INSTITUTO LATINO-AMERICANO DE TECNOLOGIA,
INFRAESTRUTURA E TERRITORIO (ILATIT)
ENGENHARIA QUIMICA

ANÁLISE DA PRODUÇÃO E OCORRÊNCIA DE FALHAS NO PROCESSO DE TINGIMENTO DE ALGODÃO UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA

CAMILA DI RIENZO FERREIRA

Foz do Iguaçu
2024

INSTITUTO LATINO-AMERICANO DE TECNOLOGIA, INFRAESTRUTURA E TERRITORIO (ILATIT)
ENGENHARIA QUIMICA

ANÁLISE DA PRODUÇÃO E OCORRÊNCIA DE FALHAS NO PROCESSO DE TINGIMENTO DE ALGODÃO UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA

CAMILA DI RIENZO FERREIRA

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Tecnologia, Infraestrutura e Território da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Engenharia Química.

Orientador: Dr. César Adolfo Rodriguez Sotomonte
Coorientador: Prof. Dr. Luís Antonio Lourenço

Foz do Iguaçu
2024

CAMILA DI RIENZO FERREIRA

ANÁLISE DA PRODUÇÃO E OCORRÊNCIA DE FALHAS NO PROCESSO DE TINGIMENTO DE ALGODÃO UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Tecnologia, Infraestrutura e Território da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Engenharia Química.

Orientador: Dr. César Adolfo Rodriguez Sotomonte
Coorientador: Prof. Dr. Luís Antonio Lourenço

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br CESAR ADOLFO RODRIGUEZ SOTOMONTE
Data: 23/10/2024 11:42:04-0300
Verifique em https://validar.iti.gov.br

_____
Orientador: Prof. Dr. César Adolfo Rodriguez Sotomonte
UNILA

Documento assinado digitalmente
GOV.BR LUIS ANTONIO LOURENCO
Data: 17/10/2024 13:51:15-0300
CPF: ***.233.099-**
Verifique as assinaturas em https://v.ufsc.br

_____
Coorientador: Prof. Dr. Luís Antonio Lourenço
Documento assinado digitalmente
gov.br MARLEI ROLING SCARIOT
Data: 21/10/2024 16:08:54-0300
Verifique em https://validar.iti.gov.br

_____
Profa. Dra. Marlei Roling Scariot
UNILA

Documento assinado digitalmente
gov.br RENATA BRAGA SOARES
Data: 17/10/2024 21:03:34-0300
Verifique em https://validar.iti.gov.br

_____
Profa. Dra. Renata Braga Soares
UFMG

Foz do Iguaçu, 26 de abril de 2024.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a meus pais, Suzana e Marcos, por seu apoio inabalável ao longo de toda esta jornada. Suas palavras de encorajamento, amor, paciência foram fundamentais para que eu persistisse nos momentos mais desafiadores. Seu apoio incondicional foi minha rocha durante todo o processo, e por isso estou profundamente grata.

A meu irmão Nathan, por todo o apoio e incentivo. Obrigado por estar ao meu lado em cada passo deste caminho.

A meus avós, meu coração transborda de gratidão por todo o amor, sabedoria e apoio que vocês me proporcionaram. Seus conselhos sábios moldaram quem sou hoje e foram uma fonte constante de motivação para alcançar meus objetivos.

A meus Professores e Orientadores, Cesar Sotomonte e Luís Antonio Lourenço, por todo conhecimento compartilhado durantes esses meses de trabalho, pela paciência que tiveram e pela disponibilidade ao longo do processo.

À minha amiga Bruna, que me esteve comigo desde o começo deste trabalho, sempre me incentivando e me fazendo acreditar que tudo isso seria possível. Sou extremamente grata pela nossa amizade e por tudo que vivemos juntas.

A meus amigos, Abel, Álvaro, Letícia e Sara por todos os momentos que compartilhamos juntos e todas as dificuldades que enfrentamos nesses anos de universidade.

RESUMO

A indústria têxtil no Brasil tem mais de 200 anos de história e é considerada a maior cadeia têxtil completa do Ocidente, com um volume de produção de 2,1 milhões de toneladas em 2022. Possui autossuficiência na produção de algodão e uma cadeia produtiva integrada. Este estudo destaca a importância do controle de qualidade nesse setor, que atualmente possui métodos tradicionais que muitas vezes resulta em custos elevados e baixa eficiência. Neste trabalho foram desenvolvidos, modelos para prever a porcentagem de defeitos em tecidos, utilizando técnicas de ciência de dados e aprendizado de máquina, baseando-se na metodologia CRISP-DM. A pesquisa começou identificando outliers nos dados de tingimento, essa prática é essencial para manter a integridade das análises estatísticas e dos modelos preditivos. Os modelos desenvolvidos precisaram passar por ajustes de hiperparâmetros e técnicas de pré-processamento, como codificação de variáveis categóricas e normalização de dados. Três modelos com algoritmo de floresta aleatória foram desenvolvidos e comparados: o Modelo 1, o Modelo 2 (com normalização de dados) e o Modelo 3 (com exclusão das variáveis de tipos de defeitos e normalização dos dados). Os resultados revelaram que o Modelo 1 obteve o melhor desempenho, apresentando o valor de 0,75 para a métrica de avaliação MAPE e o valor de 0,8462 para R^2. A exclusão das variáveis relacionadas aos tipos de defeitos no Modelo 3 resultou em uma queda significativa na capacidade preditiva, isso se comprovou pelo valor de R^2 de -0,13 e MAPE de 8,72. Em síntese, o estudo contribuiu para o avanço do conhecimento no campo da previsão de defeitos têxteis, evidenciando a importância da análise de dados e aprendizado de máquina para o desenvolvimento de modelos preditivos na indústria têxtil brasileira. As descobertas sugerem oportunidades para melhorias nas práticas industriais relacionadas à qualidade e produção têxtil, e abrem caminho para pesquisas futuras nesse campo.

Palavras-chave: Indústria Têxtil; Tingimento; CRISP-DM; Floresta Aleatória; Defeitos;

RESUMEN

La industria textil en Brasil tiene más de 200 años de historia y es considerada la mayor cadena textil completa de Occidente, con un volumen de producción de 2,1 millones de toneladas en 2022. Tiene autosuficiencia en la producción de algodón y una cadena productiva integrada. Este estudio resalta la importancia del control de calidad en este sector, que actualmente cuenta con métodos tradicionales que muchas veces resultan en altos costos y baja eficiencia. En este trabajo se desarrollaron modelos para predecir el porcentaje de defectos en tejidos, utilizando técnicas de ciencia de datos y aprendizaje automático, basados en la metodología CRISP-DM. La investigación comenzó identificando valores atípicos en los datos de teñido; esta práctica es esencial para mantener la integridad de los análisis estadísticos y los modelos predictivos. Los modelos desarrollados debían someterse a ajustes de hiperparámetros y técnicas de preprocesamiento, como codificación de variables categóricas y normalización de datos. Se desarrollaron y compararon tres modelos con un algoritmo de bosque aleatorio: Modelo 1, Modelo 2 (con normalización de datos) y Modelo 3 (con exclusión de variables de tipo de defecto y normalización de datos). Los resultados revelaron que el Modelo 1 logró el mejor desempeño, presentando un valor de 0,75 para la métrica de evaluación MAPE y un valor de 0,8462 para R^2. La exclusión de variables relacionadas con los tipos de defectos en el Modelo 3 resultó en una caída significativa en la capacidad predictiva, como lo demuestra el valor R^2 de -0,13 y MAPE de 8,72. En resumen, el estudio contribuyó al avance del conocimiento en el campo de la predicción de defectos textiles, destacando la importancia del análisis de datos y el aprendizaje automático para el desarrollo de modelos predictivos en la industria textil brasileña. Los hallazgos sugieren oportunidades para mejorar las prácticas industriales relacionadas con la calidad y la producción textil, y allanan el camino para futuras investigaciones en este campo.

Palabras clave: Industria Textil; Tintura; CRISP-DM; Bosque Aleatorio; Ciencia de los datos; Defectos.

ABSTRACT

The textile industry in Brazil has more than 200 years of history and is considered the largest complete textile chain in the West, with a production volume of 2.1 million tons in 2022. It has self-sufficiency in cotton production and an integrated production chain. This study highlights the importance of quality control in this sector, which currently has traditional methods that often result in high costs and low efficiency. In this work, models were developed to predict the percentage of defects in fabrics, using data science and machine learning techniques, based on the CRISP-DM methodology. The research began by identifying outliers in dyeing data, this practice is essential to maintain the integrity of statistical analyzes and predictive models. The developed models needed to undergo hyperparameter adjustments and pre-processing techniques, such as coding categorical variables and data normalization. Three models with a random forest algorithm were developed and compared: Model 1, Model 2 (with data normalization) and Model 3 (with exclusion of defect type variables and data normalization). The results revealed that Model 1 achieved the best performance, presenting a value of 0.75 for the MAPE evaluation metric and a value of 0.8462 for R^2. The exclusion of variables related to the types of defects in Model 3 resulted in a significant drop in predictive capacity, as demonstrated by the R^2 value of -0.13 and MAPE of 8.72. In summary, the study contributed to the advancement of knowledge in the field of predicting textile defects, highlighting the importance of data analysis and machine learning for the development of predictive models in the Brazilian textile industry. The findings suggest opportunities for improvements in industrial practices related to textile quality and production, and pave the way for future research in this field.

Keywords: Textile Industry; Dyeing; CRISP-DM; Random Forest; Data Science; Defects;

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS E TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AATCC	Associação Americana de Técnicos, Químicos e Coloristas Têxteis
CRISP-DM	*Cross Industry Standard Process for Data Mining* (Processo Padrão Inter-Indústrias para Mineração de Dados)
MAE	Erro Absoluto Médio
MSE	Erro Quadrático Médio
IA	Inteligência Artificial
ISO	*International Organization for Standardization* (Organização Internacional de Normalização)
ML	*Machine Learning*
MD	Mineração de Dados
RMSE	Raiz do Erro Quadrático Médio
AATCC	Associação Americana de Técnicos, Químicos e Coloristas Têxteis

SUMÁRIO

1 INTRODUÇÃO

A indústria têxtil no Brasil tem mais de 200 anos e é atualmente a 10ª maior indústria têxtil no mundo, com produção de quase US$ 13 bilhões no ano de 2021 (JUNIOR, 2021). Entre os anos 2017 e 2020, o Brasil foi considerado um dos maiores produtores têxteis mundiais e vem aumentando suas exportações ao longo dos anos, mesmo em um cenário de pandemia. Possui autossuficiência na produção de algodão e dispõe da maior cadeia produtiva integrada do ocidente, produzindo da fibra até o varejo (ABIT, 2022).

Esse setor engloba uma variedade de produtos como cama, mesa, banho, vestuário, entre outros e para cada item podem existir inúmeros tamanhos e cores, sendo assim, intimamente ligado ao mundo da moda e às mudanças que cada estação proporciona.

Trata-se de uma indústria absorvedora de conhecimento em que as novidades tecnológicas são exógenas, ou seja, é a indústria de máquinas e equipamentos que promove as inovações do processo produtivo (RANGEL; DA SILVA; COSTA, 2010). Mesmo assim, no cenário competitivo atual, com demandas de produtos personalizados, soluções tecnológicas e inovadoras aplicadas a processos industriais são imprescindíveis para o destaque de uma empresa no mercado mundial.

A grande quantidade de dados produzidos diariamente na indústria têxtil, faz com que seja viável que as empresas tirem proveito dessas informações para realizar previsões e verificar tendências na produção anual ou na demanda de produtos, por exemplo. A ciência de dados é uma das formas de aperfeiçoar essas informações e obter conhecimentos e novas perspectivas que ajudem na tomada de decisões.

O reconhecimento de padrões nesses dados pode ser uma vantagem para as indústrias, que podem extrair informações e interpretações dessas análises. A Inteligência Artificial é um dos campos tecnológicos que oferece modelos de apoio à decisão e ao controle se baseando em fatos reais e conhecimentos empíricos e teóricos, mesmo que fundamentado em dados incompletos.

Com base no que foi dito, este trabalho tem como objetivo a criação de um modelo de aprendizagem capaz de analisar tendencias, achar relações e realizar previsões a partir de dados de do processo de tingimento de uma indústria têxtil.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Este trabalho tem como objetivo aplicar ferramentas de ciências de dados para análise de tendências de quantidade de tecidos tingidos e ocorrência de falhas no processo de tingimento de uma indústria têxtil.

2.2 OBJETIVOS ESPECIFICOS

- Organizar os dados fornecidos pela empresa, separando os parâmetros que serão investigados e verificar a presença de outliers e dados faltantes através de *boxplots*;
- Realizar uma análise exploratória dos dados onde serão avaliadas métricas estatísticas, verificação da distribuição dos dados (histograma, assimetria, curtose) e correlação de Pearson;
- Preparação dos dados antes da implementação do modelo que tem como objetivo transformação de natureza (variáveis de tempo) e separação dos dados em treino e teste (80-20);
- Treinar um modelo de aprendizado de máquina;
- Comparar os resultados dos modelos com métricas de avaliação de regressão como: R^2, erro absoluto médio (MAE), erro quadrático médio (MSE), erro absoluto percentual médio (MAPE) e raiz do erro quadrático médio (RMSE).
- Validar o modelo, através da avaliação das métricas do algoritmo escolhido (floresta aleatória).

3 JUSTIFICATIVA

O controle de qualidade desempenha um papel importante na indústria têxtil. Entretanto, a inspeção humana tradicional pode resultar em julgamentos equivocados, aumento de custos e apresentando baixa velocidade na fabricação. Métodos tradicionais permanecem incapazes de descobrir relações globais e complexas entre dados e prever valores de dados desconhecidos para uma nova instância (ISLAM, 2020; SENAI, 2020).

Os sistemas automáticos de detecção de falhas oferecem boas alternativas para substituir inspeção de tecido humano tradicional com sistemas de visão computacional que analisam tecidos de maneira sistemática e visam um desempenho consistente. A eficiência desses sistemas automatizados, no entanto, depende de muitos parâmetros e varia de acordo com a qualidade do hardware e do algoritmo de análise (SENAI, 2020).

Algumas das problemáticas deste setor é falta de especialistas em inteligência artificial (IA), falta de conhecimento da tarefa que será automatizada, base de dados mal estruturada e de baixa qualidade e baixo poder computacional. Muitas indústrias ainda estão decifrando como captar as informações dos seus processos, impulsionadas pela crescente da indústria 4.0 e a digitalização de informação, seja com sensores monitorando a atividade das máquinas ou com o armazenamento adequado dessas informações (SENAI, 2020).

A IA já está sendo usada para automatizar determinados processos e impulsionar a eficiência, ajudar funcionários a serem mais produtivos e dedicar mais tempo às necessidades estratégicas do negócio, mas mesmo as empresas que já monitoram seus processos podem não estar preparadas para aplicar essas tecnologias, por não terem qualidade nas informações armazenadas (ISLAM, 2020).

É aconselhável que haja um monitoramento, e que as informações estejam sendo captadas com exatidão, é importante também a verificação da existência de dados que não descrevem bem a tarefa e que influenciam em tomadas de decisão tendenciosas. Essas falhas geralmente são decorrentes de tomadas de decisão da

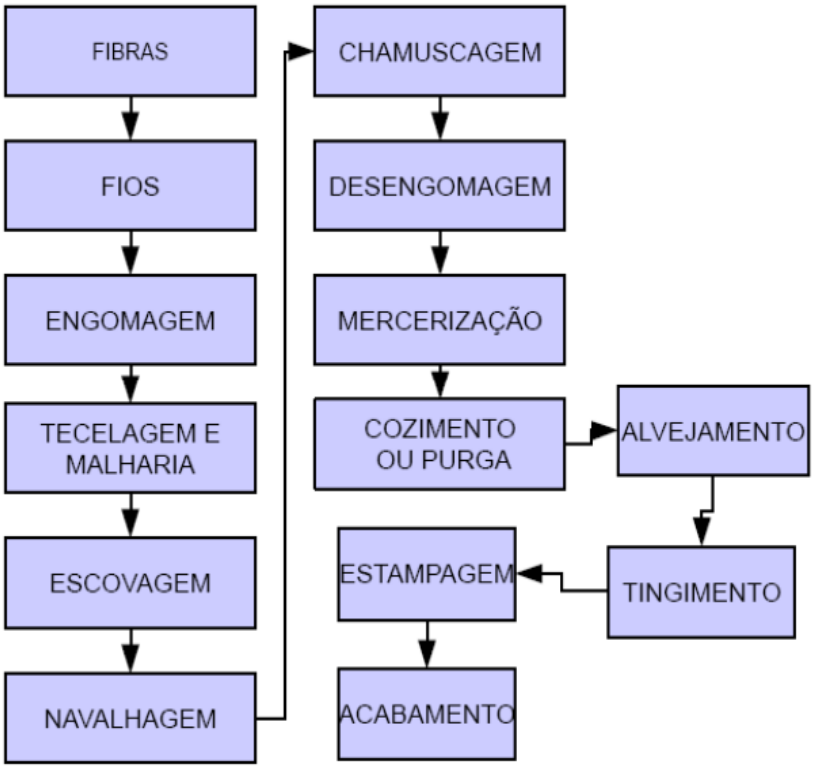
aplicação de IA sem que se tenha um objetivo bem definido. A falta de uma visão clara da finalidade do uso dessas tecnologias muitas vezes impede a empresa de usufruir dos benefícios por completo (ISLAM, 2020).

Por isso há a necessidade de processar dados brutos e explorar informações valiosas e essa demanda surge em muitas áreas da engenharia. As aplicações atuais de tecnologia da informação analisam dados e os convertem em conhecimento valioso de maneira eficiente (YILDIRIM; BIRANT; ALPYILDIZ, 2019).

4 REVISÃO BIBLIOGRÁFICA

O processo de produção na indústria têxtil é milenar e o beneficiamento têxtil é uma das mais antigas tecnologias empregadas pelo homem. O sistema de manufatura é sequencial, ondes diversas etapas se articulam de forma mecânica. Essa articulação possibilita inúmeras combinações das várias etapas de produção em uma mesma planta industrial ou em várias unidades fabris. A fabricação de produtos têxteis envolve, basicamente, as etapas de produção de fibras, fiação, tecelagem, acabamento e confecção, como mostrado na Figura 1.

Figura 1 – Fluxograma de processo para obtenção de artigos têxteis 100% algodão



Fonte: adaptado de JULIANO e PACHECO, s.d.

Essas etapas compõe o ciclo do tecido na indústria. Neste trabalho a atenção se concentrará no processo de beneficiamento, mais especificamente o beneficiamento secundário onde é feito o tingimento do tecido, essa etapa tem como função conferir cor ao substrato.

O processo de tingimento é uma das etapas fundamentais no sucesso comercial dos produtos. Além da beleza da cor, o consumidor pode exigir que o

produto tenha algumas características, como elevado grau de fixação em relação à luz, à lavagem e à transpiração e maciez, tanto inicialmente quanto após uso prolongado.

4.1 TINGIMENTO

O tingimento consiste em uma modificação física ou química do substrato têxtil. Sua execução tem como finalidade atribuir cor aos fios ou tecidos e assim agregar valor no produto final, utilizando uma ampla variedade de corantes. Esse processo consiste em três etapas e nelas acontecem os seguintes fenômenos físico-químicos: migração, absorção e difusão (fixação do corante). É possível realizar o tingimento na fibra, no fio e no tecido. O tingimento em tecido se mostra mais vantajoso em razão da obtenção de uma melhor igualização no comprimento da peça tendo assim um menor desperdício de corante, requerendo menos processos comparados ao tingimento de fibras ou fios (JULIANO; PACHECO; PEREIRA, s.d.; SALEM, 2010).

A adsorção e retenção do corante na fibra pode ser química, física ou ambas, dependendo da fibra e do corante usado. Esse processo consiste em um banho aquoso e podem ser feitos de duas formas: tingimento contínuo ou tingimento por esgotamento (não contínuo).

No tingimento contínuo, o material têxtil é alimentado continuamente em solução de corante a uma velocidade constante. Esse processo é constituído basicamente pela aplicação e fixação do corante. O tecido é espremido mecanicamente e em seguida é fixado com o auxílio de calor seco ou úmido, repouso a frio ou a quente e por fim pode-se optar por um novo banho. Nos processos contínuos, a igualização do tingimento depende das instalações mecânicas, remoção do excesso de solução de corante e da secagem. A quantidade de corante é expressa em g/L. Esse processo também é muito conhecido como foulardagem (SALEM, 2010).

Um sistema é não contínuo, quando uma operação é iniciada e terminada na mesma máquina, sendo assim, o tingimento por esgotamento consiste na diminuição gradativa dos produtos, ou seja, o corante se transfere do banho para a fibra e consequentemente há o aumento da concentração da cor no material

(JULIANO; PACHECO, s.d.). Diferentes fatores influenciam para que haja uma boa igualdade do tingimento por esgotamento como, por exemplo, o contato entre banho e substrato, a velocidade da montagem e a migração do corante (SALEM, 2010).

Ao comparar as duas formas de tingimento, é possível afirmar que a fixação do corante nas fibras ocorre mais rapidamente em tingimento contínuo que em batelada. O grau de adsorção depende de vários fatores, tais como, temperatura, pH, produtos químicos usados e tempo (JULIANO; PACHECO, s.d.; SALEM, 2010). Para cada um desses processos e substratos há um maquinário específico com condições ideais.

4.1.1 Velocidade de montagem

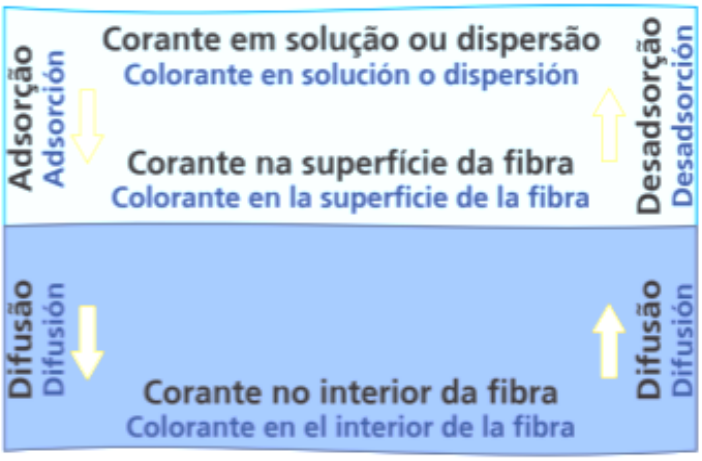
A velocidade de montagem de um corante no substrato depende das propriedades de cada corante em função de sua estrutura química. Existem, porém, fatores que podem acelerar ou retardar essa velocidade:

- pH do banho;
- Eletrólitos (sulfato ou cloreto de sódio) que aumentam a substantividade e, portanto, rendimento do corante. Em alguns tingimentos é adicionado eletrólitos para retardar a montagem;
- Agentes auxiliares são usados para acelerar ou retardar a velocidade de montagem;
- A relação de banho é muito importante no tingimento de fibras celulósicas com corantes diretos ou reativos. Quanto maior a concentração do corante no banho, maior é a substantividade. Assim, quanto mais diluirmos o banho de tingimento, menor a afinidade do corante com a fibra.

4.1.2 Migração

Na fase de equilíbrio durante o processo de tingimento, pode ocorrer o fenômeno de migração, demonstrado na Figura 2.

Figura 2 – Migração do corante para a fibra



Fonte: SALEM, 2010

Primeiramente a fibra incha com a penetração na solução preparada, seguindo por uma adsorção das moléculas de corantes que estavam na superfície da fibra, logo depois ocorre a difusão das moléculas no interior da fibra e por fim se dá a fixação das moléculas do corante no tecido.

4.2 MÁQUINAS UTILIZADAS NO PROCESSO DE TINGIMENTO

4.2.1 Processos não contínuos ou por esgotamento

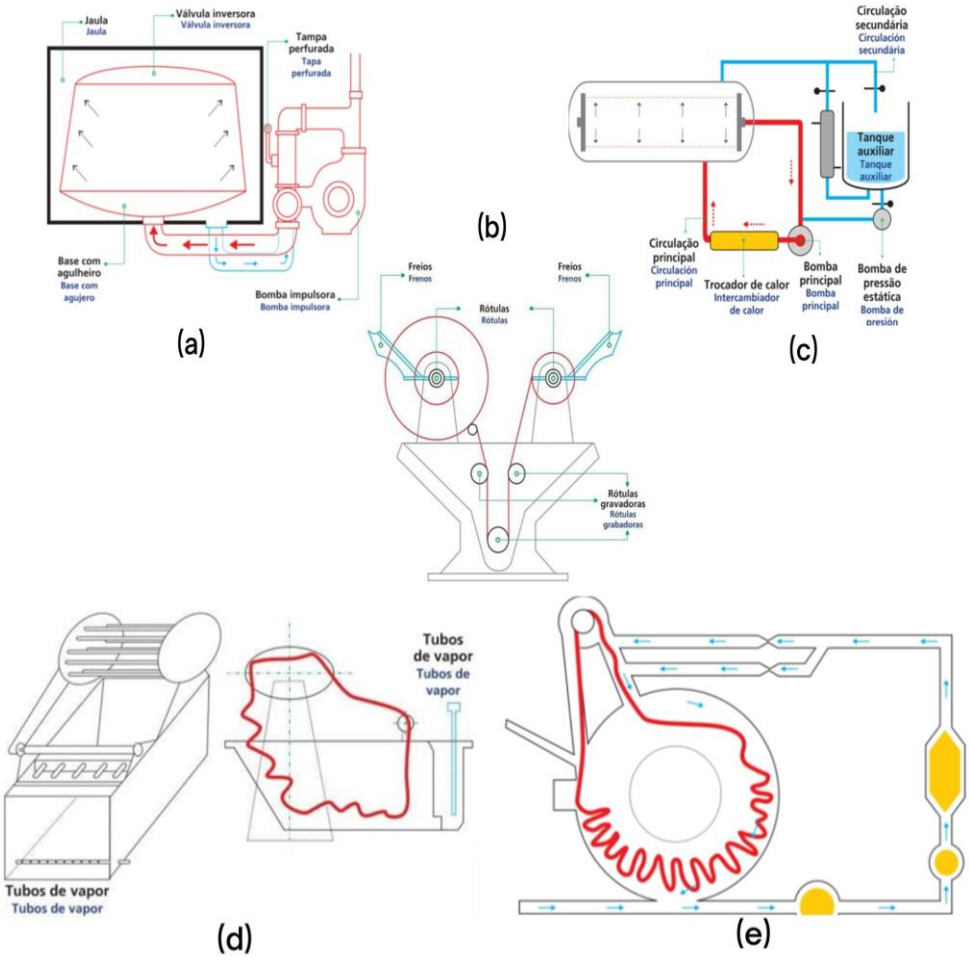
No processo de esgotamento é necessário que haja agitação o suficiente para que o corante seja distribuído homogeneamente pelo tecido. Outro parâmetro a ser controlado é o calor, que precisa estar uniformemente distribuído e de preferência ser usado vapor indireto. Esse tipo de maquinário além de ser economicamente viável também evita a contaminação do ambiente com vapores corrosivos e manchas por condensação (SALEM, 2010).

As fibras são tingidas em equipamentos com circulação de banho, podendo ser aberto ou fechado. É realizado em lotes, onde cada lote o tecido é tingido de uma cor. Esse processo possui vantagens na equalização do tingimento, na variedade de materiais que podem ser tingidos em ramas, bobinas cruzadas e peças de tecidos ou malha. Os equipamentos utilizados também não precisam de mão de obra especializada, o que faz com que esse processo seja mais barato. Entretanto, pelas suas grandes relações de banho esse método acaba consumindo muita água,

produtos químicos e em consequência dos seus ciclos demorados acabam gastando mais energia do que outras técnicas (PEREIRA, s.d.).

Os equipamentos mais conhecidos para realizar processos não contínuos são: jigger, turbo, barca e jets. Seus esquemas de funcionamento estão apresentados na Figura 3 abaixo.

Figura 3 - Sistemas de tingimento



Fonte: adaptados de SALEM, 2010

Nota: (a) Máquina de tingir rama; (b) Jigger; (c) Máquina de tingimento turbo; (d) Máquina para tingir tecidos e malhas (barca de molinete); (e) Máquina de tingimento jet.

4.2.2 Processos contínuos e semi-contínuos

Todos os sistemas contínuos de tingimento de peças iniciam-se pela impregnação em um foulard. O foulard é constituído de uma estrutura, contendo dois

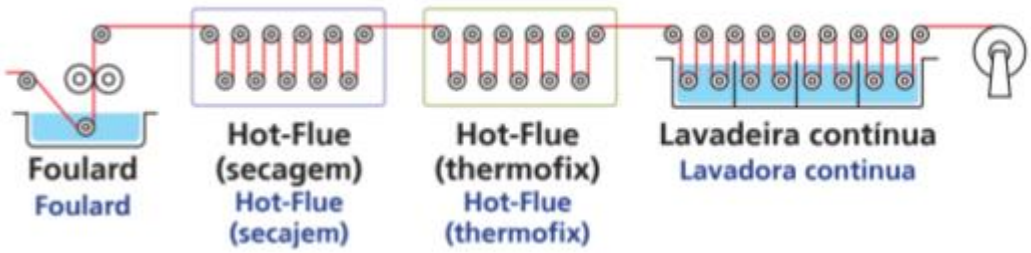
ou três rolos espremedores. Essa estrutura deve ter o menor volume possível para que seja possível uma troca rápida do banho e a pressão dos rolos deve ser igual em toda a largura para assegurar um pick-up[1] homogêneo (JULIANO; PACHECO, s.d.; SALEM, 2010).

A reação de montagem do corante na fibra é acelerada com a adição ou remoção de calor. Os banhos são curtos e renováveis, é indicado para processos de grandes quantidades de material, tendo assim uma alta produção e boa reprodutibilidade da cor.

O tecido, ao passar pelo chassi, é saturado com a solução de corante e em seguida é direcionado para os rolos onde será espremido. Durante a espremedura, uma parte da solução em excesso é direcionada no sentido contrário e retorna ao chassi, outra parte é forçada para dentro do tecido e uma pequena quantidade do banho é arrastada superficialmente pelo substrato.

Nos processos contínuos ou semi-contínuos, após a impregnação do tecido no foulard, o tingimento é fixado em operação posterior por calor seco, calor úmido, repouso a frio, repouso a quente e banho novo. Essa fixação depende das instalações mecânicas (JULIANO; PACHECO, s.d.). Para processos contínuos podemos considerar as seguintes instalações representadas da Figura 4 a 9.

Figura 4 - Pad-Thermofix

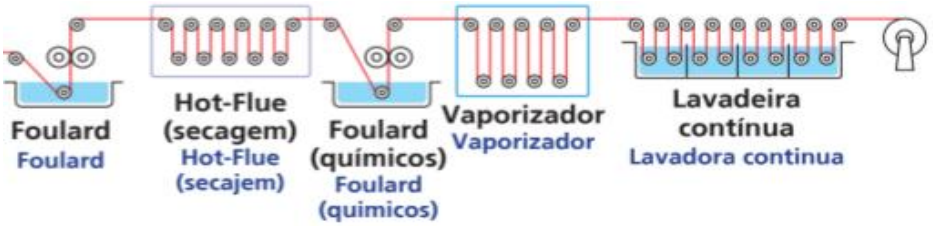


Fonte: SALEM, 2010.

Na figura 5 e 6 nota-se uma subsequente vaporização que tem como finalidade a fixação do corante no tecido.

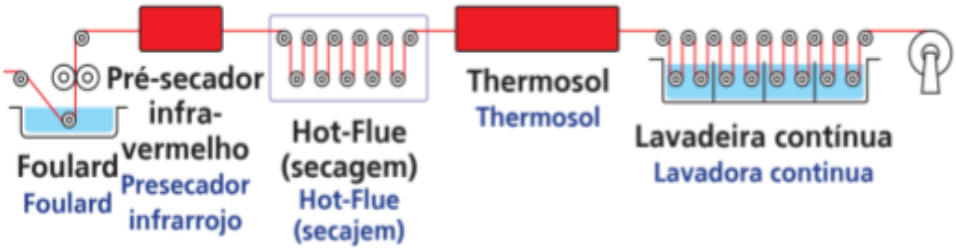
[1] Percentual de retenção do banho pelo substrato.

Figura 5 - Pad-dry/Pad-steam



Fonte: SALEM, 2010

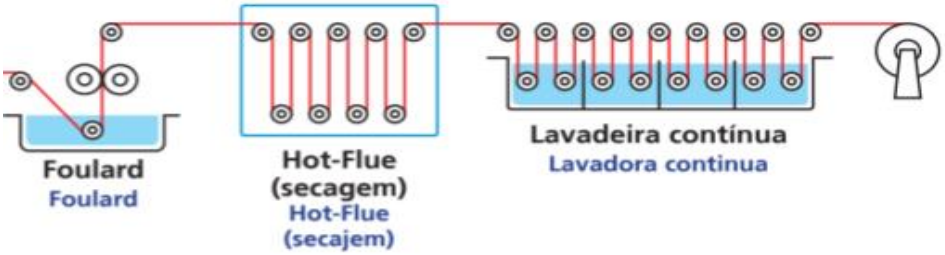
Figura 6 – Thermosol



Fonte: SALEM, 2010

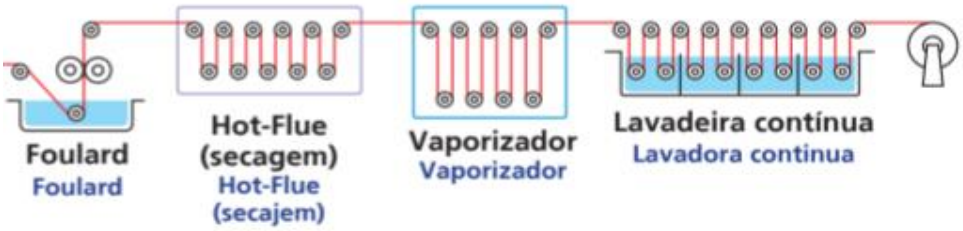
Nesse sistema (Figuras 7 e 8), após a foulardagem, o tecido passa por uma secagem por irradiação, fixando o material têxtil em câmaras quentes por insuflamento de ar.

Figura 7 - Pad-Wet-Steam



Fonte: SALEM, 2010

Figura 8 - Pad-Dry-Steam

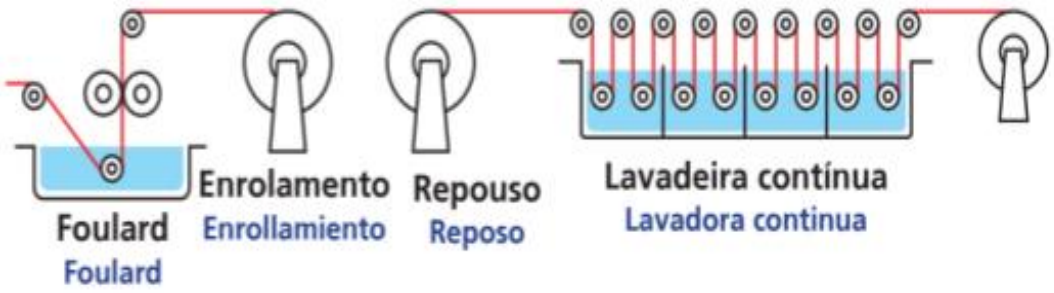


Fonte: SALEM, 2010

O pad-batch (Figura 9) é um dos métodos mais baratos na tinturaria e consiste na impregnação de tecidos planos ou malhas, seguindo por um acondicionamento em cavaletes que faz com que o material final fique protegido do contato com o ar, após isso permanece em repouso, por isso é considerado um processo semi-contínuo. A fase da reação se dá a frio, e o tecido geralmente é mantido em rotação lenta para que assim tenha uma boa migração de corante para a fibra.

De acordo com Salem (2010), o produto obtido com a implementação do Pad-Batch é de alta qualidade, possui um toque sedoso e aspecto liso, seu encolhimento é mínimo. Além disso, seu rendimento tintorial é elevado, tem excelente reprodutibilidade, baixos custos de investimento, baixo consumo energético, reduzido consumo de água e produtos químicos, necessita de pouca mão de obra e possui baixo volume de despejo nos efluentes.

Figura 9 - Pad-Batch

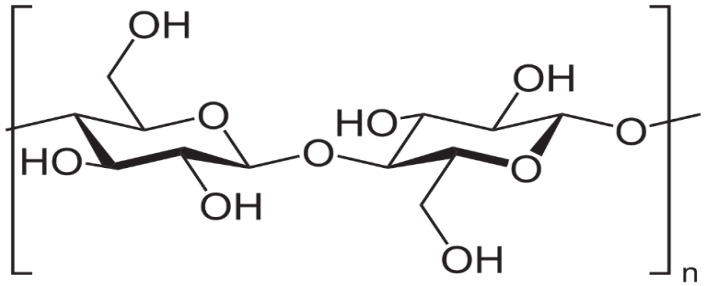


Fonte: SALEM, 2010

4.3 TINGIMENTO FIBRAS CELULÓSICAS

O foco desse trabalho é o tingimento de fibras celulósicas mais especificamente o tingimento do algodão. Essa categoria de fibras tem como componente principal a celulose. Esse polímero linear representado na Figura 10, é considerado um carboidrato constituído por 44,4% de carbono, 6,2% de hidrogênio e 49,4% de oxigênio.

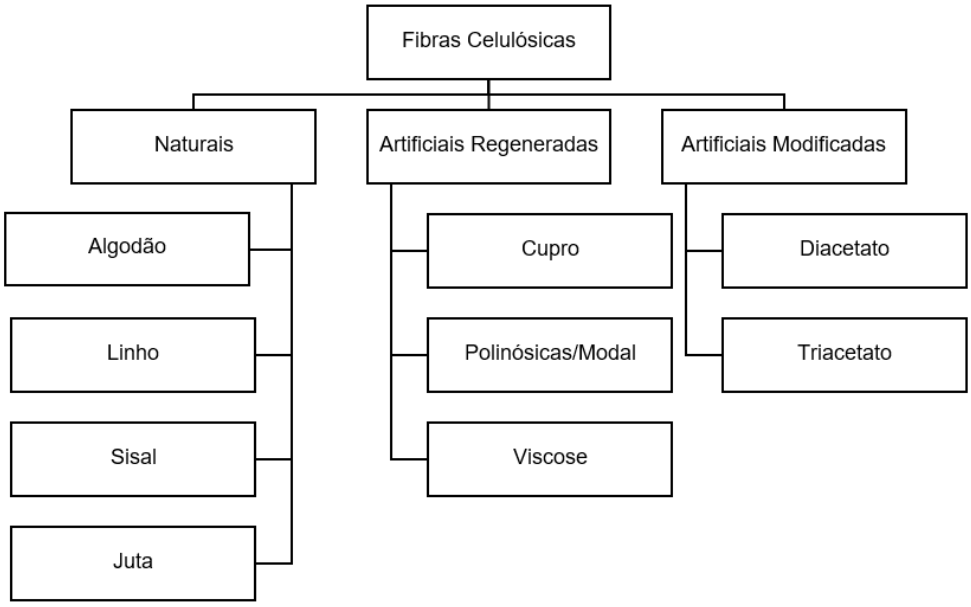
Figura 10 - Estrutura molecular da celulose



Fonte: CELULOSE, 2022.

Há algumas propriedades em comum entre essas fibras devido a sua composição predominante de celulose. Podem se queimar fácil e rapidamente desprendendo odor de papel queimado, produzem resíduo leve e cinzas que variam entre o negro e o acinzentado, apresentam ótima resistência a soluções alcalinas (KUASNE, 2008). A Figura 11 abaixo mostra como são classificadas as fibras celulósicas.

Figura 11 - Fibras têxteis compostas por celulose



Fonte: Adaptado de KUASNE, 2008.

Em relação a fibra de algodão, pode-se afirmar que quando seca, é quase inteiramente composta por celulose (de 88 a 96%) que por sua vez é responsável por dar resistência às fibras devido ao seu alto grau de polimerização e orientação molecular (FERREIRA, 2019). Além de celulose, o algodão contém proteínas, pectina, cera, cinzas, ácidos orgânicos e pigmentos (EEEP, 2015).

Quando há problemas nos processos de fiação, malharia e tecelagem, estes geralmente são revelados nos processos de beneficiamentos químicos, onde através da aplicação de cor nos substratos são visíveis (FERREIRA, 2019).

Ao escolher um corante deve-se levar em conta a afinidade entre o corante e a fibra, qual será a interação entre os dois (ligações de hidrogênio, forças intermoleculares, ligação iônica, covalente, etc.), e as propriedades químico-físicas como já mencionado anteriormente (HONORIO, 2013 apud FERREIRA, 2019). Considerando esses requisitos citados, as fibras celulósicas podem ser tingidas/combinadas com:

- Corantes ácidos;
- Corantes dispersos;
- Corantes diretos ou aniônicos;
- Corantes básicos ou catiônicos;
- Corantes reativos;
- Corantes à tina;
- Corante ao enxofre.

Para tingimento do algodão em específico, recomenda-se o uso de corantes reativos, em razão da formação de ligações covalentes, este confere boas características de solidez e brilho para o tecido (FERREIRA, 2019).

4.3.1 Corantes Reativos

Segundo Piccoli (2008), cerca de 40% dos corantes para celulose consumidos no Brasil são corantes reativos. Essas fibras celulósicas primeiro adsorvem o corante, seguida de uma reação entre a celulose e o corante formando uma ligação covalente.

A reação ocorre nos grupos hidroxílicos da celulose e para que esta aconteça, há necessidade da ionização da celulose. A ionização aumenta proporcionalmente à alcalinidade do banho. O corante reativo possui três tipos de grupos funcionais: grupo cromóforo, solubilizante e reativo (que se ligam à fibra).

Alguns corantes possuem maior reatividade que outros. Os de maior reatividade são denominados corantes a frio. Já os de menor reatividade são chamados de corantes a quente. Ter maior ou menor reatividade não significa que um corante seja melhor ou pior. A escolha deve ser feita embasada em parâmetros como substrato, processo, temperatura, pH, etc.

Nos tingimentos com corantes reativos deve-se sempre tentar conciliar a maximização de rendimento da reação com a fibra com a minimização da reação com a água do banho.

O processo de tingimento com corantes reativos pode ser dividido em duas fases, a fase do sal e a fase do álcali. Ao adicionar o sal (eletrólito) o corante se desloca para a fibra na qual ocorrerá dois fenômenos a adsorção e a difusão, ou seja, devido a presença do sal acontece a montagem do corante no tecido. Caso houver um intervalo antes da fase do álcali onde a temperatura se manterá ocorrerá um fenômeno chamado migração. Ao adicionar o álcali a etapa de fixação começa, essa fase é no qual ocorre a reação do corante com a fibra.

Utiliza-se corantes reativos a frio quando as tonalidades são muito brilhantes e não se atingem as cores com outra classe, há dificuldade de ensaboamento com os corantes a quente, há uma demanda de menor gasto de energia, e quando a tinturaria dispõe somente de máquinas para processos contínuos ou semi-contínuos, nas quais os corantes reativos a frio são mais adequados. (PICCOLI, 2008).

Os corantes reativos a quente possuem excelentes propriedades de difusão e migração e pode ser empregado quando os tecidos usados são muito compactos, os fios muito estão retorcidos, quando se tem algodão mercerizado, quando se utiliza viscose ou malhas de algodão, ou quando as máquinas de tingimento estão trabalhando com baixa circulação de banho ou baixa velocidade de substrato. (PICCOLI, 2008).

4.4 PARÂMETROS DE QUALIDADE ATUAIS

De acordo com Salem (2010), há três parâmetros de maior importância no controle de qualidade dos tingimentos: reprodutibilidade, igualização e solidez.

A reprodutibilidade pode ser avaliada visualmente ou por calorimetria, nela é avaliada a intensidade, tonalidade e pureza da cor. Quem define todas essas características é o cliente. A análise de igualização do tingimento requer pessoal capacitado e um equipamento chamado tribunal. Essa avaliação é realizada visualmente.

A solidez da cor é a capacidade que um artigo têxtil possui de manter a cor independente do seu meio, para isso são consideradas algumas variáveis como: fibra, corante, presença de auxiliares, etc. Para avaliação desses ensaios tem-se as normas ISO, normas determinadas pela Associação Americana de Técnicos, Químicos e Coloristas Têxteis (AATCC) e Associação Brasileira de Normas Técnicas (ABNT). Os ensaios de solidez podem ser classificados como:

- Determinação da solidez da cor à luz artificial;
- Determinação da solidez da cor à fricção;
- Determinação da solidez da cor a lavagem.

A primeira averiguação a se fazer é determinar qual escala foi empregada. Existem dois tipos de escalas que possuem notas e numerações diferentes. Essas normas foram descritas na ABNT NBR 8429, ABNT NBR 8430 e pela ISO 105 B02. Para cada análise obtém-se uma nota, chamada de nota de solidez, e caso haja uma mudança de tonalidade durante as testagens deve-se fazer uma observação (WALTRICK, 2020). A avaliação é feita através da exposição do produto a luz de xênon, arco voltaico e a luz solar.

Os tecidos planos são normatizados para que se tenha um padrão limite de defeitos. A norma da ABNT NBR 13484 (2004), implementa um método de classificação baseado em inspeção por pontuação de defeitos. Há inúmeras falhas e defeitos recorrentes no processo de tingimento, e o controle de qualidade deve ter a

capacidade para identificar esses defeitos e suas causas. Alguns dos defeitos mais notáveis são:

- Listras indesejadas, devido defeito nos rolos do maquinário;
- Quebra repentina do padrão do tingimento, pode ocorrer quando há uma partícula nas fendas dos cilindros;
- Marca de parada, onde há um excesso de tinta em uma região do tecido após uma parada de máquina;
- Falta de cor, pode ser devido a relação de banho, a escolha do corante, etc.
- Mancha na coloração;
- Diferentes graduações de cores, quando o tingimento é feito de forma desigual ou o tecido acaba absorvendo mais corante em determinadas regiões;

4.5 USO DE INTELIGÊNCIA ARTIFICIAL (IA) PARA CONTROLE E MONITORAMENTO DE PROCESSOS TÊXTEIS

Torna-se necessário mudanças que tragam melhores desempenhos a esse setor. A inovação começa a ser considerada como termo chave para a habilidade administrativa, principalmente na perspectiva da eficácia e do comportamento gerencial dessas empresas (KISLINGEROVA, 2018 apud CAVALCANTI; DOS SANTOS, 2021). Com novas criações e implementações contínuas de melhorias, uma empresa garante um alto nível de competitividade, pois isso permite que ela reaja com flexibilidade a mudanças.

A otimização desse setor vem por meio de estudos que envolvem a tecnologia e possuem a capacidade de reduzir incertezas e ajudar em processos de tomada de decisão. Análises obtidas através de dados possuem um imenso poder de prever comportamentos e orientar uma empresa para um futuro produtivo.

Novos elementos que facilitam inovar surgem com o uso da computação e a criação de materiais sustentáveis mostram-se, a cada dia, mais presentes no mundo

da moda, em objetos, nos esportes e nos uniformes (TIDD; BESSANT; PAVITT, 2008 apud CAVALCANTI; DOS SANTOS, 2021).

Raramente uma empresa faz um bom uso dos dados que obtém com seus processos, essas instituições não sabem quantas informação podem tirar com apenas um compilado de dados, desde previsões de produção até ideias de melhorias.

Inteligência artificial (IA) é um recurso baseado em modelos analíticos que geram previsões, regras, respostas recomendações ou resultados semelhantes. Com definição ampla, a IA compreende qualquer técnica que permita computadores imitar o comportamento humano e reproduzi-lo ou também se destacar sobre uma tomada de decisão para resolver tarefas complexas de forma independente ou com intervenção humana mínima (Russell e Norvig 2021 apud JANIESCH et al., 2021). Com a inteligência artificial é possível dar um sentido para dados que muitas vezes são ignorados.

4.5.1 Aprendizado de Máquina

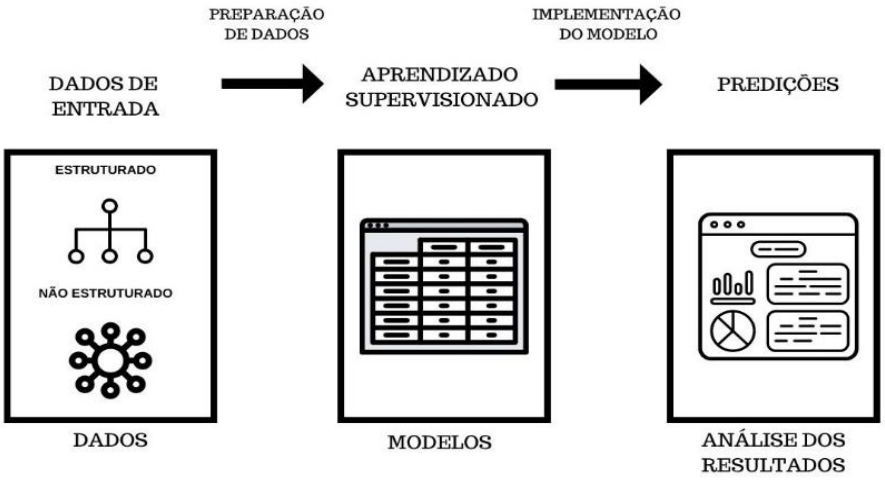
O aprendizado de máquina (*machine learning*) é uma das subáreas da inteligência artificial e visa automatizar a tarefa de construção de modelos analíticos para realizar tarefas como, detecção de objetos ou tradução de idiomas. Isto é possível através da aplicação de algoritmos que aprendem iterativamente com dados de treinamento específicos do problema, que permite que os computadores encontrem interpretações ocultas e padrões sem serem explicitamente programados (Bishop 2006 apud JANIESCH et al., 2021).

Essa ferramenta busca aprender por conta própria, relacionamentos e padrões significativos a partir de exemplos e observações (Bishop 2006 apud JANIESCH et al., 2021). Os avanços nessa área permitiram a ascensão recente de sistemas inteligentes com capacidade cognitiva semelhante à humana, uma capacidade que penetra nos negócios e vida pessoal. O aprendizado de máquina pode ser dividido em três categorias: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

O aprendizado supervisionado requer um conjunto de dados de treinamento que abrange exemplos para a entrada, bem como saídas rotuladas. Os pares de dados de entrada e saída no conjunto de treinamento são então usados para calibrar os parâmetros abertos do modelo ML. Uma vez que o modelo tenha sido treinado com sucesso, ele pode ser usado para prever a variável alvo (JANIESCH; ZSCHECH; HEINRICH, 2021). Em outras palavras, “os algoritmos ajustam parâmetros de um modelo a partir do erro medido entre respostas obtidas e esperadas” (BRUNIALT et al., 2015 p .205).

Esse método resolve problemas conhecidos e usa um conjunto de dados rotulado para treinar um algoritmo para realizar tarefas específicas, o algoritmo aprende a partir dos dados fornecidos e consequentemente aprende com o treinamento, aplicando esse aprendizado em entradas desconhecidas e levando para uma saída correta (Figura 12).

Figura 12 - Fluxo de funcionamento do aprendizado de máquina



Fonte: Adaptado de “O QUE É APRENDIZAGEM SUPERVISIONADA? **[s. d.]”.**

Diferente da aprendizagem supervisionada a aprendizagem não supervisionada ocorre quando o sistema de aprendizado deve detectar padrões sem rótulos ou especificações de aprendizado pré-existentes (JANIESCH; ZSCHECH; HEINRICH, 2021). Tem como objetivo a identificação de novos padrões e irregularidades. Os dados que são fornecidos ao modelo não são rotulados ou categorizados previamente, deste modo o computador tenta dar sentido para os dados por conta própria, descobrindo padrões ocultos (Figura 13).

Figura 13 - Etapas do aprendizado não supervisionado

Fonte: Adaptado de ALMEIDA et al., 2017.

Já no aprendizado por reforço os modelos são treinados para tomarem sequência de decisões em um ambiente incerto e complexo. Em vez de fornecer pares de entrada e saída, o estado atual do sistema é descrito, especifica-se uma meta e é fornecido uma lista de ações permitidas e suas restrições. Com isso é deixado que o modelo de aprendizado experimente o processo de atingir a meta por si mesmo, usando o princípio de tentativa e erro para maximizar uma recompensa (JANIESCH; ZSCHECH; HEINRICH, 2021).

A forma com que o programa aprende em cada uma desses diferentes tipos de aprendizagem, está associada a um conjunto de configurações previamente definidas, denominadas de hiper parâmetros (ALMEIDA *et al.*, 2017). Com essa poderosa ferramenta se tem inúmeras possibilidades de utilização. Uma delas é a implementação do aprendizado de máquina em industrias, treinando o programa para realizar previsões de comportamentos que podem afetar a empresa economicamente.

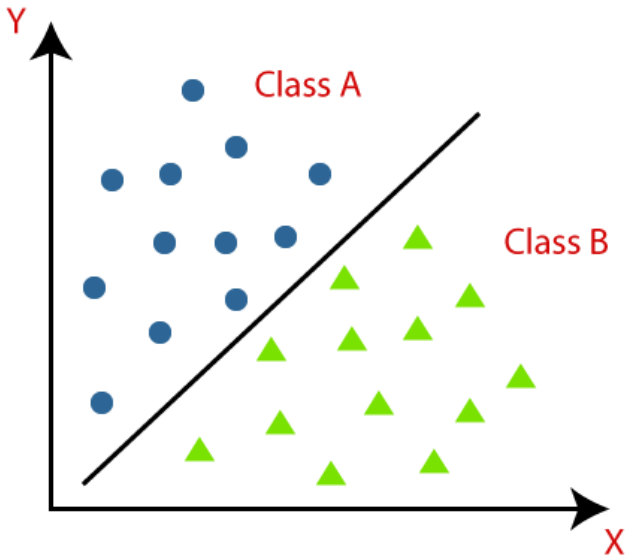
Este trabalho irá focar no aprendizado supervisionado, que se divide em duas categorias principais, a classificação e a regressão.

4.5.2 Modelos de Classificação

A classificação é um método de aprendizado de máquina supervisionado no qual o modelo busca prever os rótulos corretos de dados específicos de entrada. Nesse processo, o modelo é integralmente treinado com dados de treinamento e, posteriormente, é avaliado com dados de teste antes de ser aplicado para fazer

previsões em novos dados não observados. O objetivo primário dos algoritmos de classificação é determinar a categoria de um conjunto de dados específico, sendo esses algoritmos amplamente empregados na previsão de saídas categóricas. Os algoritmos de classificação podem ser melhor compreendidos a partir da Figura 14 (FONTANA, 2020).

Figura 14 - Algoritmo de classificação



Fonte: “Classification Algorithm in Machine Learning - Javatpoint”, s.d.

O algoritmo que implementa a classificação em um conjunto de dados é conhecido como classificador. Existem dois tipos de classificações (“Classification Algorithm in Machine Learning - Javatpoint”, s.d.):

- Se o problema de classificação tiver apenas dois resultados possíveis, ele é chamado de Classificador Binário. Exemplos: sim ou não, masculino ou feminino, spam ou não spam, gato ou cão, etc.
- Se um problema de classificação tiver mais de dois resultados, ele é chamado de Classificador Multiclasse. Exemplo: Classificações dos tipos de culturas, classificação dos tipos de música.

4.5.3 Modelos de Regressão

A regressão é uma ferramenta que busca modelar relações entre variáveis dependentes e independentes através de métodos estatísticos (Soto 2013). Sua

origem vem da correlação linear, que é a verificação da existência de um relacionamento entre duas variáveis. Ou seja, dado X e Y, quanto que X explica Y. Os métodos de regressão se utilizam dessas correlações entre as variáveis para estimar valores não existentes na amostra ou conjunto de dados. Podem ser criados a partir de diversas abordagens, desde as mais simples com poucas configurações de parâmetros e de fácil interpretação do funcionamento, até as abordagens mais complexas (GOMES, 2019; VELASQUEZ, 2020; DAMACENO, 2020).

O resultado da regressão linear é sempre um número e é utilizada adequadamente quando o conjunto de dados apresenta algum tipo de tendência de crescimento ou decrescimento constante.

Esse algoritmo pode ser utilizado em qualquer problema, onde as variáveis de entrada e saída são valores contínuos, por exemplo, prever as vendas de um determinado produto, estimar o valor de um imóvel, calcular a expectativa de vida de um país, etc. (VELASQUEZ, 2020). Um dos tipos mais usados de regressão segundo (HARRISON, 2019), é o algoritmo de Floresta Aleatória.

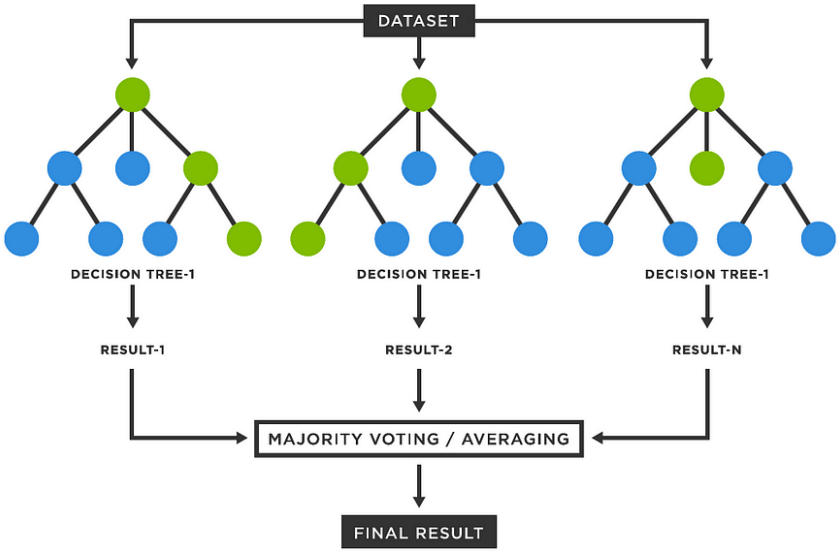
4.5.4 Floresta Aleatória

A Floresta Aleatória é um modelo de aprendizado de máquina que se baseia na construção de várias árvores de decisão durante o processo de treinamento. Cada árvore de decisão é treinada em uma subamostra aleatória do conjunto de dados original e faz suas próprias previsões. No entanto, ao contrário de uma única árvore de decisão, onde pode haver o risco de *overfitting* (ou seja, a árvore se ajusta demais aos dados de treinamento e não generaliza bem para novos dados), a Floresta Aleatória utiliza uma técnica chamada "*bagging*" (*bootstrap aggregating*) para reduzir esse risco (IBM,2023; JÚNIOR, 2018; HARRISON, 2019).

Durante o processo de treinamento da Floresta Aleatória (Figura 15), várias subamostras (ou "bolsas") são criadas a partir do conjunto de dados original usando a técnica de *bootstrap*. Cada árvore de decisão é então treinada em uma dessas subamostras, mas com uma pequena variação: em cada divisão de nó da árvore, apenas um subconjunto aleatório das características (ou atributos) é considerado para dividir o nó. Esse processo de seleção aleatória de características ajuda a garantir que

cada árvore seja diversa e reduz a correlação entre elas (IBM,2023; JÚNIOR, 2018; HARRISON, 2019).

Figura 15 - Representação do funcionamento do algoritmo floresta aleatória



Fonte: Gunay, 2023

Uma vez que todas as árvores são treinadas, as previsões de cada árvore são combinadas para produzir uma previsão final. Na classificação, isso é feito por votação majoritária, onde a classe mais frequente prevista por todas as árvores é selecionada como a previsão final. Na regressão, as previsões de todas as árvores são combinadas através de uma média para produzir o resultado final (IBM,2023; JÚNIOR, 2018; HARRISON, 2019).

O algoritmo de Floresta Aleatória é frequentemente utilizado na mineração de dados devido à sua eficácia em uma ampla gama de problemas de classificação e regressão, sua capacidade de lidar com uma grande quantidade de características e sua resistência ao *overfitting*[2].

4.6 MINERAÇÃO DE DADOS NA INDÚSTRIA TÊXTIL

A indústria têxtil possui uma das cadeias industriais mais complexas da indústria de transformação. A caracterização de seus parâmetros de operação é complexa devido à grande variedade de substratos, processos, máquinas e

[2] especialização do algoritmo no conjunto de dados

componentes utilizados. Mesmo quando se considera um produto simples como uma camiseta básica, uma grande quantidade de dados é gerada e armazenada, esses dados incluem matérias-primas, configurações da máquina e parâmetros de qualidade do produto (LEE; LIN; CHANG, 2018).

A mineração de dados (MD) consiste em explorar um conjunto de dados, de forma analítica (com técnicas estatísticas, modelos matemáticos, etc.) a fim de encontrar um padrão. Essa técnica é bem sucedida na análise de diversas situações como: quando há grande quantidade de dados disponível, quando os dados são complexos com muitas variáveis e há muitas relações não lineares e quando há necessidade de prever comportamentos ou resultados (YILDIRIM; BIRANT; ALPYILDIZ, 2019).

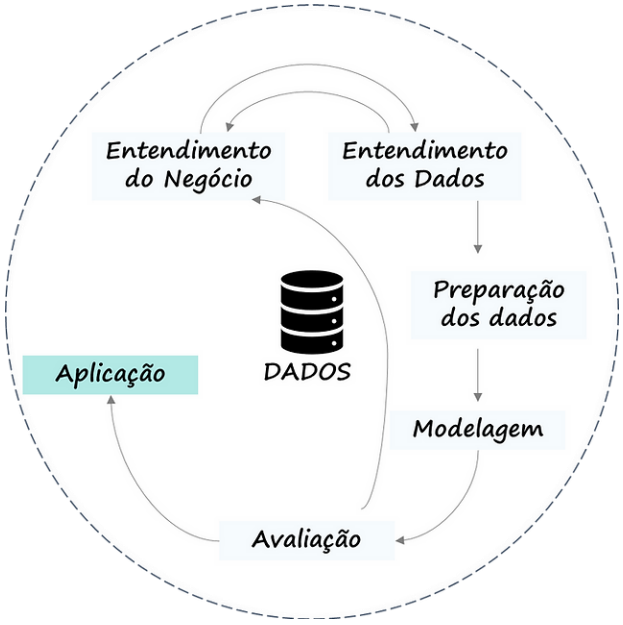
As técnicas de MD têm sido recentemente utilizadas em estudos para processar dados têxteis e convertê-los em padrões úteis a fim de obter conhecimentos valiosos e tomar as decisões corretas aumentando assim a qualidade e a produtividade.

Algumas das vantagens dos métodos de mineração de dados na indústria têxtil são (YILDIRIM; BIRANT; ALPYILDIZ, 2019):

- Predição de parâmetros esperados com base em outros parâmetros;
- Construção de modelos para reduzir o consumo de materiais têxteis, como tecidos, fios, tinturas e linhas de costura;
- Descobrir padrões que podem ser usados para produzir melhores produtos finais têxteis;
- Analisando dados têxteis para obter melhor satisfação do cliente;
- Reconhecimento e classificação de defeitos têxteis para controle de qualidade.

Uma codificação útil do processo de mineração de dados é dada pelo Processo Padrão de Indústria Cruzada para Exploração de Dados (CRISP-DM) e esse processo é ilustrado pela figura 14 (PROVOST; FAWCETT, 2016).

Figura 16 - Processo de mineração de Dados CRISP-DM



Fonte: JUNIOR, 2023

Criado em 1996, CRISP-DM é um modelo de processo que fornece uma abordagem estruturada e sistemática para o desenvolvimento de projetos de mineração de dados. Ele foi concebido para orientar os profissionais de ciência de dados em todas as etapas do ciclo de vida de um projeto de mineração de dados, desde a compreensão do negócio até a implementação das soluções. O CRISP-DM consiste em seis fases inter-relacionadas, como mostrado na Figura 14.

Inicialmente, é vital compreender o problema a ser resolvido. Isso pode parecer evidente, mas projetos de negócios raramente vêm pré-moldados como problemas claros. Reformular o problema e projetar uma solução é um processo repetitivo e o diagrama da figura 14 torna explícito o fato de que a repetição é algo recorrente durante o processo de mineração de dados (PROVOST; FAWCETT, 2016).

Se a solução do problema de negócios é o objetivo, os dados são a matéria-prima. É importante entender os pontos fortes e as limitações dos dados porque raramente há uma correspondência exata com o problema. Os dados históricos, muitas vezes são recolhidos para fins não relacionados com o problema de negócio atual ou para nenhum propósito explícito (PROVOST; FAWCETT, 2016). Nesta fase, os dados relevantes para o projeto são explorados e avaliados. Isso inclui

entender a qualidade dos dados, identificar problemas de integridade, fazer a limpeza dos dados e realizar análises exploratórias para identificar padrões preliminares.

Uma vez que os dados são compreendidos, eles precisam ser preparados para o modelo. Isso envolve integração de múltiplas fontes, seleção de variáveis relevantes, tratamento de valores nulos e transformação de dados para formatos adequados para aplicação do modelo.

Nesta etapa o modelo começa a tomar forma e pode-se ver os primeiros resultados. O tipo de modelagem a ser utilizada normalmente é definida de acordo com a necessidade do negócio e com o tipo de variável a ser analisada. Com a definição de qual modelo será utilizado, devem ser definidos quais atributos serão variáveis alvo e preditoras na construção do modelo (ROBERTO, 2022).

Com o modelo pronto, pode-se avaliar se o resultado corresponde à expectativa do projeto. Caso a resposta seja negativa ou seja considerado que há espaço para melhorias, é preciso retornar as primeiras fases do CRISP-DM para fazer as mudanças necessárias. Estas mudanças podem ter diversas formas, como a retirada de atributos estatisticamente insignificantes, correção na entrada de dados ou adequação de hiper parâmetros do modelo.

E por fim, o modelo é implantado em um ambiente operacional, onde pode ser usado para tomar decisões de negócios ou automatizar processos. Isso pode envolver integração com sistemas existentes, desenvolvimento de interfaces de usuário e implementação de procedimentos para monitorar o desempenho contínuo do modelo (ROBERTO, 2022).

É importante destacar que o CRISP-DM é um processo iterativo, o que significa que as fases podem ser revisitadas e iteradas conforme necessário para garantir que os objetivos do projeto sejam atendidos. Ele fornece uma estrutura flexível e adaptável que pode ser aplicada a uma variedade de problemas de mineração de dados em diferentes setores e domínios de negócios (ROBERTO, 2022).

5 METODOLOGIA

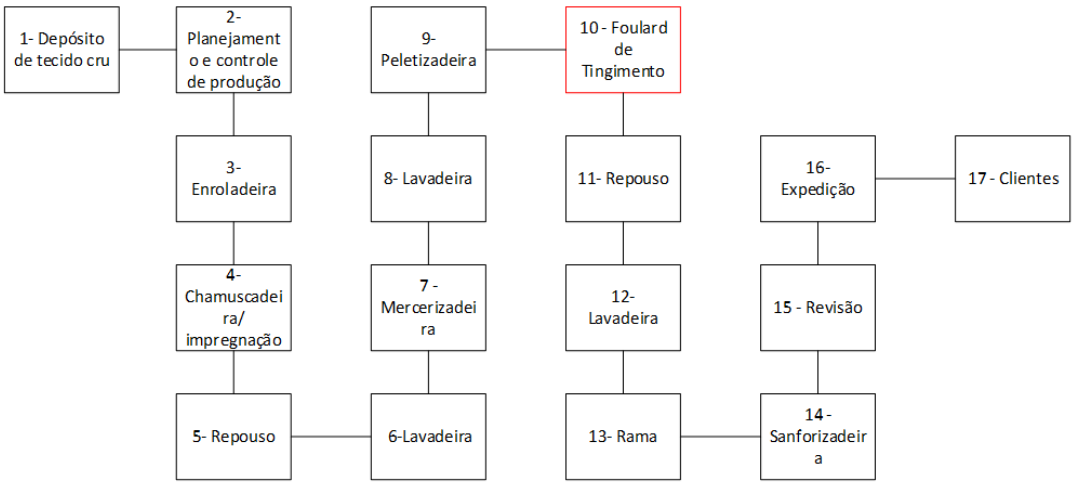
A metodologia empregada seguiu o Cross-Industry Standard Process for Data Mining (CRISP-DM), já mencionado anteriormente na revisão bibliográfica, nas quais suas etapas são apresentadas nas próximas sessões.

5.1 ENTENDIMENTO DO NEGÓCIO

Na etapa de entendimento do negócio realizou-se uma visita técnica para que houvesse uma compreensão geral de como opera o processo de produção da indústria. A empresa que é objeto de estudo é considerada referência nacional nos segmentos têxtil e confeccionista. Além disso, posiciona-se como a maior corporação verticalizada do setor têxtil paranaense e tem como matéria prima o algodão.

O processo de produção é representado pela Figura 17, que ilustra desde a chegada do algodão até o controle de qualidade onde é realizada a revisão dos tecidos já tingidos. Entretanto, apesar do processo da indústria têxtil possuir diversos estágios, este trabalho irá focar somente no processo de tingimento destacado em vermelho na figura.

Figura 17 - Diagrama de Blocos Fornecido pela Empresa



Fonte: Adaptado pela Autora, 2024.

Nesta visita foi possível conversar com os funcionários da empresa e entender melhor como é realizada cada etapa do processo, focando na revisão dos

tecidos tingidos e em como são classificados os defeitos. Por fim, a empresa disponibilizou os dados referentes a produção de metros totais de tecidos tingidos e os metros totais de tecidos com defeitos no ano de 2021. Essas foram as bases de dados utilizadas como objeto de estudo neste trabalho.

5.2 ENTENDIMENTO DOS DADOS

Neste estágio foi realizado um entendimento dos dados juntamente com a equipe técnica da empresa, na qual, analisou-se a origem, formato e estrutura dos dados disponibilizados. Os conjuntos de dados continham informações de duas diferentes fontes: setor de tingimento e controle de qualidade, destacados respectivamente no Quadro 1 e Quadro 2.

Quadro 1 - Colunas contidas nas planilhas de metros totais tingidos por dia no ano de 2021.

Coluna	Descrição
Data	Data que foi tingido o tecido
metros_maquina_1	Metros de tecido tingido na máquina 1
metros_maquina_2	Metros de tecido tingido na máquina 2
Turno	Turno de trabalho

Fonte: Elaborado pela Autora, 2024

Quadro 2 - Colunas contidas nas planilhas de metros com defeito por dia no ano de 2021.

Coluna	Descrição
data_prod	Data que foi tingido o tecido
cor	Cor do tecido
descr_item	Descrição do item
artigo	Tipo do tecido
descr_qual	Descrição da qualidade interna
metros	Metros de tecido com defeito
descr_tipo_defeito	Descrição do tipo de defeito
tipo_defeito	Código do tipo de defeito
estampado	Tecido é estampado ou não
cdpeca	Numeração da peça (etiquetado em cada rolo de tecido)
qs	Qualidade (primeira e segunda)
data_hora_revisao	Data e hora que a revisão do tecido foi feita
reprocesso	Se o tecido passou por reprocesso
numero_OB	Número da ordem de beneficiamento
turno_prod	Turno em que foi feito a revisão
nuance	Gradação de cor, tonalidade
nome_revisor	Nome de quem revisou o tecido

Fonte: Elaborado pela Autora, 2024

5.3 PREPARAÇÃO DOS DADOS

5.3.1 Limpeza Dos Dados

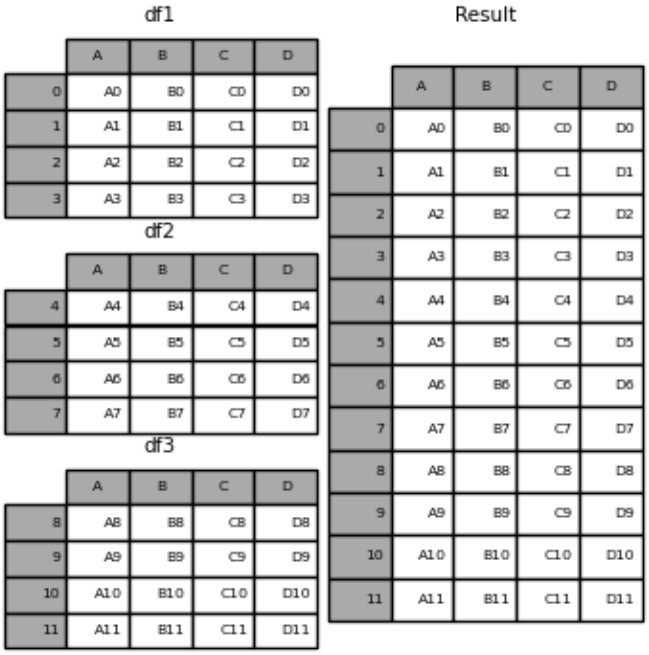
Nesta etapa é realizada uma limpeza dos dados, para garantir a qualidade e a integridade das análises. Foram verificados dados duplicados, dados ausentes, identificação de outliers, seleção de parâmetros que serão usados na análise e descarte dos que não serão necessários.

Utilizou-se algumas ferramentas para o desenvolvimento desta fase, tais como:

- A plataforma de edição de códigos Google Colaboratory, empregando a linguagem de programação Python
- A biblioteca pandas que tem como função ler, estruturar, manipular, limpar, realizar operações, visualizar, agrupar, fatiar conjuntos de dados. (“pandas - Python Data Analysis Library”, 2024).
- A biblioteca Unidecode que foi utilizada para normalização de texto e conversão de texto, como converter caracteres acentuados para caracteres sem acento (SOLC, 2024).

Os dados disponibilizados pela empresa estavam em formato de Excel para cada um dos meses do ano de 2021, por isso utilizou-se técnicas de engenharia de dados a fim de unir todos os meses em uma só planilha. Com a ferramenta *concat* da biblioteca pandas combinou-se todas as planilhas de metros totais tingidos em uma só, unindo-as verticalmente de acordo com suas colunas em comum, como mostrado na Figura 18. O mesmo processo foi feito para as planilhas de metros com defeitos.

Figura 18 - Modo de funcionamento da função concat da biblioteca pandas



Fonte: “Merge, Join, concatenate and compare - pandas 1.2.4 documentation”, 2024.

Em seguida é feito uma limpeza no conjunto de dados de metros com defeitos, nesse conjunto foram feitas a identificação das colunas relevantes, seleção dos tipos de defeitos que serão usados (mancha grave, mancha pequena média, fora de cor e parada de máquina), assim como diversos procedimentos de limpeza de dados para deixar o *dataframe* preparado para os passos seguintes.

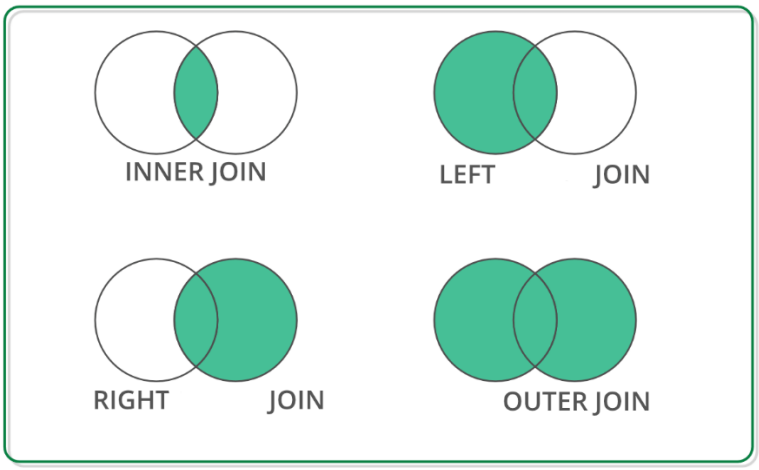
Após esses passos de limpeza, foi aplicado a função ‘*groupby()*’ da biblioteca pandas, agrupando os dados de metros de defeito por dia, mês e tipo de defeito, e então calcula-se a soma dos metros de defeito para cada combinação única desses fatores.

Para a planilha de metros totais tingidos os passos de limpeza foram somente a verificação e a remoção de dados faltantes e também utilizou-se a função ‘*groupby()*’ a fim de realizar um agrupamento (soma) de metros por turno por máquina.

Efetua-se então uma união dos dados das duas tabelas através da função '*merge()*' da biblioteca pandas. Os *DataFrames*[3] são combinados com base nas colunas que possuem em comum, que são dia e mês.

A função '*merge()*' possui alguns parâmetros que fazem com que a união entre data frames seja diferente dependendo da configuração que é escolhida, na união das duas tabelas usou-se a configuração "*inner*", ou seja, será feita uma junção interna (*inner join*). Isso significa que apenas as linhas com valores coincidentes nas colunas especificadas serão mantidas no *DataFrame* resultante. Na Figura 19 pode-se entender melhor como funciona cada uma das configurações dessa função.

Figura 19 - Tipos de junções de dados em python



Fonte: SINGH, 2020

Tem-se então um conjunto de dados com 7 colunas que estão apresentadas na Quadro 3.

Quadro 3 - Dicionário do conjunto de dados após o merge.

Coluna	Descrição
data_prod	Data do tingimento do tecido
dia	Dia do tingimento do tecido
mes	Mês do tingimento do tecido
defeitos	Tipos de defeitos
metros_defeitos	Quantidade de metros com defeitos
metros_tingidos_1	Quantidade de metros tingidos na máquina 1
metros_tingidos_2	Quantidade de metros tingidos na máquina 2

Fonte: Elaborado pela Autora, 2024.

[3] Um *DataFrame* é uma estrutura de dados bidimensional, semelhante a uma tabela ou planilha

Cria-se novas colunas a partir de metros tingidos na máquina 1 e 2, isso é melhor demonstrado na equação 1.

$$metros\_tingidos\_1 + metro\_tingidos\_2 = metros\_totais\_tingidos \quad \dots equação\ 1$$

E a partir da divisão entre metros com defeitos e metros totais tingidos obtém-se a porcentagem dos defeitos e a fração de defeitos, a equação 2 e 3 ilustram essas operações.

$$porcentagem\_defeitos = \frac{metros\_defeitos}{metros\_totais\_tingidos} * 100 \quad \dots equação\ 2$$

$$fracao\_defeitos = \frac{metros\_defeitos}{metros\_totais\_tingidos} \quad \dots equação\ 3$$

Mais 4 colunas foram criadas para saber a porcentagem e a fração que cada máquina tinge em relação ao total. As operações feitas para criar essas colunas estão demonstradas nas equações 4 a 7.

$$porcentagem\_máquina\_1 = \frac{metro\_tingidos\_1}{metros\_totais\_tingidos} * 100 \quad \dots equação\ 4$$

$$porcentagem\_máquina\_2 = \frac{metro\_tingidos\_2}{metros\_totais\_tingidos} * 100 \quad \dots equação\ 5$$

$$fracao\_máquina\_1 = \frac{metro\_tingidos\_1}{metros\_totais\_tingidos} \quad \dots equação\ 6$$

$$fracao\_máquina\_2 = \frac{metro\_tingidos\_2}{metros\_totais\_tingidos} \quad \dots equação\ 7$$

Essas 4 novas colunas serão usadas como *features*[4] posteriormente quando o modelo for criado. O novo conjunto de dados é definido pelos atributos presentes no Quadro 4.

[4] Refere-se a uma variável ou atributo específico que é usado para prever ou explicar o comportamento do alvo (ou variável de resposta)

Quadro 4 - Conjunto de Dados Após a Criação de Novas Colunas.

Coluna	Descrição
data_prod	Data do tingimento do tecido
dia	Dia do tingimento do tecido
mes	Mês do tingimento do tecido
defeitos	Tipos de defeitos
metros_defeitos	Quantidade de metros com defeitos
metros_tingidos_1	Quantidade de metros tingidos na máquina 1
metros_tingidos_2	Quantidade de metros tingidos na máquina 2
metros_totais_tingidos	Metros totais tingidos nas duas máquinas
porcentagem_defeitos	Porcentagem de defeitos
porcentagem_máquina_1	Porcentagem tingida na máquina 1
porcentagem_máquina_2	Porcentagem tingida na máquina 2
fração_máquina_1	Fração tingida na máquina 1
fração_máquina_2	Fração tingida na máquina 1

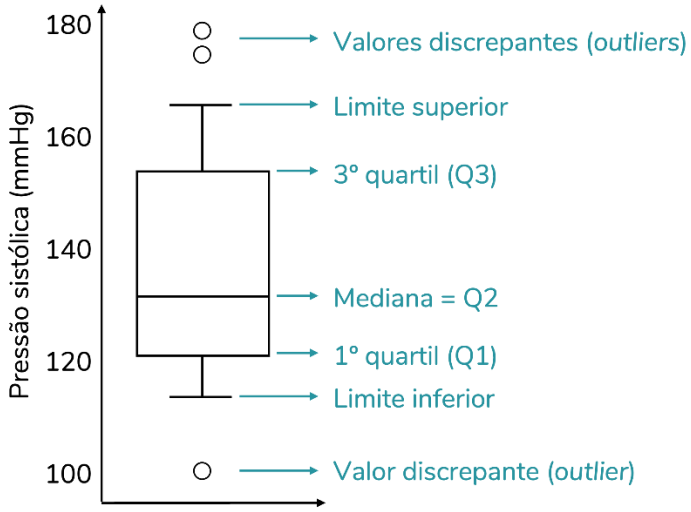
Fonte: Elaborado pela Autora, 2024.

5.3.2 Análise exploratória dos dados

A partir desse novo banco de dados verifica-se a presença de outliers. Essa verificação foi feita através de gráficos *boxplot* das colunas numéricas: metros com defeitos, metros totais tingidos e porcentagem de defeitos.

O *boxplot* é composto por seis elementos: limite inferior, quartil 1, mediana (quartil 2), quartil 3, limite superior e outliers, como mostrado na Figura 20. Os dados inconsistentes identificados como outliers foram removidos do DataFrame.

Figura 20 - Elementos que compõe um boxplot



Fonte: PERES, 2022.

Para análise da distribuição de dados foi gerado um gráfico chamado histograma, usando a biblioteca ‘*seaborn*’. O objetivo de um histograma é representar graficamente a distribuição de uma amostra de dados ou de uma população, organizando as informações de forma a facilitar a visualização dessa distribuição. Além disso, destaca a localização do valor central e a dispersão dos dados em relação a esse valor. Para complementar esse gráfico calculou-se a curtose e a assimetria da distribuição de uma variável aleatória em torno da sua média mais conhecidos como *kurtosis* e *skewness* respectivamente.

E como parte imprescindível para uma análise exploratória foram avaliados os valores de média, desvio padrão, valores mínimos, máximos e os quartis do conjunto de dados.

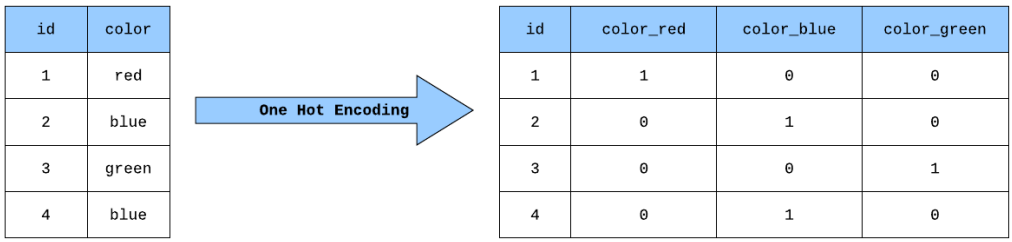
5.4 MODELAGEM

5.4.1 Pré-processamento

Na etapa de pré-processamento do modelo de *machine learning*, os dados são preparados e transformados para que possam ser utilizados de forma eficaz por algoritmos de aprendizado de máquina. Esta fase é crucial para garantir que os dados sejam adequados para o modelo e que os resultados sejam precisos e significativos.

Aplicou-se a técnica de *one-hot encoding* nas colunas de tipos de defeitos. Essa técnica foi usada para transformar as variáveis categóricas contidas na coluna tipos de defeitos em novas colunas de vetores binários. A transformação de variáveis categóricas é necessária para métodos de aprendizado de máquina que operam exclusivamente com atributos numéricos (PINHEIRO, 2021). *One-hot encoding* é uma técnica usada para representar variáveis categóricas como vetores binários como ilustrado na Figura 21.

Figura 21 - Princípio de funcionamento do *one-hot encoding*



Fonte: NOVACK, 2020

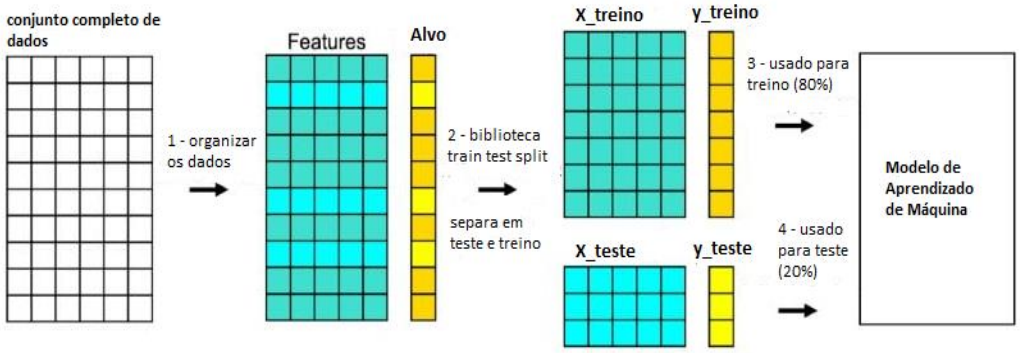
Outra mudança feita foi a transformação de natureza das colunas de dia e mês, para senos e cossenos. Transformar datas em colunas de senos e cossenos é uma técnica comumente utilizada em aprendizado de máquina quando se deseja capturar padrões sazonais ou cíclicos nos dados temporais. Isso é particularmente útil em séries temporais onde a sazonalidade é uma característica importante. Ao representar datas como senos e cossenos, é possível capturar informações sobre a periodicidade dos dados ao longo do tempo (BESCOND, 2021).

Realizou-se o *MinMaxScaler* da biblioteca *Sklearn* para normalizar os dados dentro de um intervalo específico. O banco de dados possuía colunas com dados de magnitudes muito diferentes isso pode vir a comprometer a precisão dos modelos, assim se faz necessário essa normalização. A ideia básica por trás dessa técnica é reescalar os dados de forma que todos os atributos estejam dentro de um determinado intervalo entre 0 e 1. Para cada valor em uma coluna de dados, a função subtrai o valor mínimo da coluna e divide pelo intervalo (a diferença entre o máximo e o mínimo). A fórmula para essa função é dada pela equação 8 (“sklearn.preprocessing.MinMaxScaler — scikit-learn 0.15-git documentation”, s.d.)

$$X_{norm} = \frac{X - X_{min}}{X_{máx} - X_{mín}} \quad \ldots equação\ 8$$

Para finalizar a etapa de pré-processamento é necessário dividir o conjunto de dados em teste e treino. Essa divisão é requerida para garantir que o modelo possa generalizar bem para novos dados e evitar fenômenos como o *overfitting.* Para essa tarefa foi usado o ‘*train_test_split’* importado da biblioteca *‘sklearn.model_selection’*. A divisão feita foi, 20% dos dados direcionados para teste e 80% dos dados direcionados para treino. Esse processo pode ser descrito pela Figura 22 abaixo.

Figura 22 - Procedimento de separação de dados em teste e treino



Fonte: Adaptado de GALARNYK, 2022.

5.4.2 Treinamento do modelo

Logo após os dados serem separados foi feito o treino dos modelos de aprendizado de máquina. Essa etapa tem como finalidade gerar as previsões de porcentagem de defeitos. No caso deste trabalho o modelo escolhido foi Floresta Aleatória. Esse modelo é um algoritmo que pertencem a categoria ensemble, onde múltiplos modelos são combinados parar formar um modelo mais forte e robusto.

Para que houvesse um melhor desempenho do modelo ajustou-se os hiperparâmetros de cada um dos modelos. Esse ajuste é chamado de *fine-tuning* e visa testar diversos valores de hiperparâmetros a fim de encontrar o que melhor performa no algoritmo.

Os hiperparâmetros são configurações ajustáveis que determinam o comportamento e o desempenho do modelo durante o treinamento. Ao contrário dos parâmetros do modelo, que são aprendidos durante o treinamento com base nos dados, os hiperparâmetros são definidos antes do treinamento e geralmente não são alterados durante o processo de aprendizado (HERICLIS, 2022).

No presente trabalho, foram selecionados 3 hiperparâmetros para realizar o ajuste, sendo estes: *n_estimators, max_depth* e *min_samples_split.*

O hiperparâmetro *n_estimators* refere-se ao número de árvores na floresta, em geral, quanto maior é seu valor, melhor a acurácia de predição. Porém, o aumento deste número é benéfico até certo ponto, uma vez que a partir de certa quantidade de

árvores, a melhoria na resposta combinada cessa, além do fato de um maior número de árvores consumirem uma maior quantidade de recursos computacionais (JÚNIOR, 2018).

O hiperparâmetro *max_depth* refere-se à profundidade máxima permitida para cada árvore na floresta. Em uma árvore de decisão, a profundidade representa o número de níveis de decisão, ou seja, o número de divisões que a árvore faz antes de chegar a uma folha que contém a previsão. Ao definir o *max_depth*, limita-se o crescimento da árvore, evitando que ela se torne muito complexa e se ajuste demais aos dados de treinamento (SCIKIT-LEARN, 2018; ARYA, 2022; HARRISON, 2019).

E por fim, o hiperparâmetro *min_sample_split* especifica o número mínimo de amostras necessárias para dividir um nó interno em dois outros nós durante a construção da árvore. Quando o número de amostras em um nó é menor ou igual ao *min_samples_split*, o nó não será dividido e se tornará uma folha (um nó terminal), ou seja, não haverá mais divisões abaixo dele. Isso ajuda a controlar a complexidade da árvore e a evitar que ela se torne muito detalhada o que pode levar ao sobreajuste (*overfitting*) (SCIKIT-LEARN, 2018; ARYA, 2022; HARRISON, 2019).

A performance dos modelos após o *fine-tuning* foi verificada através das métricas de avaliação.

5.5 MÉTRICAS DE AVALIAÇÃO DE REGRESSÃO

Para a avaliação de cada modelo calculou-se as métricas MAPE, RSME, MSE e MAE que são as principais métricas utilizadas para avaliar a performance de regressões. Estas serão apresentadas nos tópicos abaixo.

5.5.1 Erro quadrático médio (MSE)

É uma métrica de avaliação excelente para problemas em que grandes erros não são permitidos. Consiste na média dos erros das previsões ao quadrado. Pode-se dizer que quanto maior o valor de MSE, pior o modelo é. Apresenta valor mínimo de 0 (CAI et al., 2020; AZANK, 2020). É descrito pela equação 9.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2 \quad \dots equação\ 9$$

Onde $\widehat{y_i}$ representa o valor predito, y o valor real e m o número de amostras.

5.5.2 Raiz do erro quadrático médio (RSME)

É muito usado para melhorar a interpretabilidade da métrica. Pode ser representado pela equação 10 (AZANK, 2020).

$$RSME = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y_i})^2} \quad \dots equação\ 10$$

Onde $\widehat{y_i}$ representa o valor predito, y o valor real e m o número de amostras.

5.5.3 Erro absoluto médio (MAE)

É dado pela média das distâncias entre valores preditos e reais. É uma métrica sólida para modelos que preveem muitos dados. É descrita pela equação 11.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \widehat{y_i}| \quad \dots equação\ 11$$

Onde $\widehat{y_i}$ representa o valor predito, y o valor real e m o número de amostras.

5.5.4 Erro percentual absoluto médio (MAPE)

Essa medida exprime uma porcentagem, obtida através da divisão da diferença entre predito ($\hat{y}_i$) e real pelo valor real (y_i). O MAPE é expresso como uma porcentagem, o que facilita a interpretação. A métrica com valor de 10% indica que, em média, as previsões estão erradas em 10% em relação aos valores reais. Assim como o MSE e o MAE, quanto menor o valor, mais preciso será o modelo de regressão

(AZANK, 2020). A fórmula do erro percentual absoluto médio pode ser expressa pela equação 12.

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \quad \dots equação\ 12$$

5.5.5 Coeficiente de Determinação (R^2)

É um dos índices mais importantes para verificação da precisão do resultado de um algoritmo de regressão. Varia de 0 a 1 e pode ser representado em porcentagem. Quanto maior for o valor de $R^{2,}$ melhor será o resultado do ajuste.

Chamado também de coeficiente de determinação, essa métrica visa expressar a quantidade da variância dos dados do modelo construído, ou seja, calcula qual a porcentagem da variância que pôde ser prevista pelo modelo de regressão, dizendo o quão próximo as medidas reais estão do modelo criado (CAI et al., 2020; AZANK, 2020). Sua fórmula é dada pela equação 13.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \widehat{y_{i)}}^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2} \quad \dots equação\ 13$$

5.6 AVALIAÇÃO DE DESEMPENHO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Para esta etapa, realizou-se uma comparação das métricas de regressão dos dois modelos através de gráficos. Além da comparação entre as métricas foi feito gráficos de dispersão para análise do desempenho dos modelos.

6 RESULTADOS E DISCUSSÃO

Como explicado no capítulo de metodologia, neste trabalho foi utilizado a metodologia CRISP-DM. Realizou-se então o entendimento do negócio e o entendimento dos dados conforme especificado no capítulo anterior, e os resultados das etapas de preparação dos dados, modelagem e avaliação dos modelos são detalhados nas seções a seguir.

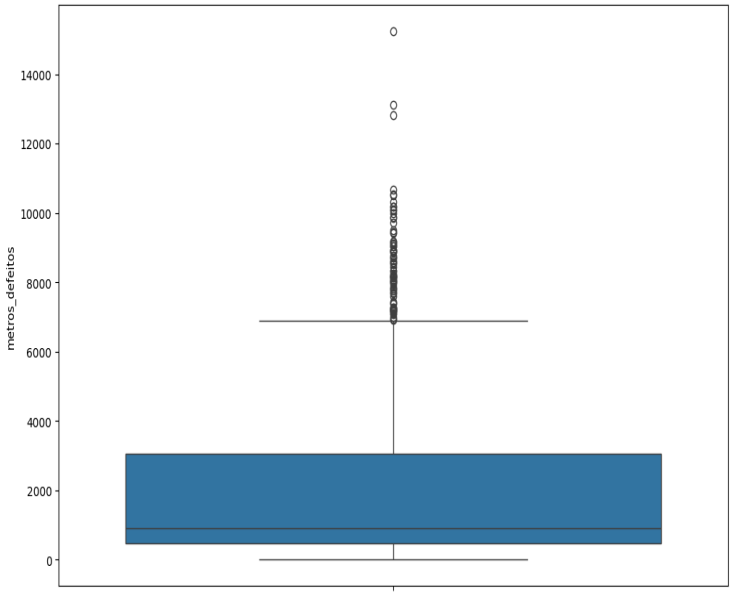
6.1 PREPARAÇÃO DOS DADOS

6.1.1 Análise Exploratória dos Dados

Iniciou-se a preparação dos dados com uma análise exploratória, começando pela avaliação das variáveis de interesse para o modelo e a identificação de possíveis outliers, para isso utilizou-se a técnica de *boxplots*.

A primeira variável analisada foi metros com defeitos, apresentada na Figura 23. Pode-se observar que esse gráfico apresenta o retângulo central deslocado para baixo, indicando uma assimetria na distribuição dos dados. Outro ponto a ser notado é a abrangência dos limites superior e inferior. O limite superior é longo, apontando uma alta dispersão dos dados, já o limite inferior é menos amplo, sugerindo que a maioria dos dados está concentrado em uma faixa relativamente estreita. Nota-se que a mediana se encontra deslocada para a parte inferior da caixa, isso indica também que a distribuição dos dados não segue a distribuição normal (MCLEOD, 2019).

Figura 23 - Boxplot da variável metros com defeitos

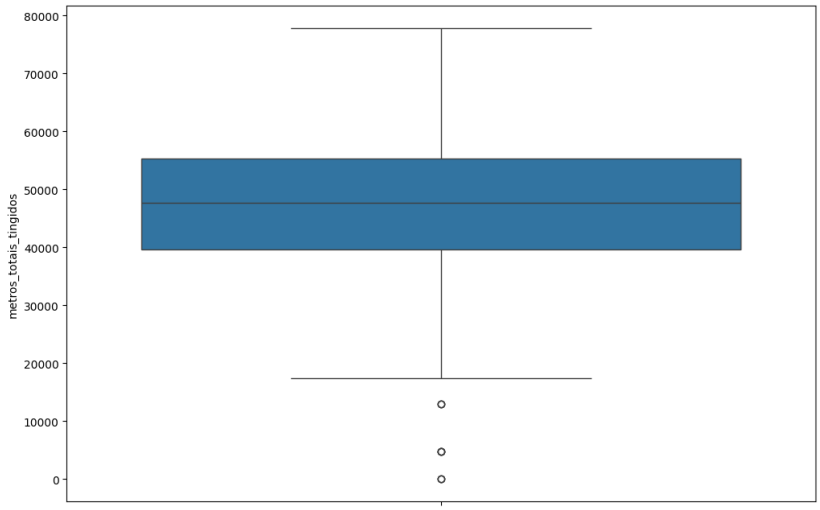


Fonte: Elaborado pela Autora, 2024.

Ao analisar os *outliers* da Figura 23 entendeu-se que esses não são erros de coleta de dados, mas sim um dia atípico na produção da empresa onde houve muitos metros de tecido com defeitos. Foi definido então, devido ao impacto que poderiam causar nos modelos, que os dados mais distantes do limite superior, mais especificamente os dados acima de 30%, seriam considerados outliers, deste modo, foram eliminados do conjunto de dados, evitando assim uma distorção nas análises estatísticas e viés nas previsões do modelo.

O *boxplot* da Figura 24 apresenta a extensão dos limites superior e inferior iguais, isso mostra que os dados de metros totais tingidos não estão concentrados em uma faixa nem apresentam uma alta dispersão. A mediana se encontra no meio da caixa, o que significa que os dados estão bem distribuídos, porém a caixa não está centralizada no meio do gráfico, indicando uma pequena assimetria na distribuição dos dados (MCLEOD, 2019).

Figura 24 - Boxplot da variável metros totais tingidos



Fonte: Elaborado pela Autora, 2024.

Nesse mesmo gráfico foram identificados 3 valores discrepantes que se encontram abaixo do limite inferior do *boxplot,* aproximadamente abaixo de 2000 metros. Ao analisar esses valores, constatou-se que esses eram dados com informações incompletas. Uma das hipóteses seria que os dados não foram coletados no dia, devido a algum problema no setor de tingimento. Esses dados faltantes foram

removidos do *dataset*[5] pois podem introduzir um viés nos resultados, impactando a performance do modelo.

A Tabela 1 apresenta as medidas de tendência e dispersão de cada coluna contida no conjunto de dados.

Pode-se notar que apenas as variáveis relacionadas ao total de defeitos têm valor de média diferente da mediana, o que é um indicativo que estes dados não seguem uma distribuição normal (isto é comprovado mais adiante com o histograma). Outra coisa que chama a atenção é que o desvio padrão dos metros com defeitos tem valor superior ao próprio valor da média.

Tabela 1 - Medidas de tendência e dispersão de cada variável.

Colunas	Média	Desvio Padrão	Mediana
Metros com defeitos	2069,96	2526,26	899,12
Metros tingidos na máquina 1	23786,02	7307,11	23901
Metros tingidos na máquina 2	23428,27	7073,22	23866,5
Metros totais tingidos	47214,29	11202,83	47690,24
Porcentagem de defeitos	4,59	5,73	1,94
Porcentagem tingida na máquina 1	50,34	12,21	50,35
Porcentagem tingida na máquina 2	49,65	12,21	49,64
Fração tingida na máquina 1	0,50	0,12	0,50
Fração tingida na máquina 2	0,49	0,12	0,49

Fonte: Elaborado pela Autora, 2024.

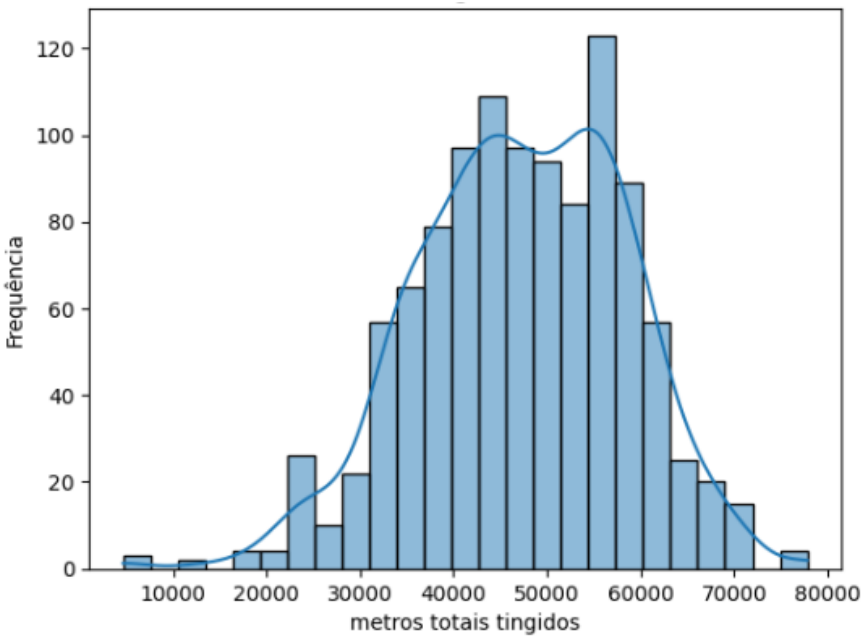
A curtose de metros totais tingidos teve um valor de 0,08, indicando uma distribuição platicúrtica, com caudas mais leves e um pico mais baixo que a distribuição normal. Isso implica que o pico da distribuição é mais baixo e mais achatado do que o pico de uma distribuição normal, sugerindo também que há uma menor concentração de dados ao redor da média e mais dispersão nos dados (TUYCHIEV, 2023).

Analisando a assimetria, obteve-se um valor -0,27 apontando uma distribuição negativamente assimétrica, com uma cauda longa a esquerda (LARSON;

[5] Conjunto de dados.

FARBER, 2006; TUYCHIEV, 2023). Ao observar o gráfico da Figura 25 é possível observar todas esses pontos apresentados.

Figura 25 - Histograma da Variável Metros Totais Tingidos

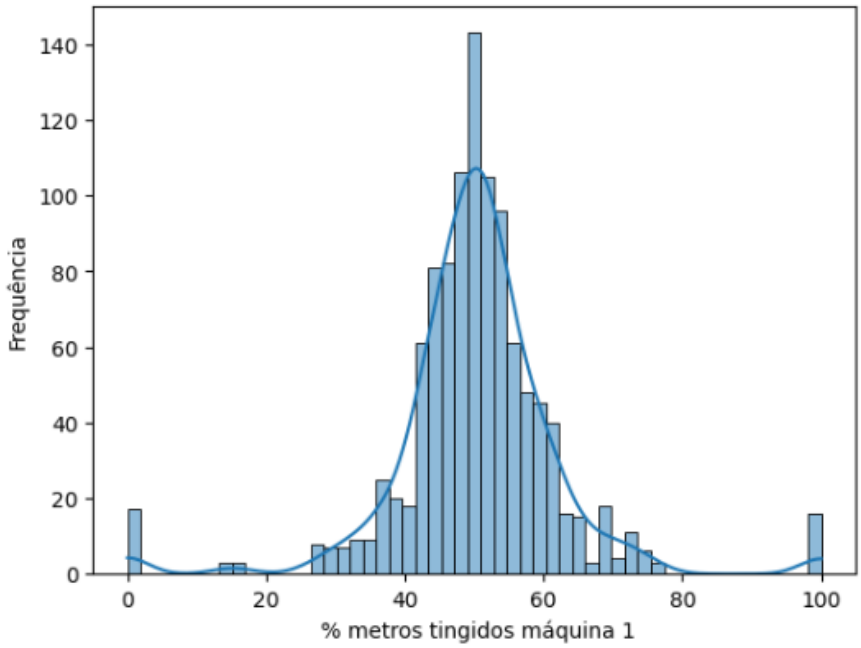


Fonte: Elaborado pela Autora, 2024.

O histograma apresentado na Figura 26 apresenta a frequência da distribuição de percentagem de metros tingidos na máquina. A curtose dessa variável é de 6,70, apontando que a distribuição é leptocúrtica com caudas mais pesadas e um pico alto e estreito que a distribuição normal indicando uma maior concentração dos dados dessa variável ao redor da média e também uma menor dispersão (TUYCHIEV, 2023).

O valor da assimetria da foi de -0,15 indicando que a distribuição está enviesada à esquerda, assim como nesses casos a média tende a ser menor que a mediana, o que pode ser provado pela tabela 1 onde a média de 50,34 e a mediana é de 50,35 (LARSON; FARBER, 2006; TUYCHIEV, 2023).

Figura 26 - Histograma da variável porcentagem de metros tingidos na máquina 1.



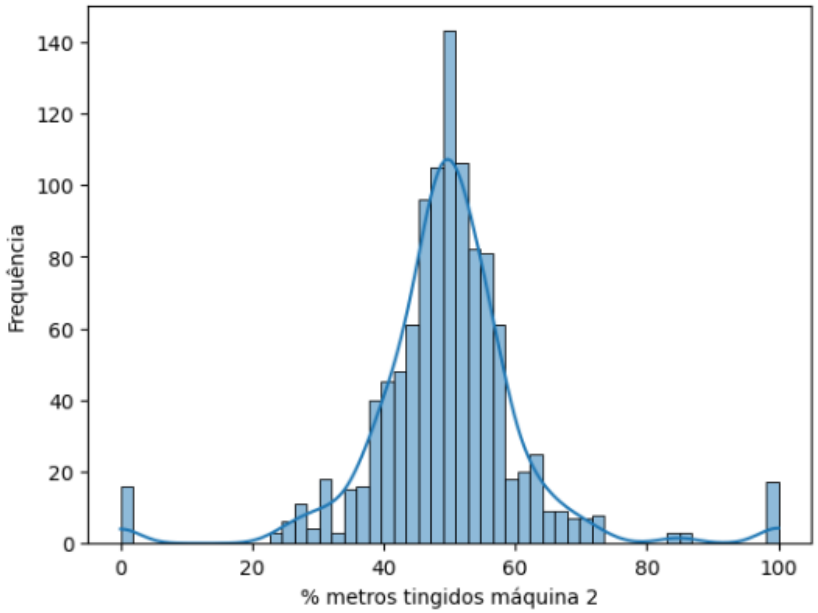
Fonte: Elaborado pela Autora, 2024.

É possível identificar nesse gráfico (figura 26), valores discrepantes na distribuição dos dados, isso é visto a partir das barras isoladas, localizadas na extremidade do gráfico nos valores de 0 e 100. Esses não foram considerados valores atípicos que necessitam ser removidos. Trata-se apenas de uma parada da máquina 1, no momento em que a máquina 2 realizava o tingimento e vice-versa, ou seja, representam valores reais do processo.

De forma similar, pode-se observar a distribuição dos dados de percentual de metros tingidos na máquina 2 (Figura 27), apresentando perfil próximo do apresentado na Figura 26, isso transparece também na curtose e assimetria, cujo os valores são 6,70 e 0,15 respectivamente. Esses valores indicam que a distribuição permanece leptocúrtica, porém a assimetria agora é positiva o que indica uma distribuição enviesada para direita (TUYCHIEV, 2023).

O gráfico dessas duas variáveis (Figura 26 e Figura 27) são espelhados e apresentam as mesmas particularidades, porém em lados opostos. Isso se dá devido ao fato de que estas duas variáveis são linearmente dependentes e complementares. Ou seja, se o percentual na máquina 1 for 30% necessariamente o percentual na máquina 2 é 70%.

Figura 27 - Histograma da variável porcentagem de metros tingidos na máquina 2



Fonte: Elaborado pela Autora, 2024.

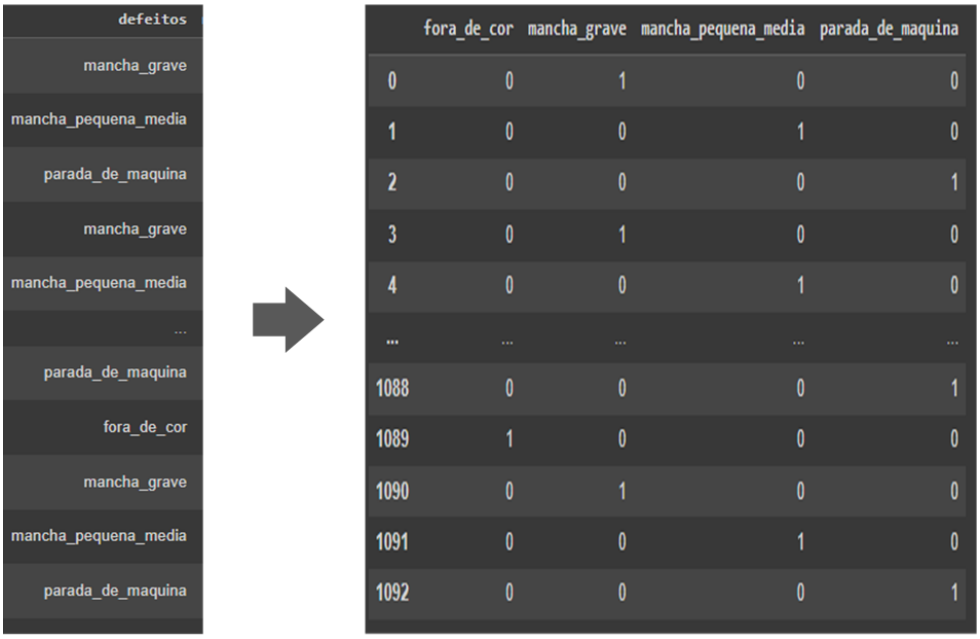
6.1.2 Pré-Processamento dos Dados

O pré-processamento de dados é um conjunto de atividades que envolvem converter dados brutos em dados preparados, ou seja, em formatos úteis e eficientes para aplicação do modelo. Além disso, é a etapa onde se escolhe quais dados devem compor o *dataset*.

Para iniciar essa etapa foi feito um estudo de quais seriam as variáveis alvo e preditoras do conjunto de dados. Como o intuito do trabalho é prever os defeitos a partir de dados de 2021, a porcentagem de defeitos foi considerada a variável alvo. Já dia, mês, metros totais tingidos, tipos de defeitos, fração de metros tingidos na máquina 1 e porcentagem de metros tingidos na máquina 1 foram consideradas variáveis preditoras.

Foi necessário então realizar algumas transformações nesse *dataset* para adequar os dados para aplicação do modelo. A primeira transformação foi a codificação da variável categórica ‘tipo de defeitos’ utilizando a técnica *one-hot encoding*. Na Figura 28 é possível observar o resultado dessa transformação.

Figura 28 - Resultado da aplicação da técnica one-hot encoding na variável 'defeitos'



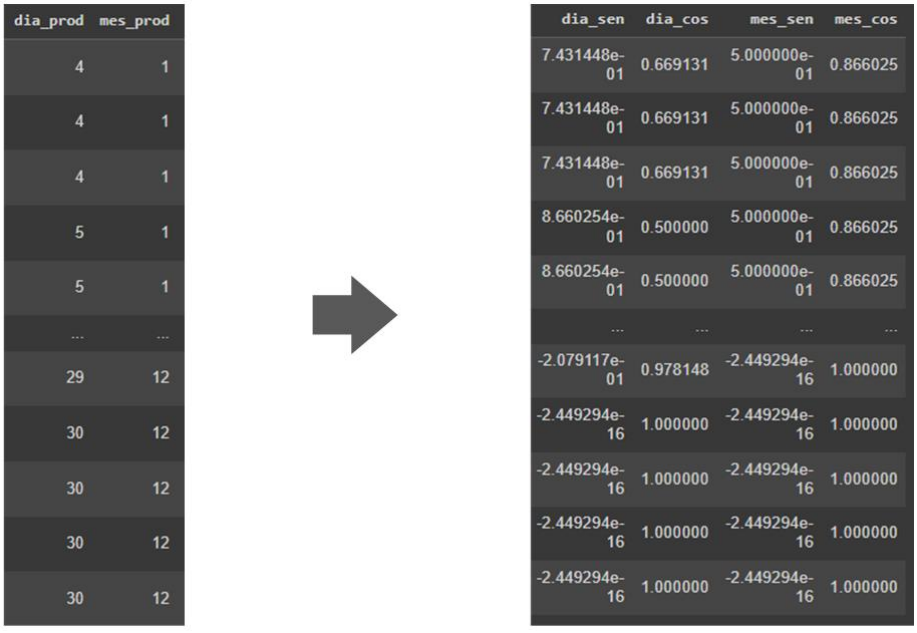
defeitos
mancha_grave
mancha_pequena_media
parada_de_maquina
mancha_grave
mancha_pequena_media
...
parada_de_maquina
fora_de_cor
mancha_grave
mancha_pequena_media
parada_de_maquina

	fora_de_cor	mancha_grave	mancha_pequena_media	parada_de_maquina
0	0	1	0	0
1	0	0	1	0
2	0	0	0	1
3	0	1	0	0
4	0	0	1	0
...	...	...	...	...
1088	0	0	0	1
1089	1	0	0	0
1090	0	1	0	0
1091	0	0	1	0
1092	0	0	0	1

Fonte: Elaborado pela Autora, 2024.

Foi preciso também fazer a transformação das colunas de datas para colunas de senos e cossenos, essa mudança de natureza está ilustrada na Figura 29.

Figura 29 - Transformação realizada nas variáveis de tempo



dia_prod	mes_prod
4	1
4	1
4	1
5	1
5	1
...	...
29	12
30	12
30	12
30	12
30	12

dia_sen	dia_cos	mes_sen	mes_cos
7.431448e-01	0.669131	5.000000e-01	0.866025
7.431448e-01	0.669131	5.000000e-01	0.866025
7.431448e-01	0.669131	5.000000e-01	0.866025
8.660254e-01	0.500000	5.000000e-01	0.866025
8.660254e-01	0.500000	5.000000e-01	0.866025
...	...	...	...
-2.079117e-01	0.978148	-2.449294e-16	1.000000
-2.449294e-16	1.000000	-2.449294e-16	1.000000
-2.449294e-16	1.000000	-2.449294e-16	1.000000
-2.449294e-16	1.000000	-2.449294e-16	1.000000
-2.449294e-16	1.000000	-2.449294e-16	1.000000

Fonte: Elaborado pela Autora, 2024.

6.2 MODELAGEM

6.2.1 Teste do Modelo 1

A Figura 30 representa o *dataset* usado para testar o modelo 1, que é composto por 1086 linhas e 11 variáveis ou colunas.

Figura 30 - Conjunto de dados usado no teste 1

	dia_sen	dia_cos	mes_sen	mes_cos	metros_totais_tingidos	porcentagem_maquina_1	fora_de_cor	mancha_grave	mancha_pequena_media	parada_de_maquina	porcentagem_defeitos
0	7.431448e-01	0.669131	5.000000e-01	0.866025	45933.00	51.348704	0	1	0	0	1.714345
1	7.431448e-01	0.669131	5.000000e-01	0.866025	45933.00	51.348704	0	0	1	0	6.104652
2	7.431448e-01	0.669131	5.000000e-01	0.866025	45933.00	51.348704	0	0	0	1	0.740535
3	8.660254e-01	0.500000	5.000000e-01	0.866025	55072.00	46.419233	0	1	0	0	0.822051
4	8.660254e-01	0.500000	5.000000e-01	0.866025	55072.00	46.419233	0	0	1	0	8.648787
...	...	...	...	...	...	...	...	...	...	...	...
1081	-2.079117e-01	0.978148	-2.449294e-16	1.000000	70016.29	51.011943	0	0	0	1	1.294542
1082	-2.449294e-16	1.000000	-2.449294e-16	1.000000	28476.00	30.116589	1	0	0	0	0.702346
1083	-2.449294e-16	1.000000	-2.449294e-16	1.000000	28476.00	30.116589	0	1	0	0	4.069883
1084	-2.449294e-16	1.000000	-2.449294e-16	1.000000	28476.00	30.116589	0	0	1	0	11.497542
1085	-2.449294e-16	1.000000	-2.449294e-16	1.000000	28476.00	30.116589	0	0	0	1	3.729983

Fonte: Elaborado pela Autora, 2024.

Após o conjunto de dados estar preparado para poder ser aplicado o modelo foi preciso separá-lo em dados de teste e treino. Para isso, primeiro foi necessário organizar os dados em variáveis alvo e preditoras (Quadro 5)

Quadro 5 - Variáveis Alvo (X) e Preditoras (y) do Modelo 1.

Variáveis Preditoras (X)	Variável Alvo (y)
dia_sen dia_cos mes_sen mes_cos metros_totais_tingidos porcentagem_maquina_1 fora_de_cor mancha_grave mancha_pequena_media parada_de_máquina	porcentagem _defeitos

Fonte: Elaborado pela Autora, 2024.

Para aplicar o modelo utilizou-se o módulo '*sklearn.ensemble'* da biblioteca 'sklearn' do qual foi importado o algoritmo floresta aleatória. A fim de melhorar o desempenho do modelo foi realizado o fine tuning, que consiste em testar os valores dos hiperparâmetros do algoritmo. Os hiperparâmetros ajustados foram *n_estimators*, *max_depth* e *min_samples_split*. Após cada ajuste calculou-se as métricas de

avaliação para verificar o desempenho do modelo. Os resultados estão dispostos na Tabela 2 abaixo.

Tabela 2 - Valores das métricas de avaliação do modelo 1.

Ajuste	Hiperparâmetros	MAPE	RSME	MSE	MAE	R^2
1	n_estimators=100 max_depth=40 min_samples_split=15	0,77	2,27	5,18	1,27	0,83
2	n_estimators=100 max_depth=30 min_samples_split=15	0,77	2,27	5,18	1,27	0,83
3	n_estimators=100 max_depth=25 min_samples_split=15	0,77	2,27	5,18	1,27	0,83
4	n_estimators=100 max_depth=40 min_samples_split=10	0,78	2,28	5,22	1,28	0,83
5	n_estimators=100 max_depth=40 min_samples_split=20	0,78	2,26	5,13	1,27	0,8409
6	n_estimators=100 max_depth=40 min_samples_split=25	0,77	2,25	5,07	1,26	0,8427
7	n_estimators=100 max_depth=40 min_samples_split=30	0,77	2,24	5,02	1,25	0,8443
8	n_estimators=100 max_depth=40 min_samples_split=40	0,75	2,22	4,96	1,24	0,8462
9	n_estimators=100 max_depth=40 min_samples_split=45	0,76	2,22	4,95	1,24	0,8465
10	n_estimators=150 max_depth=40 min_samples_split=45	0,76	2,22	4,96	1,25	0,8463

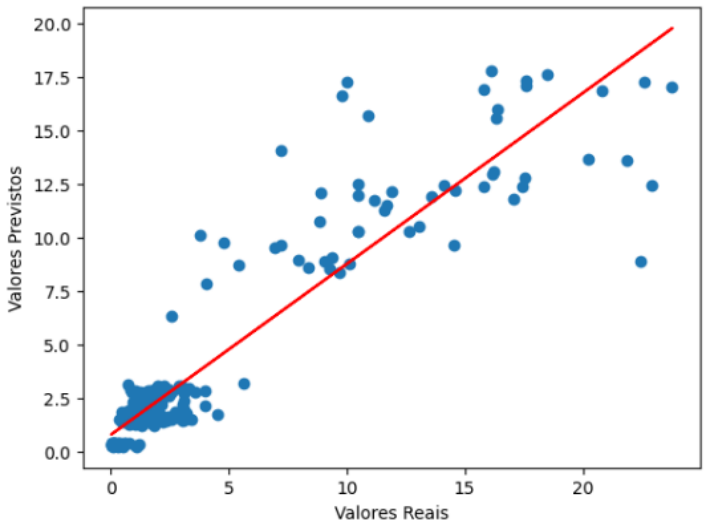
Fonte: Elaborado pela Autora, 2024.

Considerou-se o ajuste 8 da Tabela 2 o mais promissor comparado aos demais ajustes. O valor obtido de R^2 de 0,8462 no ajuste 8 significa que aproximadamente 84,62% da variabilidade dos dados é explicada pelas variáveis independentes incluídas no modelo, ou seja, o modelo está sendo eficaz em prever os dados. Analisando também a métrica MAPE, obteve-se um valor de 0,75, isso indica que as previsões do modelo estão erradas em 0,75% em relação aos valores reais.

Com a finalidade de melhor avaliar o algoritmo testado, foi gerado um gráfico de dispersão entre os valores preditos pelo modelo e os valores reais. Esse gráfico é muito útil para identificar padrões, tendências, relações, correlações e avalia a adequação do modelo. A partir desse gráfico é possível avaliar visualmente o quão bem o modelo se ajustou aos dados, comparando os pontos de dados reais com as

previsões do modelo (FROST, 2021). A Figura 31 ilustra o gráfico de dispersão do modelo.

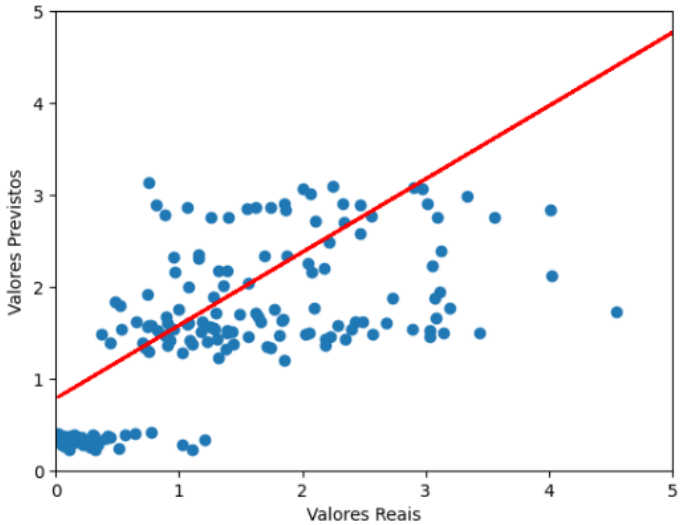
Figura 31 - Gráfico de Dispersão do Modelo 1



Fonte: Elaborado pela Autora, 2024.

O gráfico da Figura 31 mostra o quão próximo os valores preditos estão dos valores reais. Quanto mais resultados sobre a reta vermelha, melhor será a precisão do modelo. É possível notar que a maior parte dos dados são para valores de percentual de defeitos menores que 5% e os dados que estão mais distantes da diagonal são os que possuem valores mais altos de percentual de defeitos. Um zoom foi feito para poder ser observado os valores abaixo de 5% e é representado pela Figura 32.

Figura 32 - Zoom do Gráfico de Dispersão do Modelo 1

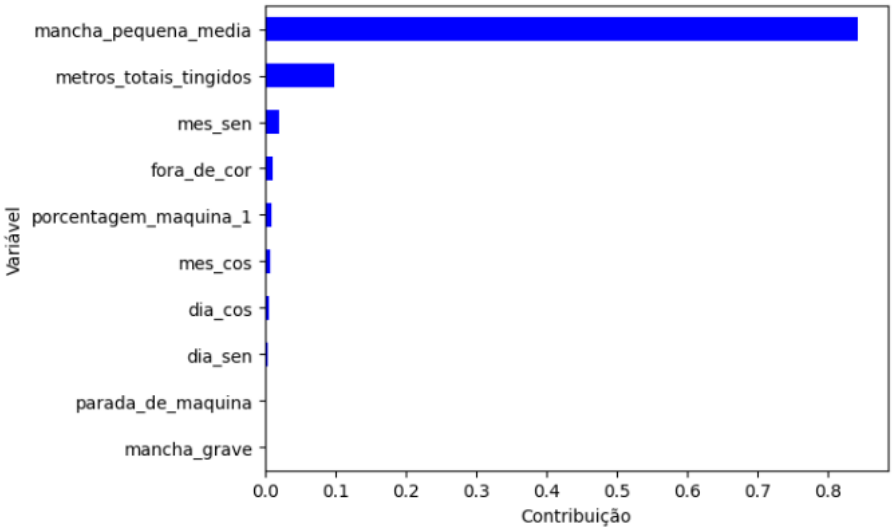


Fonte: Elaborado pela Autora, 2024.

Avaliou-se também a importância de cada variável preditora no conjunto de dados a partir da função *global feature importance.* Essa função avalia como cada recurso contribui para a capacidade do modelo de fazer previsões precisas.

Na Figura 33 é possível observar que mancha pequena média é a variável que mais contribui para as previsões feitas pelo modelo, ou seja, essa variável possui uma forte relação com a variável alvo que no caso é a porcentagem de defeitos. Seguido por mancha pequena média, tem-se metros totais tingidos que ocupa a segunda posição de contribuição para o modelo.

Figura 33 - Função Global Feature Importance Aplicada nas Variáveis Preditoras do Modelo 1

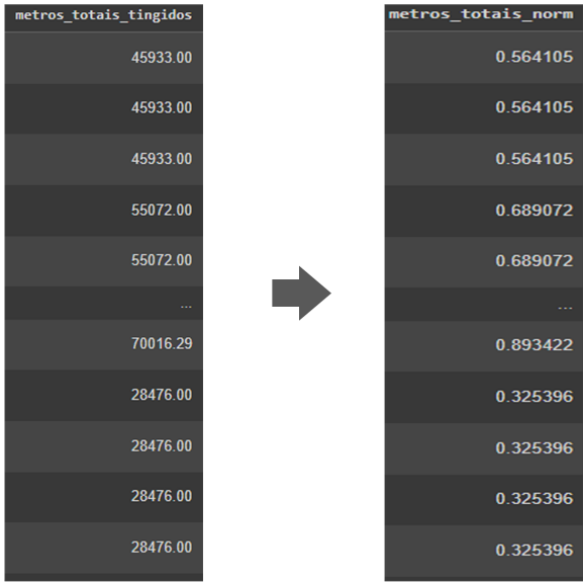


Fonte: Elaborado pela Autora, 2024.

6.2.2 Teste do Modelo 2

A fim de aprimorar o modelo 1, treinado anteriormente, buscou-se ajustar o conjunto de dados para poder realizar outro treinamento, para isso primeiro foi realizada uma técnica de pré-processamento '*MinMaxScaler'* que redimensiona os valores dos atributos em um intervalo específico entre 0 e 1. Essa técnica foi aplicada à variável metros totais tingidos devido a seus altos valores comparados aos das outras *features* do conjunto de dados. A Figura 34 mostra como ficou a variável após a normalização.

Figura 34 - Normalização da Variável Metros Totais Tingidos



metros_totais_tingidos
45933.00
45933.00
45933.00
55072.00
55072.00
...
70016.29
28476.00
28476.00
28476.00
28476.00

metros_totais_norm
0.564105
0.564105
0.564105
0.689072
0.689072
...
0.893422
0.325396
0.325396
0.325396
0.325396

Fonte: Elaborado pela Autora, 2024.

Assim, o novo conjunto de dados a ser treinado pelo modelo é dado pela Figura 35.

Figura 35 - Conjunto de Dados Usado no Modelo 2

	dia_sen	dia_cos	mes_sen	mes_cos	fora_de_cor	mancha_grave	mancha_pequena_media	parada_de_maquina	porcentagem_defeitos	fracao_maquina_1	metros_totais_norm
0	7.431448e-01	0.669131	5.000000e-01	0.866025	0	1	0	0	1.714345	0.513487	0.564105
1	7.431448e-01	0.669131	5.000000e-01	0.866025	0	0	1	0	6.104652	0.513487	0.564105
2	7.431448e-01	0.669131	5.000000e-01	0.866025	0	0	0	1	0.740535	0.513487	0.564105
3	8.660254e-01	0.500000	5.000000e-01	0.866025	0	1	0	0	0.822051	0.464192	0.689072
4	8.660254e-01	0.500000	5.000000e-01	0.866025	0	0	1	0	8.648787	0.464192	0.689072
...	...	...	...	...	...	...	...	...	...	...	...
1081	-2.079117e-01	0.978148	-2.449294e-16	1.000000	0	0	0	1	1.294542	0.510119	0.893422
1082	-2.449294e-16	1.000000	-2.449294e-16	1.000000	1	0	0	0	0.702346	0.301166	0.325396
1083	-2.449294e-16	1.000000	-2.449294e-16	1.000000	0	1	0	0	4.069883	0.301166	0.325396
1084	-2.449294e-16	1.000000	-2.449294e-16	1.000000	0	0	1	0	11.497542	0.301166	0.325396
1085	-2.449294e-16	1.000000	-2.449294e-16	1.000000	0	0	0	1	3.729983	0.301166	0.325396

Fonte: Elaborado pela Autora, 2024.

Outro ponto a ser comentado sobre esse novo conjunto de dados criado é que a variável porcentagem de metros tingidos na máquina 1 foi trocada pela variável fração de metros tingidos na máquina 1, assim o dataset é composto apenas por valores com menor ordem de grandeza.

A partir dessas alterações realizou-se a escolha das variáveis alvo e preditoras como mostrado na Quadro 6 e em seguida separou-se o dataset em dados de teste e dados de treino.

Quadro 6 - Variáveis Alvo (X) e Preditoras (y) do Modelo 2.

Variáveis Preditoras (X)	Variável Alvo (y)
dia_sen dia_cos mes_sen mes_cos metros_totais_norm fracao_maquina_1 fora_de_cor mancha_grave mancha_pequena_media parada_de_máquina	porcentagem _defeitos

Fonte: Elaborado pela Autora, 2024.

Assim como no modelo anterior utilizou-se a mesma biblioteca para importar o algoritmo de floresta aleatória e também foi realizado o *fine tuning* para obter o melhor desempenho do modelo treinado. Os ajustes dos hiperparâmetros (*fine tuning*) estão apresentados na Tabela 3.

Tabela 3 - Valores das Métricas de Avaliação do Modelo 2.

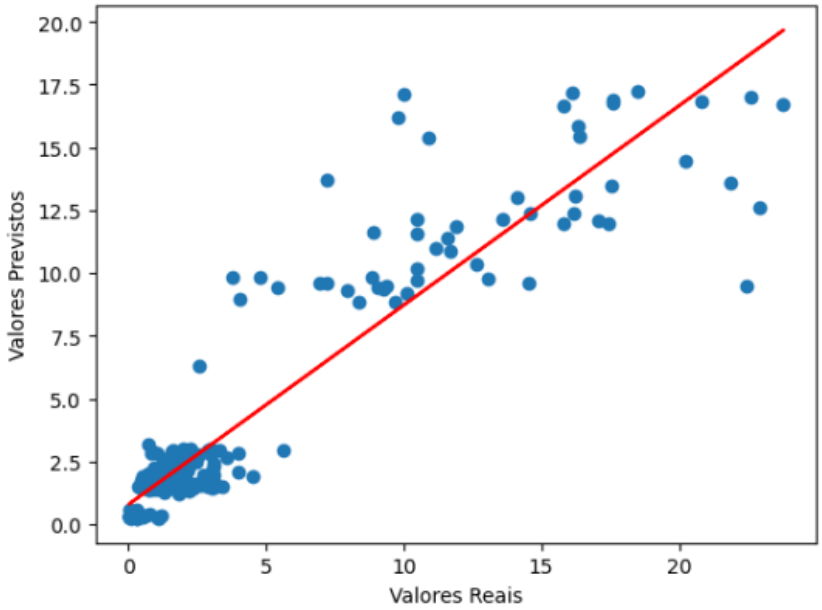
Ajuste	Hiperparâmetros	MAPE	RSME	MSE	MAE	R^2
1	n_estimators=100 max_depth=40 min_samples_split=15	0,81	2,28	5,2	1,27	0,8390
2	n_estimators=100 max_depth=30 min_samples_split=15	0,81	2,28	5,2	1,27	0,8390
3	n_estimators=100 max_depth=25 min_samples_split=15	0,81	2,28	5,2	1,27	0,8390
4	n_estimators=100 max_depth=40 min_samples_split=10	0,81	2,28	5,22	1,28	0,8383
5	n_estimators=100 max_depth=40 min_samples_split=20	0,8	2,26	5,14	1,27	0,8408
6	n_estimators=100 max_depth=40 min_samples_split=30	0,8	2,24	5,03	1,25	0,8461
7	n_estimators=100 max_depth=40 min_samples_split=40	0,79	2,22	4,96	1,25	0,8464
8	n_estimators=100 max_depth=40 min_samples_split=50	0,78	2,21	4,92	1,24	0,8476
9	n_estimators=100 max_depth=40 min_samples_split=70	0,78	2,19	4,81	1,23	0,8509
10	n_estimators=100 max_depth=40 min_samples_split=80	0,76	2,2	4,84	1,23	0,8505
11	n_estimators=150 max_depth=40 min_samples_split=70	0,78	2,19	4,81	1,23	0,8509

Fonte: Elaborado pela Autora, 2024.

De todos os ajustes testados o que teve melhor desempenho considerando todas as métricas de avaliação foi o ajuste 9 que possui um alto R^2 comparado aos ajustes 8 e 10. Observa-se também, que ao aumentar o número de árvores não causou nenhuma melhora na performance do modelo. Sendo assim o modelo foi ajustado de acordo com as configurações de hiperparâmetros dado pelo ajuste 9.

Assim como no modelo 1, também foi plotado um gráfico de dispersão (Figura 36) para verificar a performance do modelo 2.

Figura 36 - Gráfico de Dispersão do Modelo 2.

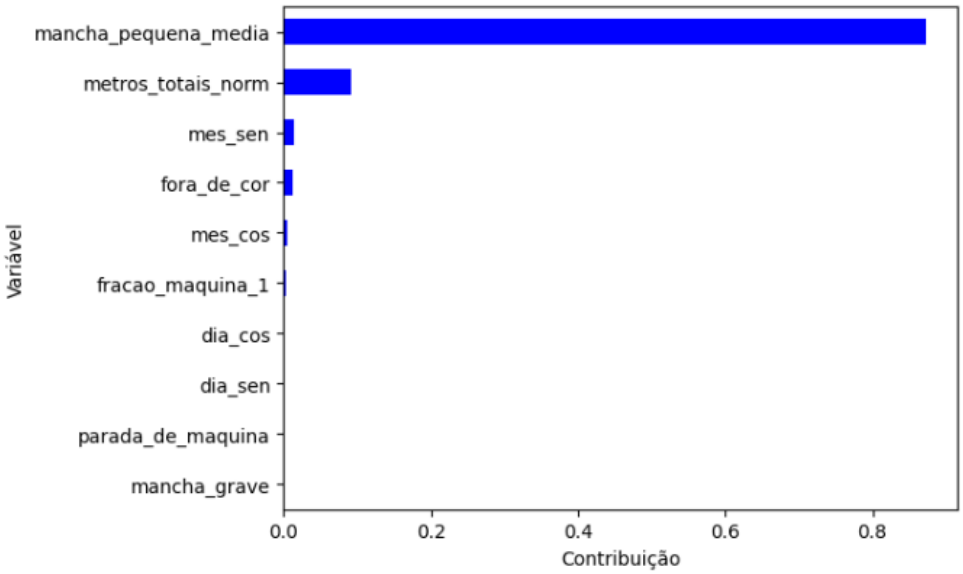


Fonte: Elaborado pela Autora, 2024

O gráfico de dispersão do modelo 2 é muito semelhante ao do modelo 1, contendo o mesmo acumulo de dados abaixo de 5% e os dados dispersos acima de 5%, isso mostra que os modelo 2 mesmo adaptado performou quase da mesma maneira que o modelo 1.

Na Figura 37 é possível observar mais uma vez que o comportamento do modelo 2 é análogo ao do modelo 1 pois as variáveis que mais contribuem são novamente mancha pequena média e metros totais normalizados.

Figura 37 - Função Global Feature Importance Aplicada nas Variáveis Preditoras do Modelo 2.



Fonte: Elaborado pela Autora, 2024

Após essas análises é possível dizer que mesmo com a normalização dos dados de metros totais tingidos o modelo 2 exibiu um desempenho quase idêntico ao do modelo 1, mas com uma queda mínima em sua performance.

6.2.3 Teste do Modelo 3

Para este modelo foram feitas algumas alterações na estrutura do *dataset* para verificar o seu desempenho em relação ao dos outros modelos criados. A primeira e principal alteração foi a retirada das variáveis que indicavam os tipos de defeitos. Essa modificação foi feita para verificar se os tipos de defeitos interferiam de alguma forma na performance do modelo.

A segunda alteração foi o uso da função *MinMaxScaler* para normalizar os dados do atributo metros totais tingidos. A Figura 38 está representado o conjunto de dados usado para treinar esse modelo 3, esse conjunto contém 1086 linhas e 7 colunas.

Figura 38 - Conjunto de Dados Usado para Treinar o Modelo 3

	dia_sen	dia_cos	mes_sen	mes_cos	porcentagem_defeitos	fracao_maquina_1	metros_totais_norm
0	7.431448e-01	0.669131	5.000000e-01	0.866025	1.714345	0.513487	0.564105
1	7.431448e-01	0.669131	5.000000e-01	0.866025	6.104652	0.513487	0.564105
2	7.431448e-01	0.669131	5.000000e-01	0.866025	0.740535	0.513487	0.564105
3	8.660254e-01	0.500000	5.000000e-01	0.866025	0.822051	0.464192	0.689072
4	8.660254e-01	0.500000	5.000000e-01	0.866025	8.648787	0.464192	0.689072
...	...	...	...	...	...	...	...
1081	-2.079117e-01	0.978148	-2.449294e-16	1.000000	1.294542	0.510119	0.893422
1082	-2.449294e-16	1.000000	-2.449294e-16	1.000000	0.702346	0.301166	0.325396
1083	-2.449294e-16	1.000000	-2.449294e-16	1.000000	4.069883	0.301166	0.325396
1084	-2.449294e-16	1.000000	-2.449294e-16	1.000000	11.497542	0.301166	0.325396
1085	-2.449294e-16	1.000000	-2.449294e-16	1.000000	3.729983	0.301166	0.325396

Fonte: Elaborado pela Autora, 2024

Após essas mudanças, procedeu-se à seleção das variáveis alvo e preditoras, conforme demonstrado no Quadro 7. Posteriormente, o conjunto de dados foi dividido em dados de teste e treinamento

Quadro 7 - Variáveis Alvo (X) e Preditoras (y) do Modelo 3.

Variáveis Preditoras (X)	Variável Alvo (y)
dia_sen dia_cos mes_sen mes_cos metros_totais_norm fracao_maquina_1	porcentagem _defeitos

Fonte: Elaborado pela Autora, 2024

Assim como nos modelos anteriores utilizou-se a classe '*RandomForestRegressor'* dentro do módulo '*sklearn.esemble'* para instanciar o algoritmo de floresta aleatória. Realizou-se o ajuste dos hiperparâmetros do modelo para encontrar a configuração com melhor desempenho. Na Tabela 4 encontram-se todos os ajustes testados.

Tabela 4 - Valores das Métricas de Avaliação do Modelo 3

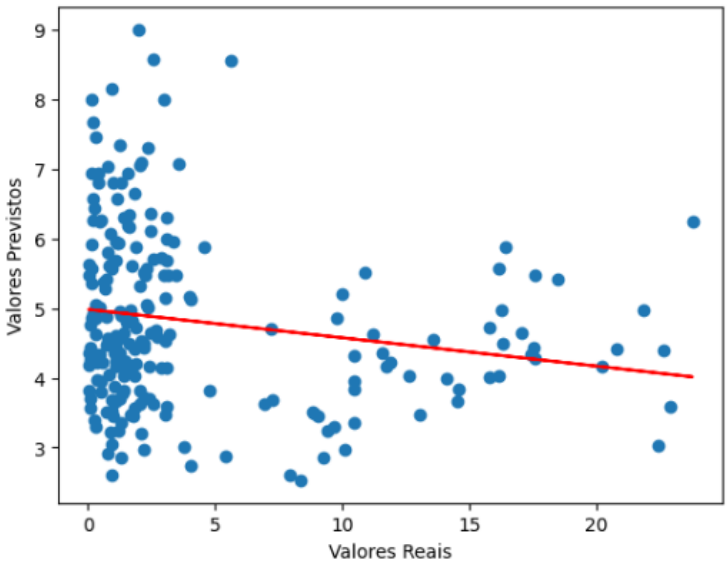
Ajuste	Hiperparâmetros	MAPE	RSME	MSE	MAE	R^2
1	n_estimators=100 max_depth=40 min_samples_split=15	9,33	6,86	47,11	5,58	-0,45
2	n_estimators=100 max_depth=40 min_samples_split=20	9,22	6,67	44,51	5,43	-0,37
3	n_estimators=100 max_depth=40 min_samples_split=30	9,02	6,42	41,25	5,23	-0,27
4	n_estimators=100 max_depth=40 min_samples_split=40	8,84	6,27	39,32	5,09	-0,21
5	n_estimators=100 max_depth=40 min_samples_split=70	8,72	6,06	36,81	4,9	-0,13

Fonte: Elaborado pela Autora, 2024

O ajuste 5 foi a configuração de hiperparâmetros que teve melhor desempenho entre os demais ajustes, isso pode ser visto através dos valores das métricas de avaliação. Entretanto esse modelo não performou bem como os outros. O valor de R^2 é negativo e o valor da métrica MAPE é muito alto, sugerindo que o modelo de regressão não está capturando a relação entre as variáveis independentes e dependentes.

A fim de se ter uma certeza sobre o desempenho do modelo foi feito um gráfico de dispersão (Figura 39) para verificar o comportamento dos valores preditos em relação aos valores reais.

Figura 39 - Gráfico de Dispersão do Modelo 3



Fonte: Elaborado pela Autora, 2024

Percebe-se através da Figura 39 que há uma grande dispersão dos dados, isso sugere novamente que não há uma relação linear entre as variáveis.

A exclusão das variáveis 'tipos de defeitos' resultou na incapacidade do modelo de se ajustar, indicando que para a performance do modelo essas variáveis são de extrema importância. Essa relevância dos atributos de tipos de defeitos pode ser observada nas Figura 33 e 37, onde o tipo de defeito ‘mancha pequena média’ é o que mais contribui para as previsões.

7 CONSIDERAÇÕES FINAIS

Com base na análise realizada utilizando a metodologia CRISP-DM, este trabalho conduziu uma investigação sobre os dados relacionados ao tingimento têxtil, com o objetivo de desenvolver modelos preditivos para previsão de porcentagem de defeitos nos tecidos. A análise exploratória revelou insights valiosos sobre a distribuição e características dos dados, destacando a presença de outliers e padrões nas variáveis investigadas.

Ao examinar os outliers, foi fundamental distinguir entre valores genuínos e dados incompletos ou anômalos. Estratégias foram adotadas para preservar a integridade dos dados, mantendo os eventos reais enquanto removendo observações que poderiam distorcer as análises estatísticas e prejudicar a precisão dos modelos preditivos.

A preparação dos dados envolveu a seleção das variáveis a serem incluídas nos conjuntos de treinamento e teste, bem como a aplicação de transformações para garantir a adequação dos dados aos algoritmos de modelagem. A normalização dos valores de metros totais tingidos, por exemplo, foi utilizada para mitigar disparidades na escala dos atributos.

Os modelos de floresta aleatória foram treinados e ajustados por meio de ajustes dos hiperparâmetros, visando maximizar o desempenho preditivo. A avaliação dos modelos revelou percepções importantes sobre sua capacidade de capturar a relação entre as variáveis independentes e dependentes.

Os resultados obtidos demonstraram que, apesar das modificações introduzidas no conjunto de dados e na estrutura dos modelos, o desempenho do modelo 1 foi melhor em relação aos modelos 2 e 3. No modelo 3, foi observado que a exclusão das variáveis relacionadas aos tipos de defeitos resultou em uma diminuição significativa na capacidade preditiva do modelo, destacando a importância desses atributos na análise.

Em suma, este estudo contribui para o avanço do conhecimento no campo da previsão de defeitos têxteis, fornecendo entendimentos sobre a importância da

análise exploratória de dados, a seleção de variáveis relevantes e a avaliação cuidadosa dos modelos preditivos. Espera-se que essas descobertas incentivem pesquisas futuras e inspirem melhorias nas práticas industriais relacionadas à qualidade e produção têxtil.

REFERÊNCIAS BIBLIOGRAFICAS

AGENCIA DINO (org.). **Indústria têxtil e de confecção faturou R$ 194 bilhões em 2021 - Bem Paraná.** [S. l.: s. n.], 2022. Disponível em: https://www.bemparana.com.br/noticia/industria-textil-e-de-confeccao-faturou-r-194-bilhoes-em-2021-271615#.Yq4u5nbMJPY. Acesso em: 18 jun. 2022.

ALMEIDA, A.; CARVALHO, F.; MENINO, F. 1 Introdução | **Introdução ao Machine Learning.** [S. l.: s. n.], 2017. Disponível em: https://dataat.github.io/introducao-ao-machine-learning/introdu%C3%A7%C3%A3o.html. Acesso em: 23 jun. 2022.

ARYA, N. **Tuning Random Forest Hyperparameters**. Disponível em: <https://www.kdnuggets.com/2022/08/tuning-random-forest-hyperparameters.html>. Acesso em: 25 mar. 2024.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA TÊXTIL (ABIT). **Perfil do Setor**. [S. l.: s. n.], 2022. Disponível em: https://www.abit.org.br/cont/perfil-do-setor. Acesso em: 23 jun. 2022.

AZANK, F. **Como avaliar seu modelo de regressão**. [*S. l.: s. n.*], 2020. Disponível em: https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regress%C3%A3o-c2c8d73dab96. Acesso em: 12 jul. 2022.

BESCOND, P. L. **Cyclical features encoding, it's about time!** Disponível em: <https://towardsdatascience.com/cyclical-features-encoding-its-about-time-ce23581845ca>. Acesso em: 20 mar. 2024.

CAI, J. et al. **Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest**. *Applied Energy*, v. 262, n. February, p. 114566, 2020. Disponível em: <https://doi.org/10.1016/j.apenergy.2020.114566>.

CAVALCANTI, A. M.; DOS SANTOS, G. F. **A INDÚSTRIA TÊXTIL NO BRASIL: uma análise da importância da competitividade frente ao contexto mundial.** Exacta, 12 maio 2021. Disponível em: <https://periodicos.uninove.br/exacta/article/view/17784>. Acesso em: 23 maio. 2022.

CLASSIFICATION ALGORITHM IN MACHINE LEARNING - JAVATPOINT. Disponível em: <https://www.javatpoint.com/classification-algorithm-in-machine-learning>. Acesso em: 25 mar. 2024.

CELULOSE. [S. l.: s. n.], 2022. Disponível em: https://pt.wikipedia.org/wiki/Celulose. Acesso em: 27 jun. 2022.

COLPANI, R. **Mineração de Dados Educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar**. Disponível em: <https://www.researchgate.net/figure/Figura-3-Fases-do-Modelo-CRISP-DM_fig2_333838915>. Acesso em: 22 fev. 2024.

DAMACENO, L. **Regressão Linear?** [S. l.: s. n.], 2020. Disponível em: https://medium.com/@lauradamaceno/regress%C3%A3o-linear-6a7f247c3e29. Acesso em: 9 jul. 2022.

DUARTE, M. **Correlação De Pearson: Uma Medida Essencial Na Análise De Dados**. Disponível em: <https://medium.com/@pe.marcos30/correla%C3%A7%C3%A3o-de-pearson-uma-medida-essencial-na-an%C3%A1lise-de-dados-a4c1c0bfbcf1>. Acesso em: 19 mar. 2024.

FONTANA, É. **Introdução Aos Algoritmos De Aprendizagem Supervisionada**. Disponível em: <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf>. Acesso em: 25 mar. 2024.

FROST, J. **Scatterplots: Using, Examples, and Interpreting**. Disponível em: <https://statisticsbyjim.com/graphs/scatterplots/>. Acesso em: 2 abr. 2024.

GALARNYK, M. **Train Test Split: What It Means and How to Use It | Built in**. Disponível em: <https://builtin.com/data-science/train-test-split>. Acesso em: 1 abr. 2024.

GOMES, P. C. T. **Regressão Linear | Aplicação do Algoritmo para Machine Learning.** [S. l.: s. n.], 2019. Disponível em: https://www.datageeks.com.br/regressao-linear/. Acesso em: 9 jul. 2022.

GOOGLE. **Google Colaboratory**. Disponível em: <https://colab.google/>.

GUNAY, D. **Random Forest**. Disponível em: <https://medium.com/@denizgunay/random-forest-af5bde5d7e1e>. Acesso em: 25 mar. 2024.

HARRISON, M. **Machine Learning – Guia De Referência Rápida**. [s.l.] Novatec Editora, 2019.

HERICLIS, S. **Hiperparâmetros - Por Quê São Importantes**. Disponível em: <https://blog.dsbrigade.com/hiperparametros-por-que-sao-importantes/>. Acesso em: 21 mar. 2024.

IBM. **What is Random Forest? | IBM**. Disponível em: <https://www.ibm.com/topics/random-

forest#:~:text=Random%20forest%20is%20a%20commonly>. Acesso em: 25 mar. 2024.

ISLAM, M. **Data Analysis: Types, Process, Methods, Techniques and Tools.** International Journal on Data Science and Technology, v. 6, n. 1, p. 10–15, 2020. Disponível em: <http://www.sciencepublishinggroup.com/j/ijdst>. Acesso em: 9 jul. 2022.

JULIANO, L. N.; PACHECO, S. M. V. **Estamparia e Beneficiamento Têxtil.** Unidade Araranguá – CEFET¬, SC., p. 84, [s.d.] disponível em: <https://wiki.ifsc.edu.br/mediawiki/images/3/30/Apostila_Estamparia_edicao_1_revisada.pdf>.

JUNIOR, O. D. **CRISP-DM o que é**. Disponível em: <https://www.dataviking.com.br/post/crisp-dm-e-util-mesmo>. Acesso em: 7 mar. 2024.

JÚNIOR, W. **MÉTODOS DE OTIMIZAÇÃO HIPERPARAMÉTRICA: UM ESTUDO COMPARATIVO UTILIZANDO ÁRVORES DE DECISÃO E FLORESTAS ALEATÓRIAS NA CLASSIFICAÇÃO BINÁRIA**. [s.l: s.n.]. Disponível em: <https://www.ppgee.ufmg.br/defesas/1461M.PDF>. Acesso em: 25 mar. 2024.

JUNIOR, B. de O. M. Setor têxtil - **Produção, comércio internacional e perspectivas para Brasil, Nordeste, Ceará e Pernambuco em 2021**. Caderno Setorial ETENE, n. 185, p. 1–10, 2021. Disponível em: <http://www.bnb.>. Acesso em: 4 jul. 2022.

KUASNE, A. **Fibras Têxteis**. Curso têxt ed. [s.l: s.n.]

LARSON, R.; FARBER, E. **Elementary Statistics**. [s.l.] Prentice Hall, 2006.

LEE, C.-Y.; LIN, J.-Y.; CHANG, R.-I. Improve Quality and Efficiency of Textile Process using Data-driven Machine Learning in Industry 4.0. **International Journal of Technology and Engineering Studies**, v. 4, n. 2, 13 abr. 2018. Acesso em: 12 set. 2019.

MCLEOD, S. **What Does a Box Plot Tell You?** Disponível em: <https://www.simplypsychology.org/boxplots.html>. Acesso em: 29 mar. 2024.

MERGE, JOIN, CONCATENATE AND COMPARE — PANDAS 1.2.4 DOCUMENTATION. Disponível em: <https://pandas.pydata.org/docs/user_guide/merging.html>. Acesso em: 8 mar. 2024.

NOVACK, G. **Building a One Hot Encoding Layer with Tensorflow**. Disponível em: <https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39>. Acesso em: 20 mar. 2024.

O QUE É APRENDIZAGEM SUPERVISIONADA? [S. l.: s. n.], [s. d.]. Disponível em: https://www.tibco.com/pt-br/reference-center/what-is-supervised-learning. Acesso em: 6 jul. 2022.

PANDAS - PYTHON DATA ANALYSIS LIBRARY. Disponível em: <https://pandas.pydata.org/about/index.html>. Acesso em: 19 mar. 2024.

PEREIRA, G. de S. **Introdução À Tecnologia Têxtil**. MINISTÉRIO DA EDUCAÇÃO SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE SANTA CATARINA UNIDADE DE ENSINO DE ARARANGUÁ, v. Módulo 2, n. CURSO TÊXTIL EM MALHARIA E CONFECÇÃO, p. 101, [s.d.]

PERES, F. **Como interpretar (e construir) um gráfico boxplot?** Disponível em: <https://fernandafperes.com.br/blog/interpretacao-boxplot/>. Acesso em: 8 mar. 2024.

PICCOLI, H. H. **Determinação do comportamento tintorial de corantes naturais extraídos da alfafa e urucum**. 2008. UNIVERSIDADE FEDERAL DE SANTA CATARINA, 2008. Disponível em: <https://core.ac.uk/download/pdf/30373211.pdf>.

PINHEIRO, N. **Pré-processamento De Dados Com Python**. Disponível em: <https://medium.com/data-hackers/pr%C3%A9-processamento-de-dados-com-python-53b95bcf5ff4>. Acesso em: 20 mar. 2024.

PROVOST, F.; FAWCETT, T. **Data Science Para Negócios**. [s.l.] Alta Books, 2016.

ROBERTO, C. **Crisp-DM: as 6 etapas da metodologia do futuro**. Disponível em: <https://blog.mbauspesalq.com/2022/04/12/crisp-dm-as-6-etapas-da-metodologia-do-futuro/>. Acesso em: 11 mar. 2024.

SALEM, V. **Tingimento Têxtil: fibras, conceitos e tecnologias**. São Paulo: Blucher: Golden Tecnologia, 2010.

SANTOS, A. **Análise Exploratória De Dados**. [s.l: s.n.]. Disponível em: <https://www.ibilce.unesp.br/Home/Departamentos/CiencCompEstatistica/Adriana/analise-exploratoria-de-dados.pdf>.

SCIKIT-LEARN. **sklearn.ensemble.RandomForestRegressor — scikit-learn documentation**. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>. Acesso em: 25 mar. 2024.

SENAI. **Como a inteligência artificial pode expandir o meu negócio?.** 2020.

Valentim, V. A. **Aula 02: O que é algoritmo? Notas de Aula.** Disponível em: < http://tscvalentin.blogspot.com/2008/02/pda-aula-02-o-que-algoritmo.html>, Acesso em: 5 jul. 2022.

SINGH, R. **Merging DataFrames with pandas | pd.merge()**. Disponível em: <https://medium.com/swlh/merging-dataframes-with-pandas-pd-merge-7764c7e2d46d>. Acesso em: 19 mar. 2024.

SKLEARN.PREPROCESSING.MINMAXSCALER — SCIKIT-LEARN 0.15-GIT DOCUMENTATION. Disponível em: <https://scikit-learn.org/0.15/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Acesso em: 23 mar. 2024.

SOLC, T. **Unidecode: ASCII transliterations of Unicode text**. Disponível em: <https://pypi.org/project/Unidecode/>. Acesso em: 19 mar. 2024.

TUYCHIEV, B. **Understanding Skewness and Kurtosis and How to Plot Them**. Disponível em: <https://www.datacamp.com/tutorial/understanding-skewness-and-kurtosis>. Acesso em: 29 mar. 2024.

VELASQUEZ, L. H. **Uma visão geral sobre machine learning - Classificação - Oper Estatística e Data Science**. [S. l.: s. n.], 2020. Disponível em: https://operdata.com.br/blog/uma-visao-geral-sobre-machine-learning/. Acesso em: 9 jul. 2022.

WALTRICK, A. C. **O que é Solidez da Cor e como avaliar uma nota de Solidez da Cor?** [S. l.: s. n.], 2020. Disponível em: https://www.linkedin.com/pulse/o-que-%C3%A9-solidez-da-cor-e-como-avaliar-uma-nota-de-ana-carla-waltrick/?originalSubdomain=pt. Acesso em: 5 jul. 2022.

YILDIRIM, P.; BIRANT, D.; ALPYILDIZ, T. Data mining and machine learning in textile industry. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 1, p. e1228, 2 out. 2019. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1228>. Acesso em: 11 mar. 2024.