERRADO, A. Interpretable regularized class association rules algorithm for classification in a categorical data space. **Information Sciences**, Elsevier, v. 483, p. 313–331, 2019. Citado na página 23.

BAUERNHANSL, T.; HOMPEL, M. T.; VOGEL-HEUSER, B. **Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung-Technologien-Migration**. [S.l.]: Springer, 2014. Citado na página 21.

BECKER, R.; THRÄN, D. Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors. **Applied energy**, Elsevier, v. 208, p. 252–262, 2017. Citado na página 25.

BERLING, P.; FARVID, M. Lead-time investigation and estimation in divergent supply chains. **International Journal of Production Economics**, Elsevier, v. 157, p. 177–189, 2014. Citado 2 vezes nas páginas 15 e 20.

BHOWMIK, S.; CHATTOPADHYAY, R.; CHATTERJEE, U. A review on security measures in data mining. **Imperial Journal of Interdisciplinary Research**, v. 208, 2016. Citado 2 vezes nas páginas 22 e 23.

BORAH, A.; NATH, B. Identifying risk factors for adverse diseases using dynamic rare association rule mining. **Expert systems with applications**, Elsevier, v. 113, p. 233–263, 2018. Citado na página 22.

BUMBLAUSKAS, D.; NOLD, H.; BUMBLAUSKAS, P.; IGOU, A. Big data analytics: transforming data to action. **Business Process Management Journal**, Emerald Publishing Limited, 2017. Citado 2 vezes nas páginas 14 e 17.

CAO, H.; BERNARD, S.; SABOURIN, R.; HEUTTE, L. Random forest dissimilarity based multi-view learning for radiomics application. **Pattern Recognition**, Elsevier, v. 88, p. 185–197, 2019. Citado na página 25.

CARDONA, D. A. B.; NEDJAH, N.; MOURELLE, L. M. Online phoneme recognition using multi-layer perceptron networks combined with recurrent non-linear autoregressive neural networks with exogenous inputs. **Neurocomputing**, Elsevier, v. 265, p. 78–90, 2017. Citado 2 vezes nas páginas 30 e 31.

ÇEBI, F.; OTAY, İ. A two-stage fuzzy approach for supplier evaluation and order allocation problem with quantity discounts and lead time. **Information Sciences**, Elsevier, v. 339, p. 143–157, 2016. Citado na página 15.

CHANG, F.-C. Heuristics for dynamic job shop scheduling with real-time updated queueing time estimates. **International Journal of Production Research**, Taylor & Francis, v. 35, n. 3, p. 651–665, 1997. Citado na página 20.

CHEN, L.; GUO, G. Nearest neighbor classification of categorical data by attributes weighting. **Expert Systems with Applications**, Elsevier, v. 42, n. 6, p. 3142–3149, 2015. Citado 2 vezes nas páginas 24 e 25.

CHENG, C.-H.; CHAN, C.-P.; SHEU, Y.-J. A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 81, p. 283–299, 2019. Citado na página 25.

CHUNG, W.; TALLURI, S.; KOVÁCS, G. Investigating the effects of lead-time uncertainties and safety stocks on logistical performance in a border-crossing jit supply chain. **Computers & Industrial Engineering**, Elsevier, v. 118, p. 440–450, 2018. Citado na página 14.

COSTA, L. B. M.; FILHO, M. G.; FREDENDALL, L. D.; GANGA, G. M. D. The effect of lean six sigma practices on food industry performance: Implications of the sector's experience and typical characteristics. **Food Control**, Elsevier, v. 112, p. 107110, 2020. Citado na página 14.

DALENOGARE, L. S.; BENITEZ, G. B.; AYALA, N. F.; FRANK, A. G. The expected contribution of industry 4.0 technologies for industrial performance. **International Journal of Production Economics**, Elsevier, v. 204, p. 383–394, 2018. Citado 3 vezes nas páginas 14, 17 e 21.

DEEPASHRI, K.; KAMATH, A. Survey on techniques of data mining and its applications. **International Journal of Emerging Research in Management &Technology**, v. 6, p. 198–201, 2017. Citado na página 22.

DJELLOULI, A.; BENYELLOUL, K.; AOURAG, H.; BEKHECHI, S.; ADJADJ, A.; BOUHADDA, Y.; ELKEDIM, O. A datamining approach to classify, select and predict the formation enthalpy for intermetallic compound hydrides. **International Journal of Hydrogen Energy**, Elsevier, v. 43, n. 41, p. 19111–19120, 2018. Citado na página 33.

DOGAN, A.; BIRANT, D. Machine learning and data mining in manufacturing. **Expert Systems with Applications**, Elsevier, p. 114060, 2020. Citado na página 22.

DOGAN, O.; OZTAYSI, B. Genders prediction from indoor customer paths by levenshtein-based fuzzy knn. **Expert Systems with Applications**, Elsevier, v. 136, p. 42–49, 2019. Citado na página 25.

ESMAEILIAN, B.; BEHDAD, S.; WANG, B. The evolution and future of manufacturing: A review. **Journal of Manufacturing Systems**, Elsevier, v. 39, p. 79–100, 2016. Citado na página 17.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996. Citado 3 vezes nas páginas 16, 22 e 34.

FRANK, A. G.; DALENOGARE, L. S.; AYALA, N. F. Industry 4.0 technologies: Implementation patterns in manufacturing companies. **International Journal of Production Economics**, Elsevier, v. 210, p. 15–26, 2019. Citado 3 vezes nas páginas 14, 16 e 21.

____. ____. **International Journal of Production Economics**, Elsevier, v. 210, p. 15–26, 2019. Citado 3 vezes nas páginas 16, 17 e 30.

FRANK, E.; HALL, M. A. **Data mining: practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2011. Citado 3 vezes nas páginas 22, 30 e 31.

GHOLAMI, A.; BONAKDARI, H.; SAMUI, P.; MOHAMMADIAN, M.; GHARABAGHI, B. Predicting stable alluvial channel profiles using emotional artificial neural networks. **Applied Soft Computing**, Elsevier, v. 78, p. 420–437, 2019. Citado 2 vezes nas páginas 23 e 31.

GONG, H.; SUN, Y.; SHU, X.; HUANG, B. Use of random forests regression for predicting iri of asphalt pavements. **Construction and Building Materials**, Elsevier, v. 189, p. 890–897, 2018. Citado 4 vezes nas páginas 25, 26, 27 e 33.

GOU, J.; MA, H.; OU, W.; ZENG, S.; RAO, Y.; YANG, H. A generalized mean distance-based k-nearest neighbor classifier. **Expert Systems with Applications**, Elsevier, v. 115, p. 356–372, 2019. Citado na página 25.

GOU, J.; QIU, W.; YI, Z.; SHEN, X.; ZHAN, Y.; OU, W. Locality constrained representation-based k-nearest neighbor classification. **Knowledge-Based Systems**, Elsevier, v. 167, p. 38–52, 2019. Citado 4 vezes nas páginas 24, 25, 27 e 28.

GYULAI, D.; PFEIFFER, A.; NICK, G.; GALLINA, V.; SIHN, W.; MONOSTORI, L. Lead time prediction in a flow-shop environment with analytical and machine learning approaches. **IFAC-PapersOnLine**, Elsevier, v. 51, n. 11, p. 1029–1034, 2018. Citado 4 vezes nas páginas 14, 15, 19 e 20.

HALIM, Z.; REHAN, M. On identification of driving-induced stress using electroencephalogram signals: A framework based on wearable safety-critical scheme and machine learning. **Information Fusion**, Elsevier, v. 53, p. 66–79, 2020. Citado 3 vezes nas páginas 26, 29 e 30.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007. Citado 2 vezes nas páginas 29 e 30.

H'NG, C. W.; LOH, W. P. A prediction of leaf mechanical properties with data mining. **Computers and Electronics in Agriculture**, Elsevier, v. 162, p. 669–676, 2019. Citado na página 24.

HU, C.; CHEN, Y.; HU, L.; PENG, X. A novel random forests based class incremental learning method for activity recognition. **Pattern Recognition**, Elsevier, v. 78, p. 277–290, 2018. Citado 3 vezes nas páginas 23, 25 e 35.

IOANNOU, G.; DIMITRIOU, S. Lead time estimation in mrp/erp for make-to-order manufacturing systems. **International Journal of Production Economics**, Elsevier, v. 139, n. 2, p. 551–563, 2012. Citado 2 vezes nas páginas 15 e 20.

JESCHKE, S.; BRECHER, C.; MEISEN, T.; ÖZDEMIR, D.; ESCHERT, T. Industrial internet of things and cyber manufacturing systems. In: **(eds) Industrial internet of things.Springer Series in Wireless Technology**. [S.l.]: Springer, 2017. p. 3–19. Citado na página 21.

JIMÉNEZ, A. A.; ZHANG, L.; MUÑOZ, C. Q. G.; MÁRQUEZ, F. P. G. Maintenance management based on machine learning and nonlinear features in wind turbines. **Renewable Energy**, Elsevier, v. 146, p. 316–328, 2020. Citado 3 vezes nas páginas 25, 27 e 28.

JUN, H.-B.; PARK, J.-Y.; SUH, H.-W. Lead time estimation method for complex product development process. **Concurrent Engineering**, Sage Publications Sage CA: Thousand Oaks, CA, v. 14, n. 4, p. 313–328, 2006. Citado na página 20.

KANG, Y.-B.; KRISHNASWAMY, S.; SAWANGPHOL, W.; GAO, L.; LI, Y.-F. Understanding and improving ontology reasoning efficiency through learning and ranking. **Information Systems**, Elsevier, v. 87, p. 101412, 2020. Citado 2 vezes nas páginas 25 e 26.

KIM, S. H.; KIM, J. W.; LEE, Y. H. Simulation-based optimal production planning model using dynamic lead time estimation. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 75, n. 9-12, p. 1381–1391, 2014. Citado 2 vezes nas páginas 19 e 20.

KONG, L.; LI, H.; LUO, H.; DING, L.; ZHANG, X. Sustainable performance of just-in-time (jit) management in time-dependent batch delivery scheduling of precast construction. **Journal of cleaner production**, Elsevier, v. 193, p. 684–701, 2018. Citado na página 14.

LEE, J.; BAGHERI, B.; KAO, H.-A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. **Manufacturing letters**, Elsevier, v. 3, p. 18–23, 2015. Citado na página 20.

LEE, S. H.; CHAN, C. S.; MAYO, S. J.; REMAGNINO, P. How deep learning extracts and learns leaf features for plant classification. **Pattern Recognition**, Elsevier, v. 71, p. 1–13, 2017. Citado na página 23.

LEPENIOTI, K.; BOUSDEKIS, A.; APOSTOLOU, D.; MENTZAS, G. Prescriptive analytics: Literature review and research challenges. **International Journal of Information Management**, Elsevier, v. 50, p. 57–70, 2020. Citado 3 vezes nas páginas 23, 24 e 27.

LI, Y.; JIANG, W.; YANG, L.; WU, T. On neural networks and learning systems for business computing. **Neurocomputing**, Elsevier, v. 275, p. 1150–1159, 2018. Citado na página 26.

LI, Y.; ZOU, C.; BERECIBAR, M.; NANINI-MAURY, E.; CHAN, J. C.-W.; BOSSCHE, P. van den; MIERLO, J. V.; OMAR, N. Random forest regression for online capacity estimation of lithium-ion batteries. **Applied energy**, Elsevier, v. 232, p. 197–210, 2018. Citado na página 26.

LINGITZ, L.; GALLINA, V.; ANSARI, F.; GYULAI, D.; PFEIFFER, A.; MONOSTORI, L. Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. **Procedia CIRP**, Elsevier, v. 72, p. 1051–1056, 2018. Citado 4 vezes nas páginas 14, 15, 19 e 20.

LIU, Y. H. **Python Machine Learning By Example**. [S.l.]: Packt Publishing Ltd, 2017. Citado 2 vezes nas páginas 28 e 29.

MAIRIZAL, A. Q.; AWAD, S.; PRIADI, C. R.; HARTONO, D. M.; MOERSIDIK, S. S.; TAZEROUT, M.; ANDRES, Y. Experimental study on the effects of feedstock on the properties of biodiesel using multiple linear regressions. **Renewable Energy**, Elsevier, v. 145, p. 375–381, 2020. Citado na página 27.

MANIKANDAN, G.; ABIRAMI, S. A survey on feature selection and extraction techniques for high-dimensional microarray datasets. In: **Knowledge Computing and its Applications**. [S.l.]: Springer, 2018. p. 311–333. Citado na página 23.

MERCADIER, M.; LARDY, J.-P. Credit spread approximation and improvement using random forest regression. **European Journal of Operational Research**, Elsevier, v. 277, n. 1, p. 351–365, 2019. Citado 2 vezes nas páginas 24 e 26.

MOONAM, H. M.; QIN, X.; ZHANG, J. Utilizing data mining techniques to predict expected freeway travel time from experienced travel time. **Mathematics and Computers in Simulation**, Elsevier, v. 155, p. 154–167, 2019. Citado na página 23.

MOURTZIS, D.; DOUKAS, M.; FRAGOU, K.; EFTHYMIOU, K.; MATZOROU, V. Knowledge-based estimation of manufacturing lead time for complex engineered-to-order products. **Procedia CIRP**, Elsevier, v. 17, p. 499–504, 2014. Citado na página 20.

NAGAHARA, S.; NONAKA, Y. Product-specific process time estimation from incomplete point of production data for mass customization. **Procedia CIRP**, Elsevier, v. 67, p. 558–562, 2018. Citado na página 17.

NETO, A. A.; PEREIRA, G. B.; DROZDA, F. O.; SANTOS, A. d. P. L. A busca de uma identidade para a indústria 4.0/the search for an industry 4.0 identity. **Brazilian Journal of Development**, v. 4, n. 4, p. 1379–1395, 2018. Citado na página 21.

NI, D.; JI, X.; WU, M.; WANG, W.; DENG, X.; HU, Z.; WANG, T.; SHEN, D.; CHENG, J.-Z.; WANG, H. Automatic cystocele severity grading in transperineal ultrasound by random forest regression. **Pattern Recognition**, Elsevier, v. 63, p. 551–560, 2017. Citado na página 26.

NING, Q.; MA, Z.; ZHAO, X. dforml (knn)-pseaac: Detecting formylation sites from protein sequences using k-nearest neighbor algorithm via chou's 5-step rule and pseudo components. **Journal of theoretical biology**, Elsevier, v. 470, p. 43–49, 2019. Citado 2 vezes nas páginas 24 e 25.

NOGUEIRA, I. B.; RIBEIRO, A. M.; REQUIÃO, R.; PONTES, K. V.; KOIVISTO, H.; RODRIGUES, A. E.; LOUREIRO, J. M. A quasi-virtual online analyser based on an artificial neural networks and offline measurements to predict purities of raffinate/extract in simulated moving bed processes. **Applied Soft Computing**, Elsevier, v. 67, p. 29–47, 2018. Citado 2 vezes nas páginas 30 e 31.

NOORI-DARYAN, M.; TALEIZADEH, A. A.; JOLAI, F. Analyzing pricing, promised delivery lead time, supplier-selection, and ordering decisions of a multi-national supply chain under uncertain environment. **International Journal of Production Economics**, Elsevier, v. 209, p. 236–248, 2019. Citado 4 vezes nas páginas 14, 15, 17 e 20.

NURUNNABI, A.; WEST, G.; BELTON, D. Outlier detection and robust normal-curvature estimation in mobile laser scanning 3d point cloud data. **Pattern Recognition**, Elsevier, v. 48, n. 4, p. 1404–1419, 2015. Citado na página 23.

OLIFF, H.; LIU, Y. Towards industry 4.0 utilizing data-mining techniques: a case study on quality improvement. **Procedia CIRP**, Elsevier, v. 63, p. 167, 2017. Citado na página 21.

PAUL, A.; MUKHERJEE, D. P. Reinforced quasi-random forest. **Pattern Recognition**, Elsevier, v. 94, p. 13–24, 2019. Citado na página 25.

PFEIFFER, A.; GYULAI, D.; KÁDÁR, B.; MONOSTORI, L. Manufacturing lead time estimation with the combination of simulation and statistical learning methods. **PROCEDIA CIRP**, Elsevier, v. 41, p. 75–80, 2016. Citado na página 15.

____. ____. **Procedia CIRP**, Elsevier, v. 41, p. 75–80, 2016. Citado na página 20.

PRODANOV, C. C.; FREITAS, E. C. de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição**. [S.l.]: Editora Feevale, 2013. Citado na página 34.

QIN, Z.; ZHANG, Y.; MENG, S.; QIN, Z.; CHOO, K.-K. R. Imaging and fusing time series for wearable sensor-based human activity recognition. **Information Fusion**, Elsevier, v. 53, p. 80–87, 2020. Citado na página 24.

RAHMAN, M. A.; ISLAM, M. Z. Application of a density based clustering technique on biomedical datasets. **Applied soft computing**, Elsevier, v. 73, p. 623–634, 2018. Citado na página 32.

RAMEZANIAN, R.; PEYMANFAR, A.; EBRAHIMI, S. B. An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in tehran stock exchange market. **Applied soft computing**, Elsevier, v. 82, p. 105551, 2019. Citado na página 23.

RISTOSKI, P.; PAULHEIM, H. Semantic web in data mining and knowledge discovery: A comprehensive survey. **Journal of Web Semantics**, Elsevier, v. 36, p. 1–22, 2016. Citado 4 vezes nas páginas 34, 35, 36 e 37.

SAGAERT, Y. R.; KOURENTZES, N.; VUYST, S. D.; AGHEZZAF, E.-H.; DESMET, B. Incorporating macroeconomic leading indicators in tactical capacity planning. **International Journal of Production Economics**, Elsevier, v. 209, p. 12–19, 2019. Citado 3 vezes nas páginas 22, 23 e 35.

SAIDI, F.; SEBAA, N.; MAHMOUDI, A.; AOURAG, H.; MERAD, G.; DERGAL, M. Structural electronic and mechanical properties of ym2 (m= mn, fe, co) laves phase compounds: First principle calculations analyzed with datamining approach. **Solid State Communications**, Elsevier, v. 274, p. 9–20, 2018. Citado na página 33.

SCHUH, G.; PROTE, J.-P.; LUCKERT, M.; HÜNNEKES, P. Knowledge discovery approach for automated process planning. **Procedia CIRP**, Elsevier, v. 63, n. 1, p. 539–544, 2017. Citado na página 14.

SETTOUTI, N.; BECHAR, M. E. A.; CHIKH, M. A. Statistical comparisons of the top 10 algorithms in data mining for classification task. **International Journal of Interactive Multimedia and Artificial Intelligence**, v. 4, n. 1, p. 46–51, 2016. Citado na página 24.

SHAO, Y.; LIU, B.; WANG, S.; LI, G. A novel software defect prediction based on atomic class-association rule mining. **Expert Systems with Applications**, Elsevier, v. 114, p. 237–254, 2018. Citado 2 vezes nas páginas 22 e 23.

SHARIFZADEH, M.; SIKINIOTI-LOCK, A.; SHAH, N. Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and gaussian process regression. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 108, p. 513–538, 2019. Citado 4 vezes nas páginas 24, 28, 29 e 31.

SHENG, J.; AMANKWAH-AMOAH, J.; WANG, X. A multidisciplinary perspective of big data in management research. **International Journal of Production Economics**, Elsevier, v. 191, p. 97–112, 2017. Citado na página 14.

SIEVERS, S.; SEIFERT, T.; FRANZEN, M.; SCHEMBECKER, G.; BRAMSIEPE, C. Lead time estimation for modular production plants. **Chemical Engineering Research and Design**, Elsevier, v. 128, p. 96–106, 2017. Citado 2 vezes nas páginas 15 e 21.

ŠIKŠNYS, L.; PEDERSEN, T.; LIU, L.; ÖZSU, M. Prescriptive analytics. **Encyclopedia of Database Systems**, Springer, p. 1–2, 2016. Citado na página 23.

SUN, T. Q.; MEDAGLIA, R. Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. **Government Information Quarterly**, Elsevier, v. 36, n. 2, p. 368–383, 2019. Citado na página 23.

TATSIOPOULOS, I.; KINGSMAN, B. Lead time management. **European Journal of Operational Research**, Elsevier, v. 14, n. 4, p. 351–358, 1983. Citado na página 20.

TRSTENJAK, M.; COSIC, P. Process planning in industry 4.0 environment. **Procedia Manufacturing**, Elsevier, v. 11, p. 1744–1750, 2017. Citado na página 21.

TSAGKRASOULIS, D.; MONTANA, G. Random forest regression for manifold-valued responses. **Pattern Recognition Letters**, Elsevier, v. 101, p. 6–13, 2018. Citado na página 26.

VANDAELE, N.; BOECK, L. D.; CALLEWIER, D. An open queueing network for lead time analysis. **IIE transactions**, Taylor & Francis, v. 34, n. 1, p. 1–9, 2002. Citado na página 20.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern recognition**, Elsevier, v. 44, n. 2, p. 330–349, 2011. Citado na página 25.

WANG, L.; TÖRNGREN, M.; ONORI, M. Current status and advancement of cyber-physical systems in manufacturing. **Journal of Manufacturing Systems**, Elsevier, v. 37, p. 517–527, 2015. Citado na página 21.

WANG, Y.; SHEN, T.; YUAN, G.; BIAN, J.; FU, X. Appearance-based gaze estimation using deep features and random forest regression. **Knowledge-Based Systems**, Elsevier, v. 110, p. 293–301, 2016. Citado na página 26.

WERKEMA, C. Introdução ás ferramentas do lean manufacturing. **Belo Horizonte: Werkema**, 2011. Citado 3 vezes nas páginas 14, 19 e 20.

WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. **Acm Sigmod Record**, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002. Citado 2 vezes nas páginas 22 e 24.

YADAV, A. K.; CHANDEL, S. Identification of relevant input variables for prediction of 1-minute time-step photovoltaic module power using artificial neural network and multiple linear regression models. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 77, p. 955–969, 2017. Citado 2 vezes nas páginas 30 e 31.

ZHANG, S.; LI, X.; ZONG, M.; ZHU, X.; WANG, R. Efficient knn classification with different numbers of nearest neighbors. **IEEE transactions on neural networks and learning systems**, IEEE, v. 29, n. 5, p. 1774–1785, 2017. Citado na página 25.

ZHANG, Y.; CAO, G.; WANG, B.; LI, X. A novel ensemble method for k-nearest neighbor. **Pattern Recognition**, Elsevier, v. 85, p. 13–25, 2019. Citado 5 vezes nas páginas 23, 24, 25, 32 e 33.

ZHAO, X.; ROSEN, D. W. A data mining approach in real-time measurement for polymer additive manufacturing process with exposure controlled projection lithography. **Journal of Manufacturing Systems**, Elsevier, v. 43, p. 271–286, 2017. Citado na página 23.

ZHAO, Y.; ZHANG, C.; ZHANG, Y.; WANG, Z.; LI, J. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. **Energy and Built Environment**, Elsevier, v. 1, n. 2, p. 149–164, 2020. Citado na página 25.