DISSERTAÇÃO DE MESTRADO

Aprendizagem de máquina aplicada à predição do tempo de fabricação de novos produtos: Um estudo exploratório com foco no tipo de material utilizado em empresa de produção mecatrônica da área médica e espacial

Roberto Canedo Rosa

Brasília, 10 de fevereiro de 2020.

UNIVERSIDADE DE BRASILIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO

Aprendizagem de máquina aplicada à predição do tempo de fabricação de novos produtos: Um estudo exploratório com foco no tipo de material utilizado em empresa de produção mecatrônica da área médica e espacial

Roberto Canedo Rosa

Dissertação submetida ao Departamento de Engenharia Mecânica da Faculdade de Tecnologia da Universidade de Brasília, como requisito parcial para obtenção do grau de Mestre em Sistemas Mecatrônicos.

Banca Examinadora

Prof. Dr. Sanderson César Macêdo Barbalho, UnB/ ENM (Orientador)

Prof. Dr. Li Weigang, UnB/ ENM (Examinador Interno)

Prof. Dr. Marcelo Becker, USP/ EESC (Examinador Externo)

Brasília, 10 de fevereiro de 2020.

Aos meus pais, Margareth e Carlos (in memorian), que sempre apoiaram as escolhas de um filho sonhador.

Agradecimentos

Gratidão a Vida e a partícula criadora inteligente que alimenta o universo com suas leis, doando aos cientistas um imenso universo de pesquisas. Agradeço a todos os Deuses e crenças que acreditam no amor como fragmento de construção, conexão e transformação social.

Agradeço a todos os meus ancestrais! Aqueles que dedicaram a sua vida pela ciência e auxiliaram na codificação das leis da natureza, sem vocês esse trabalho não seria possível.

Ao Professor Darcy Ribeiro, por seu legado como educador. Por ser um idealista que conseguiu materializar seus sonhos inconformados. Por tudo que concretizou na Universidade de Brasília (UNB) e que hoje inspira os meus caminhos acadêmicos.

Ao Professor orientador Sanderson César Macêdo Barbalho por seu entusiasmo em fazer ciência. Por sua sabedoria em fazer conexões, e por seu ponto de vista inovador da prática docente. Agradeço cada tempo doado para dar luz ao meu saber. A carreira acadêmica tornou-se mais encantadora e motivadora ao ver os seus esforços para materializar projetos de grande impacto científico e social.

Agradeço a disponibilização dos dados pela empresa e por toda a parceria no processo de entendimento e análise dos dados.

A minha mãe, Margareth, uma fonte de sabedoria e afetos que sempre me ensina. A primeira grande educadora, que me cultivou com todo amor e carinho para colher os melhores frutos. Gratidão por seu apoio, por seu amor e por sempre me embalar em seus braços. Sem você, eu nada seria.

Agradeço a todos aqueles que direta ou indiretamente me apoiaram nessa etapa. Agradeço a minha irmã Sheline por suas preocupações e afetos. Ao amigo Acássio por sua escuta sempre sensível, por seu apoio e seus ensinamentos diários. A todos os amigos que enviaram vibrações positivas para essa etapa da minha existência.

Gratidão a todos os ensinamentos adquiridos durante a pós-graduação. Aos professores do programa que me nortearam na construção de novos conhecimentos, e pela motivação para prosseguir na carreira acadêmica. A todos os amigos do Grupo de Pesquisa em Inovação, Projetos e Processos (GPIPP), por tudo compartilhado, pelas inúmeras reuniões que geraram grandes aprendizados, e por todas as parcerias.

Agradeço aos programas de apoio à ciência no Brasil que permitiram que essa dissertação fosse concretizada. Gratidão aos governantes do país que ainda lutam e enxergam a educação como uma partícula trivial para o desenvolvimento econômico, social e cultural de uma sociedade.

Agradeço as artes, ao teatro, a dança, a música, os versos e poesias que foram essenciais para produzir o texto dissertativo. A todos os estímulos sonoros e visuais que me ajudaram a encontrar caminhos diante dos desafios.

Por fim, agradeço a oportunidade de agradecer. Diante dos grandes desafios enfrentados pela sociedade contemporânea consigo compreender o privilégio de finalizar um mestrado, que eu possa retribuir para a população essa oportunidade construída por todos os brasileiros. Agradeço, ainda, ao futuro, que seja próspero e repleto de oportunidades às futuras gerações.

“O professor é, naturalmente, um artista, mas ser artista não significa que ele ou ela consiga formar o perfil e moldar seus alunos. O que um educador faz no ensino é tornar possível que os alunos se tornem eles mesmos.”

Paulo Freire

RESUMO

A indústria 4.0 é um projeto estratégico que combina inúmeras tecnologias com o intuito de transformar as cadeias de valor da indústria, da produção e dos modelos de negócio. Realizar estimativas e prever eventos futuros incorrem em desafios confrontados pela quarta revolução industrial. Predizer o *lead-time* total de fabricação de um produto ou componente na indústria torna-se uma atividade crítica e vital para um ambiente de grande competitividade enfrentado por um mercado altamente globalizado e direcionado para a inovação. O trabalho proposto apresenta um estudo com foco em predizer o *lead-time* total de fabricação de peças e componentes de novos produtos de uma empresa com produção mecatrônica na área de equipamentos médicos e espaciais. A análise preditiva e o aprendizado de máquina através dos algoritmos de Redes Neurais Artificiais (RNA), *Support Vector Machine* (SVM) e *Random Forest* (RF) delinearam a metodologia aplicada no estudo. Os modelos de predição foram aplicados em um conjunto de dados de ordens de serviço da empresa, e os resultados mostraram a eficiência do método na estimativa do *lead-time* total de fabricação. Os melhores resultados apresentaram uma taxa de acerto acima de 87% e com um erro médio absoluto menor que um (1) dia para a fabricação com o material alumínio 6061.

Palavras chaves: Aprendizagem de máquinas, análise preditiva, produção mecatrônica, *lead-time* de fabricação.

ABSTRACT

Industry 4.0 is a strategic project that combines numerous technologies in order to transform the value chains of industry, production and business models. Making estimates and predicting future events face challenges faced by the fourth industrial revolution. Predicting the total lead time for manufacturing a product or component in the industry becomes a critical and vital activity for a highly competitive environment faced by a highly globalized and innovation-driven market. The proposed work presents a study aimed at predicting the total lead time of manufacturing parts and components for new products of a company with mechatronic production in the area of medical and space equipment. Predictive analysis and machine learning through Artificial Neural Network (RNA), Support Vector Machine (SVM) and Random Forest (RF) algorithms outlined the methodology applied in the study. The prediction models were applied to a dataset of the company's work orders, and the results showed the efficiency of the method in estimating the total lead time of manufacture. The best results showed a hit rate above 87% and with an average absolute error less than one (1) day for manufacturing with 6061 aluminum material.

Key words: Machine learning, predictive analysis, mechatronic production, manufacturing lead time.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

CQ - Controle de Qualidade

EAPD - Escritório de Apoio à Pesquisa e Desenvolvimento

EMA - Erro Médio Absoluto

EQM - Erro Quadrático Médio

FN - Falsos Negativos

FP - Falsos Positivos

GI - Gerência Industrial

IoT - Internet das coisas (*Internet of Things*)

KDD - Descoberta de conhecimento em banco de dados (*Knowledge-Discovery in Databases*)

LT - Lead Time

LD - Lead time de distribuição

LFC - Lead time de Fabricação de Componentes

LM - Lead time de Montagem

LPr - Lead time de Projeto

LS - Lead time de Suprimentos

P&D - Pesquisa e Desenvolvimento

PMS - Sistemas de Manufatura Preditiva (*Predictive Manufacturing System*)

RF - Floresta Aleatória (*Random Forest*)

RNA - Redes Neurais Artificiais

SVM - Máquina de vetores de suporte (*Support Vector Machine*)

TN - Verdadeiros Negativos

TP - Verdadeiros Positivos

TS - Tratamento Superficial

UENG - Unidade de engenharia

UGD - Unidade de Gerenciamento e Documentação

LISTA DE SÍMBOLOS

Δ - Gradiente

Σ - Somatório

K - Função kernel

S - Coleção de índices dos vetores de suporte

e - Erro

exp - Exponencial

m - Número de entradas do neurônio

n - Número de iterações

w - Peso sináptico

δ - Gradiente local

η - Taxa de aprendizagem

φ - Função de ativação

∂ - Derivada parcial

1. INTRODUÇÃO

Este capítulo contempla a contextualização do tema de pesquisa buscando nortear o leitor quanto ao escopo do trabalho. A justificativa e a relevância da pesquisa são apresentadas como orientadores da importância do tema tanto na indústria, quanto na academia. O problema de pesquisa é formulado dando a sustentação prática e intelectual do trabalho dissertativo. Posteriormente, os objetivos gerais do trabalho e seus desdobramentos em objetivos específicos são expostos, e por fim, apresenta-se como a dissertação foi organizada.

1.1 CONTEXTUALIZAÇÃO

A indústria 4.0 é um termo amplamente difundido para a estratégia de base tecnológica e de mercado desenvolvida pela Alemanha que combina tecnologias de sistema de produção com processos inteligentes. O projeto da quarta revolução industrial, ou indústria 4.0, visa construir um caminho para uma nova era tecnológica com transformações nas cadeias de valor da indústria, da produção, e dos modelos de negócios. Os sistemas de manufatura são modernizados para atingir um nível inteligente, adquirindo processos flexíveis e reconfiguráveis para atingir um mercado altamente dinâmico e globalizado (ZHONG *et al*., 2017).

A manufatura inteligente tem como fundamento solucionar questões chave na fabricação como atender aos requisitos individuais do cliente, a tomada de decisão otimizada, e melhorar a eficiência dos recursos e da energia empregada nos processos (ROMEO *et al.*, 2019). Aprimorar e otimizar as relações entre produção e produto, melhorar o design, o gerenciamento, e a integração de todo ciclo de vida do produto são princípios fundamentais da manufatura inteligente. Sensores inteligentes, modelos de tomada de decisão adaptáveis, materiais avançados, dispositivos inteligentes e análise de dados são preceitos basilares desse novo modelo de fabricação (LI *et al*., 2017).

A fábrica inteligente está amparada por avanços tecnológicos que comportam a integração homem-máquina, a manufatura aditiva, as operações remotas, a internet das coisas (IoT), a computação em nuvem, o big data e os sistemas cyber-físicos (KHAN *et al.*, 2017).

Os sistemas de manufatura preditiva são também um importante desafio aplicado ao conceito de indústria 4.0, e auxiliam em características fundamentais de auto previsão, automanutenção e auto consciência. O conceito desse modelo é introduzir uma atitude antecipada a partir de previsões de estados futuros e por consequência diminuir as ações reativas da empresa. Técnicas de estatística, mineração de dados, modelagem e inteligência artificial são conceitos basilares do sistema de produção preditivo (NIKOLIC *et al.,* 2017).

Dentre os algoritmos preditivos inseridos nos fundamentos da inteligência artificial, o aprendizado de máquinas é uma abordagem que apresenta soluções promissoras para analisar dados, prever desempenho de equipamentos, gerenciar e otimizar de maneira autônoma serviços, produtos e necessidades da indústria (SUSTO *et al*., 2014).

Como tema promissor de pesquisa atrelado ao contexto de tecnologias basilares associadas ao conceito da indústria 4.0, a estimativa de *lead-time*, do tempo de desenvolvimento e fabricação de produtos no chão de fábrica representa uma característica vital. Predizer os *lead-times* na indústria afeta diretamente a gestão e os custos de um projeto (BASHIR, 2008).

O termo *lead-time* pode ser definido como tempo de entrega que corresponde ao intervalo entre o início e o fim de uma atividade (MOURTZIS *et al.,* 2014). Para Silva e Fernandes (2008) os *lead-times* de um sistema produtivo englobam o *lead-time* de projeto, o *lead-time* de suprimentos, o *lead-time* de fabricação, o *lead-time* de montagem e de distribuição, Figura 1.

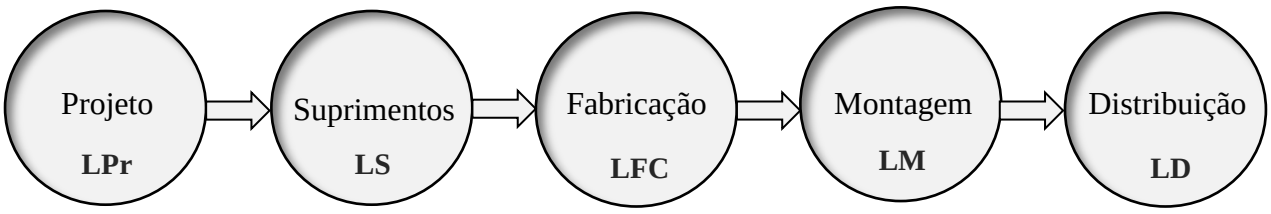

Figura 1- Lead-times de um Sistema de produção. Fonte: Silva e Fernandes (2008)

À medida que a concorrência global aumenta, os mercados pressionam para a redução do tempo de ciclo de vida do produto. A predição oportuna da estimativa do *lead-time* envolvendo todo o processo de desenvolvimento de um novo produto e o sucesso com os prazos de introdução no mercado a torna um importante ativo para as empresas de manufatura. A vantagem competitiva, desta maneira, é dada a empresa com a melhor habilidade em estimar o tempo de desenvolvimento de um novo produto, apresentando-se como uma capacidade crítica para os gestores de projeto. Entretanto, a maioria dos projetos de desenvolvimento de novos produtos são afetados amiúde por falhas no excedente de

cronograma e por suas estimativas precárias dos *lead-times* do processo (JUN *et al.*, 2006; WU, 2011).

Partindo do conceito de indústria 4.0 e de suas tecnologias basilares associadas, foi possível realizar a contextualização do tema de pesquisa que trata da predição do tempo de fabricação de novos produtos. A partir dos preceitos associados à quarta revolução industrial e de suas ferramentas de aplicação foi possível realizar um diálogo entre a necessidade de se estimar o tempo de desenvolvimento e fabricação de novos produtos, e as ferramentas e conceitos cultivados pela indústria 4.0. O próximo tópico visa elucidar sobre a relevância da pesquisa desenvolvida e justificar a sua importância.

1.2 JUSTIFICATIVA E RELEVÂNCIA

O curto prazo de entrega em um mercado extremamente competitivo melhora a imagem da empresa e o potencial para vendas futuras (ASADZADEH *et al.*, 2011). A estimativa do *lead-time* torna-se, portanto, uma importante tarefa para gestores, pois afeta diretamente as relações com os clientes e os processos de gerenciamento de operações que ocorrem no chão de fábrica (ÖZTÜRK *et al.*, 2006). Todavia, a estimativa do *lead-time* se torna uma difícil tarefa devido a sua natureza complexa, as incertezas envolvidas, as não linearidades e a pequena quantidade de padrões de dados de todo o processo de desenvolvimento de um novo produto (MOUSAVI *et al*., 2013).

Um excesso na duração do cronograma de desenvolvimento de novos produtos aumenta substancialmente o risco de obsolescência do produto e, como consequência, o crescente risco de perder a janela de mercado. Outrossim, a previsão imprecisa do tempo de desenvolvimento de um produto influencia significativamente a qualidade, a eficiência do planejamento e a programação da produção (LINGITZ *et al*., 2018).

Melhorar a estimativa dos *lead-times* que envolvem o desenvolvimento de novos produtos tornou-se uma necessidade técnica vital para gerentes de projetos. Uma estimativa de tempo precisa e confiável é imprescindível para os estágios iniciais da pesquisa industrial. Desta maneira, a previsão do tempo de um projeto torna-se essencial, tendo em vista as necessidades de alocação de recursos para os produtos em potencial e para os investimentos em P&D. Reduzir o tempo de desenvolvimento de novos produtos e melhorar a capacidade de estimar os *lead-times* de um sistema de produção dentro de um erro aceitável são esforços

corriqueiros dos gestores de projetos, com alvo na redução de tempo e custo do projeto (BASHIR, 2008; MOUSAVI *et al*., 2013).

Verifica-se, desta maneira, que estimar o tempo de desenvolvimento e fabricação de novos produtos é um tema de extrema relevância para gerentes de engenharia e de projetos que almejam garantir uma maior probabilidade de sucesso em um mercado extremamente competitivo. Como tema promissor no ambiente industrial e seus impactos que afetam diretamente a sociedade de consumo, a questão levantada torna-se um problema de pesquisa para a academia que desenvolve esforços para encontrar soluções e elucidar caminhos/recursos para modelar e controlar esse problema. Apesar dos benefícios e necessidades em estimar os tempos na indústria, poucos modelos são desenvolvidos devido à complexidade e incertezas inerentes ao processo de desenvolvimento de um novo produto (JUN *et al.*, 2005; XU e YAN, 2006).

1.3 DEFINIÇÃO DO PROBLEMA

Apesar da importância e dos impactos dos *lead-times* de desenvolvimento/fabricação de produtos nos custos de projeto, pouca pesquisa sistemática tem sido realizada. Existe uma preocupação e um grande esforço de gestores de projetos em reduzir o tempo e custo no desenvolvimento de um novo produto, mas pouca pesquisa com foco em estimar o tempo de desenvolvimento/fabricação. Com o aumento da concorrência no mercado e das complexidades do produto as empresas exigem soluções que sejam cada vez mais precisas e exatas (WU, 2011).

Métodos analíticos, experimentais e heurísticos já foram propostos para estimar o *lead-time* dos sistemas de produção, dada a natureza multifacetada do problema. Inúmeros métodos de análise de dados são aplicados para a estimativa do *lead-time*, a citar arvores de regressão (ÖZTÜRK *et al.*, 2006), raciocínio baseado em caso (MOURTZIS *et al.*, 2014), máquina de vetores de suporte (DE COS JUEZ *et al.*, 2010), as redes bayesianas (MORI e MAHALEC, 2015) e as redes neurais artificiais (ASADZADEH *et al.*, 2011). Atualmente os métodos mais robustos para a predição do *lead-time* na indústria envolvem técnicas de inteligência artificial e aprendizado de máquina (MOURTZIS *et al.*, 2014).

Diante da relevância, dos impactos, e dos métodos propostos para estimar o *lead-time* nos sistemas de produção, apresentam-se como problemas de pesquisa traduzidos como perguntas:

1. Como os métodos de aprendizagem de máquina podem ser úteis na estimativa do *lead-time* de fabricação de componentes para protótipos em uma empresa de base tecnológica com foco no desenvolvimento de produtos mecatrônicos da área médica e espacial?
2. De que forma é possível utilizar as ordens de serviço de uma empresa com foco no desenvolvimento de produtos mecatrônicos para predizer o tempo de fabricação de componentes dessa empresa estudo de caso?

De acordo com Gil (2002) o problema de pesquisa além de ser formulado como pergunta deve ser claro, preciso, empírico, suscetível à solução e delimitado a uma dimensão viável. O problema de pesquisa é delimitado ao estudo de caso da coleta de dados de uma empresa que desenvolve produtos mecatrônicos, o que torna o problema disponível para a investigação. Além disso, o problema de pesquisa é empírico e suscetível de solução por apresentar soluções similares na bibliografia através de aplicação do método de aprendizado de máquinas.

A solução do problema de pesquisa revela sua importância bilateral. Na academia apresenta-se como um estudo de caso específico de uma empresa de base tecnológica que desenvolve produtos médicos e espaciais, sendo um trabalho original que avalia as ordens de serviço da empresa específica para encontrar padrões por métodos de aprendizado de máquina. Na indústria mostra-se como uma importante metodologia de pesquisa para amparar gestores de projetos para tomadas de decisões com foco em estimar o tempo de fabricação de componentes, que tem impactos diretos nos custos e na gestão do projeto de desenvolvimento de produtos.

1.4 OBJETIVOS

O objetivo geral do trabalho compreende estimar/predizer o tempo de fabricação de peças e componentes de novos produtos de uma empresa de base tecnológica, com foco em desenvolvimento de novos produtos com características mecatrônicas da área de equipamentos médicos e espaciais. O delineamento metodológico aplicado à pesquisa para atingir o objetivo do trabalho inclui a análise preditiva e o aprendizado de máquina. A metodologia e os resultados são validados a partir de métricas de análise de desempenho dos modelos de aprendizagem propostos.

O objetivo geral se desdobra nos objetivos específicos:

- Realizar uma contextualização da relação entre a indústria 4.0, a predição do tempo de desenvolvimento de novos produtos e inteligência artificial;
- Identificar conceitos e métodos da análise preditiva e do aprendizado de máquinas;
- Verificar os métodos com melhores resultados a partir de análises de desempenho do modelo;
- Evidenciar os resultados encontrados, pela aplicação dos métodos de aprendizagem de máquinas e análise preditiva, frente a discussões de pesquisa.

1.5 ORGANIZAÇÃO DA DISSERTAÇÃO

Após considerar os pressupostos de contextualização, a justificativa e relevância do tema, os objetivos gerais e específicos, a organização da dissertação é apresentada:

- Primeiramente, foi redigida uma breve contextualização com o objetivo de nortear a respeito da indústria 4.0, e como o problema de pesquisa, com foco em predizer o tempo de fabricação de componentes de novos produtos, está inserido dentro deste contexto da quarta revolução industrial (Capítulo 1). Busca-se nessa introdução apresentar a importância da pesquisa desenvolvida e como o método proposto (aprendizagem de máquina) e seus impactos auxiliam os gerentes de projeto na construção do modelo de uma fábrica inteligente. Os objetivos gerais e específicos são apresentados buscando orientar e nortear o processo de pesquisa.
- O capítulo 2 descreve os conceitos associados à indústria 4.0 e como esse novo modelo de fábrica inteligente tem impactado a gestão e os processos na indústria. Um foco é dado para os modelos preditivos da manufatura e seus métodos de aplicação, sendo um dos temas basilares para construção de uma fábrica com autoconsciência e auto previsão. Neste capítulo também são apresentados os conceitos de *lead-time* junto às pesquisas desenvolvidas com o objetivo de predizer o tempo de desenvolvimento e fabricação de produtos na indústria.
- O capítulo 3 expõe o processo de construção de um modelo preditivo, a técnica de aprendizagem de máquina e suas etapas fundamentais. Um foco é dado para três métodos, Redes Neurais Artificiais (RNA) e *Support Vector Machine* (SVM), e *Random Forest* (RF) por apresentarem bons resultados com problemas similares ao exposto pela dissertação. As métricas de desempenho junto aos modelos de

validação são apresentadas, além das ferramentas de aplicação da aprendizagem de máquina.

- O capítulo 4 descreve a metodologia aplicada à pesquisa tendo como norte os modelos de análise preditiva e de aprendizagem de máquinas. O capítulo visa elucidar sobre os caminhos trilhados para solucionar o problema de pesquisa.
- O capítulo 5 apresenta os resultados e discussões a partir dos procedimentos de análise realizados, tendo como base a fundamentação teórica da dissertação, os objetivos de pesquisa e a metodologia aplicada. O capítulo tem como norte apresentar os resultados do modelo preditivo aplicado expondo como as questões direcionadoras da dissertação foram resolvidas.
- Por fim, apresentam-se as Conclusões da pesquisa, nas quais são exibidas as principais contribuições, os pontos relevantes do conhecimento adquirido na elaboração do trabalho, as dificuldades encontradas e as sugestões para trabalhos futuros. Em seguida, expõe-se o Referencial bibliográfico.

2. INDÚSTRIA 4.0 E MODELOS PREDITIVOS

Com o objetivo de contextualizar a pesquisa realizada e traçar considerações conceituais acerca da indústria 4.0 e como os modelos preditivos são inseridos nesse contexto, o capítulo apresentado vislumbra tecer diálogos a respeito do projeto estratégico e tecnológico da quarta revolução industrial, as potencialidades da inteligência artificial para a implantação desse novo modelo, e apresentar conceitos e resultados de pesquisa que utilizaram ferramentas aplicadas à indústria 4.0 para determinar o tempo de desenvolvimento/fabricação de novos produtos.

2.1 INDÚSTRIA 4.0 E FABRICAÇÃO INTELIGENTE

A Indústria 4.0 é um termo e projeto tecnológico/estratégico originado na Alemanha que visa criar fábricas inteligentes controladas de forma autônoma a partir de conceitos que envolvem a fusão dos níveis físico e digital (sistemas cyber-físicos), a descentralização dos sistemas de manufatura, a individualização tanto dos sistemas de compra e distribuição quanto do desenvolvimento de produtos e serviços, a adaptação dos novos sistemas às necessidades humanas, e a responsabilidade social corporativa com foco na sustentabilidade e eficiência dos recursos (LASI *et al*., 2014). A indústria 4.0 combina várias tecnologias de sistemas de produção e processos inteligentes para uma transformação das cadeias de valor da indústria, da produção e dos modelos de negócios (ZHONG *et al*. 2017).

O termo indústria 4.0 assume conceitos da integração da automação, troca de dados e tecnologias modernas na fabricação. Entre as expressões comumente associadas à definição de indústria 4.0 pode-se citar a produção inteligente, a integração homem-máquina, manufatura aditiva, operações remotas, Internet das coisas (IoT), computação em nuvem, big data e os sistemas cyber-físicos. Não há uma definição amplamente utilizada e difundida para o termo indústria 4.0, mas o conceito é baseado em seis princípios (KHAN *et al.,* 2017):

- Interoperabilidade: A capacidade de comunicação entre os sistemas cyber-físicos, dispositivos IoT, e entre as fábricas e os seres humanos;
- Virtualização: A habilidade de virtualizar processos físicos através da criação de cyber-gêmeos do mundo físico, para simulação e modelagem de plantas virtuais;

- Descentralização: A capacidade do sistema cyber-físico tomar decisões independentes sem nenhum comando central;
- Capacidade em tempo real: A habilidade de coleta e análise de dados para detectar falhas e encontrar soluções para lidar com o problema com o foco na produção rápida;
- Serviços orientados: A utilização de serviços cyber-físicos no contexto da arquitetura orientada a serviços com a finalidade de facilitar os gerentes nas tomadas de decisões, assim como operadores e clientes;
- Modularidade: A adição de maneira fácil e rápida de novas máquinas, módulos e outros sistemas cyber-físicos sem alterar os módulos já existentes como o foco na atualização das fábricas.

A fabricação inteligente, conceito difundido pela indústria 4.0, utiliza de tecnologias cibernéticas de alto nível com o objetivo de desenvolver processos dinâmicos e flexíveis nos negócios e na engenharia a partir de capacidades de autoconsciência, auto comparação, auto reconfiguração e automanutenção (LEE *et al*., 2014). Resolver questões fundamentais na fabricação como atender aos requisitos do cliente, e a tomada de decisão otimizada e eficiente dos recursos e energia, é um processo chave do conceito de fábrica inteligente (ROMEO *et al*., 2019). Nesse contexto, os sistemas de produção preditiva fornecem as habilidades necessárias para solucionar esses desafios (NIKOLIC *et al*., 2017).

Os seis pilares da manufatura inteligente (Figura 2) são a tecnologia de manufatura e processos, os materiais, os dados, a engenharia preditiva, a sustentabilidade e o compartilhamento de recursos e redes.

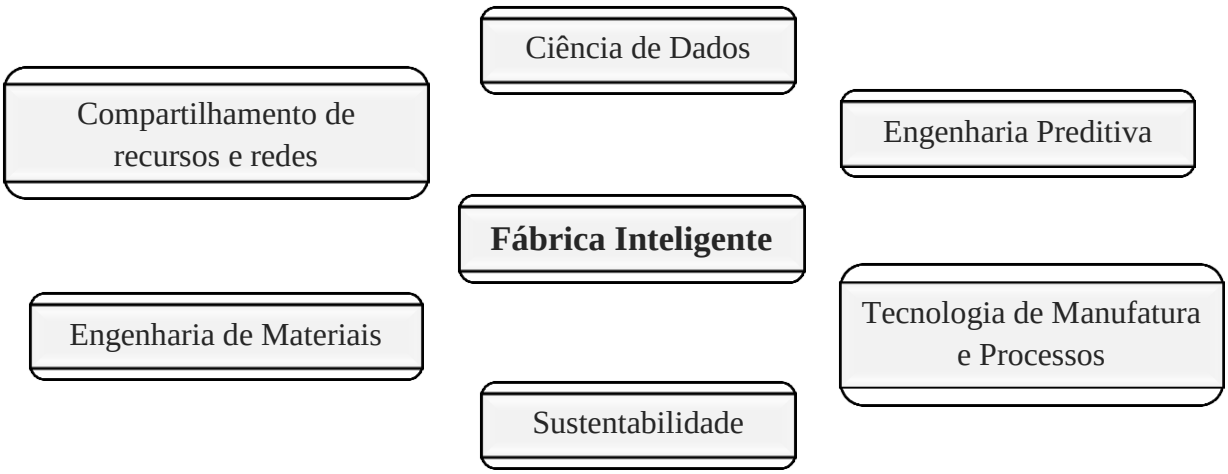

Figura 2- Os pilares da manufatura inteligente. Fonte: Adaptado de Kusiak (2018)

A tecnologia de manufatura e processos na manufatura inteligente refere-se aos avanços nas maneiras de projetar e integrar as operações de fabricação, como, por exemplo, o advento da manufatura aditiva que levou ao desenvolvimento de novos materiais, impactou o design e abriu portas para novas aplicações como a biomanufatura. Uma maior integração dos processos dos novos materiais, do design do produto e dos processos de fabricação são questões fundamentais para a construção da indústria inteligente. Os materiais inteligentes, orgânicos e os biomateriais são elementos promissores que também compõem um pilar para a manufatura inteligente (KUSIAK, 2018).

A influência das tecnologias de informação nos sistemas de manufatura leva a uma coleta crescente de uma grande quantidade de dados (*Big data*). Essa grande quantidade de dados deriva de canais distintos como sensores, áudio, vídeo, web, mídias sociais. A coleta, o processamento, o transporte, a integração, a transformação, o armazenamento, a computação e a extração de conhecimento do big data são desafios da manufatura inteligente no contexto da indústria 4.0 (KHAN *et al.*, 2017).

A engenharia preditiva é um conceito adicionado ao espaço de soluções integrativas da manufatura inteligente em que a empresa provê uma atitude antecipada e não mais reativa do sistema de produção. A engenharia preditiva oferece um novo paradigma para a construção de modelos com alta fidelidade dos fenômenos de interesse com o objetivo de apoiar decisões às futuras condições de mercado e dos processos de fabricação (KUSIAK, 2018).

Além das contribuições ambientais, a partir de alocação de recursos (produtos, materiais, energia e água) de maneira eficiente através de módulos inteligentes de criação de valor, a manufatura inteligente oferece uma oportunidade para realizar uma fabricação sustentável nos níveis econômico, social e ambiental. Oportunidades de fabricação sustentável incluem a adaptação dos equipamentos de fabricação, o aumento da eficiência da formação dos trabalhadores bem como sua motivação, a alocação eficiente de produtos, o design sustentável dos processos, e a reutilização e remanufatura do produto (STOCK e SELIGER, 2016).

Com a digitalização e virtualização da manufatura muitas atividades criativas e de tomadas de decisão ocorrem em ambientes digitais. A manufatura inteligente se beneficia do sucesso dos modelos de aplicativos para realizar o compartilhamento de recursos, redes, equipamentos, software, conhecimento e de espaços colaborativos de modelagem e criatividade (KUSIAK, 2018).

Os sistemas de manufatura preditiva (PMS) revelam ser habilidades importantes para os desafios da produção na criação de sistemas inteligentes aplicados ao conceito da indústria 4.0. Os sistemas devem ter auto previsão, automanutenção e autoconsciência, e para solucionar esse desafio, o PMS combina técnicas de estatística, mineração de dados, modelagem e métodos de inteligência artificial visando converter dados em informações relevantes com a finalidade de realizar previsões (NIKOLIC *et al.*, 2017). Com o objetivo de tecer considerações mais detalhadas sobre os sistemas de manufatura preditiva, um tópico foi desenvolvido visando nortear seus conceitos para o alcance dos objetivos da dissertação, sendo apresentado a seguir.

2.2 SISTEMAS DE MANUFATURA PREDITIVA

Os sistemas inteligentes capazes de prever eventos e estados e sugerir soluções ideias para problemas futuros no ambiente de fabricação são definidos como sistemas de manufatura preditiva (NIKOLIC *et al.*, 2017). Esses sistemas de manufatura dependem tanto dos avanços nos campos da ciência da computação, das tecnologias de informação, da estatística, da modelagem e simulação, quanto do desenvolvimento da ciência e tecnologia de fabricação (MONOSTORI, 2014).

Os sistemas de manufatura preditiva são um campo da ciência com pesquisas emergentes que lida com muitas incertezas que se relacionam com a produtividade, a eficiência, a flexibilidade e a segurança dos processos produtivos. Para tratar essas incertezas é necessário quantifica-las com o objetivo de determinar uma estimativa objetiva para a capacidade de fabricação, o que é definido como transparência. Para tanto, tecnologias para análises preditivas como a mineração de dados, *big data*, a internet das coisas (IoT) e as redes de sensores inteligentes, são necessárias para converter os dados adquiridos em informações relevantes para a solução dos problemas futuros e armazená-los na nuvem, onde podem ser acessados a qualquer momento. A aplicação dessas tecnologias visa reduzir custos, melhorar a eficiência operacional e aumentar a qualidade do produto (LEE, 2013).

Nicolic *et al.* (2017) apresenta um *framework* de tecnologias que suportam o desenvolvimento do sistema de manufatura preditiva e suas interconectividades, Figura 1. O termo KDD refere-se ao processo de extração de conhecimento útil a partir de dados, e a mineração de dados (*Data mining*) é uma etapa específica desse processo em que algoritmos específicos extraem padrões dos dados. A evolução do KDD e da mineração de dados está

apoiada por cruzamentos de pesquisas nas áreas de banco de dados, inteligência artificial, reconhecimento de padrões, estatística e os sistemas inteligentes (FAYYAD *et al.*, 1996), temas potenciais para a aplicação na manufatura preditiva como observado pelo *framework* da Figura 3.

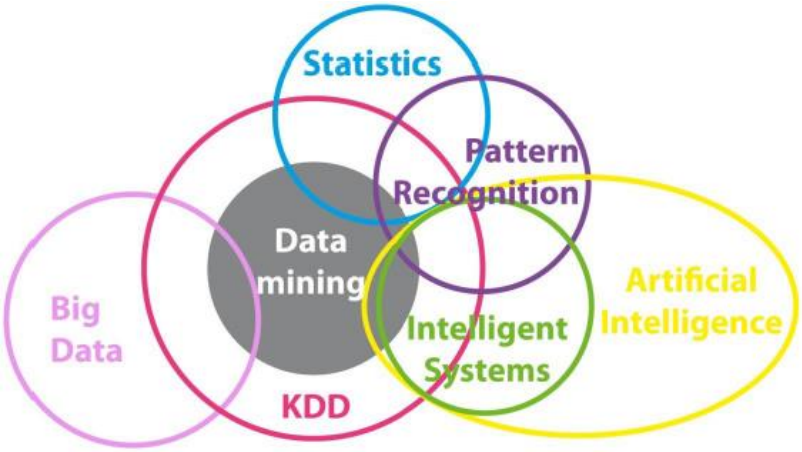

Figura 3- Estrutura de tecnologias aplicadas à manufatura preditiva. Fonte: Nikolic *et al.* (2017)

Um desafio da indústria 4.0 é aplicar os métodos, algoritmos e ferramentas adequadas, diante do escopo de tecnologias apresentadas, com a finalidade de extrair os conhecimentos necessários e transformá-los em informação útil, considerando todas as incertezas invisíveis que envolvem os sistemas de manufatura preditiva (NICOLIC *et al.*, 2017).

Os sistemas inteligentes, que buscam prever estados futuros para a redução de problemas desnecessários, são um pilar para os sistemas preditivos aplicados à indústria 4.0. O aprendizado de máquinas como parte integrante dos sistemas inteligentes é amplamente aplicado em múltiplas áreas da manufatura como otimização, controle e qualidade. O objetivo da técnica de aprendizado de máquinas é a detecção de padrões que descrevem relacionamentos e a estrutura entre os elementos dos dados (WUEST *et al*., 2016).

A técnica de aprendizado de máquina oferece fortes argumentos para soluções primorosas de problemas da fabricação capazes de aprender e adaptar-se às mudanças. Características como a capacidade de lidar com problemas com alta dimensionalidade, a competência em reduzir a natureza complexa dos problemas, a adaptação às mudanças no ambiente com esforço e custo razoáveis, o aprimoramento do conhecimento existente com o aprendizado dos próprios resultados, a habilidade de trabalhar com os dados da fabricação sem a exigência de requisitos especiais e a capacidade de identificar intra e inter-relações relevantes, fazem do

aprendizado de máquina uma técnica completa para a aplicação nos sistemas preditivos aplicados na manufatura (WUEST *et al*., 2016).

A rápida adaptação ao ambiente dos algoritmos de aprendizado de máquina para a aquisição de conhecimento futuro de um sistema apoiam gestores na tomada de decisões nos processos de manufatura com o objetivo de melhorar o desempenho dos processos de fabricação. A detecção de padrões e o reconhecimento de regularidades das relações existentes dos dados são características potenciais do aprendizado de máquina. Entretanto, é importante salientar que não é recomendável basear resultados de decisões de um algoritmo de aprendizado de máquina apenas por comparações. Cada problema possui variáveis diferentes de análise e o desempenho de cada algoritmo depende dos dados disponíveis no banco de dados, das etapas de pré-processamento e da seleção das configurações dos parâmetros (WUEST *et al*., 2016).

A partir das considerações realizadas a respeito dos sistemas preditivos aplicados a manufatura e das potencialidades de seus métodos, o próximo tópico pretende afunilar esses conceitos para aplicar ao tema de pesquisa da dissertação que pondera sobre a predição do tempo de fabricação de componentes de novos produtos. Para tanto, alguns conceitos sobre *lead-time* aplicados à manufatura são apresentados, além de pesquisas com escopos de trabalho direcionados ao tempo de desenvolvimento de produtos. Outro capítulo será destinado para a compreensão detalhada sobre a metodologia de análise preditiva e dos métodos de aprendizagem de máquina.

2.3 MODELOS PREDITIVOS APLICADOS A PREDIÇÃO DO TEMPO DE DESENVOLVIMENTO E FABRICAÇÃO DE PRODUTOS

O termo *lead-time* pode ser definido ou traduzido como o tempo de entrega (MOURTZIS *et al.,* 2014) e influencia de maneira significativa a qualidade, a eficiência do planejamento e a programação da produção (LINGITZ *et al.,* 2018). Para Yang (2009) *lead-time* é o período fixo de tempo para a conclusão de uma ordem de produção.

A estimativa de *lead-time* no chão de fábrica é uma característica crítica e vital, pois afeta intensamente a relação com os clientes. Em um ambiente de grande competitividade entre empresas que necessitam de rápida adaptação às necessidades do mercado, estimativas curtas e exatas dos *lead-times* são ideais, melhoram a imagem dos fabricantes e o potencial para futuras vendas. (ÖZTÜK *et al.,* 2006).

O *lead-time* é basicamente composto pelo tempo de atravessamento ou de fluxo de fabricação, compreendendo o intervalo entre o início de uma atividade produtiva e o seu término (ÖZTÜK *et al.,* 2006). Esse tempo é compreendido essencialmente por duas grandezas: (i) relacionado aos suprimentos que corresponde ao tempo de reposição de materiais, e (ii) relacionado a produção, que corresponde ao atendimento as demandas dos clientes (SELITTO e WALTER, 2008).

O *lead-time* de atravessamento de ordens de fabricação, que corresponde ao tempo que a manufatura gasta desde o aceite comercial da ordem até a disponibilidade final do produto, e posterior transporte para o cliente, compreende os seguintes componentes: (i) emissão da ordem de fabricação, que corresponde a compra e a coleta de materiais ao chão de fábrica; (ii) transporte até a primeira atividade; (iii) espera em fila e até atingir o tamanho do lote; (iv) processamento e *setup* de operações; (v) inspeção e eventuais retrabalhos; (vi) transporte até a próxima atividade; e (vii) recorrência até a última atividade (SELITTO e WALTER, 2008).

Os principais componentes do *lead-time* de fabricação incluem o tempo de fila, de processamento e de transporte, sendo medidas críticas para o desempenho da fabricação. Os prazos de entrega são influenciados por inúmeros fatores como a capacidade, o carregamento, o lote, a programação, custos e controle (MOURTZIS *et al.,* 2014).

Diante de todas as variáveis que envolvem o *lead-time*, a sua estimativa torna-se, portanto, algo extremamente complexo em função de não linearidades e pequenas quantidades de padrões dos dados (MOUSAVI *et al.,* 2013). As incertezas inerentes na estimativa do *lead-time* dependem de vários fatores e caracterizam a natureza multifacetada do problema (ASADZADEH *et al.,* 2011).

Vários métodos foram propostos para a estimativa do *lead-time* como simulação, teoria das filas, curvas operacionais logísticas, estatística, análise estocástica, inteligência artificial, e métodos híbridos (MOURTZIS *et al.,* 2014). De acordo com Mourtzis *et al.* (2014) atualmente um dos métodos mais robustos para a estimativa de *lead-time* são os métodos de inteligência artificial, que incluem a mineração de dados, os sistemas especialistas, as redes neurais artificias, os algoritmos genéticos, a lógica fuzzy, o raciocínio baseado em caso e os modelos híbridos.

Uma das características fundamentais da inteligência artificial inclui a capacidade de modelar não linearidades e complexidades, sendo um método amplamente utilizado para a estimativa do *lead-time* na indústria (ASADZADEH *et al.*, 2011). Com o intuito de identificar

trabalhos que objetivaram estimar o *lead-time* na indústria o Quadro 1 foi desenvolvido, tendo como norte a identificação dos autores e seus métodos de pesquisa.

Quadro 1– Identificação de autores e métodos de pesquisa com foco na predição do lead-time na indústria.

Autor	**Método**
Jun *et al.* (2005)	Modelo analítico
Öztürk *et al.* (2006)	Mineração de dados: Árvore de regressão
Jun *et al.* (2006)	Algoritmo heurístico
Alenezi *et al.* (2008)	Regressão de Vetores de Suporte
Bashir (2008)	Modelo paramétrico
Ruben e Mahmoodi (2010)	Modelo heurístico
De Cos Juez *et al.* (2010)	Máquinas de Vetor de Suporte
Wu (2011)	Máquina de regressão de vetores de suporte
Asadzadeh *et al.* (2011)	Redes Neurais Artificiais
Lin (2011)	Redes Neurais Artificiais
Susanto *et al.* (2012)	Redes Neurais Artificiais
Mousavi *et al.* (2013)	Algoritmo imperialista e regressão de vetores de suporte
Mourtzis *et al.* (2014)	Raciocínio baseado em caso
Mori e Mahalec (2015)	Redes Bayesianas
Pfeiffer *et al.* (2016)	Regressão multivariada
Lingitz *et al.* (2018)	*Random Forests*
Gyulai *et al.* (2018)	*Random Forests*

Em análise do Quadro 1, verifica-se que existe uma grande expressividade de métodos aplicados ao aprendizado de máquina que incluem algoritmos de redes neurais artificiais, os modelos de vetores de suporte, e do *random forests*. Comprova-se, desta maneira, a importância do aprendizado de máquina na modelagem preditiva do *lead-time* na indústria. Para aprofundar o conhecimento dos trabalhos identificados pelo Quadro 1, transcrevem-se pontos chaves das pesquisas realizadas.

Um modelo analítico foi desenvolvido por Jun *et al.* (2005) com o objetivo de estimar o tempo de ciclo de um processo complexo de desenvolvimento de produtos. O modelo captura sete padrões do processo de desenvolvimento de produtos e a estimativa de tempo é realizada para cada padrão e respectivamente para todo o tempo de desenvolvimento do produto.

Öztürk *et al.* (2006) explorou a mineração de dados para estimar o *lead-time* na fabricação sob encomenda. O método de mineração escolhido foi o da árvore de regressão. A partir de simulação de quatro tipos de lojas foi possível adquirir os dados e aplicar métodos de mineração de dados e comparar seus resultados. Os resultados empíricos demonstraram que a mineração superou outros métodos como a regressão linear nesse tipo de problema. A análise

do erro absoluto médio, entre as variáveis preditas e as reais, foi a métrica utilizada para a análise de desempenho do modelo.

No trabalho desenvolvido por Jun *et al.* (2006) um algoritmo heurístico foi desenvolvido com o objetivo de estimar o *lead-time* do processo de desenvolvimento de produtos complexos. Como modelo de avaliação do algoritmo proposto, um estudo de caso foi realizado associado a experimentos computacionais, o que demonstrou a eficiência do método.

Alenezi *et al.* (2008) avalia o tempo do sistema de produção sob encomenda a partir do modelo de regressão de vetores de suporte e compara com modelos de redes neurais artificiais. O modelo realiza a previsão de tempo de fluxo em tempo real em sistemas de múltiplas fontes e produtos. Os resultados mostraram que o modelo de regressão do vetor de suporte apresenta menor erro de previsão além de ter maior robustez.

No trabalho desenvolvido por Bashir (2008) um modelo paramétrico foi desenvolvido a partir de três fatores (complexidade do produto, envolvimento de parceiros e velocidade do gerador) para estimar o tempo necessário para o desenvolvimento de futuros geradores hidrelétricos. Os fatores foram encontrados a partir de uma análise fatorial. A partir de uma avaliação de desempenho dada pela magnitude média do erro relativo entre os valores reais e previstos foi possível encontrar uma boa acurácia para a estimativa do tempo de desenvolvimento do gerador hidrelétrico.

A estimativa do *lead-time* é realizada a partir heurísticas de programação por Ruben e Mahmoodi (2010) com base em dados de simulação. A partir do desvio padrão, como medida de desempenho do método aplicado, foi possível avaliar e verificar a sua eficácia.

De Cos Juez *et al.* (2010) aplica o modelo de máquina de vetor de suporte (SVM) para prever se um lote de componentes metálicos de motores aeroespaciais será finalizado no tempo previsto. A validade do modelo foi verificada a partir da análise de uma amostra com diferentes componentes semelhantes. O modelo SVM demonstrou ter um bom desempenho.

A casa de qualidade *fuzzy* associada à máquina de regressão de vetores de suporte foi a proposta de modelo para a estimativa do tempo de desenvolvimento de produtos aplicada por Wu (2011). O método de estimativa é aplicado e mostra-se eficaz para o caso de design do molde de injeção. A média do erro absoluto entre as variáveis reais e preditas foi utilizada como métrica de desempenho do modelo.

O método de um algoritmo neuro-fuzzy flexível é proposto por Asadzadeh *et al.* (2011) para avaliar o melhor método de aprendizagem de máquina para estimar o *lead-time* de fabricação. Como modelo de validação, o método foi aplicado a uma fabricante de produtos eletrônicos. Os resultados mostraram que o modelo de redes neurais artificiais é superior aos métodos de regressão e regressão nebulosa para os dados aplicados da empresa.

O tempo de ciclo de uma máquina de molde, com dados coletados de uma empresa de embalagens de circuitos integrados, foi avaliado por Lin (2011). O método de redes neurais foi adotado com o objetivo de aplicar ajustes nos parâmetros de entrada da máquina e no tempo de ciclo. Os resultados mostraram a viabilidade do método empregado.

Susanto *et al.* (2012) propõe uma metodologia baseada em redes neurais artificiais para estimar o *lead-time* do produto na produção de tecidos. Os resultados mostraram que as redes neurais foram capazes de estimar com êxito o *lead-time* com um erro percentual absoluto médio abaixo de 8%. O autor comprova que as redes neurais apresentaram uma abordagem estruturada para formular o *lead-time* de produtos na indústria têxtil, e pode auxiliar na tomada de decisões na empresa.

Uma integração dos modelos regressão de vetores de suporte e do algoritmo imperialista competitivo é a proposta de trabalho desenvolvida por Mousavi *et al.* (2013) para estimar o tempo de projetos de desenvolvimento de novos produtos. Para avaliar o desempenho do método um estudo de caso é realizado em um projeto da indústria manufatureira. Os resultados experimentais da pesquisa demonstraram que o modelo apresenta alta acurácia e precisão efetiva para as estimativas de tempos de projetos de desenvolvimento de novos produtos.

Mourtzis *et al.* (2014) avaliou o modelo de raciocínio baseado em caso para predizer o tempo de fabricação sob encomenda de produtos complexos. O modelo utiliza do método de avaliação da similaridade entre casos passados e novos, a partir do cálculo da distância euclidiana, para estimar o *lead-time* de fabricação. Os resultados preliminares demonstraram que houve uma redução significativa no número de iterações entre o departamento de engenharia e os clientes comprovando uma melhor precisão do *lead-time* de fabricação de moldes da empresa.

No trabalho desenvolvido por Mori e Mahalec (2015) modelos de redes bayesianas são empregados com o intuito de prever as distribuições de probabilidades para tempos de produção da indústria siderúrgica. Uma comparação entre as redes bayesianas, os vetores de

suporte (SVM) e as redes neurais artificias, é realizada e concluiu-se que todos métodos são capazes de prever com precisão as variáveis de saída ou as distribuições de probabilidade.

A partir do método de regressão multivariada Pfeiffer *et al*. (2016) realizou a predição do tempo de fabricação aplicado em dados simulados de um sistema de programação da produção *flow-shop* de pequeno tamanho. O erro quadrático médio do modelo previsto foi utilizado como métrica de desempenho do método aplicado.

Lingitz *et al.* (2018) avalia o tempo de fabricação na indústria de semicondutores buscando confrontar métodos de aprendizado de máquina com o intuito de encontrar aqueles de melhor acurácia. Alguns métodos populares de aprendizado de máquina foram utilizados como as redes neurais artificiais, algoritmo *support vector machine* (SVM), e o dos *k-nearest neighbor* (kNN). A partir da média do erro absoluto das variáveis preditoras foi possível identificar que o algoritmo *Random Forest* apresentou melhores resultados para o modelo proposto.

Gyulai *et al.* (2018) verificou a acurácia de métodos de aprendizado de máquina para prever o *lead-time* de fabricação em um ambiente *flow-shop*. Os métodos avaliados incluem a regressão linear, a árvore de regressão, *Random Forests*, e a regressão do vetor de suporte. O erro quadrático médio da raiz normalizada verificou que o melhor resultado foi do algoritmo *Random Forests*.

Identificado o problema de pesquisa, que engloba a predição do tempo de fabricação de componentes pela indústria, a contextualização do problema, que está inserido nas metodologias aplicadas a indústria 4.0, e a identificação dos métodos potenciais aplicados para a resolução do problema; o próximo capítulo tem como objetivo aprofundar os conhecimentos técnicos da análise preditiva e dos métodos de aprendizagem de máquinas. Esse estudo apresentará os procedimentos cruciais da análise preditiva, os algoritmos aplicados a essa dissertação, as métricas de desempenho dos modelos propostos e as ferramentas utilizadas para a aplicação dos métodos.

3. ANÁLISE PREDITIVA

Este capítulo apresenta os métodos aplicados à análise preditiva. O aprendizado de máquinas aplicado à análise preditiva é apresentado assim como seus procedimentos técnicos. A etapa de pré-processamento e as técnicas de aprendizado de máquina que incluem o algoritmo de redes neurais artificiais (RNA), *Support Vector Machine* (SVM) e *Random Forest* (RF) são detalhados. As métricas de desempenho dos modelos expostos assim como suas ferramentas de aplicação, são apresentadas como instrumentos efetivos de avaliação e execução.

3.1 APRENDIZADO DE MÁQUINA E ANÁLISE PREDITIVA

O aprendizado de máquina juntamente com as tecnologias de *Big Data* e computação de alto desempenho foram implementadas com o objetivo de desenvolver novas possibilidades para desvendar, quantificar, e entender uma grande quantidade de dados estruturados e não estruturados advindos na era da tecnologia moderna. O aprendizado de máquinas foi uma evolução de um subcampo da inteligência artificial com a finalidade de desenvolver algoritmos de auto aprendizado para obter conhecimento a partir de dados e fazer previsões (NILSSON, 1998; LIAKOS, 2018).

Tendo em vista os conceitos de Fayyad *et al.* (1996) que define o conhecimento de dados como um processo, não trivial, de extração de informações previamente desconhecidas e potencialmente úteis a partir de dados armazenados em um banco de dados, o aprendizado de máquina tem como alvo capturar esse conhecimento com o intuito de melhorar o desempenho de modelos preditivos e tomar decisões orientadas pelos dados (RASCHKA e MIRJALILI, 2017).

As metodologias que envolvem a aprendizagem de máquina buscam a aquisição de conhecimento com o objetivo de aprender com a experiência para a execução de uma tarefa, e para isso, compartilha de um campo interdisciplinar com a estatística, a teoria da informação, a teoria de jogos e da otimização (SHALEV e BEN, 2014). Para capturar e utilizar o conhecimento adquirido, a aprendizagem de máquina utiliza de tarefas que são classificadas em duas categorias principais definidas como aprendizagem supervisionada e não supervisionada (JAMES *et al.*, 2013).

No aprendizado supervisionado os dados são apresentados como entradas e saídas que se correspondem, visando construir uma regra ou modelo que possa mapear as entradas em saídas. O principal objetivo do aprendizado supervisionado é aprender um modelo a partir dos dados de treinamento que permita realizar previsões sobre dados futuros. Nesse tipo de aprendizado o conjunto de amostras de dados de saída desejados já é conhecido (LIAKOS *et al.*, 2018). Desta maneira, no aprendizado supervisionado, para cada observação *i, i=1,2,3,...,n* existe um vetor de medida preditora *Xi,* com uma medida de resposta associada *Yi* (JAMES *et al.*, 2013).

No aprendizado não supervisionado decorre uma situação um pouco mais desafiadora uma vez que não há distinção entre os conjuntos de dados de treinamento e de teste, seu objetivo principal é processar os dados de entrada e descobrir seus padrões ocultos (Liakos, 2018). Neste cenário a máquina trabalha de certo modo às cegas uma vez que as variáveis de resposta, dados de saída, são desconhecidas o que não permite supervisionar a análise. Desta maneira, para cada observação *i, i=1,2,3,...,n* observa-se um valor de medida *Xi,* mas sem nenhuma resposta associada *Yi* (JAMES *et al.*, 2013).

A aplicação da aprendizagem de máquina pressupõe o conhecimento do tipo das variáveis no banco de dados em valores dos preditores e das classes (saídas). As entradas e saídas podem ser definidas como variáveis quantitativas ou qualitativas. Alguns métodos são definidos de forma intensiva às variáveis quantitativas, outros para as qualitativas, e outros para ambos. Para o caso de aprendizado supervisionado o método de regressão é comum a variáveis quantitativas, e o de classificação para as variáveis categóricas ou qualitativas (HASTIE *et al*., 2009).

Um algoritmo desenvolvido para uso de aprendizado de máquina tem um fluxo típico de construção para a modelagem preditiva, Figura 4. O roteiro aplicado dispõe de procedimentos cruciais e indispensáveis para um desempenho ideal de um algoritmo de aprendizagem que são as etapas de pré-processamento, o aprendizado, a predição e a avaliação do modelo aplicado.

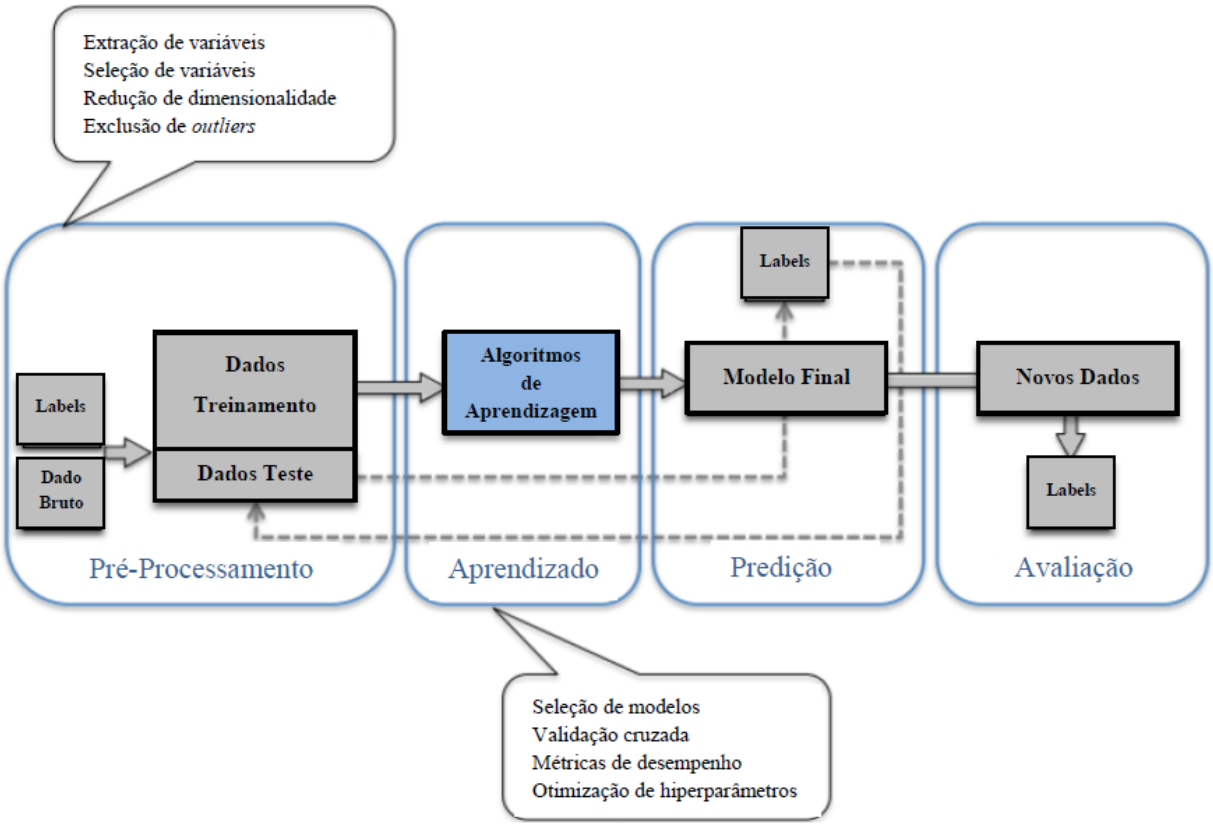

Figura 4– Diagrama de aplicação do aprendizado de máquina para análise preditiva. Fonte: Adaptado de Raschka e Mirjalili (2017)

A primeira etapa, o pré-processamento é uma das etapas mais cruciais no processo de implementação do algoritmo de aprendizagem de máquina, uma vez que os dados brutos raramente são adquiridos na forma correta para um desempenho ideal. Nesta etapa, é realizada a divisão aleatória do conjunto de dados em treinamento e teste. As divisões comumente utilizadas são 60:40, 70:30 ou 80:20 dependendo do tamanho da base de dados inicial. Entretanto, outras proporções de dados de treinamento e teste podem ser utilizadas em *datasets* maiores como 90:10 ou 99:1 (RASCHKA e MIRJALILI, 2017).

A aprendizagem é apresentada como a segunda etapa do roteiro de aplicação, sendo de grande relevância e que dá nome ao processo. Nessa etapa os algoritmos de aprendizado de máquina são aplicados, juntamente com suas métricas de desempenho e otimização. Posteriormente, desenvolvem-se as duas etapas finais definidas como avaliação e predição, em que o modelo encontrado a partir dos dados de treinamento pode ser avaliado, enquanto a sua performance, e aplicados nos dados de teste para avaliação de sua performance. Uma métrica comumente utilizada, como exemplo, para avaliação dos algoritmos de classificação é

a acurácia, que define as proporções de classificação de instâncias corretas (RASCHKA e MIRJALILI, 2017).

Mais detalhes sobre cada etapa do processo de aplicação do aprendizado de máquinas serão explorados ao longo do texto.

É importante destacar que, além do processo aplicado pelo diagrama da Figura 4, é imprescindível para o sucesso da implementação de um modelo preditivo o conhecimento especializado das informações contidas no banco de dados por um especialista, e o conhecimento do contexto do problema. Esse conhecimento especializado deve ser aplicado com o intuito de obter os dados verdadeiramente relevantes para os objetivos da pesquisa, identificando as informações não importantes, eliminando ruídos prejudiciais e aumentando o desempenho do modelo preditivo (KUHN e JOHNSON, 2013).

O próximo tópico apresentado tece maiores profundidades metodológicas acerca do pré-processamento, uma etapa de grande relevância para o sucesso do algoritmo de aprendizagem. Neste tópico, serão abordadas práticas importantes para realizar um pré-processamento seguro além de problemas comumente encontrados na aplicação da modelagem preditiva como o *overffiting*, uma das grandes causas de falhas e ineficiências dos algoritmos de predição.

3.2 PRÉ-PROCESSAMENTO

As técnicas de processamento de dados de maneira geral referem-se à adição, exclusão ou transformação dos dados dispostos no conjunto de treinamento. Essas transformações auxiliam a reduzir o impacto das distorções ou das discrepâncias no conjunto de dados, que podem levar a melhorias significativas no desempenho do modelo preditivo. Uma característica fundamental que deve ser considerada para a construção do algoritmo de predição é a relação entre o número de amostras e o número de preditores, sendo ideal que um conjunto de amostras seja potencialmente maior que o número de preditores (saídas). Outra característica essencialmente importante é a consideração do tipo de variável de entrada e de saída, se categórica (nominal ou ordinal) ou numérica (contínua ou discreta) (KUHN e JOHNSON, 2013).

3.2.1 TRANSFORMAÇÃO DOS DADOS

A transformação de dados revela ter uma relação positiva para a melhora do desempenho de muitos algoritmos preditores. A normalização e a padronização são abordagens

comumente utilizadas para a modificação da escala de uma variável. A primeira, refere-se a um redimensionamento para uma escala que varia de 0 a 1, e a segunda a centralização das variáveis em uma média 0 com desvio padrão 1. A transformação de dados também é útil na correção de distribuições assimétricas. Identifica-se, primeiro, as variáveis que apresentam a característica de assimetria a partir da razão entre o maior e menor valor da variável preditora, resultados acima de 20 podem ser indicativos de distribuição assimétrica. Algumas transformações logarítmicas, de raízes quadradas ou valor inverso podem contribuir com a correção dessa assimetria (KUHN e JOHNSON, 2013).

A identificação dos conjuntos de dados preditores (entradas) com alto índice de correlação, e que, portanto, apresentam a mesma informação, pode colaborar para que o processo de aprendizagem não resulte em modelos instáveis. Uma abordagem utilizada para lidar com esse conjunto de variáveis de entrada que apresentam informações correlacionadas consiste em remover um dos preditores com alta colinearidade. Existem vantagens potenciais na remoção de preditores antes da modelagem, como uma redução do tempo e da complexidade computacional, a diminuição do risco de distribuições degeneradas, além de facilitar a construção de um modelo mais parcimonioso e interpretável (KUHN e JOHNSON, 2013; RASCHKA e MIRJALILI, 2017).

Uma abordagem mais heurística para lidar com esse problema de colinearidade de variáveis preditoras é remover um número mínimo de preditores para garantir que todas as correlações aos pares estejam abaixo de um certo limite, o que pode ter um efeito significativo no desempenho do modelo. O algoritmo consiste em calcular a matriz de correlação dos preditores, determinar os dois preditores associados com à maior correlação absoluta em pares (A e B), determinar a correlação média entre os preditores selecionados (A e B) e as demais variáveis da base, se o preditor ‘A’ tiver uma correlação média maior, deve-se remove-lo, caso contrário remova o preditor ‘B’ (KUHN e JOHNSON, 2013).

Outra classe de transformação de preditores são as técnicas de redução de dados. Esses métodos visam reduzir o tamanho do conjunto de dados gerando um número menor de preditores com o objetivo de capturar a maioria das informações das variáveis originais. Assim, menos variáveis são utilizadas e ao mesmo tempo o modelo fornece uma fidelidade razoável aos dados originais. Para a grande maioria das técnicas de redução de dados os novos preditores são funções dos preditores originais, portanto, todos os preditores originais são necessários para a criação das novas variáveis substitutas. A essa classe de métodos dá-se o nome de extração de sinal ou técnicas de extração de recursos (KUHN e JOHNSON, 2013).

3.2.2 EXCLUSÃO DE OUTLIERS

Define-se *outliers* como amostras que são excepcionalmente distantes dos valores da maioria representativa do conjunto de dados. Quando uma ou mais amostras são suspeitas de serem discrepantes, a primeira etapa é garantir que os valores sejam cientificamente válidos e que não incorra em erros de registro de dados. Uma atenção deve ser dada para não remover ou alterar dados de forma apressada, principalmente quando o tamanho da amostra for pequeno. Após a análise, a exclusão desses preditores no conjunto de dados de treinamento pode resultar em melhora da performance do modelo. Ressalta-se que essa decisão de exclusão dos *outliers* dever ser realizada antes da aplicação de técnicas para a redução de dimensionalidade (KUHN e JOHNSON, 2013).

3.2.3 TRATAMENTO DE DADOS FALTANTES

Os dados faltantes aparecem frequentemente relacionado às variáveis preditoras da amostra. Esses dados faltantes aparecem de forma corriqueira em amostras de dados, sendo estruturalmente ausentes ou não determinados no momento da construção dos dados. Primeiro, é importante entender o motivo, o padrão e a frequência dos valores faltantes, para que esse dado seja adequadamente considerado na etapa de análise. Deve-se considerar se o valor ausente representa uma lacuna informativa que pode apresentar algum significado ao modelo. Em alguns casos, a porcentagem de dados ausentes é substancialmente suficiente para remover esse preditores das atividades de modelagem, em conjuntos de dados grandes e assumindo que a falta não seja informativa. Em outros casos, há um preço alto na remoção de amostras sendo necessária abordar alternativas diferentes da remoção dos dados (KUHN e JOHNSON, 2013).

Alternativamente as técnicas de remoção, técnicas de interpolação podem ser utilizadas, como a imputação do dado faltante. Nesse caso, podem-se usar informações nos preditores do conjunto de treinamento para estimar os valores dos outros preditores, a partir da média, mediana ou valor mais frequente de um preditor ou técnicas de imputação múltipla. Isso equivale a um modelo preditivo dentro de outro modelo preditivo. Utiliza-se como exemplo modelos de regressão e do algotitmo *k-Nearest Neighbors* (KNN). Torna-se importante observar, entretanto, que essa camada de modelos extras adiciona incertezas, sendo essa imputação incorporada à reamostragem, o que aumenta o tempo computacional para a construção dos modelos, mas pode fornecer estimativas honestas de desempenho (RASCHKA e MIRJALILI, 2017).

3.2.4 ADIÇÃO DE PREDITORES

Quanto um preditor é do tipo categórico é comum decompor o preditor em um conjunto de variáveis mais específicas. Os dados categóricos são recodificados em *bits* de menor informação chamados de variáveis *dummy*. Cada categoria obtém sua própria variável *dummy* a partir de um indicador ‘0’ ou ‘1’. No entanto, a decisão de incluir todas as variáveis fictícias dependerá da escolha do modelo de aprendizagem, nos modelos como a regressão linear simples, por exemplo, o uso das variáveis *dummy* pode incluir problemas numéricos. Se o modelo de aprendizagem não for sensível ao uso das variáveis dummy, seu uso ajuda a melhorar a sua interpretação (KUHN e JOHNSON, 2013).

3.2.5 OVERFITTING

O problema de *overffing* ocorre quando um classificador de um modelo construído se adapta de forma precisa aos dados de treinamento, porém não pode ser utilizado ou generalizado para os dados de teste ou novos dados. Diz-se que o modelo, então, está muito ajustado e pode codificar apenas peculiaridades aleatórias ou detalhes específicos da base de treinamento, o que pode resultar em uma diminuição da taxa de acerto. Para evitar o *overfitting* é importante que o classificador tenha a maior simplicidade possível. Sendo assim, será dada uma maior importância às regularidades presentes nos dados e as de menor importância são penalizadas (DOMINGOS, 2012).

Para evitar o *overffiting* é importante ajustar corretamente os parâmetros do modelo e avaliar o seu desempenho. O método de validação cruzada auxilia a combater o problema de sobreajuste. Os dados são divididos em vários conjuntos de treinamento e teste, os quais demonstram frequentemente encontrar parâmetros de ajuste mais ideais e fornecem uma representação mais precisa do desempenho preditivo do modelo. Além da validação cruzada existem outros métodos para o controle do *overffiting*, como adicionar um termo de regularização à função de avaliação para penalizar classificadores ou realizar testes de significância estatística como o qui-quadrado (KUHN e JOHNSON, 2013; DOMINGOS, 2012).

O método de validação cruzada além de auxiliar a evitar os problemas com o sobreajuste do modelo, também é uma das principais técnicas de reamostragem, utilizada quando o pesquisador não dispõe de dados suficientes para dividir os dados em treinamento, teste e

validação. Aproxima-se ao conjunto de validação através da reutilização das observações do conjunto de treinamento original (HASTIE *et al*., 2009).

3.2.6 VALIDAÇÃO CRUZADA

A validação cruzada é uma clássica abordagem e um dos métodos de reamostragem mais populares para estimar o desempenho da generalização de modelos de aprendizagem de máquina. No método de validação cruzada o conjunto de dados inicial é dividido em um conjunto de treinamento e outro para teste, sendo o primeiro utilizado para o treinamento do modelo e o segundo para estimar seu desempenho. No entanto, se ocorrer a reutilização do mesmo conjunto de dados de teste repetidamente durante a seleção do modelo, ele se tornará parte dos dados de treinamento e, portanto, será provável que o modelo sofra de sobreajuste (*overfitting*) (KUHN e JOHNSON, 2013; RASCHKA e MIRJALILI, 2017).

Para solucionar esse problema de construção de modelos muito ajustados, a técnica de validação cruzada *k-fold*, Figura 5, é uma das mais utilizadas.

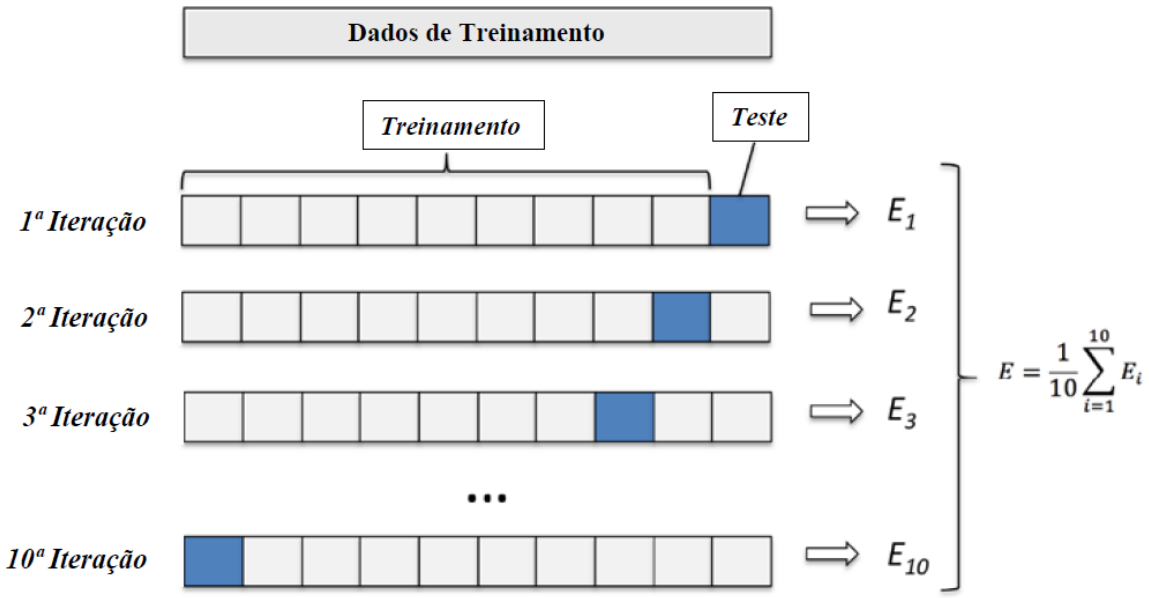

Figura 5 – Método de validação cruzada k-fold. Fonte: Adaptado de Raschka e Mirjalili (2017)

Utiliza-se essa técnica de validação para o ajuste do modelo, encontrando os melhores valores de hiperparâmetro que produz um desempenho de generalização satisfatória. Nessa técnica, divide-se aleatoriamente o conjunto de dados em *k* partes de tamanhos aproximadamente iguais, em que *k-1* serão utilizados para os dados de treinamento para

adequação do modelo preditivo e o restante dos dados é reservado para a estimativa do seu desempenho. Esse procedimento é repedido *k* vezes até que todas as partes participem tanto do treinamento como da validação. Realiza-se o cálculo do desempenho médio do modelo com base nas diferentes e independentes etapas com o objetivo de obter uma estimativa menos sensível de desempenho (KUHN e JOHNSON, 2013; RASCHKA e MIRJALILI, 2017).

O valor padrão para a repetição *k* na validação cruzada *k-fold* é 10, sendo uma escolha razoável para a maioria das aplicações. Entretanto, se o conjunto de dados for relativamente pequeno pode ser útil aumentar o tamanho das iterações. Com o aumento do valor de *k* mais dados de treinamento são utilizados em cada iteração, o que pode resultar em uma melhor performance do modelo. No entanto, grandes valores para *k* aumentam o tempo de execução do algoritmo e produz estimativas com maiores variações. Por outro lado, se o conjunto de dados for grande, podem-se utilizar menores valores para k e mesmo assim obter uma estimativa precisa do desempenho médio do modelo, reduzindo o custo computacional (RASCHKA e MIRJALILI, 2017).

3.3 TÉCNICAS DE APRENDIZAGEM

3.3.1 REDES NEURAIS ARTIFICIAIS (RNA)

Redes Neurais Artificiais (RNAs) são um processador paralelo e distribuído composto por unidades de processamento simples capazes de armazenar conhecimento, a partir de um processo de aprendizado, e disponibiliza-lo para realizar tarefas diversas como reconhecer padrões, tomar decisões e fazer previsões em ambiente dinâmico. O algoritmo de aprendizado é o procedimento usado para realizar o processo de aprendizagem cuja função é modificar os pesos sinápticos da rede de maneira ordenada para atingir um objetivo desejado. As RNAs oferecem propriedades e capacidades úteis como a não linearidade, o mapeamento de entrada em saídas, a adaptabilidade, a tolerância ao erro e a uniformidade da análise e design, sendo um grande atrativo para a resolução de problemas (HAYKIN, 2008; TAYLOR e SMITH, 2006).

Um neurônio é uma unidade de processamento de informação fundamental para a operação de uma RNA. A Figura 6 apresenta o modelo comumente utilizado para a representação de um neurônio artificial. O neurônio recebe um conjunto de estímulos de entrada, realiza a transformação destes em sinais e gera um estímulo de saída. Três elementos

são básicos para a construção do modelo de um neurônio artificial: o conjunto de sinapses, o somador e a função de ativação.

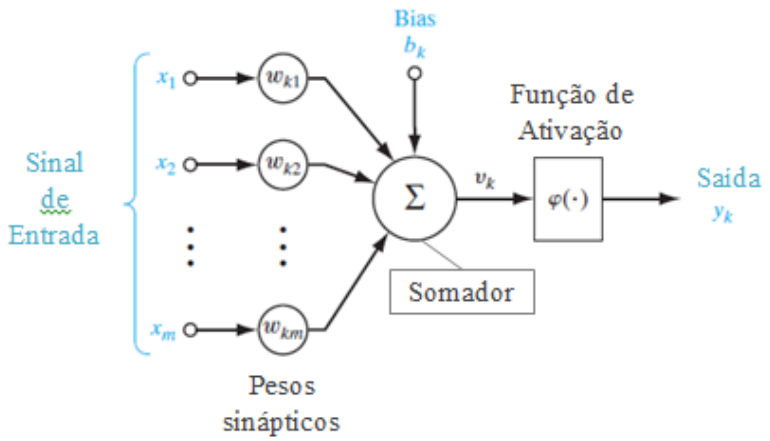

Figura 6- Modelo de neurônio artificial. Fonte: Adaptado de Haykin (2008)

Os sinais chegam e partem de cada neurônio por meio do conjunto de sinapses, que são as conexões. Cada sinapse apresenta um peso associado que corresponde à informação efetivamente armazenada no neurônio e na rede. O elemento somador tem por função somar os sinais de entrada, ponderando as respectivas forças sinápticas do neurônio a partir de uma combinação linear. A função de ativação ou função de esmagamento delimita a amplitude de saída do neurônio para um valor finito (HAYKIN, 2008).

O Multilayer Perceptron (MLP) ou percéptron de múltiplas camadas é uma das topologias amplamente implementadas das redes neurais, sendo capaz de aproximar funções arbitrárias em importantes problemas de estudo em dinâmica não linear e mapeamento de funções. Duas importantes características do MLP são a não linearidade dos seus elementos de processamento, e a massiva interconectividade (PRINCIPE *et al.*, 2000).

Uma rede MLP, Figura 7, é composta basicamente por três elementos: a camada de entrada, que transmite os sinais de entrada para a camada intermediária ou camada oculta; a camada intermediária, que projeta os sinais de entrada a partir de um tratamento não linear; e a camada de saída, que recebe as ativações dos neurônios das camadas intermediárias e produz as saídas da rede neural.

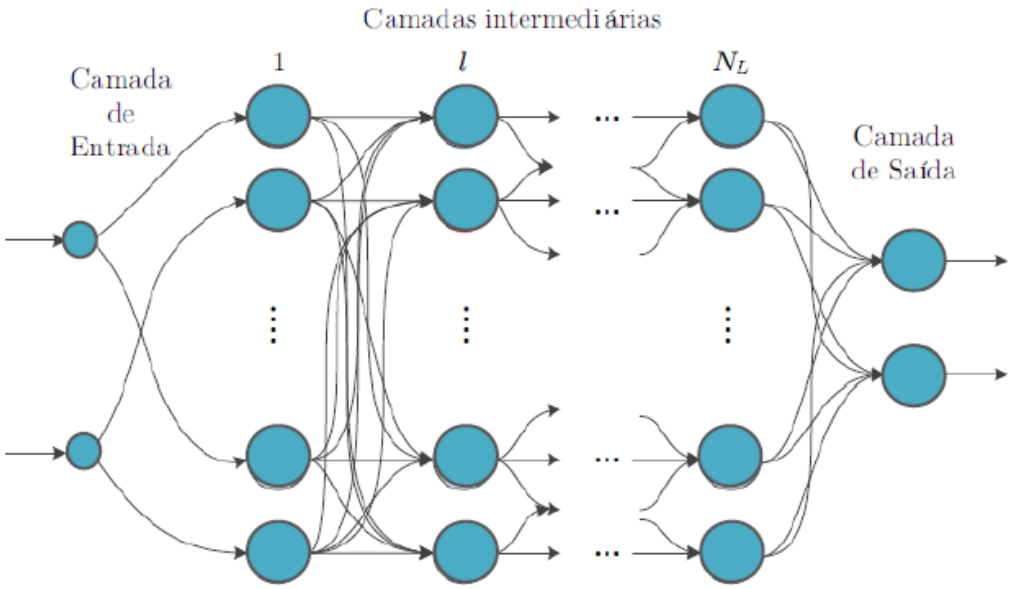

Figura 7– Arquitetura da rede MLP. Fonte: Adaptado de Boccato (2013).

Cada neurônio da camada intermediária ou da camada de saída em um MLP é projetado para executar basicamente dois cálculos. O primeiro é o cálculo do sinal de função que aparece na saída de cada neurônio, sendo expresso como uma função não linear contínua do sinal de entrada junto aos pesos sinápticos associados a esse neurônio. O segundo cálculo é o da estimativa do vetor gradiente, necessária para a passagem reversa pela rede (HAYKIN, 2008).

De acordo com Cybenko (1989) uma camada intermediária é suficiente para aproximar qualquer função contínua, e duas camadas intermediárias são suficientes para aproximar qualquer função matemática em uma MLP. A utilização de um grande número de camadas não é recomendada uma vez que o erro medido durante o treinamento é propagado para as camadas anteriores.

O número de neurônios da camada intermediária em geral é definido empiricamente dependendo de vários fatores como o número do conjunto de dados de treinamento, a quantidade de ruídos, a complexidade da função a ser aprendida e a distribuição estatística dos dados de treinamento. Alguns métodos propostos para auxiliar na escolha do número de neurônios da camada intermediária são: definir o número de neurônios em função do número de entradas e saídas; ou utilizar um número de conexões dez vezes menor que o número de exemplos (BRAGA *et al.*, 2007).

A rede MLP é normalmente treinada pelo algoritmo *backpropagation*. Esse algoritmo supervisionado tem como princípio a propagação dos erros pela rede permitindo a adaptação

dos elementos ocultos. Os vetores de entrada são propagados de camada a camada, até a camada de saída, onde é então comparado ao vetor desejado e encontrado o erro. O erro encontrado é retropropagado pela rede, com a finalidade de ajustar os pesos sinápticos e reduzir o erro de saída até um valor aceitável, em que ocorre a condição de parada do treinamento (PRINCIPE *et al.*, 2000). O modelo da Figura 8 é utilizado para descrever o algoritmo de *backpropagation*.

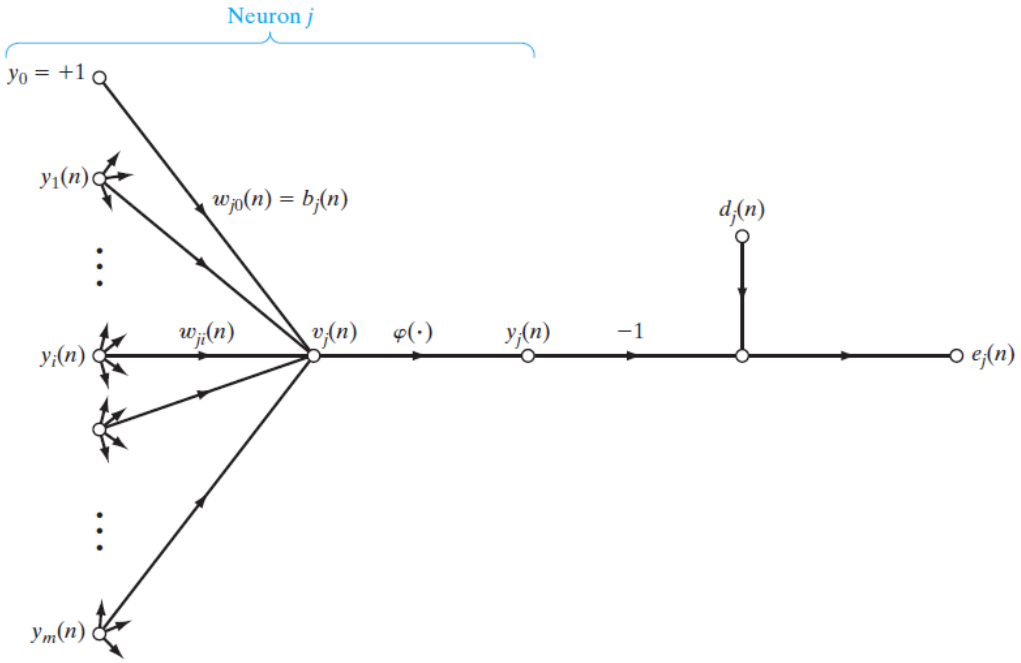

Figura 8– Modelo de treinamento do algoritmo de backpropagation. Fonte: Haykin (2008)

O modelo e seu equacionamento são apresentados a partir da proposta exposta por Haykin (2008). O valor de $v_j(n)$ produzido na entrada da função de ativação (φ) associado ao neurônio j, aos pesos sinápticos $w_{ji}(n)$ e as entradas $y_i(n)$ é definido por:

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \tag{1}$$

onde m é o número total de entradas aplicadas ao neurônio j. A função sinal $y_i(n)$ que aparece na saída do neurônio j com n iterações é:

$$y_i(n) = \varphi_j(v_j(n)) \tag{2}$$

O algoritmo de retropropagação aplica a correção $\Delta w_{ji}(n)$ ao peso sináptico, que corresponde ao cálculo da derivada parcial. De acordo com a regra da cadeia, tem-se que:

$$\frac{\partial \mathcal{E}(n)}{\partial w_{ji}(n)} = \frac{\partial \mathcal{E}(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)} \tag{3}$$

A derivada parcial apresentada representa um fator de sensibilidade, determinando a direção do modelo no espaço de pesos sinápticos w_{ji}. Diferenciando os dois lados da equação em relação ao termo $e_j(n)$ tem-se:

$$\frac{\partial \mathcal{E}(n)}{\partial e_j(n)} = e_j(n) \tag{4}$$

Diferenciando a equação para o erro,$e_j(n) = d_j(n) - y_j(n)$ em ambos os lados pelo respectivo fator $y_j(n)$:

$$\frac{\partial e_j(n)}{\partial y_j(n)} = -1 \tag{5}$$

Diferenciando a Equação 2 em relação a $y_j(n)$:

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi_j{}'(v_j(n)) \tag{6}$$

Diferenciando a Equação 1, em relação a $w_{ji}(n)y_i$:

$$\frac{\partial v_j(n)}{\partial w_{ji}(n)} = y_i(n) \tag{7}$$

Utilizando o resultado das Equações 4, 5, 6 e 7 na Equação 3, tem-se:

$$\frac{\partial \mathcal{E}(n)}{\partial w_{ji}(n)} = -e_j(n)\varphi_j{}'(v_j(n))y_i(n) \tag{8}$$

A correção $\Delta w_{ji}(n)$ aplicada a $w_{ji}(n)$ é definida pela regra de delta:

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial w_{ji}(n)} \tag{9}$$

onde o parâmetro η corresponde a taxa de aprendizagem do algoritmo *backpropagation*. O sinal negativo (-) corresponde a descida do gradiente no espaço de pesos. Por consequência das Equações 8 e 9, tem-se a correção dos pesos:

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_i(n) \tag{10}$$

onde o gradiente local δ_jé definido por:

$$\delta_j(n) = e_j(n) \varphi_j{}'(v_j(n)) \tag{11}$$

O gradiente local aponta para as mudanças necessárias nos pesos sinápticos. De acordo com a Equação 11, o gradiente local para o neurônio de saída j é igual ao produto do erro correspondente para o neurônio $e_j(n)$ e a sua derivada da função de ativação associada $\varphi_j{}'(v_j(n))$.

Um aspecto relevante relacionado ao projeto de redes multicamadas é o tipo de função de ativação utilizada. O cálculo do gradiente para cada neurônio requer o conhecimento da derivada da função de ativação, e para que a derivada exista é imprescindível que a função seja contínua, sendo esse um requisito que a função de ativação deve atender (HAYKIN, 2008). Inúmeras funções de ativação são aplicadas para funções não lineares e diferenciáveis, mas uma das mais utilizadas é a função sigmoide, definida pela Equação (12) em que $a > 0$.

$$\varphi_j\left(v_j(n)\right) = \frac{1}{1 + e^{(-a v_j(n))}} \tag{12}$$

Outro importante parâmetro de avaliação para o algoritmo de *backpropagation* é a taxa de aprendizagem do modelo, que avalia a medida de rapidez com que o vetor de pesos será atualizado. A taxa de aprendizagem varia entre o intervalo [0,1] e quanto menor o parâmetro

da taxa de aprendizado, menores serão as alterações nos pesos sinápticos (BRAGA *et al.*, 2007).

A melhoria da taxa de aprendizagem, a partir da sua suavização na trajetória dos espaços de pesos, é alcançada ao custo de uma taxa mais lenta de aprendizado. Se a taxa de aprendizagem, por outro lado, for muito grande para acelerar o processo de aprendizagem, as grandes alterações nos pesos sinápticos podem tornar a rede instável. Um método simples e amplamente utilizado para aumentar a taxa de aprendizado e evitar a instabilidade da rede é modificar a regra delta incluindo a variável *momentum* (HAYKIN, 2008).

3.3.2 SUPPORT VECTOR MACHINE (SVM)

O algoritmo de aprendizagem *Support Vector Machine* (SVM) ou máquina de vetores de suporte são uma classe de técnicas de modelagem com alto poder e flexibilidade. O SVM representa uma generalização do algoritmo classificador de margem máxima, que por sua vez acomoda a fronteiras não lineares em problemas de classificação. O objetivo do algoritmo SVM é maximizar a margem, definida como a distância entre o hiperplano de separação e as amostras de treinamento mais próximas desse hiperplano, que são os vetores de suporte (RASCHKA e MIRJALILI, 2017; JAMES *et al.*, 2013).

Na Figura 9 é possível visualizar um classificador SVM de margem máxima para o caso separável quando há apenas dois preditores (X_1 e X_2). A reta contínua central representa o hiperplano de margem máxima e a margem é representada pelas distâncias entre o hiperplano e as linhas pontilhadas, definidas como vetores de suporte.

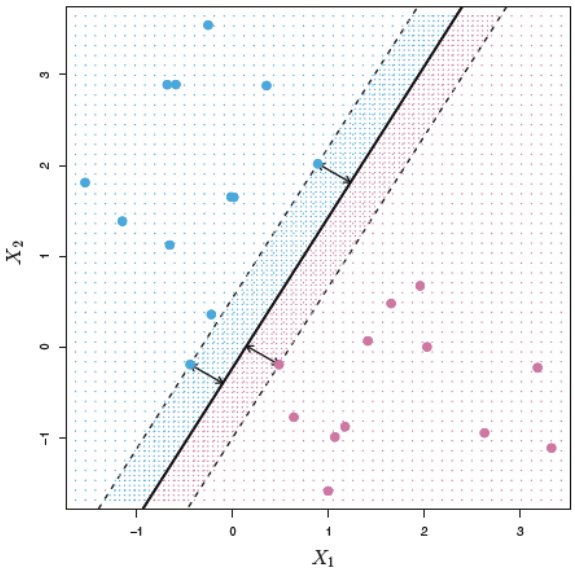

Figura 9- Modelo linear para o classificador SVM. Fonte: James et al. (2013)

O classificador de margem flexível do vetor de suporte, Figura 10, busca considerar que não ocorra uma separação perfeita entre duas classes com o objetivo de maior robustez às ações individuais e a melhor classificação para a maioria das observações do treinamento. Ao invés de buscar a maior margem possível, de modo que todas as classificações estejam corretas, permite-se que algumas estejam do lado incorreto do hiperplano. Observações que se encontram diretamente na margem ou no lado errado da margem de sua classe são denominadas como vetores de suporte, e afetam o classificador de vetores de suporte (JAMES *et al.*, 2013).

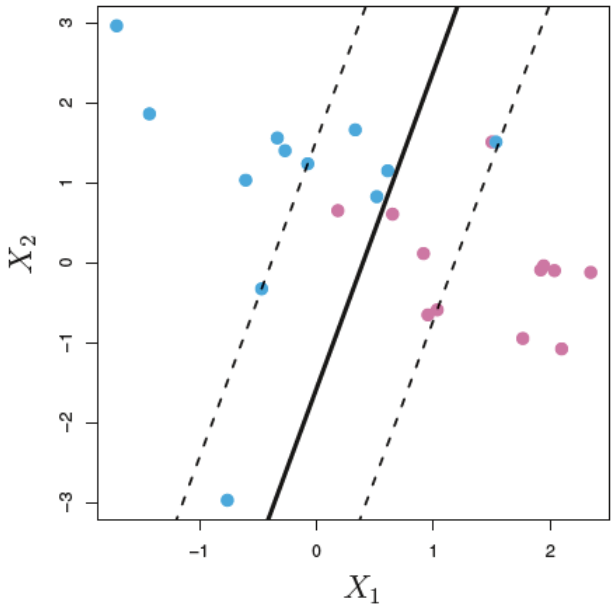

Figura 10- Modelo para o classificador de margem flexível do SVM. Fonte: James et al. (2013)

Um classificador de vetor de suporte é combinado a uma *kernel* não linear visando mais flexibilidade dos limites de decisão. A função *kernel* quantifica a similaridade entre duas observações. Para o classificador *kernel* do tipo linear essa quantificação da similaridade é feita a partir da correlação de Pearson dos preditores padronizados. O classificador para o vetor de suporte adicionado da função *kernel* é apresentado:

$$f(x) = \beta_0 + \sum_{i \epsilon S} \alpha_i K(x, x_i) \qquad (13)$$

onde S corresponde à coleção de índices dos vetores de suporte, K representa a função *kernel*, e β_0 e α são coeficientes encontrados a partir do produto interno entre todos os pares

de observações de treinamento. A função *kernel* que agrega uma generalização para o produto interno (x, x_i) pode assumir várias formas de acordo com o modelo. Nas Equações 14, 15 e 16 são representadas respectivamente as funções *kernel* linear, polinomial e radial, as mais utilizadas nos modelos classificadores SVM.

$$K(x_i, x_{i'}) = \sum_{j=i}^{p} x_{ij} x_{i'j} \tag{14}$$

$$K(x_i, x_{i'}) = (1 + \sum_{j=i}^{p} x_{ij} x_{i'j})^d \tag{15}$$

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=i}^{p} (x_{ij} - x_{i'j})^2 \tag{16}$$

A função *kernel* é que determina a similaridade entre duas observações para o classificador de vetor de suporte. A função do tipo linear, por exemplo, quantifica a similaridade entre as observações através da correlação de Pearson dos preditores padronizados. A escolha da função kernel ideal depende do problema, sendo necessária a inclusão de parâmetros extras dependendo da função escolhida, como o grau polinomial '$\boldsymbol{d}$' na função kernel polinomial. Esses parâmetros junto ao valor do custo constituem os parâmetros de ajuste do modelo de aprendizagem por SVM (KUHN e JOHNSON, 2013).

3.3.3 RANDOM FOREST

O algoritmo de aprendizagem *Random Forest* (RF) ou floresta aleatória é um conjunto de árvores de classificação e regressão construídas a partir de um subconjunto de dados, baseado em uma premissa de que um conjugado de classificadores detém melhores resultados que classificadores individuais. O algoritmo *Random Forest* é executado com eficiência em grandes conjuntos de dados, possui método para tratar dados ausentes com precisão, fornece estimativa de quais atributos são importantes para a classificação e possui uma maior robustez ao ruído se comparado a outros métodos de classificação. (BREIMAN, 2001).

O Random Forest é um classificador bem sucedido com base em algoritmos de aprendizagem de conjunto proposto por Breiman (2001). O *Random Forest* cria uma quantidade de árvores de decisão a partir de um conjunto de dados utilizando o *bagging*, um

meta-algoritmo para melhorar os modelos de classificação de acordo com a estabilidade e precisão da classificação. O uso do *bagging* no treinamento reduz a variância e evita o *overfitting*. Este procedimento extrai casos aleatórios do conjunto de dados de treinamento, sendo esses utilizados para cada uma das árvores de decisão do algoritmo. Cada árvore classificadora é descrita como um preditor. O algoritmo constrói a sua decisão por meio da contagem de votos dos componentes preditores em cada classe e seleciona a classe vencedora por votação acumulada (SON *et al.*, 2009).

O Random Forest é um conjunto de B árvores $\{T_1\ (x), \ldots, T_B(x)\}$, onde $x \in \{x_1, \ldots, x_m\}$ é um vetor m-dimensional de variáveis de um objeto classificado. O conjunto produz B saídas $\{\hat{y}_1 = T_1(x), \ldots, \hat{y}_B = T_B(x)\}$, onde $\hat{y}_B$, $b \in \{1, 2, \ldots, B\}$, é a previsão de um objeto classificado pela b-ésima árvore. As saídas de todas as árvores de decisão são agregadas para definir uma previsão final $\hat{y}$. No caso de problemas de classificação a saída $\hat{y}$ é a classe prevista para a maioria das árvores, e para a regressão, é a média das previsões individuais de cada árvore (SVETNIK et al., 2003). Na Figura 11 apresenta-se o framework da arquitetura do algoritmo *Random Forest*.

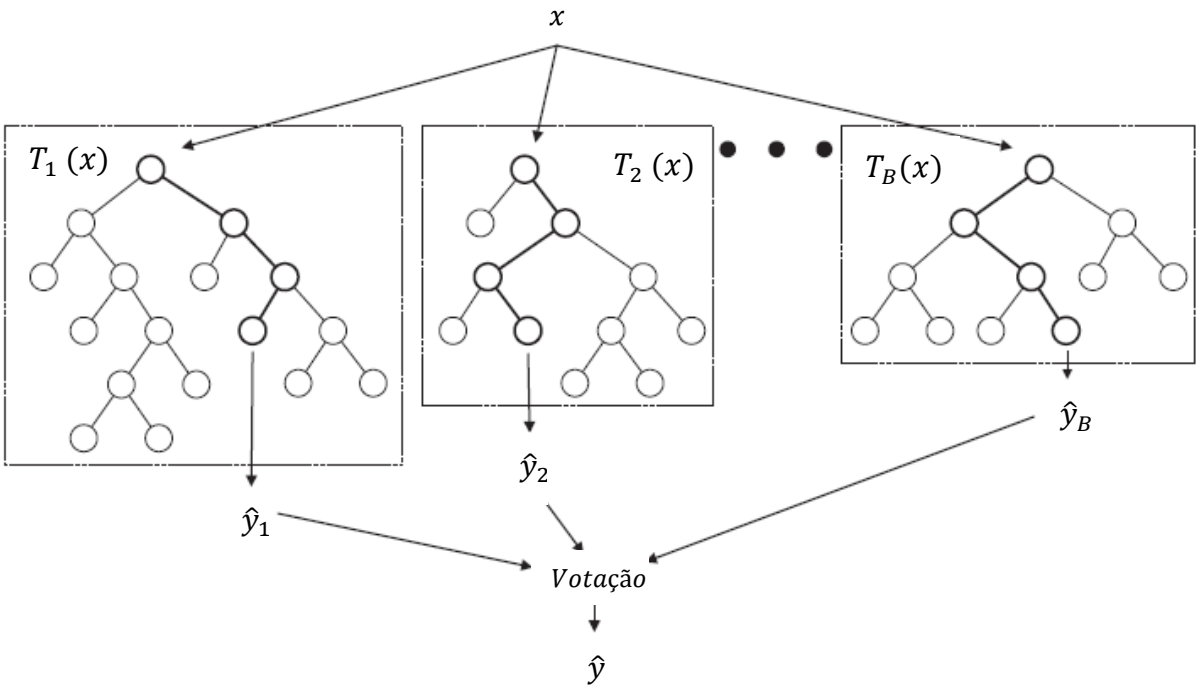

Figura 11- Arquitetura do algoritmo Random Forest. Adapatado de Verikas (2011)

Breiman (2001) comprovou que o método de aprendizagem por florestas aleatórias são protegidas a problemas de *overffiting*, o que revela que o modelo não sofrerá com adversidades recorrentes se um número excessivo de árvores foi construído para a floresta.

Entretanto quando maior o número de árvores maior o custo computacional para treinar e construir o modelo.

3.4 MÉTRICAS DE DESEMPENHO

A avaliação do desempenho de um modelo de classificação é uma das tarefas mais importantes e críticas para o aprendizado de máquinas (JURMAN *et al*., 2012). Uma análise estatística rigorosa e correta dos resultados deve ser desenvolvida para ser aceita como válida pela comunidade científica de aprendizado de máquinas, já que a validação estatística dos resultados em documentos publicados é indispensável (DEMŠAR, 2006**;** GARCIA e HERRERA, 2008).

Em um documento típico de aprendizado de máquina existem etapas muito bem estabelecidas como o pré-processamento, a seleção dos dados para treinamento e teste, a execução dos algoritmos de classificação, e a posterior análise do modelo de classificação, que geralmente é a precisão (*accuracy*). A etapa restante envolve a verificação estatística da hipótese de desempenho dos aprimoramentos realizados nas etapas de pré e pós-processamento (DEMŠAR, 2006).

Um método comum para descrever o desempenho de um modelo de classificação é a matriz de confusão, que é uma tabulação das classes observadas e previstas para os dados. A matriz de confusão codifica o número de previsões corretas e incorretas para cada classe. As células da diagonal principal da matriz de confusão apresentam os casos em que existe a previsão correta das classes enquanto os demais valores ilustram o número de erros para cada caso possível (KUHN e JOHNSON, 2013; BALLABIO *et al*., 2018).

Para uma classificação binária os valores da matriz de confusão recebem designações específicas. O número de exemplos das classes reconhecidos corretamente são os verdadeiros positivos (TP), o número de exemplos reconhecidos corretamente que não pertencem a classe são os verdadeiros negativos (TN), exemplos atribuídos incorretamente às classes são os falsos positivos (FP), e os que não foram reconhecidos como exemplos de classe são os falsos negativos (FN), Tabela 1 (SOKOLOVA e LAPALME, 2009).

Tabela 1- Matriz de confusão classes binárias.

Classe de dados	**Classificação positiva**	**Classificação negativa**
pos	Verdadeiro positivo (tp)	Falso negativo (fn)
neg	Falso positivo (fp)	Verdadeiro negativo (tn)

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

A Tabela 2 apresenta as métricas de desempenho comumente utilizadas (Acurácia Média, Taxa de Erro, Precisão, Recuperação e *F-measure*) para avaliação dos algoritmos de aprendizagem, a partir de um exemplo multiclasse.

Tabela 2– Métricas de desempenho para os algoritmos de aprendizagem.

Métrica de Desempenho	**Modelo Matemático**
Acurácia média	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$
Taxa de Erro	$\frac{\sum_{i=1}^{l} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$
Precision	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fp_i)}$
Recall	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fn_i)}$
F-measure	$\frac{(\beta^2 + 1) Precision \cdot Recall}{\beta^2 Precision + Recall}$

A métrica mais simples, amplamente utilizada e com interpretação direta é a taxa de precisão (*accuracy*), que reflete a concordância entre as classes observadas e previstas (KUHN e JOHNSON, 2013). Também apresentam como métricas comumente aplicadas a recuperação e o *F-measure*. A precisão de um classificador é a fração de previsões corretas no conjunto de testes, quanto maior a precisão, melhor o classificador. A recuperação ou sensibilidade (*recall*) específica de cada classe é a fração de previsões corretas sobre o número de exemplos positivos dos dados. Quanto maior o *recall*, melhor o classificador. A medida *F-measure* da classe tenta equilibrar os valores de precisão e recuperação a partir do cálculo da média harmônica. Quanto maior o valor de *F-measure* melhor o classificador (ZAKI *et al*., 2014).

3.5 FERRAMENTAS DE APLICAÇÃO

Com o objetivo de tecer considerações relevantes acerca das plataformas e linguagens de programação utilizadas pela pesquisa desenvolvida nesta dissertação, dados relevantes são apresentados acerca da linguagem de programação Python e da plataforma de mineração de dados e aprendizado de máquina Weka. A linguagem Python foi de grande relevância para o sucesso da pesquisa, sendo a grande norteadora para as implementações, e a plataforma Weka apresentou um relevante papel coadjuvante, mas não menos importante.

3.5.1 PYTHON

Com uma vasta biblioteca de programação com suporte para todas as áreas de ciências da computação, o Python (Figura 12) se estabeleceu como uma linguagem de programação gratuita e universalmente disponível e uma das mais populares da computação científica. Graças a um caráter de altamente interativo e um ecossistema amadurecido de suas bibliotecas, é uma opção de grande atratividade em análise exploratória de dados (DUBOIS, 2007). Amplamente utilizado em centros acadêmicos e na indústria, com implementações de códigos compactos e legíveis, o Python oferece uma boa combinação de funcionalidades e pacotes especializados com algoritmos de aprendizado de máquina, sendo em geral melhor que linguagens estatísticas especializadas como o R ou SAS (BOWLES, 2015).

Figura 12- Apresentação visual utilizada pelo Python.

Algumas características que justificam a vanguarda da linguagem de programação Python, são a licença de código aberto, a execução de tarefas em multiplataformas, a sintaxe limpa mas com construções sofisticadas, uma poderosa interatividade, a capacidade de extensão do código compilado e incorporação em aplicativos, o grande número de bibliotecas instaladas, a ligação a todos os kits de ferramenta GUI, a forte comunidade de desenvolvedores e um repositório de módulos que simplificam o gerenciamento de software (OLIPHANT, 2007).

Para tarefas de programação em aprendizado de máquina a biblioteca Scikit-learn é uma das mais populares e acessíveis na atualidade. O Scikit-learn utiliza um ambiente rico que possibilita implementações de ponta mantendo uma interface fácil de usar. O Scikit-learn é distribuído sob a licença BSD, incorpora o código compilado para eficiência, concentra-se na programação imperativa e incorpora bibliotecas de outras linguagens para implementações de referência e com licenças compatíveis, que são características que o diferem de outras caixas de ferramentas de dados (PEDREGOSA *et al*., 2011; RASCHKA e MIRJALILI, 2017).

3.5.2 WEKA

O software Weka, Figura 13, é uma coleção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados em que os algoritmos podem ser aplicados diretamente a um conjunto de dados, ou chamados a partir do próprio código Java (AHER e LOBO, 2013).

Figura 13- Apresentação visual utilizada pelo Weka.

O Weka é um software de mineração de dados de código aberto desenvolvido pela Universidade de Waikato, Nova Zelândia, no ano de 1997. Ele utiliza a Licença Pública Geral GNU (GPL). O software é escrito na linguagem Java e contém uma GUI para interagir com os arquivos de dados e produzir resultados visuais. O software possui uma API geral que possibilita incorporar outros aplicativos e outras bibliotecas (PREDIC *et al.*, 2018).

O Weka utiliza o formato *.arff* para seu conjunto de dados. O software possibilita muitas abordagens em programação orientada a objeto, incluindo classificadores (SVMs árvore de decisão, métodos de aprendizagem) e métodos de clusterização. O software oferece uma interface gráfica com o usuário e outra por linha de comando. Quanto aos métodos de validação, utilizam-se as técnicas padrões e a validação cruzada. Além disso, o software permite a visualização e avaliação estatística dos resultados (GEWEHR *et al.*, 2007).

4. METODOLOGIA

Este capitulo apresenta a metodologia aplicada à pesquisa com o norte do problema de pesquisa e dos objetivos apresentados. Primeiramente apresenta-se o delineamento metodológico a partir do enquadramento metodológico de pesquisa e posteriormente são apresentados os procedimentos metodológicos visando encontrar as soluções do problema, embasado pela fundamentação teórica.

5.1 DELINEAMENTO METODOLÓGICO

A partir da definição do problema de pesquisa e dos objetivos apresentados, um delineamento metodológico se faz necessário como meio de justificar os fundamentos teóricos da pesquisa. O delineamento metodológico expõe o desenvolvimento da pesquisa com foco nos procedimentos técnicos da análise e da coleta de dados (GIL, 2002).

Com o objetivo de classificar a pesquisa científica a partir do enquadramento metodológico, apresenta-se sua qualificação do ponto de vista dos procedimentos técnicos de acordo com Gil (2002), e Marconi e Lakatos (2003):

- Quanto à natureza: a pesquisa apresentada é definida como aplicada, com o objetivo de produzir conhecimento científico para a aplicação prática voltada para soluções de problemas concretos.
- Quanto aos objetivos: é classificada como pesquisa exploratória, buscado analisar as relações existentes entre as variáveis do problema.
- Quanto aos procedimentos: enquadra-se na fonte de pesquisa documental valendo-se de relatórios técnicos da empresa estudada. A pesquisa documental baseia-se em materiais que não receberam ainda um tratamento analítico ou que podem ser reelaborados de acordo com os objetivos de pesquisa.
- Quanto à abordagem do problema: define-se como pesquisa quantitativa, o que significa traduzir em números as informações dos documentos com o objetivo de classificação e análise.
- Quanto às características: caracteriza-se como pesquisa *Ex Post Facto*, que define um estudo realizado a partir de um fato passado, e como pesquisa de estudo

de caso. O estudo de caso consiste de um estudo profundo de um ou mais objetos, de maneira que permita o seu amplo e detalhado conhecimento.

5.2 PROCEDIMENTOS METODOLÓGICOS

Definido o delineamento metodológico da pesquisa faz-se necessário apresentar as atividades desenvolvidas com vistas a cumprir os objetivos estabelecidos. Com o foco na predição do *lead-time* de fabricação de componentes para protótipos de uma empresa de base tecnológica os procedimentos metodológicos foram guiados pela análise preditiva. As atividades desenvolvidas foram:

1. Estudo da logística de fabricação da empresa;
2. Identificação das variáveis preditoras;
3. Transformação e limpeza dos dados;
4. Exclusão de *outliers*;
5. Análise das variáveis preditoras e da classe;
6. Aplicação dos algoritmos de aprendizagem
7. Avaliação e comparação do desempenho estatístico dos algoritmos

A primeira atividade de pesquisa foi guiada pelo estudo de logística de fabricação da empresa estudada. Essa etapa foi necessária para compreender os fluxos de informações e materiais da empresa com foco na compreensão das variáveis que estavam inseridas na tabela de dados coletada. A compreensão desses dados e do comportamento de fabricação da empresa auxiliou na identificação de variáveis preditoras potenciais de acordo com a escolha da saída de interesse (*lead-time* total).

Após o estudo da logística de fabricação da empresa foi possível delimitar as variáveis de interesse da pesquisa. Uma primeira limpeza da tabela de dados coletada foi realizada a partir de exclusão de colunas de dados que não apresentavam informações relevantes para a predição do tempo de fabricação do componente. Identificada as variáveis de controle iniciou-se a etapa de pré-processamento da tabela.

Uma limpeza dos dados foi necessária identificando valores de variáveis que não eram cientificamente válidos. Variáveis com valores quantitativos negativos foram excluídas por não corresponderem a valores reais. Variáveis numéricas que ao invés de números apresentavam caracteres, e vice-versa, também foram excluídas da tabela de dados. Uma série de dados faltantes também foram identificados e excluídos.

Uma transformação foi realizada nas variáveis preditoras com o interesse em reduzir os erros de preenchimento da tabela. Os erros de digitação foram eliminados e por consequência ocorreu uma redução massiva do escopo das variáveis preditoras categóricas. Uma transformação de variáveis numéricas no formato de datas para o formato de variável categórica binária também foi realizada visando uma melhor compreensão dos dados pelos algoritmos de aprendizagem.

Uma nova análise da Tabela de dados foi realizada buscando identificar *outliers* que são amostras excepcionalmente distantes dos valores da maioria representativa dos dados. A partir de análise gráfica foi possível identificar e excluir esses *outliers*. Além disso, nessa etapa foi possível limitar a amplitude do escopo para cada variável preditora e para a classe. Uma nova avaliação foi realizada nessa etapa buscando identificar valores cientificamente inválidos como um valor de *lead-time* total menor que outras variáveis preditoras quantitativas.

Após a etapa de pré-processamento dos dados uma etapa de análise das variáveis preditoras e da classe foi realizada com foco em sua compreensão e mapeamento. A partir da análise das médias, desvios, dispersões, histogramas, e diagramas de extremos e percentis foi possível identificar o comportamento das variáveis preditoras e da classe.

A etapa seguinte constituiu-se de aplicar os algoritmos de aprendizagem de máquinas nos dados da tabela pré-processada. Para aplicação dos métodos de aprendizagem as variáveis categóricas foram transformadas em variáveis '*dummy*' e optou-se em utilizar uma transformação escalar nas variáveis preditoras quantitativas com o foco em distribuir e equilibrar seus pesos de importância para o processo de treinamento.

Os valores de parâmetros para os métodos aplicados de redes neurais e do algoritmo SVM foram aplicados. Nas redes neurais, o número de neurônios e camadas ocultas foram identificados a partir de testes de valores e verificação da acurácia. O método de validação utilizado foi o *k-fold* com valor de $k = 10$.

A avaliação dos resultados do modelo de aprendizado foi aplicada pela análise da acurácia e dos valores do erro médio absoluto e do erro médio quadrático entre os valores de treino e validação. Quatro topologias de análise foram estruturadas visando acompanhar o desempenho dos algoritmos a partir da redução de variáveis categóricas com alto escopo. A primeira topologia foi com a tabela completa na etapa de pré-processamento. Na segunda topologia, o valor da variável 'conjunto' foi reduzida a partir do diagrama de Pareto. Na terceira topologia, os dados com maior volume de dados da variável 'descrição do material'

foram selecionados, também, pelo diagrama de Pareto. Na quarta topologia os materiais e conjuntos mais representativos foram avaliados pelo diagrama de Pareto. Tais informações ficarão mais palpáveis à medida que as análises forem apresentadas no decorrer do trabalho.

Uma segunda etapa de análise do desempenho dos algoritmos de aprendizagem de máquina foi realizada visando verificar os quatro materiais que mais apareciam nas ordens de serviço. Para esses materiais a acurácia, e os erros absolutos e quadráticos foram aplicados como métrica de desempenho para duas topologias, com a variável ‘conjunto’ completa e com sua redução pelo diagrama de Pareto.

Uma comparação entre o desempenho dos algoritmos foi realizada com o objetivo de avaliar o desempenho dos modelos de aprendizagem construídos. A partir dos testes de Friedman e Nemenyi foi possível identificar a existência de diferença estatística significativa entre os testes de aprendizagem encontrando os testes com melhores resultados.

5. RESULTADOS E DISCUSSÕES

Este capitulo apresenta os resultados obtidos na pesquisa a partir do delineamento e procedimentos metodológicos apresentados no capítulo anterior. Primeiro, realiza-se um estudo da logística de fabricação da empresa buscando compreender os fluxos internos de materiais e informações. Identificam-se as variáveis preditoras e a variável de saída. Posteriormente a etapa de pré-processamento dos dados é realizada através dos métodos de transformação, limpeza e exclusão dos *outliers*. Realiza-se uma análise das variáveis do problema e aplicam-se os métodos de aprendizagem de máquina por Redes Neurais Artificiais (RNA), *Support Vector Machine* (SVM) e *Random Forest* (RF).

5.1 ESTUDO DA LOGÍSTICA DE FABRICAÇÃO DA EMPRESA

A empresa do estudo de caso é uma empresa brasileira do ramo de tecnologia com atuações nas áreas médica, industrial, aeroespacial e de defesa. A empresa é pioneira no desenvolvimento de tecnologia de ponta com especialidade no desenvolvimento de produtos mecatrônicos. O tipo de fabricação da empresa é o MTO (*Make To Order*) para produtos médicos e industriais, e ETO (*Engineering To Order*) para produtos das áreas de espaço, sendo o norte de pesquisa avaliar a documentação das ordens de serviço para a fabricação de componentes dos protótipos dos produtos MTO e para os entregáveis dos produtos ETO.

Em primeiro momento, é necessário explorar e conhecer os dados disponibilizados para compreender a estrutura organizacional e os fluxos de informações e de materiais que envolvem a logística de fabricação dos componentes, para posteriormente, aplicar os métodos de modelagem preditiva.

Um estudo do processo de criação dos dados adquiridos foi realizado com o objetivo de entender os processos envolvidos na fabricação dos componentes, as siglas utilizadas pela empresa para designar etapas, as funções dos setores envolvidos e os fluxos de informação necessários para a execução das ordens de serviço. A partir de um processo de melhoria dos indicadores de desempenho com foco no aumento da capabilidade dos processos de fabricação de componentes para novos produtos, a empresa passou por uma reestruturação dos seus fluxos de informação e de materiais.

Uma unidade organizacional no setor de engenharia foi inserida, denominada Unidade de Gerenciamento e Documentação (UGD), com a função de realizar a interface com a manufatura. A UGD passou a ser responsável pela entrega de documentação do projeto e o recebimento de peças, partes e componentes a serem integrados nos protótipos e produtos ETO. Em uma segunda etapa, um comitê de acompanhamento da manufatura das partes, peças e componentes, foi criado com o intuito de realizar um mapeamento dos processos e os obstáculos para o cumprimento dos prazos de projeto, Figura 14. O termo UENG representa os engenheiros responsáveis pelo projeto de peças e componentes fabricados (BARBALHO e TORRES, 2008).

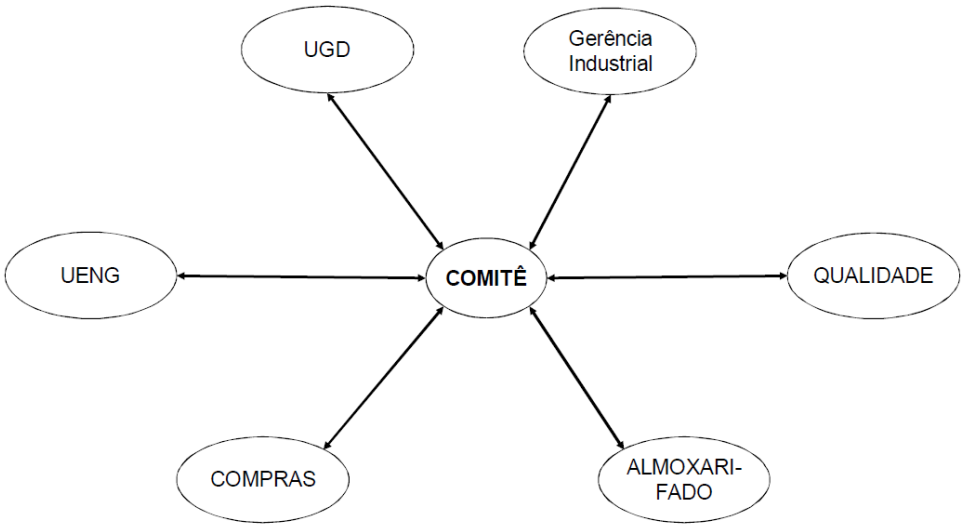

Figura 14- Estrutura organizacional do comitê de manufatura. Fonte: Barbalho e Torres (2008)

A partir do conhecimento e aprendizado adquirido pela empresa no processo de melhoria de desempenho com o foco no aumento da capabilidade do processo de fabricação, uma reestruturação da logística para a fabricação dos protótipos foi realizada, Figura 15. A UENG elabora as especificações técnicas dos projetos e repassa para a UGD que registra os documentos e os transfere para a gerência industrial e, posteriormente, para os setores de fabricação interna (FABRIC), para o setor de COMPRAS para o caso de peças e partes terceirizadas, ou para o almoxarifado (ALMOX) para o caso de peças e componentes disponíveis no estoque. Com as peças já fabricadas, caso necessário, uma etapa de inspeção é realizada pelo departamento de qualidade. Por fim, a UGD registra a entrada das peças e libera para a montagem (BARBALHO e TORRES, 2008).

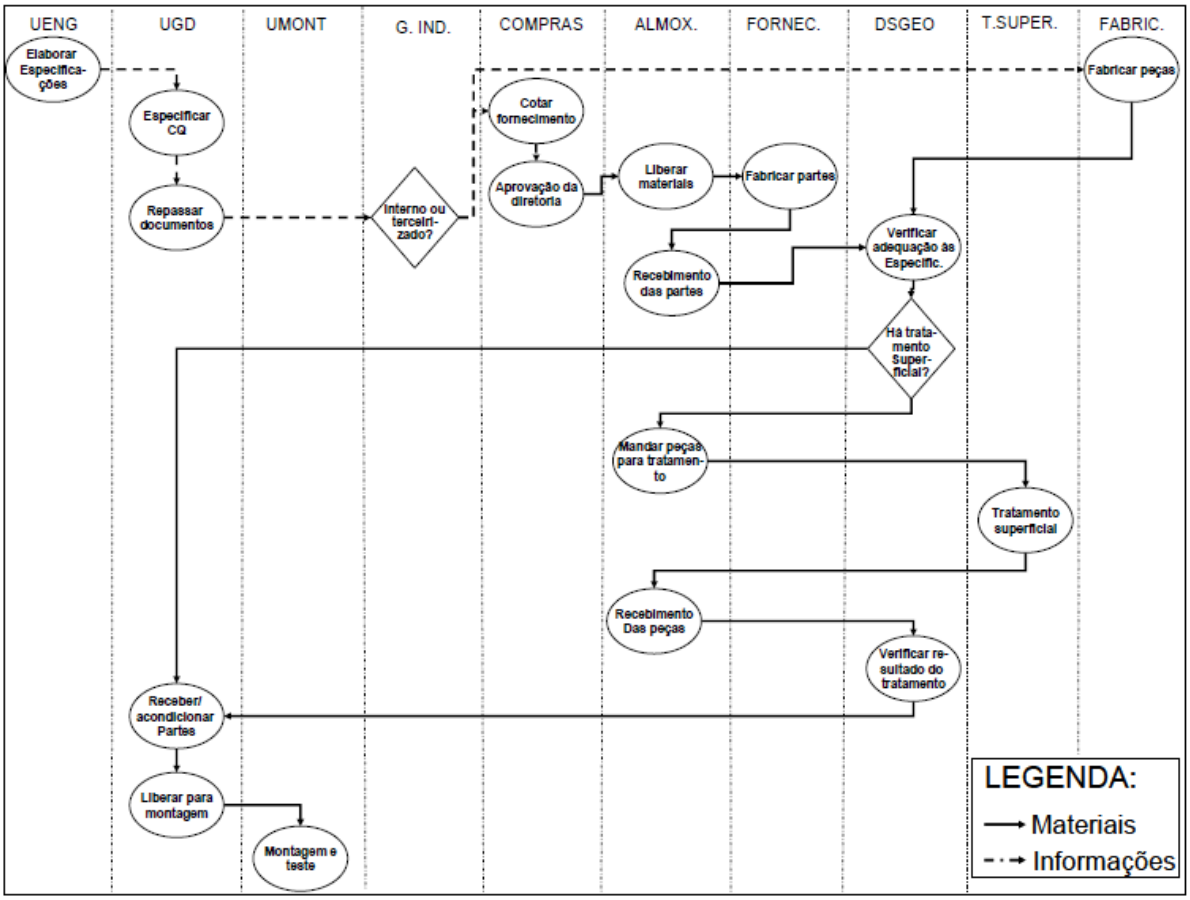

Figura 15- Logística para a fabricação de componentes da empresa. Fonte: Barbalho e Torres (2008)

Uma planilha de rastreabilidade para acompanhar as peças foi desenvolvida a partir do fluxo de logística da Figura 15, com o objetivo de facilitar o fluxo de informações dentro da empresa. A ferramenta é composta por várias abas que representam as etapas do processo da Figura 15. O desenho é entregue a gerência industrial que alimenta a aba e destina ao setor responsável. A oficina da entrada na planilha com a data de início e fim da fabricação. Os departamentos de qualidade, almoxarifado, tratamento superficial e pintura identificam a data de início e fim de suas atividades, dependendo do tipo de projeto. Ao final do fluxo, a peça é entregue a UGD onde é marcada a data de recebimento e disponibilização da peça ao engenheiro que a projetou (BARBALHO e TORRES, 2008).

Essa planilha com o fluxo de informações e materiais foi disponibilizada para o estudo de caso do projeto de pesquisa de dissertação com o objetivo de identificar padrões das ordens de serviço e estimar o *lead-time* entre o início e o fim de uma atividade de fabricação de componentes para protótipos e produtos ETO. Como verificado, além do tempo de fabricação, a execução da ordem de serviço envolve um processo de logística de informações e de materiais que torna complexa a estimativa do *lead-time* pela empresa. Uma estimativa do

lead-time de execução das ordens de serviço pode gerar impactos significativos no controle de custos e tempos da empresa estudada. Estudos anteriores demonstraram impacto significativo do *lead-time* de peças para protótipos sobre o tempo de ciclo de um projeto de desenvolvimento de produtos (BOLAÑOS, 2017).

5.2 IDENTIFICAÇÃO DAS VARIÁVEIS PREDITORAS

A planilha de informações disponibilizada inicialmente pela empresa dispunha de 18.329 linhas e 81 colunas, totalizando um total de 1.484.649 células de dados. Trata-se de um histórico de dados armazenados durante seis anos da empresa que corresponde o período entre 2006 e 2011. A partir de diálogos estabelecidos com um especialista e gerente responsável da empresa foi possível identificar as possíveis variáveis preditoras do modelo. Uma transformação inicial de variáveis foi realizada a partir de diálogo com o gerente da empresa do estudo de caso visando facilitar a extração do conhecimento e o input das variáveis preditoras. As variáveis identificadas e suas transformações realizadas são identificadas:

- Conjunto: trata-se de uma variável categórica que corresponde ao conjunto, na estrutura de produto, de um projeto em desenvolvimento ou produto ETO.
- Quantidade: corresponde à quantidade de componente que será fabricado para cada ordem de serviço específica.
- Decisão fabricar ou comprar (*make or buy*): variável categórica binária que corresponde se a produção será feita internamente ou por terceirização.
- Descrição do material: variável categórica que descreve o material específico para a fabricação do componente.
- CQ-Entrada: variável no formato de data que corresponde aos valores de entrada e saída do componente pelo controle de qualidade. A variável foi transformada em categórica binária (Sim/Não) que avalia se o produto necessita ou não de passar pelo controle de qualidade após a fabricação.
- Envio para TS: variável no formato de data que corresponde aos valores de entrada e saída do componente pelo tratamento superficial, um processo terceirizado pela empresa. A variável foi transformada em categórica binária (Sim/Não) que verifica a necessidade de tratamento superficial.
- CQ-Entrada1: variável no formato de data que corresponde aos valores de entrada e saída do componente pelo segundo controle de qualidade, um controle

específico para os processos de tratamento superficial. A variável foi transformada em categórica binária (Sim/Não) e avalia se o produto passou ou não pelo segundo controle de qualidade.

- Repasse GI-EAPD: valor numérico que contabiliza os dias necessários para o repasse de informação da ordem de serviço do Escritório de Apoio à Pesquisa e Desenvolvimento (EAPD) para a Gerência Industrial (GI).
- Material liberado: variável numérica que corresponde ao tempo em dias necessários para a liberação do material para a produção. A variável foi transformada em categórica verificando se o material será liberado com atraso ou sem atraso. Para a fabricação externa a variável adquiriu o valor 'Falso'.
- Tempo de espera: variável numérica que corresponde ao tempo de espera da ordem de serviço para a fabricação interna por conta da fila de produção. Para a fabricação externa a variável adquiriu o valor 'Falso'.
- LT produção: variável numérica que corresponde aos dias necessários para a fabricação interna ou externa.
- LT necessidade: corresponde ao tempo em dias, valor numérico, de necessidade da peça fabricada. Esse tempo corresponde ao tempo entre o repasse da unidade de engenharia para o EAPD, e a data de necessidade do componente fabricado. Para essa variável específica designou-se o valor 'falso' quando não houve a data de necessidade do componente.
- LT total: variável numérica que corresponde ao período em dias que a ordem de serviço levou para ser executada a partir dos fluxos de materiais e de informações. A variável corresponde ao valor da classe ou a saída para a aplicação dos modelos de aprendizagem preditiva.

5.3 TRANSFORMAÇÃO E LIMPEZA DE DADOS

A partir da identificação das variáveis preditoras (entrada) e da classe (saída), detectou-se uma série de dados faltantes que foram excluídos para não oferecer desvios no modelo de aprendizagem. Dados não correspondentes ou com informações inválidas, como um valor numérico no lugar de uma variável categórica, também foram eliminados. Dados numéricos negativos que não correspondiam a valores cientificamente válidos também foram excluídos da planilha de dados.

Além dos dados faltantes, verificou-se que em duas colunas específicas, ‘descrição de material’ e ‘conjunto’, detinham variáveis com um extenso número de preditores distintos. A variável ‘descrição do material’ era composta por 660 materiais distintos, e uma grande maioria estava redigida de inúmeras maneiras. A palavra alumínio, por exemplo, apresentava um valor maior a 10 grafias distintas. A partir da correção dos dados redigidos de maneira errônea o total de materiais distintos reduziu 86,5%. O mesmo procedimento foi realizado para a variável ‘conjunto’ que passou por uma redução 45,13% para categorias de projetos distintos.

Verificou-se que a variável tipo de fabricação mecânica detinha uma grande expressividade na empresa correspondendo a 76,85% do total. Dessa maneira, a análise de fabricação mecânica foi escolhida e norteou os estudos da planilha de dados. Os demais tipos de fabricação de ótica, filmes finos e seus retrabalhos apesar de essenciais para a empresa apresentavam um conjunto de dados muito pequenos, e com um menos padronização em seus processos de fabricação de acordo com o especialista gerente da empresa, se comparados à fabricação mecânica.

Um estudo da correlação entre os preditores quantitativos foi realizado com o objetivo de identificar as entradas com alto índice de correlação e, por conseguinte, colaborar para diminuir a probabilidade de resultar modelos instáveis no processo de aprendizagem. Na Figura 16 é possível visualizar um gráfico de calor com os valores de correlação pelo coeficiente de Pearson entre as variáveis quantitativas.

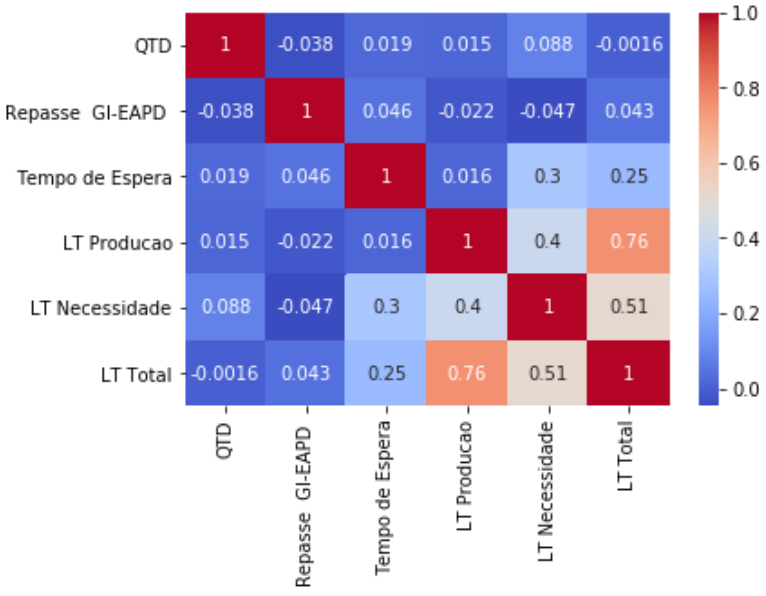

Figura 16- Correlação entre as variáveis preditoras quantitativas.

Verifica-se, a partir da visualização da Figura 16, que as variáveis de maiores correlações são o ‘LT produção’ com o ‘LT Total’, e o ‘LT de Necessidade’ com o ‘LT Total’. Como as variáveis preditoras identificadas estão fortemente correlacionadas com a variável classe (saída), e não com outras variáveis preditoras, não há justificativa para a exclusão de qualquer variável quantitativa da planilha de dados. Com o objetivo de verificar o comportamento de distribuição, a dispersão, e a regressão entre as variáveis com maior colinearidade da Figura 16, o gráfico da Figura 17 foi construído.

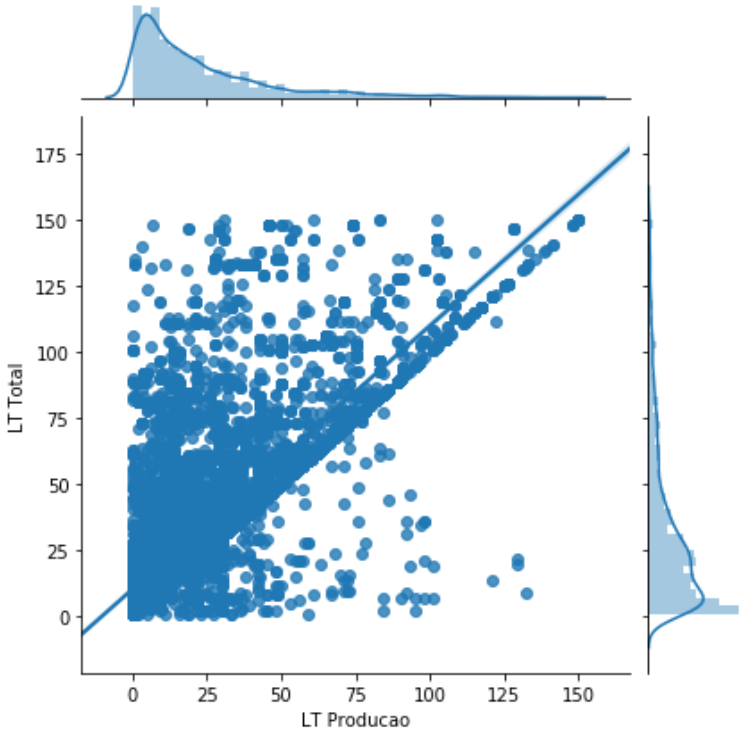

Figura 17- Gráfico de dispersão, distribuição e regressão das variáveis com maior colinearidade.

A Figura 17 exibe o gráfico das variáveis com maior correlação com a interseção do conjunto de dados das variáveis ‘LT produção’ e ‘LT Total’, a reta de regressão linear que corresponde ao comportamento habitual e aproximado entre as variáveis, e o modelo de distribuição a partir de histograma destacando a média e o desvio padrão, de maneira visual. Verifica-se que as distribuições das variáveis não são normais ou gaussianas, apresentam não linearidades, com médias concentradas em valores pequenos e alto desvio padrão tanto para o ‘LT produção’ quanto para o ‘LT total’. Maiores profundidades sobre as variáveis preditoras e da classe serão abordadas no tópico análise das variáveis preditoras e da classe.

5.4 EXCLUSÃO DE OUTLIERS

Com o objetivo de identificar as amostras excepcionalmente distantes dos valores da maioria representativa do conjunto de dados e certificar que os valores tabelados são cientificamente válidos o gráfico da Figura 18 foi construído buscando identificar os *outliers* das variáveis quantitativas da planilha de dados.

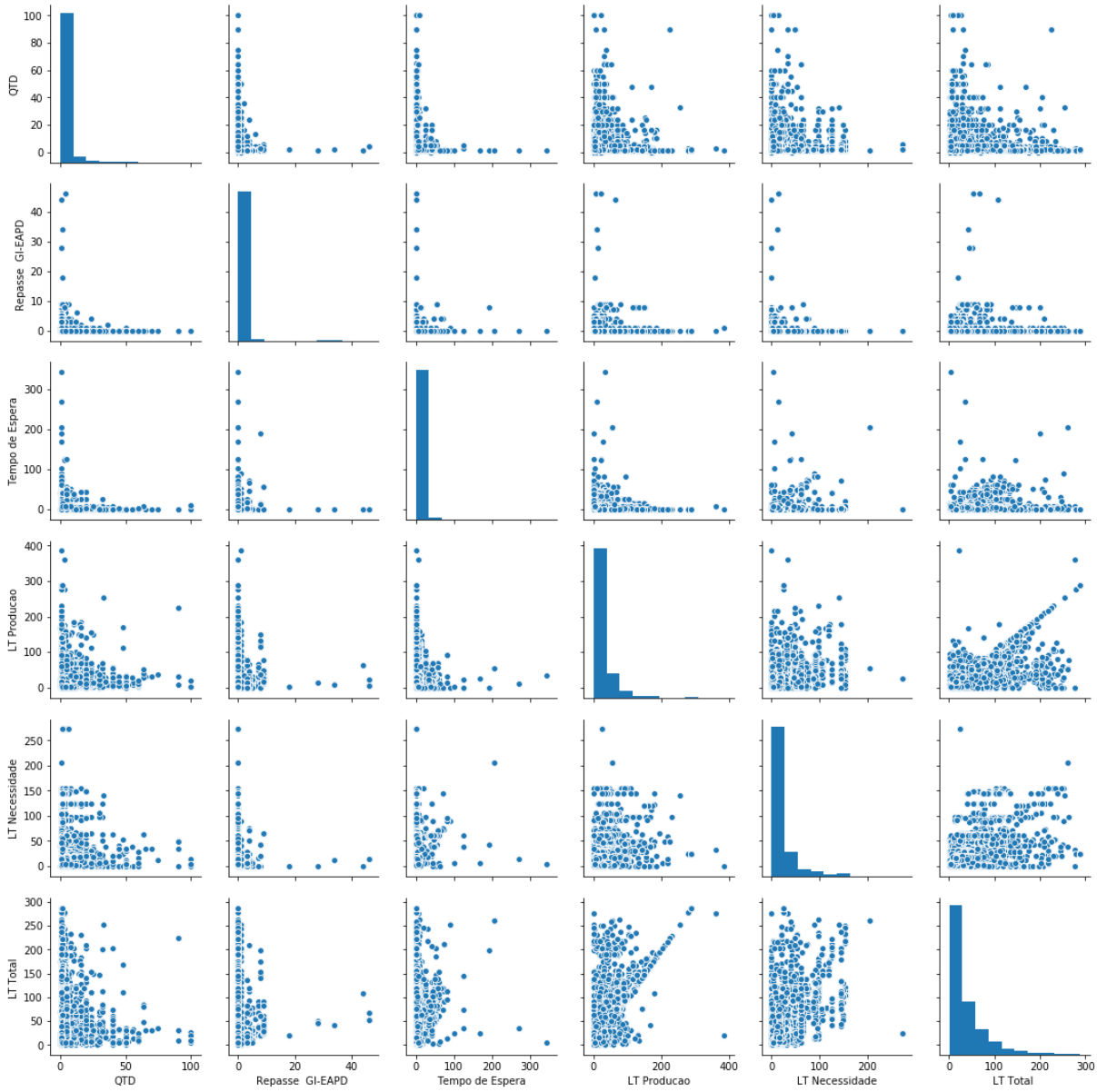

Figura 18- Gráfico de dispersão das variáveis quantitativas da planilha de dados.

A partir da análise dos gráficos de dispersão das variáveis quantitativas (Figura 18) verifica-se que existe um grande número de *outliers* concentrados, principalmente, em valores quantitativamente altos. Com o objetivo de aumentar a probabilidade de melhorar a

performance do modelo de aprendizagem, uma avaliação foi realizada buscando a exclusão desses valores muito discrepantes e concentrados, sobretudo, em valores numéricos altos. Na Figura 19 é possível visualizar o comportamento das variáveis após a exclusão desses *outliers*. Visualiza-se uma maior distribuição da densidade de dados ao longo do escopo das variáveis quantitativas. Dados que não são válidos cientificamente, como um valor menor de *lead-time* de produção que o *lead-time* total, também foram excluídos.

Algumas faixas de valores também foram filtradas na Figura 19, de acordo com diálogos com o especialista gerente da empresa. Novas considerações e filtros para as variáveis preditoras serão realizados posteriormente, com o objetivo de comparar o desempenho do algoritmo de aprendizagem a partir de variações no escopo das variáveis.

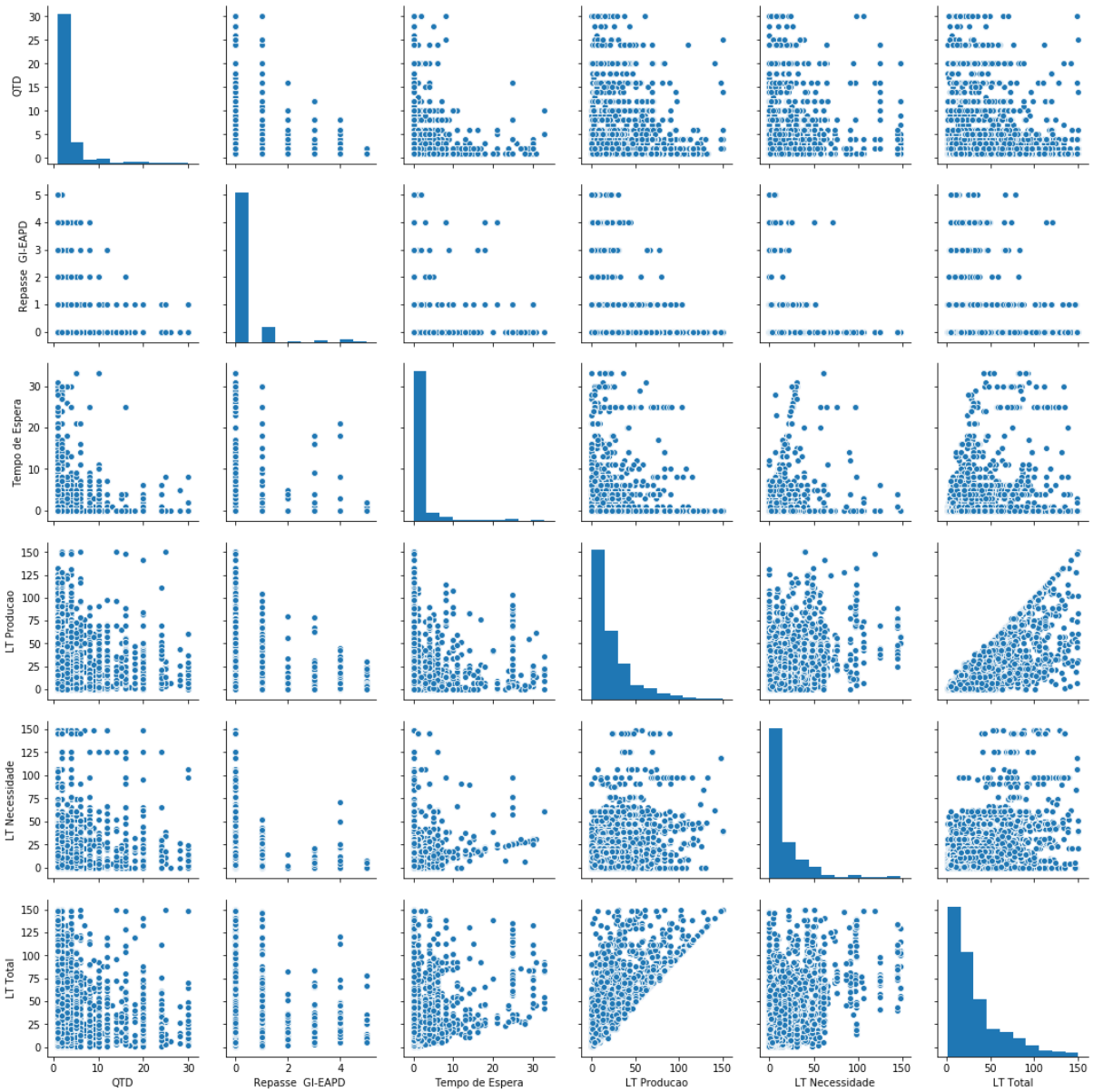

Figura 19- Gráfico de dispersão das variáveis quantitativas após estudo e exclusão de *outliers*.

Após o tratamento inicial dos dados a partir de transformação, limpeza e exclusão de *outliers* houve uma redução de 2.729 linhas de dados, o que corresponde a um percentual de perda de 19,36% dos dados da planilha referentes à fabricação do tipo mecânica, ficando a planilha final com 11.356 linhas. A próxima sessão dos resultados avaliará as variáveis numéricas e categóricas da planilha de dados após os tratamentos realizados.

5.5 ANÁLISE DAS VARIÁVEIS

Uma análise gráfica é realizada nessa seção para compreender o comportamento das variáveis numéricas e categóricas da planilha de dados disponibilizada pela empresa após o pré-processamento. Os resultados serão plotados e suas considerações são discutidas em sequência. O primeiro conjunto de dados avaliados são as variáveis preditoras numéricas. Os valores de média e desvio padrão para as variáveis quantitativas são apresentadas na Tabela 3. Os valores apresentados ajudarão na sustentação de discussões dos gráficos de contagem das variáveis preditoras.

Tabela 3- Média e desvio das variáveis preditoras.

Variável preditora (x)	**Média ($\bar{x}$)**	**Desvio padrão (σ)**
QTD	2,60	3,56
Repasse GI-EAPD	0,20	0,68
Tempo de Espera	1,26	4,11
LT Necessidade	15,46	24,63
LT Produção	22,25	23,65

Na Figura 20 identificam-se os gráficos de quantidade e repasse GI-EAPD. A partir do gráfico de quantidades é possível corroborar que existe uma maior concentração de fabricação de componentes com poucas quantidades, destacando-se valores abaixo de cinco componentes por ordem de serviço e uma média de 2,60, de acordo com a Tabela 3. Infere-se, a partir do gráfico de quantidades de componentes por ordem de fabricação, que como se trata de uma logística voltada para a fabricação de protótipos de uma empresa com alta tecnologia, existe uma fabricação direcionada a produtos originais com componentes de alta especificidade, alta diferenciação e baixo volume.

O gráfico que apresenta o repasse GI-EAPD mostra que existe uma concentração para esse fluxo de informação de até um dia. Com uma grande parte das ordens de serviço concentrada em um valor abaixo de um dia, com uma média de 0,20 dias (Tabela 3) para o repasse GI-

EAPD, demonstra-se que existe uma preocupação e uma boa performance para o fluxo dessa informação na empresa estudada.

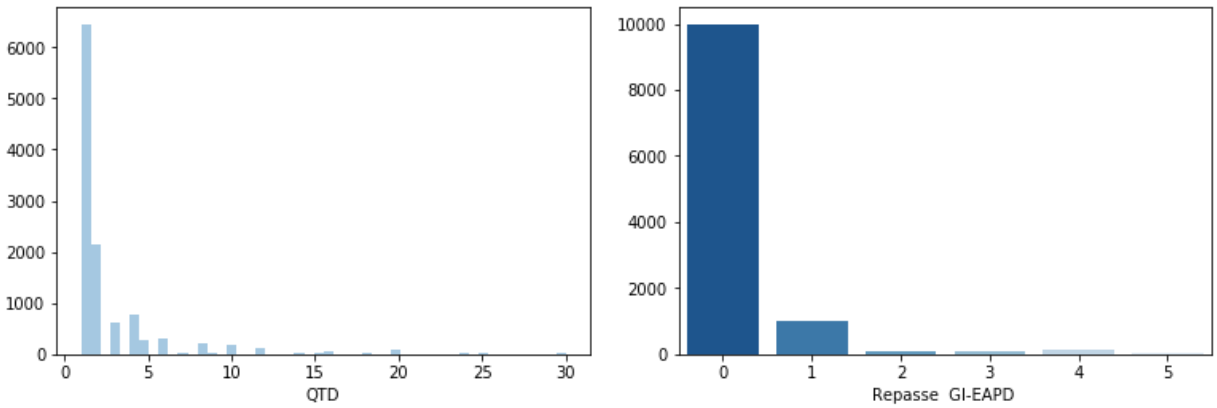

Figura 20- Histograma das variáveis numéricas ‘QTD’ (esquerda) e Repasse GI-EAPD (direita).

Na Figura 21 é possível visualizar o gráfico do tempo de espera, que corresponde ao tempo de atraso na fabricação de um componente devido à fila interna e a indisposição de materiais e equipamentos. Importante ressaltar que essa variável é definida para o tipo de fabricação interna.

Visualiza-se na Figura 21 que existe uma concentração de informações do tempo de espera para a fabricação interna, aproximado, de até oito dias, com uma média de 1,26 dias e um desvio padrão de 4,11 dias (Tabela 3). O tempo de espera com maior volume de dados correspondente foi de um dia ou menos. A partir da análise do gráfico é possível inferir que a empresa estudada dispõe de equipamentos que suprem a sua demanda de ordens de serviço, mas que podem ser melhoradas.

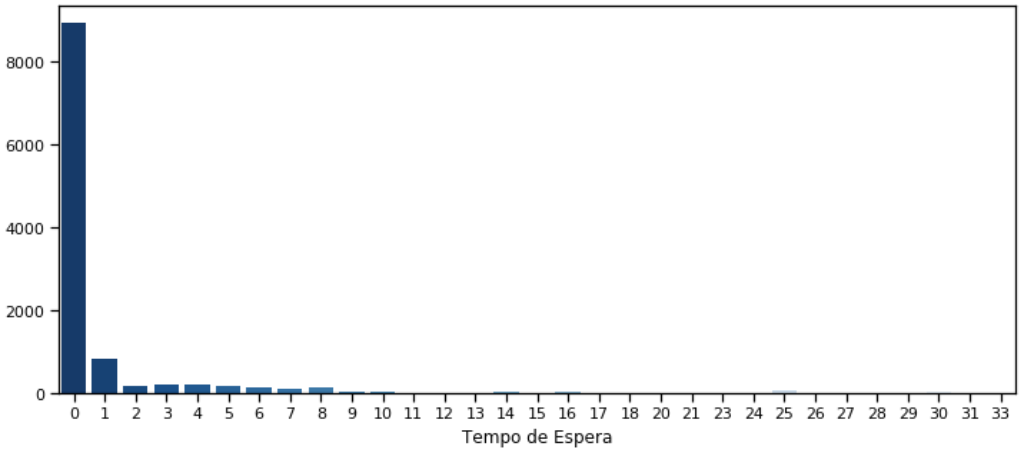

Figura 21- Histograma da variável tempo de espera.

A Figura 22 plota o gráfico do *lead-time* de necessidade da empresa. Essa variável é preenchida quando a unidade de engenharia repassa a ordem de serviço para a equipe de produção e define o tempo ideal, da equipe de projeto, para a fabricação do componente. A grande maioria dos dados para essa variável concentra-se em valores abaixo de 21 dias. Com uma alta centralização do *lead-time* no valor de um (1) dia, infere-se que a equipe de projeto define uma demanda com fluxo intenso de manufatura, concentrado em curto prazo de entrega.

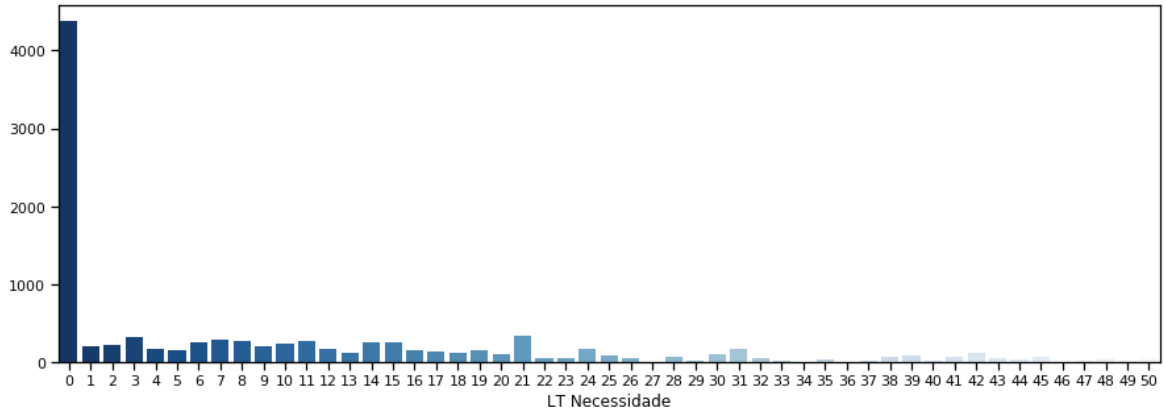

Figura 22- histograma da variável LT Necessidade.

A próxima variável preditora quantitativa de análise é o *lead-time* de produção que corresponde ao tempo de fabricação do componente de produto (Figura 23). Um gráfico de dispersão foi realizado pelo tamanho do escopo da variável, facilitando a visualização dos dados. Verifica-se que valores altos de *lead-time* de produção apresentam pouca representatividade. A média aproximada do *lead-time* de produção de acordo com a Tabela 4 foi de 22,25 dias com um alto desvio padrão em torno de 23,65 dias.

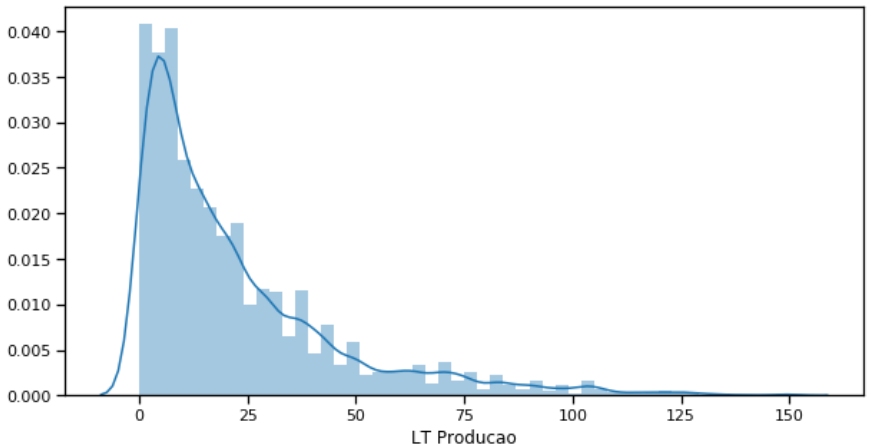

Figura 23- Gráfico de dispersão do lead-time de produção.

Avaliando as médias e desvio do *lead-time* de necessidade ($15{,}46 \pm 24{,}63$) e do *lead-time* de produção ($22{,}25 \pm 23{,}65$) verifica-se que existe uma diferença considerável entre o tempo de necessidade para finalizar a fabricação, preenchido pela equipe de projeto, e o tempo real de fabricação do componente. Sendo um dado que pode significar que há falha no planejamento ou na execução de atividades de projeto que, por sua vez, demoram mais que o planejado. Desta maneira, uma melhor precisão na estimativa desses *lead-times* podem auxiliar na aproximação do valor médio dessas variáveis e por consequência melhorar índices de desempenho da empresa.

A próxima atividade de análise concentra-se no diagnóstico das variáveis categóricas da planilha de dados. Na Figura 24 visualiza-se a contagem de dados para a variável descrição do material.

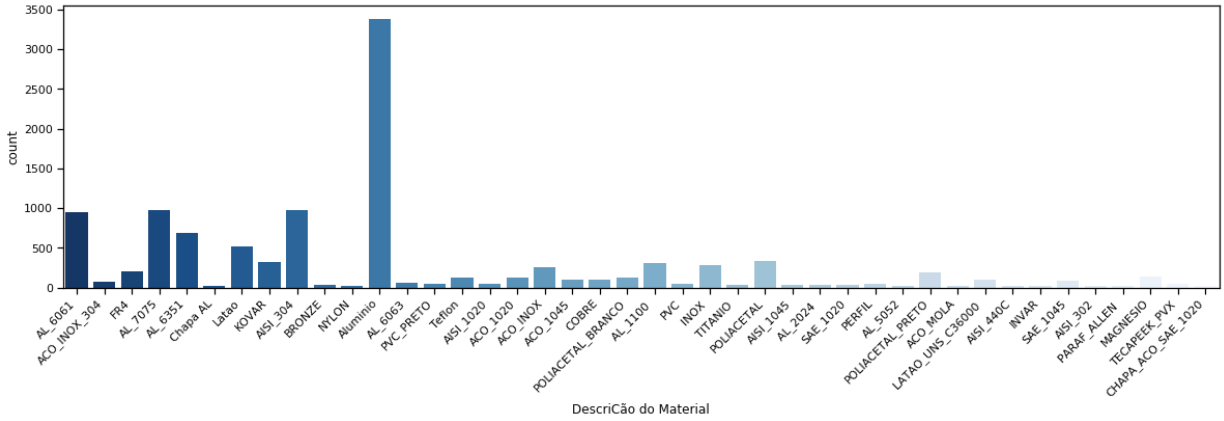

Figura 24- Histograma da variável descrição do material.

Verifica-se na Figura 24 que os materiais que mais ocorrem na fabricação dos componentes da empresa são respectivamente: alumínio comum (3382), AISI 304 (980), AL_7075 (973), e AL_6061 (950). Esses quatro materiais totalizam 55,35% dos materiais utilizados na fabricação dos componentes mecânicos, sendo os materiais de base para a empresa estudada. Na Figura 25 apresenta-se o gráfico de contagem para a variável conjunto. Por questão de sigilo dos conjuntos de projetos da empresa, os nomes dos conjuntos foram substituídos por nomes genéricos. Para melhor visualização do gráfico os conjuntos menos expressivos foram retirados.

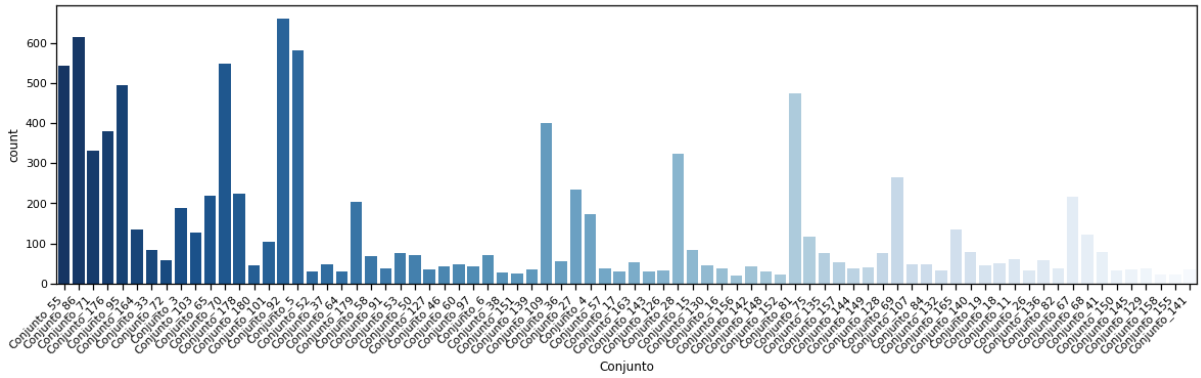

Figura 25- Histograma da variável conjunto.

Na variável conjunto existe uma maior variabilidade de dados se comparada ao gráfico de materiais. Verifica-se que os conjuntos mais representativos e, portanto, os projetos com maior empenho da estrutura de fabricação da empresa são, respectivamente, Conjunto_92 (660), Conjunto_86 (614), Conjunto_5 (582), Conjunto_70 (548), Conjunto_55 (543), Conjunto_95 (493) e Conjunto_81 (474). Na Figura 26 visualizam-se os gráficos da ordem de produção e do controle de qualidade.

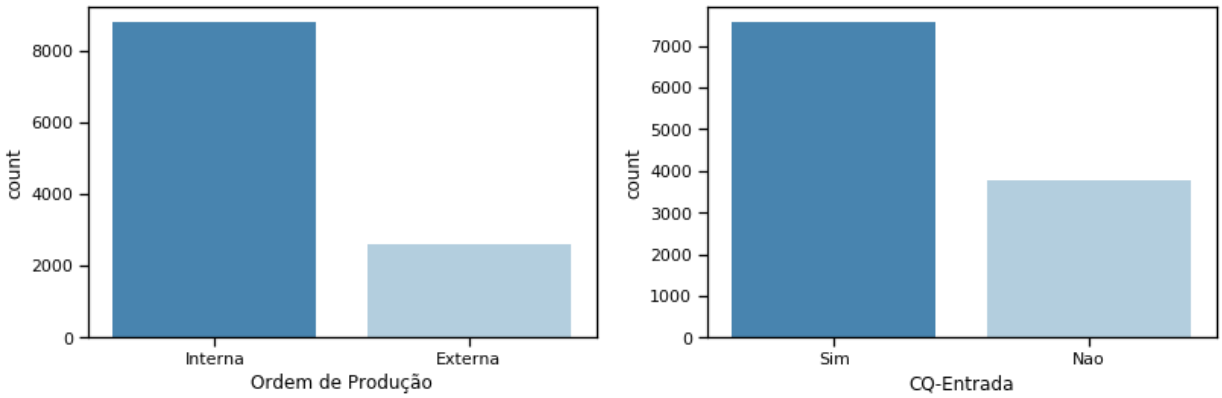

Figura 26- Histograma das variáveis: Ordem de Produção (esquerda) e CQ-Entrada (direita).

Verifica-se em análise da Figura 26 que a ordem de produção interna apresenta uma grande representatividade para a empresa correspondendo a um percentual de 77,30% das ordens de produção. O valor baixo de produção externa comprova que a empresa detém um chão de fábrica robusto e efetivo para a concretização das ordens de serviço. Na Figura 26 verifica-se que a grande maioria dos componentes fabricados passa pelo controle de qualidade, o que representa um percentual de 66,61% do volume total. As variáveis ‘Envio para TS’ e ‘CQ-Entrada1’ foram plotadas na Figura 27.

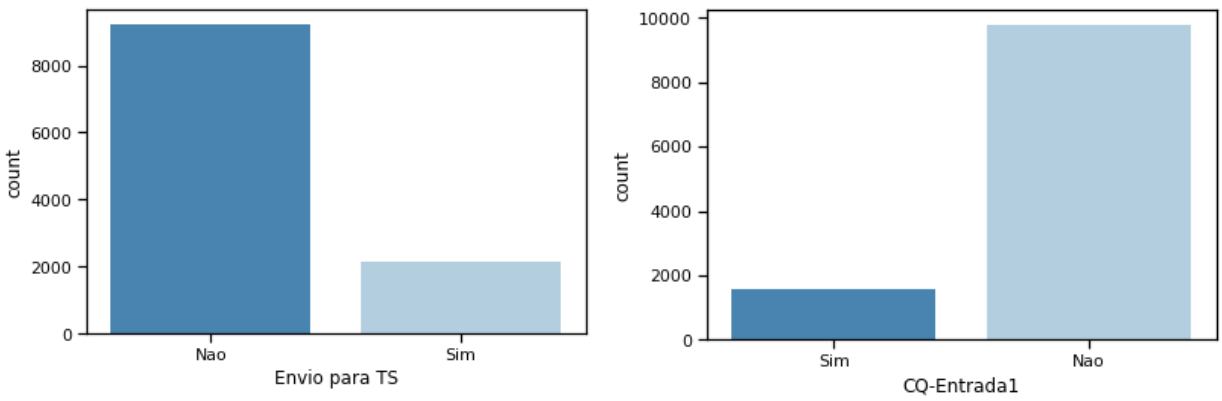

Figura 27- Histograma das variáveis: Envio para TS (esquerda) e CQ-Entrada1 (direita).

Na Figura 27 é possível identificar que um menor número de componentes passa pelo processo de tratamento superficial na empresa. Um percentual de 18,84% dos componentes fabricados necessita passar por esse tratamento de pintura química. O segundo controle de qualidade (CQ-Entrada1) é realizado após o tratamento superficial e corresponde a 13,94% dos componentes fabricados. Na Figura 28 visualiza-se a contagem da variável 'Material Liberado'.

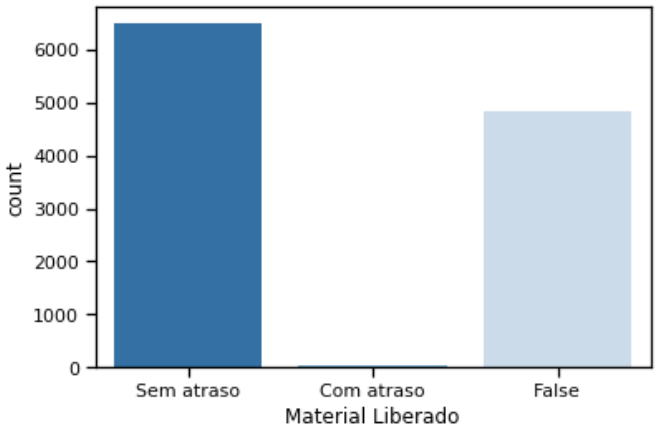

Figura 28- Histograma da variável material liberado.

Verifica-se, em análise da Figura 28, que para a ordem de produção interna o material geralmente é liberado sem atraso, o que confirma que a empresa não tem grandes problemas com prazos referentes a compra de materiais e estoque. O termo 'False' é utilizado para ordens de produção externa e ordens de produção interna que não tiveram o campo preenchido.

Após análise por meio de contagem das variáveis preditoras, avalia-se a classe ou a variável de saída para o modelo preditivo. Na Figura 29 o gráfico de dispersão da classe *lead-time* total pode ser visualizado.

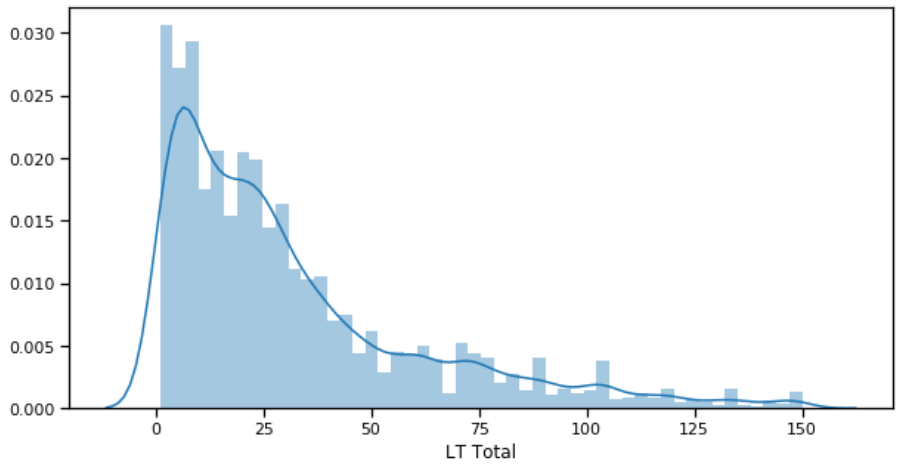

Figura 29- Gráfico de dispersão da classe LT total.

Com uma média de 31,98 dias e desvio padrão de 30,51 dias verifica-se que os dados para o LT total de fabricação dos componentes (Figura 29) não apresenta grandes padronizações. A curva de dispersão não é normal com média concentrada em valor quantitativo baixo se comparado ao escopo da variável e com alto desvio padrão.

Uma segunda análise realizada para avaliar o comportamento das variáveis preditoras e da classe foi a plotagem de diagramas de extremos e quartis. Os gráficos foram construídos com o objetivo de mapear o comportamento de algumas variáveis preditoras a partir do relacionamento com a variável classe.

Identificam-se os valores máximos e mínimos, o centro de distribuição com a mediana (linha no centro da caixa), o valor da média (triângulo representado no terceiro quartil), os quartis de distribuição dos dados, e os *outliers* (pontos fora dos padrões e desenhados acima do diagrama de caixa). Os gráficos apresentados foram resultados que exibiram informações relevantes e sem prejuízos de visualização pelo tamanho das variáveis.

Na Figura 30 visualizam-se os diagramas de extremos e quartis das variáveis (diagrama de caixa) ‘Ordem de Produção’ e ‘CQ-Entrada’.

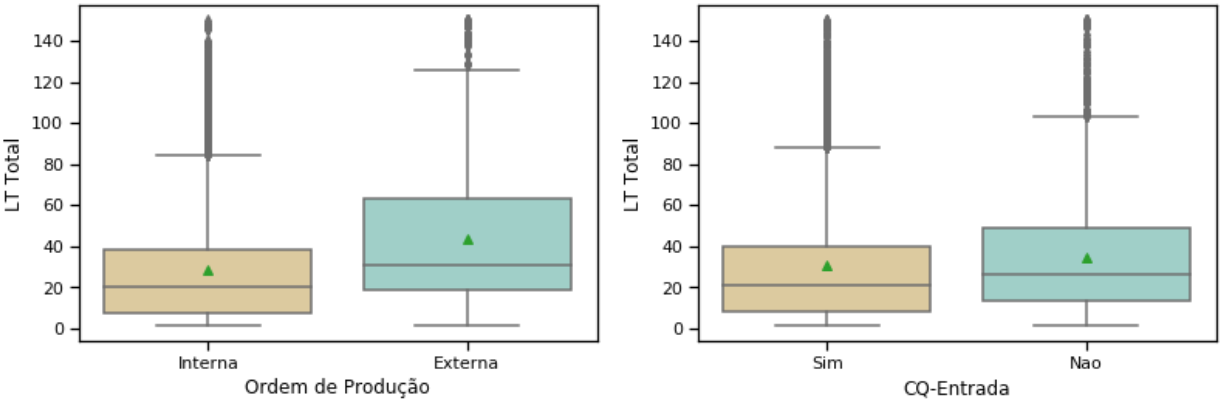

Figura 30- Diagrama de caixa ‘Ordem de Produção’ (Esquerda) e ‘CQ-Entrada’ (Direita).

Em análise da Figura 30 é possível identificar que a média de dias para o cumprimento da ordem de serviço de fabricação interna é menor que para a externa. Para a ordem de produção interna existe um padrão de dias de fabricação aproximado entre o intervalo de 0 a 80 dias, identificando uma faixa de *outliers* maior que para a produção externa, com intervalo aproximado de 0 a 120 dias.

Visualizando o gráfico para o controle de qualidade (Figura 30) verifica-se que existe certa padronização do *lead-time* total para componentes que passam pelo controle de qualidade e aqueles que não passam por essa etapa. Com médias, medianas e intervalos muito próximos é possível inferir que o controle de qualidade não é uma etapa que interfere substancialmente no *lead-time* total de fabricação de um componente.

Na Figura 31 são plotados os gráficos de extremos e quartis das variáveis ‘Envio para TS’ e ‘CQ-Entrada1’.

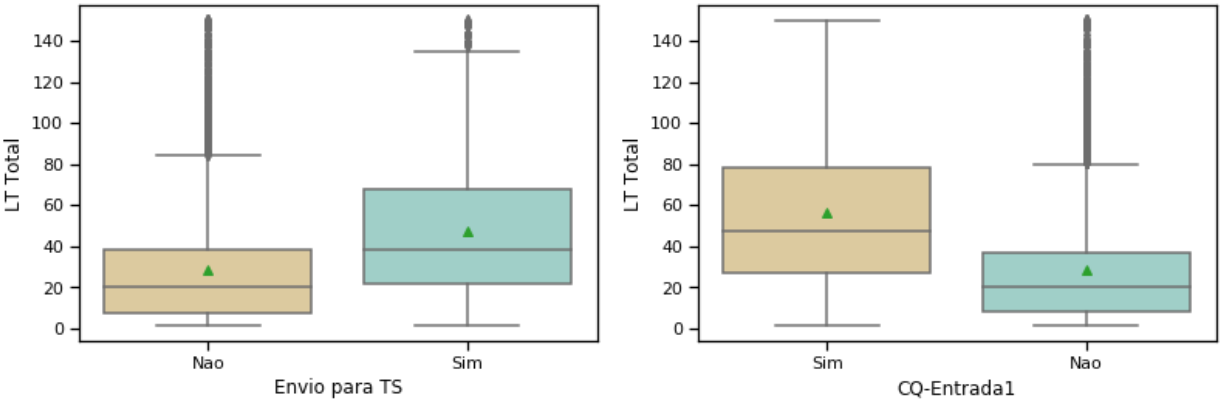

Figura 31- Diagrama de extremos e percentis das variáveis ‘Envio para TS’ (Esquerda) e ‘CQ-Entrada1’ (Direita).

Constata-se que o tratamento superficial interfere no *lead-time* total de fabricação do componente com média muito maior para peças que passam por esse procedimento. Um menor valor para a faixa de *outliers* foi encontrado para os componentes que são tratados quimicamente o que pode corroborar para uma maior padronização do *lead-time* total desses componentes.

Analisando o gráfico para o segundo controle de qualidade de componentes fabricados (Figura 31) verifica-se que os componentes que passam por esse procedimento alcançam maiores valores de *lead-time* total, informação compreensível por se tratar de um controle de qualidade posterior ao tratamento superficial. Importante se faz ressaltar a inexistência de valores *outliers* para os componentes que passam por esse controle de qualidade, e para os que não passam existe um tempo estimado, para a execução da ordem de serviço, compreendido entre os valores de 0 a 80 dias.

5.6 APLICAÇÃO DOS ALGORITMOS DE APRENDIZAGEM

Após a análise das variáveis da planilha de dados e da determinação das variáveis preditoras, a próxima análise da pesquisa consiste em aplicar os métodos de aprendizagem por redes neurais artificiais (RNA), do algoritmo SVM, e do *Random Forest* para prever o tempo de fabricação dos componentes para protótipos da empresa estudada. Ressalta-se que os três algoritmos utilizados foram selecionados pela incidência em pesquisas científicas já expostas na fundamentação teórica. Os parâmetros para o projeto dos algoritmos e o modelo de validação são apresentados, assim como as métricas de desempenho através da acurácia, do erro médio absoluto e do erro quadrático médio.

Para avaliar o desempenho do algoritmo de aprendizagem por redes neurais é importante, primeiro, estabelecer alguns parâmetros da rede neural. Alguns desses parâmetros foram selecionados de acordo com recomendações da literatura e outros por testes com os dados do problema proposto.

O número de camadas ocultas, assim como, a quantidade de neurônios para cada camada foram avaliados a partir de testes. A partir de avaliação da mesma divisão do bloco de dados pelo método de validação *k-fold* ($k = 10$) foi possível identificar o desempenho do número de camadas e a quantidade de neurônios para a solução do problema proposto a partir da acurácia. Como não é recomendado o uso de grandes quantidades de camadas escondidas, os testes foram limitados a um máximo de duas camadas ocultas. A variação de valores para o

número de neurônios foi de 50 até um limite máximo de 500 neurônios para cada camada. Na Tabela 4 é possível verificar os parâmetros selecionados para compor o treinamento e a validação da rede neural.

Tabela 4- Valores dos parâmetros para o projeto de redes neurais.

Parâmetro	**Condição**
Tipo de Treinamento	Supervisionado
Tipo de problema	Classificação
Arquitetura	Perceptron de múltiplas camadas
Entrada	[conjunto, QTD, Ordem de Produção, Descrição do Material, CQ- Entrada, Envio para TS, CQ-Entrada1, Repasse GI-EAPD, Material Liberado, Tempo de espera, LT produção]
Classe (saída)	LT Total
Algoritmo para otimização do peso	'adam'
Número de camadas ocultas	2
Número de neurônios nas camadas ocultas	[300,300]
Número máximo de épocas	1000
Tolerância para a função de perda	0,0001
Função de ativação	'relu'
Taxa de aprendizagem	0,001
Fator *momentum*	0,3
Método de validação	*k-fold* (*k*=10)

O segundo algoritmo de aprendizado avaliado foi o SVM. Na Tabela 5 é possível verificar os parâmetros e as condições utilizadas para o projeto do algoritmo SVM. Parâmetros não apresentados foram usados como o padrão (*default*) utilizado pela biblioteca do algoritmo na linguagem python.

Tabela 5- Valores dos parâmetros para o projeto do algoritmo SVM.

Parâmetro	**Condição**
Tipo de Treinamento	Supervisionado
Tipo de problema	Classificação
Entrada	[conjunto, QTD, Ordem de Produção, Descrição do Material, CQ- Entrada, Envio para TS, CQ-Entrada1, Repasse GI-EAPD, Material Liberado, Tempo de espera, LT produção]
Classe (saída)	LT Total
kernel	'linear'
Tolerância para a função de perda	0,0001
Função de decisão	'ovr'
Método de validação	*k-fold* (*k*=10)

O terceiro algoritmo de aprendizagem utilizado foi o *Random forest*. O número de árvores para o algoritmo *Random forest* foi escolhido a partir de testes que variaram de 0 a 100 com

intervalo de 5 arvores. O melhor resultado está apresentado pela Tabela 6, onde os parâmetros e condições para o algoritmo *Random forest* foram apresentados. Parâmetros não apresentados foram usados como o padrão (*default*) utilizado pela biblioteca do algoritmo na linguagem python.

Tabela 6- Valores dos parâmetros para o projeto do algoritmo Random Forest.

Parâmetro	**Condição**
Tipo de Treinamento	Supervisionado
Tipo de problema	Classificação
Entrada	[conjunto, QTD, Ordem de Produção, Descrição do Material, CQ- Entrada, Envio para TS, CQ-Entrada1, Repasse GI-EAPD, Material Liberado, Tempo de espera, LT produção]
Classe (saída)	LT Total
Número de árvores	70
Critério	Entropia
Método de validação	*k-fold* (k=10)

Para avaliar o desempenho do modelo de aprendizagem em predizer o tempo de fabricação de componentes da empresa quatro topologias de análises foram avaliadas. A primeira topologia de análise utilizou as informações dos dados a partir do pré-processamento já exposto pelo trabalho.

Nas demais topologias de análise os parâmetros com maior número de variáveis (Conjunto e Descrição do Material) foram reduzidos a partir de avaliação do diagrama de Pareto, em que avalia os 20% dos itens da planilha que representam cerca de 80% do total de dados. Essa redução dos dados teve por objetivo avaliar o comportamento de aprendizagem da rede neural, do algoritmo SVM e do *Random Forest* a partir dos dados restritos dessas variáveis com maior representatividade nas ordens de serviço da empresa.

Na Figura 32 é possível avaliar o diagrama de Pareto para a variável preditora ‘Descrição do Material’. Verifica-se que 81,57% dos materiais da planilha correspondem aos materiais alumínio, AISI 304, AL 7075, AL 6061, AL 6351, Latão, Poliacetal, kovar, AL 1100, e Inox. Esses foram os materiais selecionados para compor a segunda topologia de análise.

Na terceira topologia de análise o mesmo procedimento foi realizado para a variável preditora ‘Conjunto’. O gráfico de Pareto foi analisado e os conjuntos que correspondem a 80% do total de dados foram filtrados. Optou-se por não apresentar o gráfico de Pareto pela quantidade de informações visuais que foram geradas.

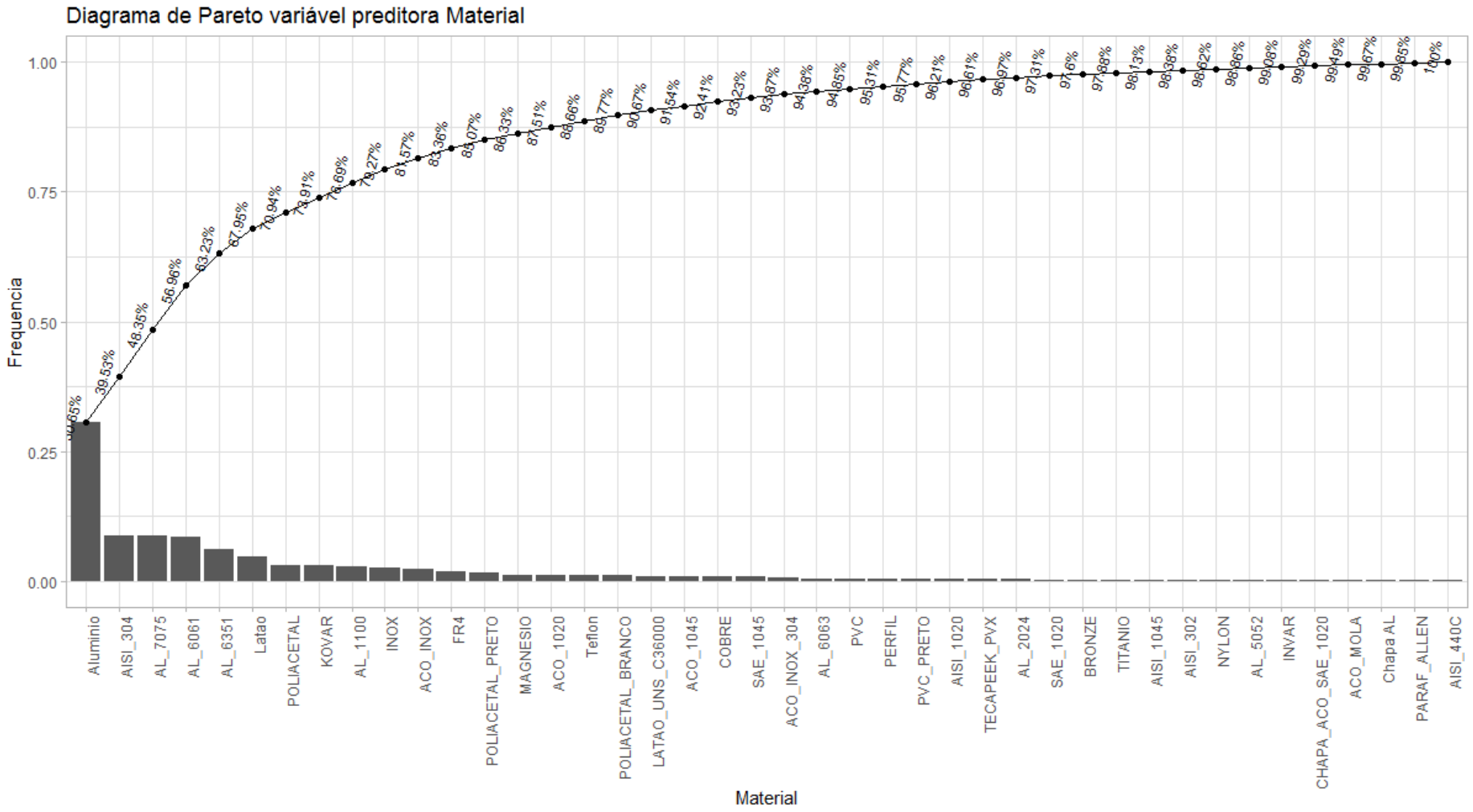

Figura 32- Diagrama de Pareto para a variável preditora ‘Descrição do Material’.

.

A quarta topologia de análise para avaliação do comportamento de aprendizagem da rede neural foi uma junção dos conjuntos e materiais mais representativos para o processo de fabricação da empresa avaliada. Na Tabela 7 são apresentadas as taxas de acerto (acurácia), e seus respectivos desvios (em dias), o erro médio absoluto (EMA), e o erro quadrático médio (EQM) da Rede Neural Artificial (RNA), do *Support Vector Machine* (SVM) e do algoritmo *Random forest* (RF) para cada topologia avaliada.

Tabela 7- Taxas de acerto da rede neural e do SVM para cada topologia analisada.

Topologia	**RNA**			**SVM**			**RF**		
	Acurácia	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**
1	71,67(0,98)	3,50	118,17	71,46(0,57)	3,44	111,75	71,38(1,16)	4,08	146,46
2	71,74(1,22)	3,02	85,43	70,96(1,23)	3,02	82,46	72,89(1,34)	3,50	116,75
3	70,61(1,94)	3,74	133,03	70,23(1,93)	3,43	106,04	70,39(1,40)	4,21	156,52
4	69,58(1,57)	3,15	85,45	69,02(1,67)	3,11	82,69	69,23(1,01)	3,78	122,02

Verifica-se a partir dos valores da acurácia e dos erros entre os valores de treinamento e teste que houve uma proximidade entre o comportamento das redes neurais, do algoritmo SVM e do *Random forest*. Uma taxa de acerto aproximada de 70% foi encontrada e um erro médio absoluto próximo a três dias. Os desvios para o valor da acurácia ficaram em torno de 1% o que comprova uma estabilidade do modelo para a avaliação dos dados da planilha analisada.

Na Tabela 8 é possível visualizar o erro relativo que corresponde ao percentual do erro médio absoluto em relação ao valor do *lead-time* total para as topologias analisadas. Verifica-se que o erro médio absoluto dos algoritmos de treinamento correspondeu a um valor entre 10,46% (SVM) e 13,16% (*Random forest*) do valor do *lead-time* total de fabricação.

Tabela 8- Erro relativo para cada topologia analisada.

Algoritmo	**Topologia**	**% correspondente do LT total**
RNA	1	10,94%
	2	10,80%
	3	11,41%
	4	11,40%
SVM	1	10,76%
	2	10,80%
	3	10,46%
	4	11,26%
RF	1	12,76%
	2	10,94%
	3	13,16%
	4	11,82%

Com o intuito de verificar se existe diferença significativa entre os algoritmos empregados aplicou-se o teste de Friedman e o teste *post-hoc* de Nemenyi. O teste de Friedman é um teste estatístico não paramétrico que classifica os algoritmos para cada conjunto de dados com o melhor desempenho. Quando a hipótese nula é rejeitada $p - valor < 0{,}05$ (nível de significância) conclui-se que existe pelo menos um (1) modelo com desempenho significativamente diferente dos demais (Demsar, 2006). Quando a hipótese nula é rejeitada pelo teste de Friedman, com uma confiança de 95%, verifica-se que existem diferenças significativas. Torna-se necessário, portanto, realizar os testes *post-hoc* para identificar o valor e comportamento estatístico dessas diferenças (Pereira *et al.*, 2015).

Os resultados do teste de verificação de diferença estatística *post-hoc* de Nemenyi são plotados em um diagrama com o ranking médio dos algoritmos e sua diferença crítica. Na Figura 33 apresenta-se o resultado para os testes de Friedman e Nemenyi, desenvolvidos com o pacote '*tsutils*' da linguagem R, verificando a diferença estatística entre os algoritmos RNA, SVM e *Random forest* para o conjunto de dados da planilha de testes para a Topologia 1 a partir de uma verificação de 30 testes com variação entre os conjuntos de treinamento e testes da base de dados. Verifica-se que a hipótese nula foi rejeitada o que confirma que pelo menos um (1) modelo tem desempenho diferenciado dos restantes.

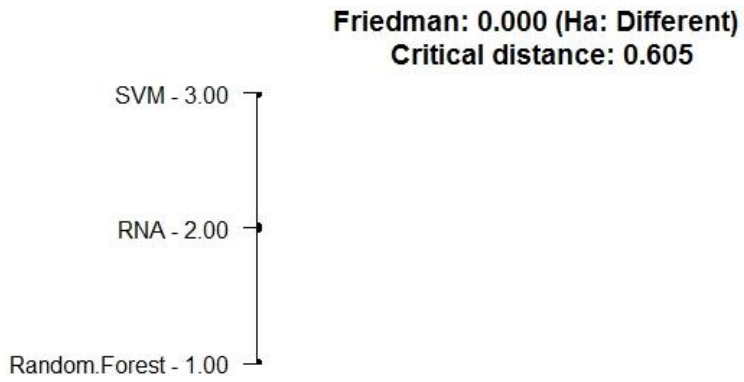

Figura 33- Diagrama para o teste de Friedman e Nemenyi para a tabela de dados completa.

Verifica-se em análise da Figura 33 que existe diferença significativa entre o resultado dos três algoritmos avaliados a partir da verificação da distância crítica de $0{,}605$. O algoritmo *Random Forest* apresentou o melhor resultado em contrate com os piores erros médios absolutos encontrados. Em segundo lugar de desempenho ficou o algoritmo RNA e por último o algoritmo SVM.

Com o objetivo de avaliar o comportamento da aprendizagem da rede neural, do SVM e do algoritmo *Random forest* diante das informações de dados disponibilizados pela empresa, realizou-se uma avaliação para os quatro tipos de materiais mais recorrentes no processo de fabricação (Aluminio, AISI_304, AL_7075, AL_6061).

5.6.1 ALGORITMOS DE APRENSIZAGEM APLICADOS AO MATERIAL ALUMÍNIO COMUM

A primeira análise de materiais da planilha de dados foi o alumínio, que revelou ser o material com maior representatividade na fabricação de componentes para protótipos e produtos ETO da empresa. Na Tabela 9 é possível visualizar as médias e os desvios padrões das variáveis quantitativas para a fabricação com alumínio.

Tabela 9- Média e desvio das variáveis quantitativas para o material alumínio.

Variável preditora (x)	**Média ($\bar{x}$)**	**Desvio padrão (σ)**
QTD	1,96	1,68
Repasse GI-EAPD	0,19	0,73
Tempo de Espera	0,40	1,23
LT Necessidade	12,24	14,66
LT Produção	16,17	18,94
LT Total	22,62	21,05

Comparando a Tabela 9 que expõe as médias e seus desvios para a tabela de dados filtrada pelo material alumínio e a tabela de dados sem esse filtro, verifica-se que, de maneira geral, que os dados apresentaram valores menores para as médias, com maiores diferenças para o *lead-time* de necessidade com uma redução de 68,25% e o *lead-time* total com uma redução de 29,27%.

Duas topologias de dados foram utilizadas para verificar o desempenho dos algoritmos de aprendizagem com o filtro de material alumínio. As melhores condições para os parâmetros para a arquitetura da RNA foi de uma (1) camada com 500 neurônios e para o Random Forest 60 árvores de decisão. As demais características dos algoritmos foram idênticas ao problema anterior. Para o algoritmo SVM os mesmos parâmetros foram utilizados.

Em uma primeira avaliação utilizou-se apenas o filtro de material ‘Alumínio’ na planilha de dados já investigada. Posteriormente, uma análise pelo diagrama de Pareto, Figura 34, foi realizada buscando limitar a quantidade da variável ‘conjunto’. Na segunda topologia os

dados utilizados para a aprendizagem dos algoritmos foram reduzidos a partir da filtragem de 80% dos conjuntos mais representativos (Figura 34).

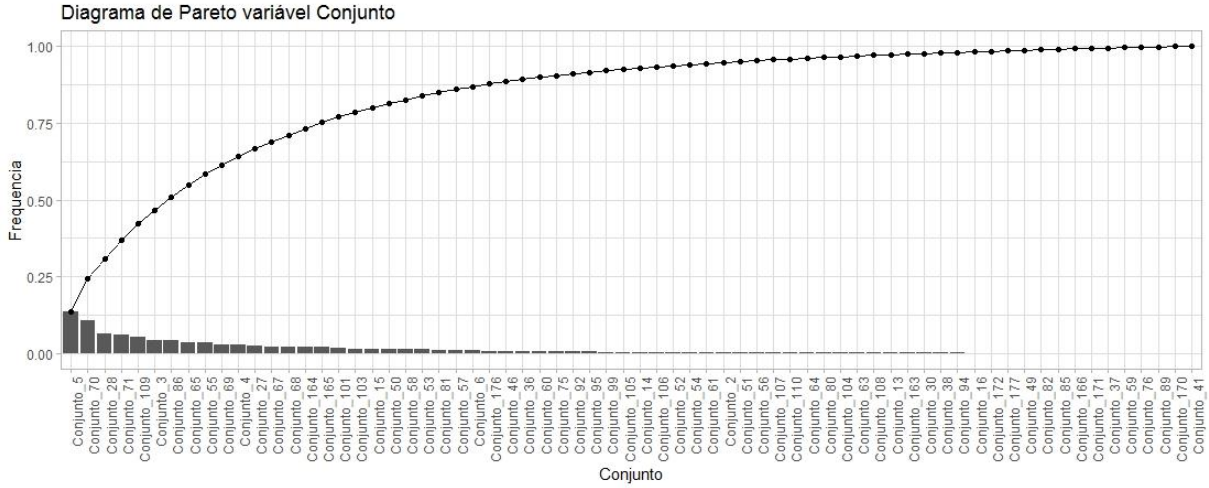

Figura 34- Diagrama de pareto para a variável preditora 'conjunto' do material alumínio.

O resultado para o desempenho dos algoritmos de Redes Neurais Artificiais (RNA), *Support Vector Machine* (SVM), e do *Random Forest* (RF) para o filtro do material alumínio com os resultados da acurácia, seu desvio padrão, o erro médio absoluto (EMA), e o erro quadrático médio (EQM), são apresentados na Tabela 10.

Tabela 10- Taxas de acerto e erros para a RNA e SVM para o material alumínio.

Topologia	**RNA**			**SVM**			**RF**		
	Acurácia	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**
1	69,91(1,82)	2,74	59,61	69,94(2,02)	2,76	59,76	68,54(2,90)	3,42	91,96
2	70,33(2,26)	2,49	48,08	70,68(2,31)	2,56	53,49	68,42(2,51)	3,34	89,63

Verifica-se que a taxa de acerto dos algoritmos RNA e SVM para o filtro do material alumínio ficaram muito próximas à taxa de acerto para a tabela de dados completa (Tabela 7). Uma menor acurácia foi encontrada para o algoritmo *Random Forest*.

Correspondendo a um percentual de 30,65% do total de materiais utilizados pela empresa, apesar de uma taxa de acerto similar apresentou um erro absoluto médio, de predições erradas, abaixo do encontrado para a tabela de dados completa (sem o filtro para o material alumínio).

Para verificar o erro relativo, valor percentual correspondente do erro para o valor do *lead-time* de fabricação total do material alumínio, a Tabela 11 foi construída. Destaca-se o valor

percentual de 8,93% para a topologia 1 do algoritmo SVM que apresentou o menor percentual. Maiores percentuais de erro foram encontrados para o algoritmo *Random Forest*.

Tabela 11- Erro relativo por tipo de algoritmo e suas topologias para a fabricação com o material alumínio comum.

Algoritmo	Topologia	% correspondente do LT Total
RNA	1	12,11%
	2	10,79%
SVM	1	8,93%
	2	11,09%
RF	1	15,12%
	2	14,76%

Para verificar se existe diferença estatística significativa entre os algoritmos apresentados para o tipo de material alumínio recorreu-se aos testes de Friedman e Nemenyi. O diagrama da Figura 35 apresenta os resultados encontrados após realização de trinta testes. Verifica-se que a hipótese nula foi rejeitada o que comprova a diferença significativa entre a acurácia dos algoritmos testados.

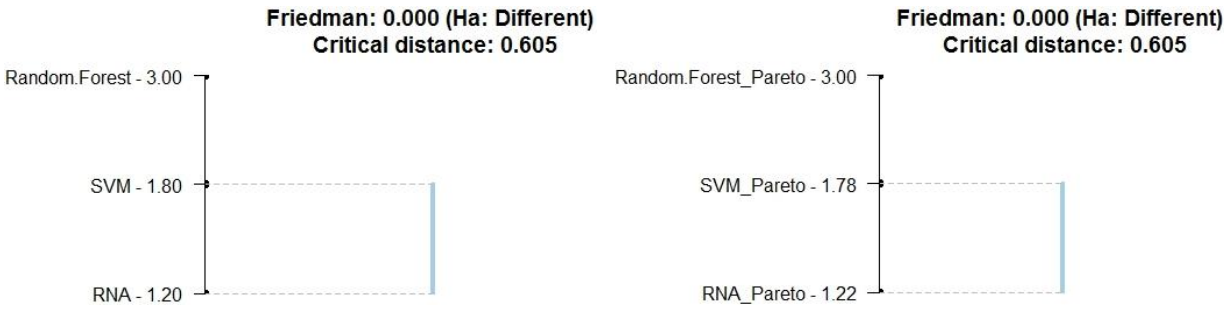

Figura 35- Diagrama para o teste de Friedman e Nemenyi da topologia 1 (esquerda) e topologia 2 (direita) do material alumínio.

Verifica-se, em análise da Figura 35, que o melhor desempenho encontrado foi dos algoritmos RNA e SVM, não tendo diferença estatística entre os dois métodos. O algoritmo *Random Forest* (RF) apresentou os piores resultados para o tipo de material filtrado.

5.6.2 ALGORITMOS DE APRENDIZAGEM APLICADOS AO MATERIAL AISI_304

A próxima análise realizada foi para o filtro de materiais AISI_304, o segundo material com maior representatividade para a fabricação dos componentes de protótipo da empresa. Na Tabela 12 os dados da média e desvio padrão das variáveis quantitativas para a fabricação com AISI_304 são apresentados.

Tabela 12- Média e desvio das variáveis quantitativas para o material AISI_304.

Variável preditora (x)	Média ($\bar{x}$)	Desvio padrão (σ)
QTD	2,69	3,57
Repasse GI-EAPD	0,25	0,72
Tempo de Espera	0,76	3,10
LT Necessidade	12,29	20,49
LT Produção	19,62	20,21
LT Total	25,92	24,45

Comparando os resultados da Tabela 12 que expõe as médias e seus desvios para a tabela de dados filtrada pelo material AISI_304 e a tabela de dados sem esse filtro, verifica-se que, os dados que tiveram maiores diferenças percentuais, para os valores das médias, foram o tempo de espera com redução de 39,68% e o repasse GI-EAPD com um acréscimo de 25%. O *lead-time* total teve uma queda percentual de 18,95% para o valor da média se comparado à tabela de dados sem o filtro de material AISI_304.

Duas topologias de dados foram utilizadas para verificar o desempenho dos algoritmos de aprendizagem com o filtro de material AISI_304. O diagrama de Pareto para análise da segunda topologia do material AISI_304 é apresentado pela Figura 36.

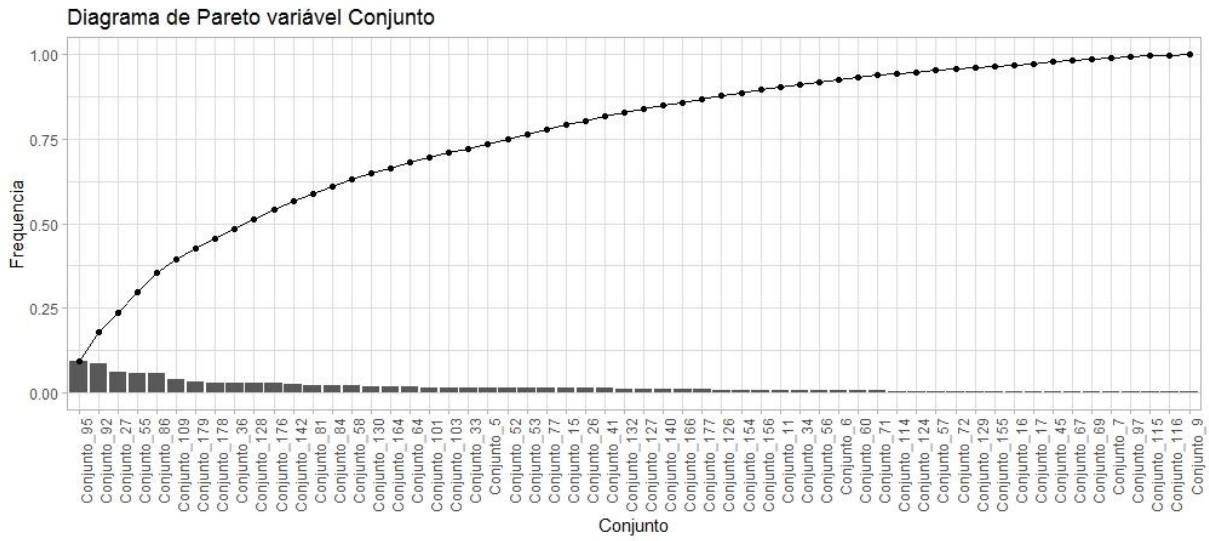

Figura 36- Diagrama de Pareto para a variável preditora ‘conjunto’ do material AISI_304.

Os melhores resultados para o RNA foi de uma camada escondida com 300 neurônios, e para o *Random Forest* (RF) 60 árvores de decisão. Para o algoritmo SVM os mesmos parâmetros foram utilizados. O resultado para o desempenho dos algoritmos de RNA, SVM e RF para o filtro do material AISI_304 com os resultados da acurácia, seu desvio padrão, o

erro médio absoluto (EMA), e o erro quadrático médio (EQM), são apresentados na Tabela 13.

Tabela 13- Taxas de acerto e erros da RNA e do SVM para o material AISI_304.

Topologia	RNA			SVM			RF		
	Acurácia	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**
1	69,15(5,29)	3,07	61,80	71,45(4,61)	2,82	62,65	69,03(4,36)	3,77	97,75
2	71,26(3,60)	2,76	52,76	73,69(4,37)	2,54	47,44	70,83(3,55)	3,13	70,71

Verifica-se que as taxas de acerto dos algoritmos RNA e SVM ficaram próximas ao encontrado para as demais análises realizadas. O algoritmo SVM merece destaque para a análise do material AISI_304 por encontrar melhores acurácias e um erro médio absoluto menor que o algoritmo RNA.

Para verificar o erro relativo, valor percentual correspondente do erro para o valor do *lead-time* de fabricação total do material AISI_304, a Tabela 14 foi construída. Visualiza-se que os percentuais encontrados ficaram próximos aos encontrados da tabela sem o filtro de materiais AISI_304.

Tabela 14- Erro relativo por tipo de algoritmo e suas respectivas topologias para a fabricação com o material AISI_304.

Algoritmo	**Topologia**	**% correspondente do LT Total**
RNA	1	11,84%
	2	11,55%
SVM	1	10,88%
	2	10,63%
Random Forest	1	14,54%
	2	12,08%

Analisando os resultados encontrados pelos algoritmos de treinamento e o percentual do erro médio absoluto correspondente da variável de saída, verifica-se que estimar o *lead-time* dos materiais AISI_304 e do alumínio comum, por RNA, SVM ou *Random Forest*, corresponde a obter estimativas de *lead-time* total de fabricação dos componentes com acurácia muito semelhantes. Destaca-se da Tabela 14 os maiores percentuais dos erros médios absolutos do algoritmo *Random Forest*, o que comprova maiores erros nas estimativas erradas de *lead-time*.

Para verificar se existe diferença estatística significativa entre os algoritmos apresentados para o tipo de material AISI_304 recorreu-se aos testes de Friedman e Nemenyi. O diagrama da Figura 37 apresenta os resultados encontrados após realização de trinta testes. Verifica-se

que a hipótese nula foi rejeitada o que comprova a diferença significativa entre a acurácia dos algoritmos testados.

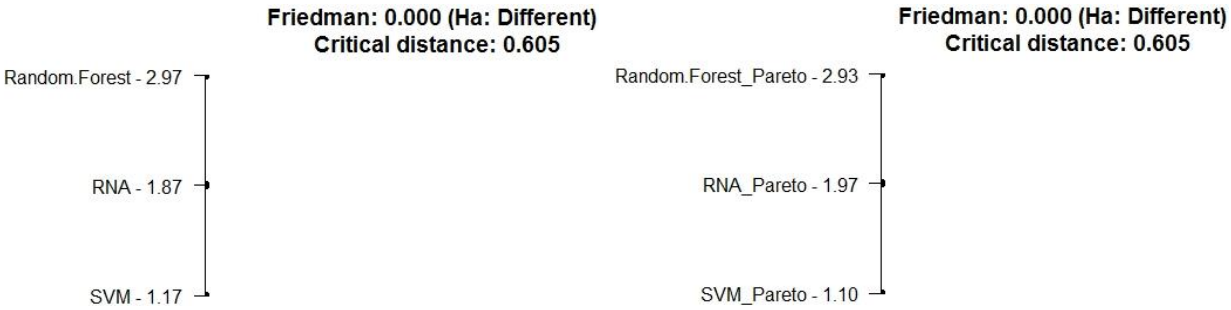

Figura 37- Diagrama para o teste de Friedman e Nemenyi da topologia 1 (esquerda) e topologia 2 (direita) do material AISI_304.

Verifica-se, em análise da Figura 37, que os algoritmos apresentaram diferença estatística significativa nos trinta testes. O algoritmo SVM teve o melhor desempenho, seguido das redes neurais artificiais (RNA), e por último o *Random Forest* com os piores resultados.

5.6.3 ALGORITMOS DE APRENDIZAGEM APLICADOS AO MATERIAL AL_7075

A próxima análise realizada foi para o filtro de materiais AL_7075, o terceiro material com maior representatividade para a fabricação dos componentes de protótipo da empresa. Na Tabela 15 os dados da média e do desvio padrão para as variáveis quantitativas da fabricação com AL_7075 são apresentados.

Tabela 15- Média e desvio das variáveis quantitativas para o material AL_7075.

Variável preditora (x)	**Média ($\bar{x}$)**	**Desvio padrão (σ)**
QTD	1,70	2,35
Repasse GI-EAPD	0,09	0,44
Tempo de Espera	1,00	4,07
LT Necessidade	24,51	37,66
LT Produção	32,28	28,13
LT Total	54,15	40,39

Comparando os resultados da Tabela 15 que expõe as médias e seus desvios (com o filtro do material AL_7075) e a tabela de dados sem esse filtro, verifica-se que, os dados que tiveram maiores diferenças percentuais foram o *lead-time* total com um acréscimo na média de 69,32%, o *lead-time* de necessidade com um acréscimo de 58,54% e o repasse GI-EAPD com uma redução de 55,00%.

Uma nova avaliação para os parâmetros dos algoritmos foi realizada com o intuito de obter os melhores resultados para o tipo de material AL_7075. Para o algoritmo RNA verificou-se o número de camadas e neurônios, e para o *Random Forest* o número de árvores da floresta. Para o algoritmo SVM os mesmos parâmetros foram utilizados. Verificou-se que uma (1) camada escondida com 100 neurônios foi o suficiente para chegar a bons resultados para o algoritmo RNA, sendo o valor utilizado para a construção do modelo. No caso do *Random Forest* um total de 60 árvores obteve os melhores resultados.

Duas topologias de dados foram utilizadas com os parâmetros e condições já expostos. O diagrama de Pareto para análise da segunda topologia é apresentado pela Figura 38.

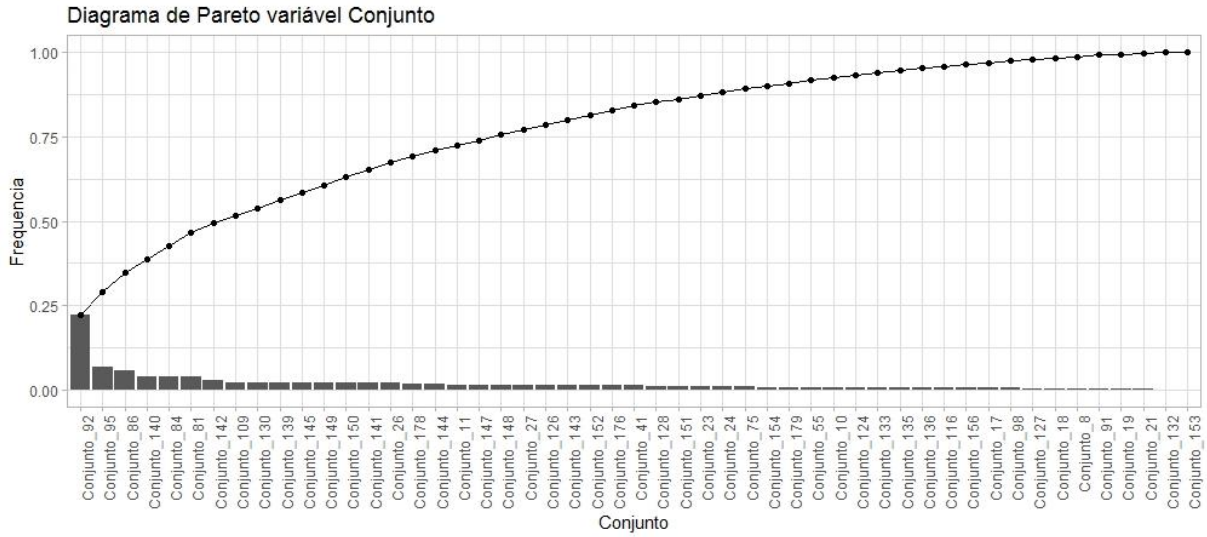

Figura 38- Diagrama de pareto para a variável preditora ‘conjunto’ do material AL_7075.

O resultado para o desempenho dos algoritmos RNA, SVM e *Random Forest* (RF) para o filtro do material AL_7075 com os resultados da acurácia, desvio padrão, erro médio absoluto, e o erro médio quadrático, são apresentados na Tabela 16.

Tabela 16- Taxas de acerto e erros da RNA e do SVM para o material AL_7075.

Topologia	**RNA**			**SVM**			**RF**		
	Acurácia	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**
1	82,19(3,75)	2,33	62,34	82,76 (3,76)	2,19	49,69	81,91(2,69)	4,03	180,48
2	82,51(8,46)	1,88	39,70	82,68(8,13)	2,10	46,66	81,14(8,43)	4,07	166,24

Verifica-se uma melhora significativa dos resultados encontrados para a taxa de acertos dos algoritmos RNA, SVM e *Random Forest* na avaliação de conjuntos fabricados com o

material AL_7075. Valores maiores que 80% para a taxa de acerto foram encontrados, verificando uma maior padronização do *lead-time* total de entrega da ordem de serviço para o material AL_7075. Desta maneira, pode-se afirmar que, para o material AL_7075, é possível predizer o tempo de fabricação total com uma maior precisão dos resultados.

Para verificar o erro relativo, valor percentual correspondente do erro para o valor do *lead-time* de fabricação total do material AL_7075, a Tabela 17 foi construída. Verifica-se que porcentagens maiores foram encontradas para o algoritmo *Random Forest* corroborando maiores erros em predições erradas deste algoritmo.

Tabela 17- Erro relativo por tipo de algoritmo e suas respectivas topologias para a fabricação com material AL_7075.

Algoritmo	**Topologia**	**% correspondente do LT Total**
RNA	1	4,30%
	2	3,60%
SVM	1	4,04%
	2	4,03%
Random Forest	1	7,44%
	2	7,52%

Visualiza-se, nas Tabelas 16 e 17, que os percentuais para o tipo de material AL_7075 apresentaram resultados muito positivos em relação às demais análises realizadas, possibilitando, dessa maneira, afirmar que é possível prever o tempo de fabricação dos componentes fabricados com AL_7075 com maiores precisões e melhores taxas de acerto. Para esse conjunto de dados, o algoritmo *Random Forest* apresentou os maiores valores para o erro médio absoluto. Importante ressaltar também a diferença significativa nos desvios padrões para a segunda topologia.

Com o intuito de verificar se existe diferença estatística significativa entre os algoritmos apresentados recorreu-se aos testes de Friedman e Nemenyi. O diagrama da Figura 39 apresenta os resultados encontrados após realização de trinta testes. Verifica-se que a hipótese nula foi rejeitada o que comprova a diferença significativa entre a acurácia dos algoritmos testados.

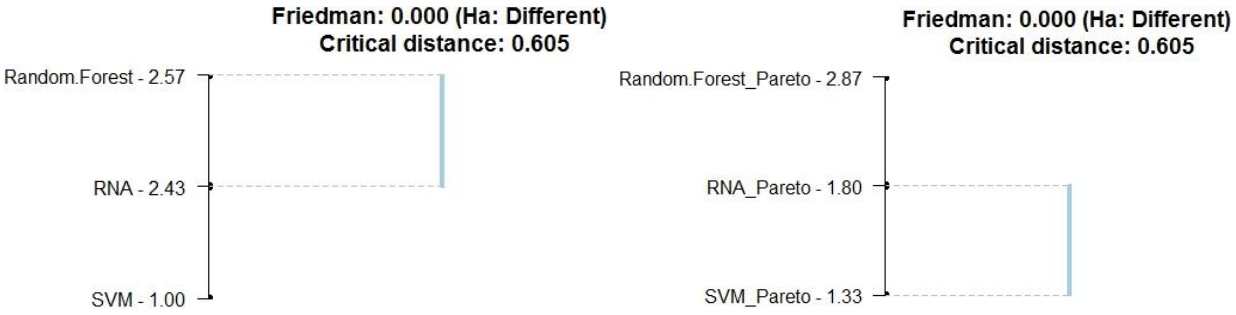

Figura 39- Diagrama para o teste de Friedman e Nemenyi da topologia 1 (esquerda) e topologia 2 (direita) do material AL_7075.

Verifica-se, em análise da Figura 39, que o algoritmo SVM teve o melhor desempenho para a topologia 1 com diferença de $1{,}43$ $(2{,}43 - 1{,}00)$ para o segundo colocado RNA. Para esta topologia não houve diferença significativa entre o algoritmo RNA e o *Random forest*. Para a topologia 2 os algoritmos SVM e RNA tiveram melhores desempenhos, e destacaram-se do *Random forest* (RF).

5.6.4 ALGORITMOS DE APRENDIZAGEM APLICADOS AO MATERIAL AL_6061

A última análise realizada foi para o filtro de materiais AL_6061, o quarto material com maior representatividade para a fabricação dos componentes de protótipo da empresa. Na Tabela 18 os dados da média e do desvio padrão para as variáveis quantitativas para a fabricação com AL_6061 são apresentados.

Tabela 18- Média e desvio das variáveis quantitativas para o AL_6061.

Variável preditora (x)	**Média ($\bar{x}$)**	**Desvio padrão (σ)**
QTD	1,37	1,04
Repasse GI-EAPD	0,31	0,85
Tempo de Espera	1,81	5,96
LT Necessidade	13,38	22,78
LT Produção	21,80	19,08
LT Total	38,09	27,01

Comparando os resultados da Tabela 18 que expõe as médias e seus desvios com o filtro do material AL_6061 e a tabela de dados sem esse filtro, verifica-se que, os dados que tiveram maiores diferenças percentuais para os valores da média foram o repasse GI-EAPD com um acréscimo de 55,00% e a quantidade de materiais com uma redução de 47,31%. Se comparado os valores das médias, o *lead-time* total teve um acréscimo de 19,10%.

Duas topologias de dados foram utilizadas com os melhores parâmetros e condições encontrados para os algoritmos de aprendizagem. Para as RNA, (1) camada escondida com 100 neurônios foi o suficiente para chegar a bons resultados. No caso do *Random Forest* um total de 60 árvores obteve os melhores resultados. Para o algoritmo SVM os mesmos parâmetros foram utilizados.

O diagrama de Pareto para análise da segunda topologia do material AL_6061 é apresentado pela Figura 40.

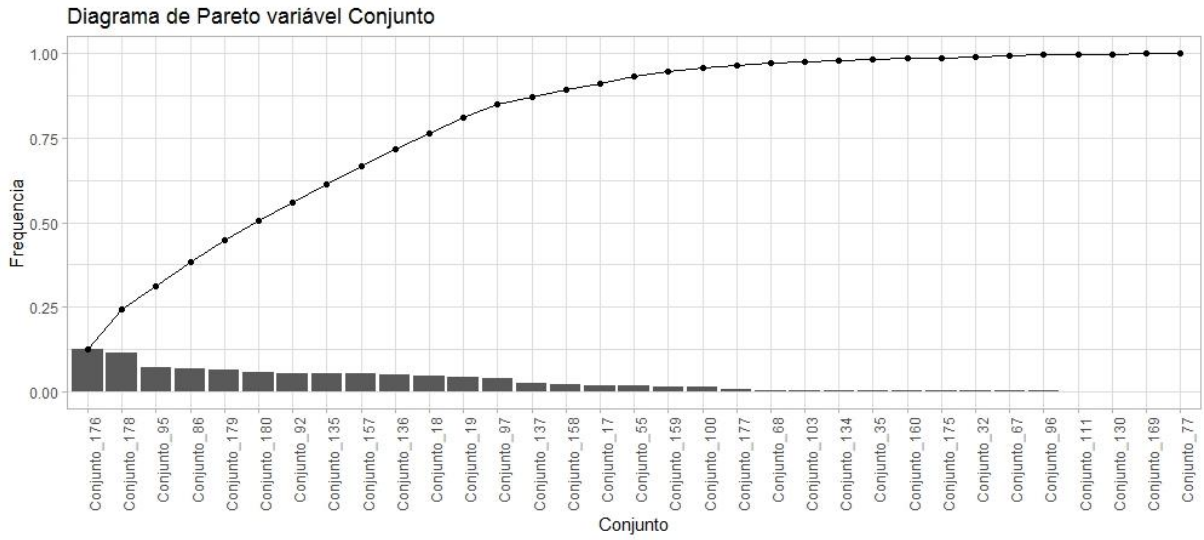

Figura 40- Diagrama de pareto para a variável preditora 'conjunto' do material AL_6061.

O resultado para o desempenho dos algoritmos de Redes Neurais Artificiais (RNA), *Support Vector Machine* (SVM), e do *Random Forest* (RF) para o filtro do material AL_6061 com os resultados da acurácia, seu desvio padrão, o erro médio absoluto, e o erro médio quadrático, são apresentados na Tabela 19.

Tabela 19- Taxas de acerto e erros da RNA e do SVM para o material AL_6061.

Topologia	**RNA**			**SVM**			**RF**		
	Acurácia	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**	**Acurácia**	**EMA**	**EQM**
1	85,21(5,17)	1,31	31,47	86,11(3,93)	1,02	19,78	86,88(4,22)	1,17	24,15
2	87,09(4,00)	0,72	10,05	87,87(4,18)	0,72	10,55	87,24(4,26)	0,89	15,99

Em análise da Tabela 19, verifica-se que o valor da taxa de acerto dos algoritmos de aprendizagem foi bastante promissor. Com valores médios acima de 85%, os modelos de aprendizagem apresentaram a melhor predição para esse tipo de material.

Destaca-se o valor do erro médio absoluto entre os valores preditos e os reais abaixo de um (1) dia para a segunda topologia. Isso significa que é possível prever o tempo de fabricação dos componentes de protótipos de material AL_6061 com taxa de acerto média de 87% e erros menores que um (1) dia.

Esses resultados corroboram a eficácia do método aplicado tornando-se uma ferramenta importante para o uso de gestores de engenharia da empresa, com foco em um maior controle do tempo total de fabricação para esse tipo de material.

Para verificar o erro relativo, valor percentual correspondente do erro para o valor do *lead-time* de fabricação total do material AL_6061, a Tabela 20 foi construída.

Tabela 20- Erro relativo por tipo de algoritmo e suas respectivas topologias para a fabricação com o material AL_6061.

Algoritmo	**Topologia**	**% correspondente do LT Total**
RNA	1	3,44%
	2	1,87%
SVM	1	2,68%
	2	1,87%
Random Forest	1	3,07%
	2	2,34%

Constata-se, em análise da Tabela 15 e 14, que os algoritmos de previsão conseguiram prever de maneira positiva o *lead-time* de fabricação total dos componentes para o material AL_6061, tornando o material com maiores padronizações de produção pela empresa estudada e, por consequência, o material com melhores resultados para predições de *lead-time.*

Com o intuito de verificar se existe diferença estatística significativa entre os algoritmos apresentados recorreu-se aos testes de Friedman e Nemenyi. O diagrama da Figura 41 apresenta os resultados encontrados. Verifica-se que a hipótese nula foi rejeitada o que comprova a diferença significativa entre os algoritmos testados.

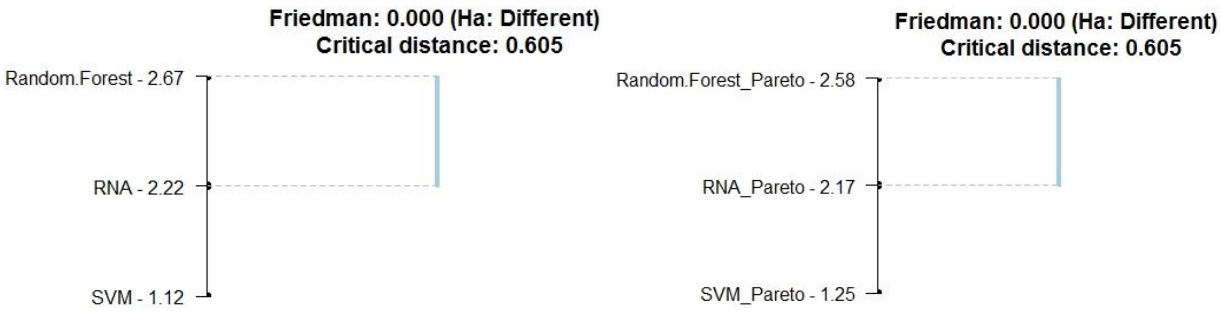

Figura 41- Diagrama para o teste de Friedman e Nemenyi da topologia 1 (esquerda) e topologia 2 (direita) do material AL_6061.

Verifica-se (Figura 41) que o comportamento de desempenho foi semelhante para as duas topologias analisadas. O algoritmo SVM teve melhor desempenho estatístico, já o RNA e *Random forest* (RF) não tiveram diferenças significativas entre si.

5.7 DISCUSSÕES

Os resultados apresentados, a partir dos dados disponibilizados da ordem de serviço de fabricação de componentes da empresa com foco no desenvolvimento de produtos mecatrônicos, demonstraram que as ordens de serviço apresentam baixa quantidade de componentes por ordem de serviço o que comprova uma tipologia de fabricação personalizada.

As variáveis que continham dados de fluxo de informação (Repasse GI-EAPD) na empresa corroboraram a eficiência da empresa nesse quesito após a reestruturação de sua logística de fabricação com o uso da planilha de rastreabilidade das peças fabricadas. Os baixos tempos de espera para a fabricação demostraram que a empresa dispõe de um chão de fábrica robusto para a fabricação dos itens de produção interna, mas que podem ser aprimorados para atingir uma excelência em cumprir o *lead-time* de necessidade.

O *lead-time* de necessidade apresentou valores de atraso médio de 6,79 dias em relação ao *lead-time* de produção, o que comprova uma necessidade em melhores estimativas do *lead-time* de produção e por consequência do *lead-time* total de fabricação.

Comprovou-se que existe um maior *lead-time* total para peças que são produzidas externamente e que passam pelo tratamento superficial, que por sua vez ocorre por terceirização. Peças que necessitam passar pelo primeiro controle de qualidade não apresentaram valores significativos de interferência no *lead-time* total do produto.

Os modelos de aprendizagem de máquina proporcionaram estimativas positivas e que podem ser utilizados pelos gerentes da empresa com o objetivo de melhorar o desempenho no processo de fabricação. Para a planilha de dados completa, assim como para o filtro de materiais ‘alumínio comum’ e ‘AISI_304’, o modelo construído apresentou uma taxa de acerto na estimativa do tempo de fabricação total aproximada de 70%.

A partir de diálogos com o gerente especialista da empresa constatou-se que as peças fabricadas com alumínio comum são menos críticas, não apresentam função estrutural ou em equipamentos ópticos, e abrange uma grande variedade em termos de complexidade de desenho. Já as peças fabricadas com AISI_304 apresentam uma geométrica muito variável, são difíceis de usinar e com passo de usinagem muito pequeno. O material AISI_304 é utilizado para materiais muito específicos do produto e para dispositivos de fabricação e teste. Essas especificidades de fabricação para esses dois tipos de materiais (Alumínio comum e AISI_304) auxilia na compreensão dos menores resultados para as taxas de acerto encontradas, se comparado com os resultados encontrados para outros materiais avaliados (AL_7075 e AL_6061).

Os materiais de destaque com maiores taxas de acerto e menores erros encontrados pelos algoritmos de aprendizagem constituíram de fabricação com os materiais AL_7075 e o AL_6061. Para o AL_7075 uma taxa de acerto aproximada de 82,35% foi encontrada para as redes neurais artificiais, 82,72% para o algoritmo SVM, e 81,52% para o *Random Forest*. Erros médios absolutos próximos a dois (2) dias foram encontrados nas predições dos algoritmos RNA e SVM, e para o RF o erro aproximado foi de 4 dias.

As peças fabricadas com o material alumínio 7075 são utilizadas para componentes de estrutura, ou seja, peças que suportarão esforço mecânico para o lançamento de câmeras de satélite e peças de suporte mecânico para placas eletrônicas. São peças com geometrias complexas, com grande comprimento e críticas, em termos de tolerância geométrica. As peças com o tipo de material AL_7075 são fabricadas, geralmente, por terceiros que foram qualificados para fabricar equipamentos aeroespaciais. Verifica-se, assim, que os melhores resultados encontrados para a fabricação com o alumínio 7075 comprovam a maior padronização e controle de fabricação para esse tipo de material.

O grande destaque do trabalho apresentado foram os resultados obtidos com a predição do *lead-time* total do material AL_6061. Uma taxa de acerto média aproximada de 86,53% para a RNA, 87% para o algoritmo SVM, e de 86,15% para o *Random Forest*, com erros absolutos médios menores que um (1) dia. A partir desses resultados é possível afirmar que existe uma

maior padronização na fabricação do AL_6061 o que torna esse material ter predições muito próximas do real.

A partir de diálogos com o gerente especialista da empresa, constatou-se que o material AL_6061 era muito crítico e usado em projetos espaciais em peças com interface com lentes e por onde passa o fluxo de lux para geração de imagem. Essas peças são mais similares entre si, todas demandam trabalhos no torno CNC (Controle Numérico Computadorizado), serviços de fresa, tratamento térmico e usinagem por interferência entre peças. O processo de fabricação das peças fabricadas por esse tipo de material passou por qualificação para uso aeroespacial. Desta maneira, os resultados encontrados com melhores taxas de acerto dos algoritmos de aprendizagem comprovam o maior controle de fabricação com o material alumínio 6061 pela empresa estudo de caso.

A partir dos testes de Friedman e Nemenyi foi possível comparar o desempenho dos algoritmos na construção dos modelos de aprendizagem. O algoritmo *Random Forest* obteve os melhores resultados na avaliação do conjunto de dados com maior quantidade de informações, sem nenhum filtro de material aplicado. Entretanto, não conseguiu ter bons resultados nos demais conjunto de dados com o filtro de materiais alumínio comum, AISI_304, AL_7075 e AL_6061. O algoritmo SVM teve o melhor desempenho ficando oito (8) vezes na liderança nos testes de avaliação de diferença significativa entre os algoritmos. O aprendizado por RNA ficou na liderança em três (3) testes.

Como houve uma mudança fluida de desempenho dos algoritmos para cada situação avaliada não é possível afirmar que existe uma superioridade de um método em relação ao outro na solução desse problema, com as variáveis e os dados específicos da empresa estudada. Entretanto, é possível confirmar que o SVM ficou na frente dos outros algoritmos por mais vezes.

A partir dos resultados apresentados comprova-se que é possível aplicar os algoritmos de aprendizagem de máquinas para realizar predições do *lead-time* total de fabricação de componentes para protótipos e ETO da empresa estudo de caso. As ordens de serviço mostraram-se uma base de dados potencial para realizar predições do tempo de fabricação com resultados bastante promissores para os engenheiros gestores.

Com o preenchimento dos dados da planilha de rastreabilidade associado aos modelos preditivos de aprendizagem de máquina os gestores engenheiros responsáveis podem adquirir uma maturidade em estimar o *lead-time* total de fabricação dos componentes a partir do

conhecimento gerado tanto pelo preenchimento da planilha, como na verificação da efetividade dos algoritmos no dia a dia. Verificaram-se alguns pontos que podem ser melhorados e estudados pela empresa como as grandes diferenças entre o *lead-time* de necessidade da peça e o *lead-time* real de fabricação, e o uso do método proposto pode auxiliar na aproximação entre os valores previstos e os valores reais.

5. CONCLUSÕES

Neste capítulo as conclusões do texto dissertativo são apresentadas. A partir do problema de pesquisa, dos objetivos, da metodologia e dos resultados, são determinadas as conclusões geradas pelo trabalho desenvolvido, as limitações do estudo e as sugestões para futuros trabalhos como meio motivador de pesquisa científica.

5.1 CONCLUSÕES

No texto dissertativo apresentado métodos de aprendizagem de máquina (RNA, SVM e *Random forest*) foram aplicados para solucionar um problema de estimativa do *lead-time* total de fabricação de componentes em uma empresa de base tecnológica. O problema de pesquisa foi apresentado como norteador do estudo, que buscou avaliar o desempenho dos algoritmos em encontrar respostas precisas na estimativa do tempo total de fabricação dos componentes para protótipos de produtos na empresa.

Os objetivos do trabalho foram gerados como modelo orientador da pesquisa dissertativa. A fundamentação teórica da pesquisa foi desenvolvida como base científica para amparar a metodologia e os resultados. A metodologia foi construída a partir do seu delineamento e procedimentos, levando em consideração os modelos de análise preditiva apresentados na fundamentação teórica.

Os resultados foram apresentados a partir de toda construção intelectual e em etapas da pesquisa evidenciando resultados promissores para os gerentes de fabricação da empresa estudada. Os melhores resultados de predição do tempo de fabricação total para componentes de protótipos conseguiram estimar com uma acurácia média maior que 86% com um erro médio absoluto abaixo de um dia de fabricação para o tipo de material alumínio 6061. Esses resultados comprovam a eficácia do método proposto, tornando-se uma ferramenta complementar para os gestores da empresa.

Comparando-se os métodos de aprendizagem de máquinas aplicados no trabalho dissertativo (RNA, SVM, e *Random Forest*), verificou-se que o algoritmo SVM teve melhor desempenho destacando-se oito (8) vezes, ou 66,67% dos testes, dos demais algoritmos. As redes neurais foram melhores em três (3) vezes, ou 25% dos testes. O algoritmo *Random Forest* apresentou melhores resultados para um conjunto de dados maiores da planilha de

rastreabilidade. Importante ressaltar, também, que maiores erros médios absolutos foram encontrados para o algoritmo *Random Forest*. A partir dos resultados apresentados não é possível indicar um melhor método de aprendizagem de máquinas generalista para o problema estudado.

Previsões mais precisas do *lead-time* total de fabricação podem gerar resultados efetivos na empresa no que tange à organização dos cronogramas de produção, afetando custos e os tempos de cumprimento das ordens de serviço. O método apresentado pode, também, ser utilizado por outras empresas visando estudar as taxas de acertos do modelo em aplicações distintas para a logística de fabricação.

Além das previsões no tempo de fabricação dos componentes, que apresentaram resultados promissores e aplicáveis, o estudo das variáveis que realizam a gestão da planilha de rastreabilidade revelou que o tempo de fabricação externa e o tratamento superficial interferem vertiginosamente no *lead-time* de fabricação, enquanto o primeiro controle de qualidade não gera tanta interferência. A análise dessas variáveis mapeou os fluxos de informações e materiais da empresa gerando resultados importantes que podem ser reveladores para a melhoria de desempenho da empresa.

Por fim, é importante reforçar que, visando a interlocução com o problema de pesquisa, os métodos de aprendizagem de máquina podem ser efetivos na estimativa do *lead-time* de fabricação de componentes. Para o caso da empresa estudada melhores taxas de acerto foram encontradas para tipos específicos de materiais. Apesar de que em algumas ocasiões estimativas não muito precisas foram geradas, elas podem, também, ser utilizadas e melhoradas com a inserção de novos dados de componentes de produtos com o passar dos anos. Além disso, a empresa pode verificar amiúde a eficácia do método, tornando-se mais uma ferramenta para esse fim. O estudo das ordens de serviço de rastreabilidade de peças mostrou-se um conjunto de dados eficiente para a análise com modelos preditivos por métodos de aprendizado de máquinas.

5.2 LIMITAÇÕES DA PESQUISA

A dissertação apresentada foi disponível e suscetível para a investigação devido ao relacionamento parceiro com o gestor da empresa estudada. Algumas limitações de pesquisa foram direcionadas ao processamento de dados e ao custo computacional exigido pelos algoritmos de aprendizagem de máquinas ao criar o modelo de predição. Com uma demanda

de horas de processamento para gerar resultados, ocorreram entraves em relação à amplitude de dados e as possibilidades de avaliação da planilha de dados, mas que foram solucionadas com o decorrer da pesquisa, principalmente com a mudança da ferramenta de aplicação.

Como se trata de um estudo de caso específico de uma empresa de base tecnológica que desenvolve produtos mecatrônicos para a área médica e espacial o estudo desenvolvido não pode ser generalizado a outras empresas. Entretanto, o método aplicado pode orientar pesquisas futuras que almejam estudar problemas de pesquisa semelhantes.

Um processo iterativo foi realizado durante as etapas de pré-processamento e aplicação das técnicas de aprendizagem no trabalho apresentado, buscando encontrar as melhores variáveis preditoras e taxas de acerto. Durante esse processo verificou-se que a estimativa do *lead-time* total teve melhores desempenhos ao incluir a variável preditora *lead-time* de produção, um relacionamento bastante compreensível.

Alguns testes foram realizados com o objetivo de realizar predições, também, do *lead-time* de produção, entretanto, não foram obtidos resultados promissores, ou seja, as taxas de acerto ficaram muito baixas. Em diálogos com o gerente especialista da empresa, verificou-se que é possível realizar uma previsão do *lead-time* de produção, dado esse que entraria como entrada do modelo, entretanto não existe, ainda, uma precisão dessa estimativa pelos gestores da empresa.

Portanto, o estudo apresentado com as melhores taxas de acerto encontradas, após análise de todas as variáveis da planilha de rastreabilidade, construiu um modelo com boas previsões, mas que depende de uma boa estimativa do *lead-time* de produção para ter resultados práticos promissores. A partir dessas considerações realizam-se sugestões para trabalhos futuros.

5.3 SUGESTÕES PARA TRABALHOS FUTUROS

A partir das limitações de pesquisa apresentadas, sugere-se, um estudo com foco em identificar outras variáveis preditoras da empresa que possam auxiliar a previsão do *lead-time* de produção, uma variável de entrada para o modelo proposto pela dissertação. Como sugestões, indica-se o estudo da complexidade de desenho das peças fabricadas, que podem correlacionar-se de maneira positiva com o *lead-time* de produção. A inclusão dessa variável pode melhorar o desempenho tanto da estimativa do *lead-time* de produção com para o *lead-time* total.

Sugere-se, também, a aplicação de métodos mistos de aprendizagem de máquinas como meio de observação do desempenho dos algoritmos, o que pode melhorar as taxas de acerto e diminuir os erros encontrados.

Além disso, sugerem-se avaliações similares com empresas do mesmo ramo ou com o mesmo potencial tecnológico da empresa estudada, a título de pesquisa comparativa com o trabalho apresentado.

REFERENCIAS BIBLIOGRAFICAS

AHER, S. B.; LOBO, L. M. R. J. Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. **Knowledge-Based Systems**, vol. 51, p. 1-14, 2013.

ALENEZI, A.; MOSES, S. A.; TRAFALIS, T. B. Real-time prediction of order flowtimes using support vector regression. **Computers & Operations Research**, v. 35, n. 11, p. 3489-3503, 2008.

ASADZADEH, S. M.; AZADEH, A.; ZIAEIFAR, A. A neuro-fuzzy-regression algorithm for improved prediction of manufacturing lead time with machine breakdowns. **Concurrent Engineering**, v. 19, n. 4, p. 269-281, 2011.

BALLABIO, D.; GRISONI, F.; TODESCHINI, R. Multivariate comparison of classification performance measures. **Chemometrics and Intelligent Laboratory Systems**, v. 174, p. 33-44, 2018.

BARBALHO, S. C. M.; TORRES, L. Melhoria de indicadores de desempenho em desenvolvimento de produtos por meio do enfoque no aumento da capabilidade dos processos: o caso da fabricação de protótipos de novos produtos. In: XV SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, 2008, São Paulo. **Anais**. São Paulo, SP: Unesp, 2008. p. 1-12.

BASHIR, H. A. Modeling of development time for hydroelectric generators using factor and multiple regression analyses. **International Journal of Project Management**, v. 26, n. 4, p. 457-464, 2008.

BOCCATO, L. **Novas propostas e aplicações de redes neurais com estados de eco**. 2013. Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2013.

BOLAÑOS, R.D.S. (2017). Proposta metodológica para estimar o tempo de desenvolvimento de projetos de desenvolvimento de produtos: estudo exploratório em empresa de base tecnológica. Publicação ENM.DM-128/17, Departamento de Engenharia Mecânica, Universidade de Brasília, Brasília, DF, xvii, 102p.

BOWLES, M. **Machine learning in Python**: essential techniques for predictive analysis. John Wiley & Sons, 2015.

BRAGA, A. P.; FERREIRA, A. C. L.; LUDERMIR, T. B. **Redes neurais artificiais**: teoria e aplicações. Rio de Janeiro, Brazil: LTC Editora, 2007.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of control, signals and systems**, v. 2, n. 4, p. 303-314, 1989.

DE COS JUEZ, F. J. et al. Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model. **Mathematical and Computer Modelling**, v. 52, n. 7-8, p. 1177-1184, 2010.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine learning research**, v. 7, n. Jan, p. 1-30, 2006.

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

DUBOIS, P. F. Guest Editor's Introduction: Python--Batteries Included. **Computing in Science & Engineering**, v. 9, n. 3, p. 7, 2007.

FAYYAD, U.; PIATETSKY, S. G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

GARCIA, S.; HERRERA, F. An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. **Journal of Machine Learning Research**, v. 9, n. Dec, p. 2677-2694, 2008.

GEWEHR, J. E.; SZUGAT, M.; ZIMMER, R. BioWeka extending the Weka framework for bioinformatics. **Bioinformatics applications note**, vol. 23, p. 651-653, 2007

Gil, A. C. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

GYULAI, D. et al. Lead time prediction in a flow-shop environment with analytical and machine learning approaches. **IFAC-PapersOnLine**, v. 51, n. 11, p. 1029-1034, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. Prentice Hall, 2008.

JAMES, G. et al. **An introduction to statistical learning**. New York: Springer, 2013.

JUN, H. B.; AHN, H. S.; SUH, H. W. On identifying and estimating the cycle time of product development process. **IEEE Transactions on Engineering Management**, v. 52, n. 3, p. 336-349, 2005.

JUN, H. B.; PARK, J. Y.; SUH, H. Lead time estimation method for complex product development process. **Concurrent Engineering**, v. 14, n. 4, p. 313-328, 2006.

JURMAN, G.; RICCADONNA, S.; FURLANELLO, C. A comparison of MCC and CEN error measures in multi-class prediction. **PloS one**, v. 7, n. 8, p. e41882, 2012.

KHAN, M. et al. Big data challenges and opportunities in the hype of Industry 4.0. In: **2017 IEEE International Conference on Communications (ICC)**. IEEE, 2017. p. 1-6.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. New York: Springer, 2013.

KUSIAK, A. Smart manufacturing. **International Journal of Production Research**, v. 56, n. 1-2, p. 508-517, 2018.

LASI, H. et al. Industry 4.0. **Business & information systems engineering**, v. 6, n. 4, p. 239-242, 2014.

LEE, J. et al. Recent advances and trends in predictive manufacturing systems in big data environment. **Manufacturing letters**, v. 1, n. 1, p. 38-41, 2013.

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

LI, B. H. et al. Applications of artificial intelligence in intelligent manufacturing: a review. **Frontiers of Information Technology & Electronic Engineering**, v. 18, n. 1, p. 86-96, 2017.

LIAKOS, K. G. et al. Machine learning in agriculture: A review. **Sensors**, v. 18, n. 8, p. 2674, 2018.

LIN, Y. J. Application of extracted rules from a multilayer perceptron network to moulding machine cycle time improvement. **IEEE Transactions On Components, Packaging And Manufacturing Technology**, v. 1, n. 3, p. 436-445, 2011.

LINGITZ, L. et al. Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. **PROCEDIA CIRP**, v. 72, p. 1051-1056, 2018.

MARCONI, M. A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MONOSTORI, L. AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. **Engineering applications of artificial intelligence**, v. 16, n. 4, p. 277-291, 2003.

MORI, J.; MAHALEC, V. Planning and scheduling of steel plates production. Part I: Estimation of production times via hybrid Bayesian networks for large domain of discrete variables. **Computers & Chemical Engineering**, v. 79, p. 113-134, 2015.

MOURTZIS, D. et al. Knowledge-based estimation of manufacturing lead time for complex engineered-to-order products. **Procedia CIRP**, v. 17, p. 499-504, 2014.

MOUSAVI, S. M. et al. A new support vector model-based imperialist competitive algorithm for time estimation in new product development projects. **Robotics and Computer-Integrated Manufacturing**, v. 29, n. 1, p. 157-168, 2013.

NIKOLIC, B. et al. Predictive manufacturing systems in industry 4.0: trends, benefits and challenges. **Annals of DAAAM & Proceedings**, v. 28, 2017.

NILSSON, N. J. **Introduction to Machine Learning**: An Early Draft of a Proposed Textbook. 1998.

OLIPHANT, T. E. Python for scientific computing. **Computing in Science & Engineering**, v. 9, n. 3, p. 10-20, 2007.

ÖZTÜRK, A.; KAYALIGIL, S.; ÖZDEMIREL, N. E. Manufacturing lead time estimation using data mining. **European Journal of Operational Research**, v. 173, n. 2, p. 683-700, 2006.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825-2830, 2011.

PEREIRA, D. G.; AFONSO, A.; MEDEIROS, F. M. Overview of Friedman's test and post-hoc analysis. **Communications in Statistics-Simulation and Computation**, v. 44, n. 10, p. 2636-2653, 2015.

PFEIFFER, A. et al. Manufacturing lead time estimation with the combination of simulation and statistical learning methods. **Procedia CIRP**, v. 41, p. 75-80, 2016.

PREDIC, B. et al. Data mining based tool for early prediction of possible fruit pathogen infection. **Computers and Electronics in Agriculture**, vol. 154, p. 314-319, 2018.

PRINCIPE, J. C.; EULIANO, N. R.; LEFEBVRE, W. C. **Neural and adaptive systems**: fundamentals through simulations. New York: Wiley, 2000.

RASCHKA, S.; MIRJALILI, V. **Python machine learning**. Birmingham-UK: Packt Publishing Ltd, 2017.

ROMEO, L. et al. Machine learning-based design support system for the prediction of heterogeneous machine parameters in industry 4.0. **Expert Systems with Applications**, v. 140, p. 112869, 2019.

RUBEN, R. A.; MAHMOODI, F. Lead time prediction in unbalanced production systems. **International Journal of Production Research**, v. 38, n. 7, p. 1711-1729, 2000.

SELLITTO, M. A.; WALTER, C. Medição e controle do tempo de atravessamento em um sistema de manufatura. **Measurement and control of lead-time in a manufacturing system**, p. 135-147, 2008.

SHALEV, S. S.; BEN, D. S. **Understanding machine learning**: From theory to algorithms. Cambridge university press, 2014.

SILVA, F.; FERNANDES, F. C. F. Proposta de um sistema de controle da produção para fabricantes de calçados que operam sob encomenda. **Gestão & Produção**, v. 15, n. 3, p. 523-538, 2008.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information processing & management**, v. 45, n. 4, p. 427-437, 2009.

SON, Jong-Duk et al. Development of smart sensors system for machine fault diagnosis. **Expert systems with applications**, v. 36, n. 9, p. 11981-11991, 2009.

STOCK, T.; SELIGER, G. Opportunities of sustainable manufacturing in industry 4.0. **Procedia Cirp**, v. 40, p. 536-541, 2016.

SUSANTO, S.; TANAYA, P. I.; SOEMBAGIJO, A. S. Formulating standard product lead time at a textile factory using artificial neural networks. In: **2012 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering**. IEEE, 2012. p. 99-104.

SUSTO, G. A. et al. Machine learning for predictive maintenance: A multiple classifier approach. **IEEE Transactions on Industrial Informatics**, v. 11, n. 3, p. 812-820, 2014.

SVETNIK, Vladimir et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. **Journal of chemical information and computer sciences**, v. 43, n. 6, p. 1947-1958, 2003.

TAYLOR, B. J.; SMITH, J. T. **Methods and Procedures for the Verification and Validation of Artificial Neural Networks**. Fairmont, WV: Springer Science+Business Media, Inc., 2006.

VERIKAS, Antanas; GELZINIS, Adas; BACAUSKIENE, Marija. Mining data with random forests: A survey and results of new tests. **Pattern recognition**, v. 44, n. 2, p. 330-349, 2011.

WU, Q. Fuzzy measurable house of quality and quality function deployment for fuzzy regression estimation problem. **Expert Systems with Applications**, v. 38, n. 12, p. 14398-14406, 2011.

WUEST, T. et al. Machine learning in manufacturing: advantages, challenges, and applications. **Production & Manufacturing Research**, v. 4, n. 1, p. 23-45, 2016.

XU, D.; YAN, H. S. An intelligent estimation method for product design time. **The International Journal of Advanced Manufacturing Technology**, v. 30, n. 7-8, p. 601-613, 2006.

YANG, L. A standard lead time calculator based on optimization technique. In: **2009 International Conference on Machine Learning and Cybernetics**. IEEE, 2009. p. 3561-3565.

ZAKI, M. J.; MEIRA JR, W.; MEIRA, W. **Data mining and analysis**: fundamental concepts and algorithms. Cambridge University Press, 2014.

ZHONG, R. Y. et al. Intelligent manufacturing in the context of industry 4.0: a review. **Engineering**, v. 3, n. 5, p. 616-630, 2017.