UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MARLON ALVES BOMFIM

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA DETERMINAR A VELOCIDADE DE PRODUÇÃO EM MÁQUINAS COEXTRUSORAS

LONDRINA

2021

MARLON ALVES BOMFIM

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA DETERMINAR A VELOCIDADE DE PRODUÇÃO EM MÁQUINAS COEXTRUSORAS

APPLICATION OF MACHINE LEARNING TECHNIQUES TO DETERMINE THE PRODUCTION SPEED IN EXTRUSION MACHINES

Trabalho de conclusão de curso de graduação apresentada como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador(a): Dr. Rafael Henrique Palma Lima

LONDRINA

2021

MARLON ALVES BOMFIM

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA DETERMINAR A VELOCIDADE DE PRODUÇÃO EM MÁQUINAS COEXTRUSORAS

Trabalho de Conclusão de Curso de Graduação para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 23/novembro/2021

_____

Rafael Henrique Palma Lima
Doutor
Universidade Tecnológica Federal do Paraná

_____

Bruno Samways Dos Santos
Doutor
Universidade Tecnológica Federal do Paraná

_____

Gislaine Camila Lapasini Leal
Doutora
Universidade Estadual de Maringá

AGRADECIMENTOS

Aos amigos, que em todos os momentos auxiliaram com ideias e com apoio psicológico durante a execução deste trabalho.

Ao grupo de pesquisa em otimização e mineração de dados, que formaram a primeira banca não oficial e ajudaram a encorpar este trabalho.

Aos professores, que tornaram esta empreitada e este sonho possível de ser realizado, e pelas orientações que permitiram chegar até aqui.

RESUMO

Devido à necessidade de se otimizar processos, recursos e decisões, as indústrias têm investido em sistemas de informação para reduzir erros e perdas em seus processos produtivos. Ter maior confiabilidade e precisão nas informações recebidas é vital na tomada de decisões, e a utilização de técnicas de aprendizado de máquina (do inglês, *Machine Learning* - ML) tem auxiliado as indústrias, e por este motivo, a implementação destes algoritmos contribuem para o ganho de resultados. Com este foco, este trabalho teve como objetivo, a aplicação a aplicação de técnicas de ML de classificação e regressão para predizer a velocidade de produção necessária para a confecção de materiais coextrusados, com base em dados históricos fornecidos. Os dados foram pré-processados e ao todo foram utilizadas 12 técnicas de aprendizado de máquina no estudo de modelos de classificação e regressão, cada um com configurações distintas. O primeiro método buscou classificar entre os valores históricos que a base de dados tem para definição de velocidade de produção, e o segundo método visou determinar a velocidade aproximada usando a base de dados como fonte de cálculo. Os resultados foram satisfatórios, com destaque para as técnicas de Árvores de Decisão e *Random Forest* que obtiveram um índice médio de acurácia de 77% e 78% em classificação e 79% e 82% em regressão, respectivamente.

Palavras-chave: Aprendizado de Máquina. Indústria de Embalagens. Extrusão. Classificação. Regressão.

ABSTRACT

Due to the need to optimize processes, resources and decisions, industries have invested in information systems errors and losses in their production process. Having a greater reliability accuracy in the information received is vital in decision making, and the use of machine learning has helped the industries, and for this reason, the implementation of Machine Learning (ML) methods contribute to the gain of results in the data automation and reliability. With this focus, this paper reports the application of ML classification and regression techniques to predict the production speed needed for the manufacture of coextruded materials, based on historical data provided To study this problem, consolidated ML techniques were chosen. The data was pre-processed and we used 12 machine learning techniques in the study of classification and regression models altogether, each one with distinct configurations. The first method aims to classify among historical values which the data base has to define the production speed and setup, and the second method aims to determine the approximate speed using the database as calculation source. The results were satisfactory, with emphasis on the Decision Trees and Random Forest techniques, which obtained an average accuracy rate of 77% and 78% in classification and 79% and 82% in regression, respectively.

Keywords: Machine Learning. Packaging Industry. Extrusion. Classification. Regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ANN	Redes Neurais Artificiais
DT	Árvore de Decisão
IA	Inteligência Artificial
KNN	K-Vizinhos Mais Próximos
LR	Regressão Logística
MAE	Erro Médio Absoluto
MAPE	Erro de porcentagem média absoluta
ML	Aprendizado de Máquina
MLP	*MultiLayer Perceptron*
MSE	Erro Quadrático Médio
NB	Naïve Bayes
RMSE	Raiz do erro quadrático médio
R^2	R-quadrado
RF	Floresta Aleatória
SVM	Máquina de vetores de suporte

SUMÁRIO

1. INTRODUÇÃO

Em um cenário competitivo em que as empresas estão em constante busca por inovações, as indústrias têm investido em meios que retornem melhorias em seus processos produtivos e aumente a qualidade de seus produtos. Com o avanço da tecnologia e dos métodos produtivos, e aliados a modelos matemáticos e à inteligência computacional, a Indústria 4.0 mostra a busca constante das indústrias em otimizar as suas produções, e ter os dados em tempo real para tomada de decisões.

Gomes (2010) busca descrever que organizações que adotaram o uso da Inteligência Artificial (IA) notam que há um potencial benéfico, independente do segmento em que estão inseridos. Isso se dá ao fato de que a utilização da IA vai além da automação mecânica, abrangendo processos cognitivos, que são capazes de gerar aprendizado. Assim, um sistema de IA além de executar atividades repetitivas, numerosas e manuais, também fica responsável por demandas de análise e tomada de decisão.

Dos principais benefícios que a IA oferece, Albuquerque (2021) descreve:

- **Melhora na tomada de decisão**: por ser capaz de organizar e conferir maior clareza das informações;
- **Comodidade e escalabilidade**: já que são feitos de maneira mais rápida e simplificada;
- **Aumento da automação de atividades**: sendo elas lógicas, analíticas e cognitivas, gerando maior velocidade no tratamento de informações;
- **Redução de erros, riscos e custos operacionais**: por ter a capacidade de identificar gargalos, falhas e outros pontos falhos nos processos da empresa.

Primeiramente, deve-se encontrar a estratégia ideal para a utilização de IA, estabelecendo-se objetivos e ter um bom planejamento, e sabendo-se qual método melhor se encaixa perante cada situação.

Entre os principais recursos da IA utilizadas estão *Machine Learning* (ML), que envolve métodos de avaliação de dados que automatiza padrões analíticos em desenvolvimento, *Deep Learning* (DL), que realça a utilização de redes neurais artificiais com várias camadas de abstração, com maior aplicação no reconhecimento de padrões, e o Processamento de Linguagem Natural (PLN), que busca estudar

maneiras de reproduzir processos de desenvolvimento ligados ao funcionamento da linguagem humana (COSSETTI, 2019).

Diversos segmentos da indústria têm usado técnicas de ML para auxiliar na tomada de decisão. Por exemplo, aplicações são encontradas nas áreas de extração, de transformação, energética, alimentícia, construção civil, embalagens, dentre muitas outras. Saber qual método de ML utilizar para cada uma é essencial para que a implementação destas tecnologias seja totalmente efetiva.

Por este motivo, este trabalho enfatiza a análise de uso de ML para predição de velocidade de produção na etapa de extrusão em uma empresa do ramo de embalagens, em que busca demonstrar que ela se torna ainda mais necessário, pois quanto mais precisas são as informações, menores são as chances de que um produto saia defeituoso para os clientes, o que poderia gerar impacto na conservação dos alimentos.

Reforçando a importância da segurança alimentar que as embalagens fornecem, esta pode ser abordada em duas perspectivas diferentes: em uma delas a embalagem exerce a função de proteção e na conservação do produto ou alimento, contribuindo para a segurança; e na outra, a embalagem não poder se transformar em algo que afete a segurança e qualidade do produto, uma vez que sua produção pode ter composições de natureza diversa, e em contato direto com os alimentos, possam resultar em contaminação física, química ou microbiológico (MOREIRA; POÇAS, 2003).

Por dentro do processo produtivo de uma embalagem, a composição de materiais extrusados é de extrema importância, por agregar novas características e propriedades ao produto final, e para sua confecção é importante definir quais as resinas que geram o material, quanto tempo é necessário para programar a extrusora para produção (*setup*), qual é a taxa de vazão e produção do extrusado pela máquina, quais os testes de qualidade necessários para aprovação e liberação do produto, entre outros.

Na construção do material extrusado, a combinação da velocidade em que o filme é puxado, o método de sopro do balão e a regulagem do resfriamento permite alcançar a melhor razão de propriedades do material a ser produzido, e qual velocidade de extrusão pode ser influenciada por parâmetros como pressão e temperatura, além de fatores específicos do produto (SOUZA; ALMEIDA, 2015).

A falta de treinamento, ou a configuração errada de um produto impacta

diretamente a correta produção dos materiais, gerando retrabalho, e desperdício de materiais, visto que a reutilização de resinas de polietileno não é muito usual pelas industrias na América Latina. Uma possibilidade de orientação para as empresas quanto a redução de erros operacionais é a aplicação de técnicas de Inteligência Computacional, a partir de processos como *Data Mining* (DM) para definir como uma segunda forma de classificação dos dados.

Por este motivo, esta pesquisa discute sobre a aplicação das técnicas de aprendizado de máquina no momento de definir qual é a velocidade necessária para produzir um filme extrusado e seu tempo de *setup*, mediar a eficiência e eficácia das técnicas de ML utilizadas, e sugerir melhorias para a medição das informações obtidas. Para tanto, utilizou-se como metodologia a natureza quanti-qualitativa por haver extração de dados e informações para analisar as faixas de velocidades e de tempo de setup necessárias para produção, com apoio de modelagem matemática e simulação computacional.

1.1 Objetivos

1.1.1 Objetivo Geral

Aplicar modelos de classificação e regressão para a predição da configuração de velocidade de máquinas extrusoras, testando a eficiência desses modelos utilizando os dados fornecidos.

1.1.2 Objetivos Específicos

- Obter e preparar a base de dados para a aplicação de ML;
- Implementar as técnicas de uma linguagem computacional capaz de predizer as velocidades de máquina, podendo ser utilizado futuramente na prática;
- Comparar a eficiência do modelo aplicado a partir dos resultados obtidos após a implementação da metodologia e efetuar comparações quanto as métricas das técnicas de classificação e regressão com o intuito de definir o mais indicado para a predição;
- Discutir as contribuições acadêmicas das aplicações de aprendizado de máquina no ambiente industrial;

- Validar os resultados obtidos pelas técnicas por meio de métricas pré-estabelecidas.

1.2 Estrutura do trabalho

As etapas metodológicas foram divididas em quatro capítulos. O Capítulo 2 apresenta o referencial teórico, em que consta as fundamentações teóricas das técnicas de aprendizado de máquina supervisionada que foram utilizadas para análises dos dados, os trabalhos correlatos e como funciona a processo de confecção de materiais coextrusados.

O Capítulo 3 descreve o método de obtenção das informações, a estrutura dos dados utilizados, a descrição dos métodos de análise, assim como os meios utilizados para implementação das técnicas de aprendizado de máquina.

O Capítulo 4 apresenta a prévia dos resultados obtidos com os dados dos materiais coextrusados. O Capítulo 5 compreende a análise descritiva dos dados e a comparação e análise dos resultados obtidos das técnicas de ML.

O Capítulo 6 teve como propósito apresentar as contribuições sobre o conteúdo estudado e as conclusões obtidas pela a atual pesquisa, bem como foram dadas sugestões para pesquisas futuras sobre o tema.

2. REFERENCIAL TEÓRICO

Este capítulo aborda os principais conceitos necessários para o entendimento do trabalho desenvolvido, como aspectos de aprendizado de máquina, trabalhos correlatos e coextrusão de embalagens.

2.1 Machine Learning

O termo *Machine Learning* (ML) comumente traduzido para o português como “aprendizado de máquina”, é um termo que foi mencionado pela primeira vez por um pioneiro em jogos de computador e inteligência artificial, Arthur Lee Samuel, especificando a ciência que utiliza os computadores para trabalhar autonomamente sem uma programação explícita (SYAM; SHARMA, 2018). De maneira sucinta, Jonsson *et al*. (2016) definem que o aprendizado de máquina é um campo de estudo onde os programas de computador podem aprender e melhorar a execução de tarefas específicas por meio do treinamento em dados históricos.

Já Kucak e Dambic (2018) descrevem que implementar um algoritmo de aprendizado de máquina significa implementar um modelo que produza informações corretas, desde que sejam fornecidos os dados de entrada, e que haja o tratamento das informações.

As técnicas de ML são reconhecidas como uma subdivisão da Inteligência Artificial (RUSSEL; NORVIG, 2013), e é desenhada por algoritmos que aprendem a partir de dados utilizados como modelos. Seus princípios são utilizados em outras áreas de IA e sua utilização é vista em muitos campos, contribuindo tanto para a resolução de problemas, quanto para o aperfeiçoamento de atividades (SCHÜSSLER; BASTIANI; BUSSLER, 2018). Para ser inteligente, um sistema que está em um ambiente em mudança deve ter a capacidade de aprender, se o sistema pode aprender e se adaptar a tais mudanças, o projetista do sistema não precisa prever e fornecer soluções para todas as situações possíveis (ALPAYDIN, 2010).

Roza (2016) afirma que o objetivo principal de ML é ir além dos exemplos existentes no conjunto de treinamento, pois independe da quantidade de dados disponíveis. É muito pouco provável que durante os testes, exatamente os mesmos exemplos sejam executados.

Entretanto, a quantidade de dados disponíveis em seu conjunto de análise se torna indispensável, caso contrário, a péssima definição poderá ocasionar problemas na estruturação dos modelos de aprendizado, que será agravado devido ao fato de que este problema só se torna visível quando inseridas novas instâncias desconhecidas no modelo (CARVALHO, 2014).

Segundo Hein (2021), há uma variedade de métodos e ferramentas de aprendizado de máquina disponíveis para estudo, e o primeiro passo para desenvolver qualquer novo aplicativo é escolher o método mais adequado, pois a escolha depende do tipo de dados, o número de pontos de dados disponíveis para treinar o sistema e a saída desejada. Logo, realizar escolhas erradas pode resultar em falsas correlações sendo feitas durante o treinamento e modelos preditivos ineficazes.

Bhavsar *et al*. (2017) explicam que os métodos de aprendizado de máquina podem ser caracterizados, ou categorizados com base no tipo de aprendizado. Para isto, os métodos de aprendizado de ML podem ser divididos em quatro grupos, sendo eles: supervisionado, não supervisionado, semissupervisionado e por reforço.

No aprendizado supervisionado avalia-se que o ganho estatístico supervisionado envolve a construção de um modelo estatístico para prever ou estimar uma saída com base em uma ou mais entradas (JAMES *et al.*, 2013). Em outras palavras, tem-se que na aprendizagem supervisionada, para um conjunto de variáveis preditoras no conjunto de dados (*input*), o valor correto da variável dependente é fornecido ao algoritmo, ou seja, existe uma variável de saída correspondente a um conjunto de variáveis de entrada (SYAM; SHARMA, 2018).

Já no aprendizado não supervisionado, a estrutura de dados é utilizada sem o conhecimento das respostas corretas. Ludermir (2021) descreve que o algoritmo deve agrupar os exemplos a partir de suas familiaridades em seus atributos e deverá analisar a base de dados fornecida e tentar determinar se os dados podem ser agrupados, formando agrupamentos ou *clusters*. Em seguida, é feita a avaliação dos resultados, para determinar o que cada agrupamento significa dentro dos métodos determinados no problema analisado.

O aprendizado semissupervisionado geralmente é usado para as mesmas aplicações que o aprendizado supervisionado, combinando com o modo não supervisionado, utilizando exemplos rotulados e exemplos não rotulados (RUSSEL; NORVIG, 2013).

No algoritmo de aprendizado por reforço, o algoritmo é estruturado para receber um sinal de reforço, que pode ser recompensa ou punição, e não uma resposta correta, desta maneira, o algoritmo formula uma hipótese com base nos exemplos, para poder tomar uma decisão, sendo esta boa ou ruim (LUDERMIR, 2021).

Figura 1 - Diagrama dos tipos de aprendizado em machine learning.

Fonte: Adaptado de Almeida, Carvalho e Menino (2017).

As subseções a seguir descatam o aprendizado supervisionado e o não supervisionado, por representarem a maioria das técnicas de aprendizado de máquina pesquisadas.

2.1.1 Aprendizagem de Máquina Supervisionada

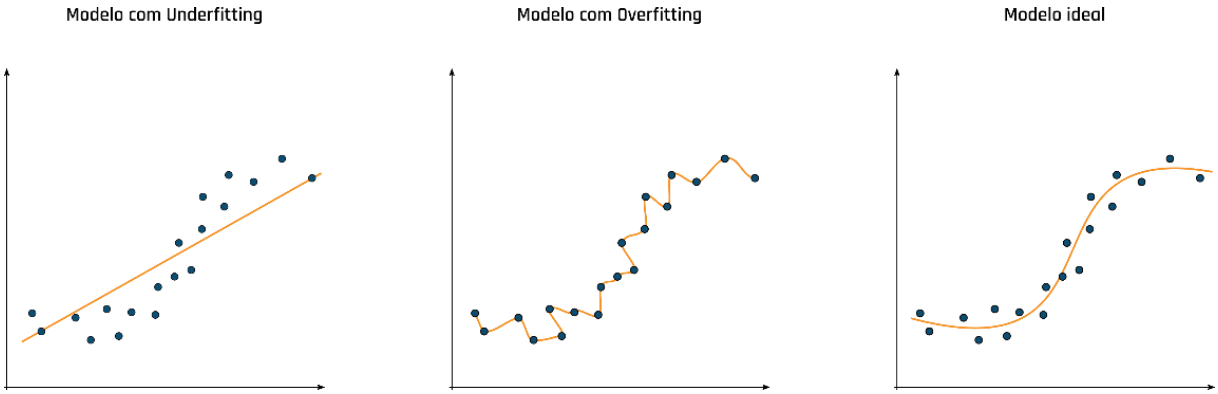
O método de aprendizado de máquina supervisionado, pode ser descrito como um processo de usar a experiência para ganhar especialização ou conhecimento. A aprendizagem supervisionada realiza um cenário em que a experiência contém informações significativas que estão faltando nos exemplos de teste e invisíveis aos quais a experiência aprendida deve ser aplicada. Nesse cenário, a experiência adquirida deverá prever as informações que faltam para os dados

separados para teste, e analogamente pode-se pensar neste método como um ambiente em que um professor supervisiona o aluno, fornecendo as informações extras ou necessárias (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Neves (2018) simplifica descrevendo que a base de dados utilizada para estes métodos possui *flags* ou marcadores, que representam a resposta desejada ou esperada, e a partir disto, o sistema se molda fazendo com que dada uma entrada, seja emitido como saída o mesmo valor do supervisor.

Nas técnicas de ML, o ajuste desbalanceado da complexidade dos modelos pode gerar problemas de *underfitting* (sub-ajuste) ou *overfitting* (superajuste). Almeida, Carvalho e Menino (2017) explicam que o problema de *underfitting* está vinculado à falta de capacidade do modelo se adaptar à representação dos dados (etapa de treinamento), enquanto no *overfitting* o modelo se adequa muito aos dados de treinamento e perde a sua capacidade de generalizar com novos dados (etapa de testes), resultando em erros muito altos. A Figura 2 mostra um exemplo com diferentes ajustes do modelo aos dados:

Figura 2 - Diferentes ajustes do modelo aos dados



Fonte: Almeida, Carvalho e Menino (2017).

Almeida, Carvalho e Menino (2017) explicam que os hiper-parâmetros dos modelos de aprendizado supervisionado devem ser configurados com o propósito de regular o nível de assertividade e precisão. Essas características estão vinculadas ao viés, ou *bias* em inglês e à variância do modelo.

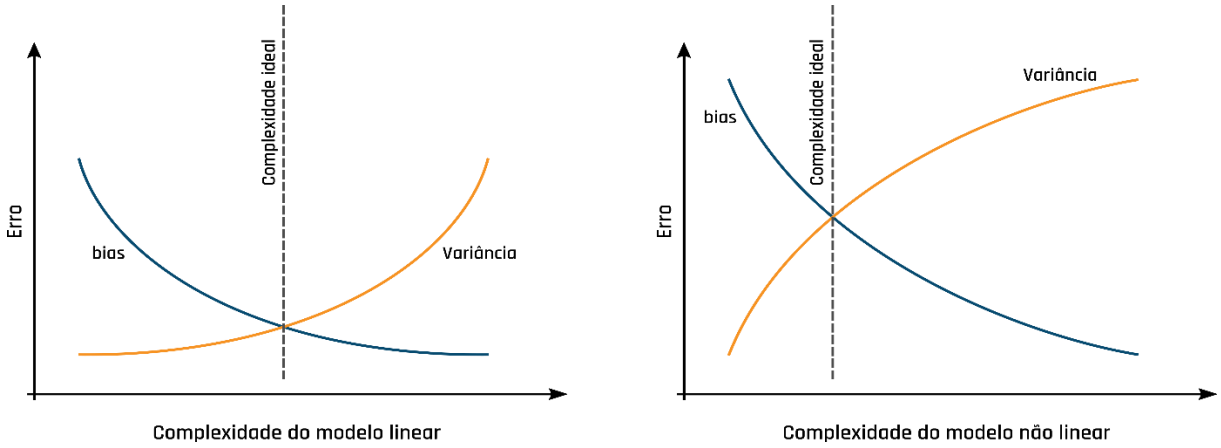
O viés avalia o quanto o algoritmo está errado, desconsiderando o efeito de amostras variáveis, enquanto a variância avalia o quanto o algoritmo flutua em torno do valor esperado, conforme a amostra varia. Em um cenário ideal, ambos devem ser baixos (ALPAYDIN, 2010). Complementando, Almeida, Carvalho e Menino (2017)

dizem que o viés corresponde à capacidade do algoritmo se ajustar aos dados apresentados na etapa de treinamento e a variância é a variabilidade das previsões do modelo.

O modelo ideal é aquele que apresenta a melhor compensação entre o viés e a variância. Se houver um alto viés, mas baixa variância, implica que a classe de modelo conterá *underfitting*. Se houver alta variação, mas baixo viés, a classe do modelo estará com *overfitting* em que terá abrangência geral e também aprenderá com dados ruidosos. Quanto à variância, também depende do tamanho do conjunto de treinamento, a variabilidade devido à amostra diminui à medida que o tamanho da amostra aumenta (ALPAYDIN, 2010).

A Figura 3 mostra um esquema para a complexidade ideal em modelos lineares e não lineares:

Figura 3 - Esquema do *trade-off* no aprendizado supervisionado.



Fonte: Almeida, Carvalho e Menino (2017).

2.1.1.1 Classificação

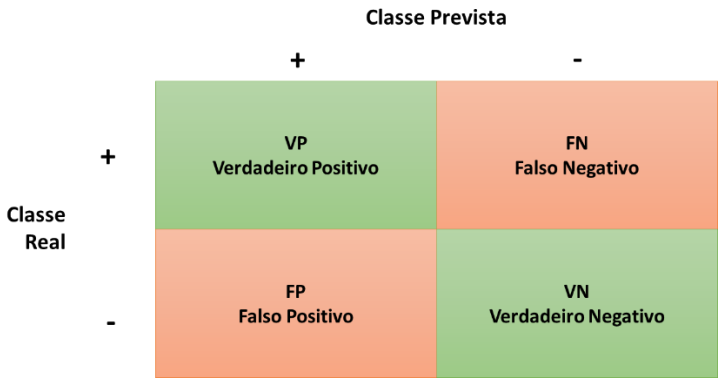
Para se trabalhar com um problema de classificação, o algoritmo deve ter como objetivo de aprendizado de máquina, categorizar ou classificar as entradas fornecidas com base no conjunto de dados de treinamento (BHAVSAR *et al.*, 2017).

Jonsson *et al.* (2015) reforça que os dados utilizados podem ser na forma de texto, números ou outros tipos de valores nominais. O conjunto de dados de treinamento em um problema de classificação inclui um conjunto de pares de entrada e saída categorizados em classes, e muitos problemas de classificação são binários, como exemplo, a representação da resposta em verdadeiro e falso (BHAVSAR *et al.*,

2017).

Amidi e Amidi (2018) descrevem que as métricas são importantes para avaliar a desempenho do modelo. Para isso, a matriz de confusão é usada para se ter um cenário mais amplo e completo quando se está avaliando o desempenho de um modelo. Ela é definida conforme a Figura 4:

Figura 4 - Matriz de confusão.



Fonte: Adaptado de Amidi e Amidi (2018).

A métrica utilizada para a avaliação do desempenho das técnicas de classificação do aprendizado de máquina é conhecida como acurácia, dada pela equação 1:

$$A = \frac{V_p + V_n}{V_p + F_p + V_n + F_n} \quad (1)$$

Para a definição da métrica de precisão, para mensurar o número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe (VP), é dada pela equação 2:

$$P = \frac{V_p}{V_p + F_p} \quad (2)$$

A especificidade avalia a capacidade do método de detectar resultados negativos, e é definida pela equação 3:

$$E = \frac{V_n}{V_n + F_p} \quad (3)$$

A sensibilidade (também conhecida como *recall*) avalia a capacidade do método de detectar com sucesso resultados classificados como positivos, e é definida pela equação 4:

$$S = \frac{V_p}{V_p + F_n} \quad (4)$$

O indicador *score* F1, é uma média harmônica entre precisão e *recall* pois ambos avaliam Vp, mas de acordo com referências diferentes, dada pela equação 5:

$$F1 = \frac{2V_p}{2V_p + F_p + F_n} \; OU \; F1 = 2 * \frac{S * P}{S + P} \quad (5)$$

2.1.1.2 Regressão

Na regressão deseja-se encontrar alguma correlação padrão entre as variáveis, como uma relação funcional entre os componentes X e Y dos dados (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Já Bhavsar *et al*. (2017) complementa que, para um problema de regressão o objetivo do algoritmo de ML é o desenvolvimento de um relacionamento entre saídas e entradas, utilizando uma função contínua para auxiliar as máquinas a compreender como as saídas estão mudando ou se comportando para determinadas entradas. A regressão é uma ferramenta que busca modelar relações entre variáveis dependentes e independentes por meio de métodos estatísticos (SOTO, 2013).

Almeida, Carvalho e Menino (2017) descrevem que o método de regressão, pode ser visto por uma variável independente, que caracteriza uma grandeza que está sendo manipulada em um experimento e que não sofre influência de outras variáveis, enquanto a variável dependente, caracteriza valores associados diretamente à variável independente.

Para a avaliação dos algoritmos de regressão, Silveira (2019) descreve que as medidas de erros mais comuns para avaliar a precisão e o desempenho de modelos preditivos são o *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *R-Squared* (R^2) e o *Mean Absolute Percentage Error* (MAPE).

O MAE é uma das principais métricas para avaliação de modelos preditivos e seu cálculo está definido pela equação 6:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (6)$$

Em que y_i é o valor de saída esperado (ou real) e $\hat{y}_i$ o valor de saída predito pelo modelo. O MAE estima a média do erro de previsão do modelo.

Silveira (2019) continua descrevendo que o MSE, é uma das métricas mais utilizadas para avaliação de modelos preditivos, e que nessa métrica é calculado o erro quadrado médio das previsões, onde quanto maior esse valor, pior é o modelo. O MSE eleva os erros obtidos ao quadrado, tornado os valores positivos e os erros maiores penalizam mais o modelo. Por isso, é necessário um tratamento prévio dos dados para que *outliers* não influenciem de negativamente um modelo com bom desempenho.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{7}$$

A métrica RMSE é calculada como a raiz quadrada das médias das diferenças quadradas entre a previsão e o dado real. Em outras palavras, é a raiz quadrada do MSE, demonstrada na equação 8:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{MSE} \tag{8}$$

Silveira (2019) comenta que o problema ao utilizar o MSE e RMSE é que é difícil saber se o modelo é bom o suficiente apenas olhando para o seu valor. Para contornar essa situação, recomenda-se utilizar o coeficiente de determinação R². O R² está relacionado ao MSE e apresenta a vantagem de ser livre de escala, quando seu valor é negativo, representa que o modelo é pior do que usar a média como previsão. Esta situação está especificada na Equação 9:

$$R^2 = 1 - \frac{MSE(modelo)}{MSE(base)} \tag{9}$$

Onde o MSE (base) é calculado substituindo-se o $\hat{y}_i$, valor predito, por $\bar{y}$, média dos valores observados. Logo, R² é a relação entre o quão bom é o modelo quando comparado ao modelo que considera a média.

O *Mean absolute percentage error* (MAPE), ou erro percentual absoluto médio, é a média dos erros percentuais absolutos das previsões, porém, se o valor real for zero, esta equação não pode ser utilizada. O erro é definido como o valor real menos o valor previsto, conforme consta na Equação 10:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \tag{10}$$

Silveira (2019) complementa que, como são utilizados erros percentuais

absolutos, evita-se o problema dos erros positivos e negativos que se anulam mutuamente. Esta medida é fácil de entender porque fornece o erro em termos de porcentagens, e por isso o MAPE tem apelo gerencial e é uma medida comumente utilizada em previsões. Quanto menor a MAPE, melhor a previsão.

2.2 Técnicas de *Machine Learning*

As subseções a seguir apresentam as técnicas utilizadas.

2.2.1 *Naïve Bayes*

Naïve Bayes é uma técnica simples que prevê resultados com base no teorema Bayesiano. O treinamento do classificador *Naïve Bayes* é rápido em comparação com outros modelos computacionalmente intensivos, pois ele classifica um determinado ponto de dados, com base na probabilidade condicional de estar em uma classe, dados os valores de seus escalares constituintes, sem depender de nenhum parâmetro adicional. A classe que tem a maior probabilidade de ocorrência, dadas as entradas, será a classe prevista (ZEINEDDINE *et al.*, 2020).

Conforme definido por Amaral (2016), na etapa de treinamento é definido uma tabela de valores e atribuído um peso para os atributos individualmente em cada uma das classes de classificação. Ao submeter uma nova instância para classificação, o modelo somará os pesos atribuidos em cada uma das classes, e a classe que somar o maior peso será a classe do novo item.

Para Zhang (2014), do ponto de vista da probabilidade, de acordo com a Regra Bayesiana, a probabilidade de um exemplo $E = (x_1, x_2, \dots, x_n)$ ser da classe *c* é de acordo com a Equação 11:

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)} \tag{11}$$

É classificado como a classe *C* = +, se e somente se:

$$f_b(E) = \frac{p(C = +|E)}{p(C = -|E)} \geq 1, \tag{12}$$

Onde $f_b(E)$ é chamado de classificador Bayesiano.

Suponha que todos os atributos sejam independentes, dado o valor da variável de classe, ou seja:

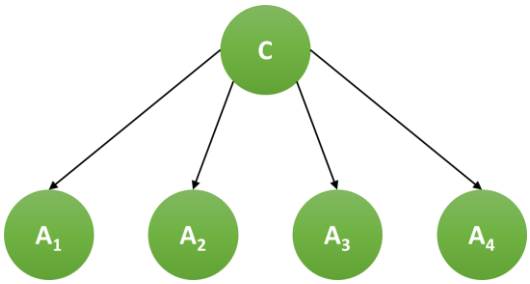
$$p(E|c) = p(x_1, x_2, \dots, x_n|c) = \prod_{i=1}^{n} p(x_i|c), \tag{13}$$

O classificador resultante é então:

$$f_{nb}(E) = \frac{p(C = +|E)}{p(C = -|E)} \prod_{i=1}^{n} \frac{p(x_i|C = +)}{p(x_i|C = -)}, \tag{14}$$

A função $f_{nb}(E)$ é chamada de classificador *Naïve Bayes*, ou simplesmente *Naïve Bayes* (NB). A Figura 5 mostra um exemplo do classificador. Em *Naïve Bayes*, cada nó de atributo não tem pai, exceto o nó de classe:

Figura 5 - Exemplo de Naïve Bayes



Fonte: Zhang (2014).

De acordo com Leal (2018), o classificador *Naive Bayes* apresenta bom desempenho mesmo se o conjunto de dados para o treinamento for pequeno. Isto é uma vantagem, pois o classificador é parametrizado pela média e variância de cada atributo independente de outros, em casos de utilização do método Gaussiano.

2.2.2 Regressão Logística

A regressão logística quantifica a relação entre um resultado categórico dependente e uma ou mais variáveis preditoras independentes. A regressão logística fornece probabilidades previstas para cada categoria. Esta é uma técnica paramétrica relativamente simples, amplamente utilizada em auditoria clínica (STYLIANOU *et al.*, 2015).

A regressão logística, apesar do nome, é um modelo linear para classificação em vez de regressão. A regressão logística também é conhecida na literatura como regressão *logit*, classificação de entropia máxima ou classificador log-linear. Neste

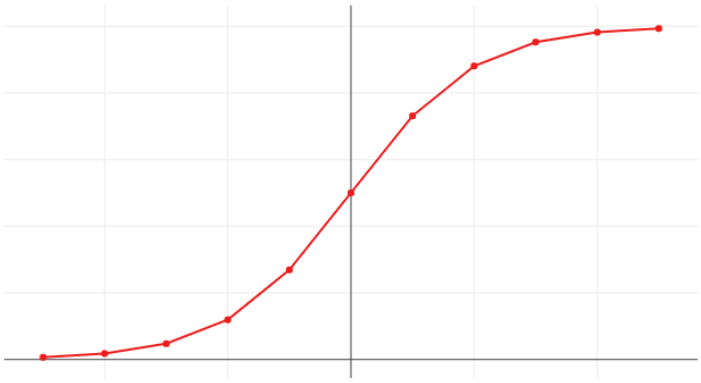
modelo, as probabilidades que descrevem os resultados possíveis de um único ensaio são modeladas usando uma função logística (BISHOP, 2006).

Os autores Shalev-Shwartz e Ben-David (2014), supõem que a regressão logística é usada para tarefas de classificação e que se pode interpretar a função *h(x)* como a probabilidade de que o rótulo de *x* seja igual a 1. A classe de hipótese associada à regressão logística é a composição de uma função sigmoide com intervalo de 0 a 1. Em particular, a função sigmoide usada na regressão logística é a função logística, definida na equação 15:

$$\emptyset_{sig}(z) = \frac{1}{1 + \exp(-z)} \tag{15}$$

O nome “sigmoide” significa “em forma de S”, referindo-se ao gráfico desta função, apresentado na Figura 6:

Figura 6 - Representação do Gráfico da função sigmoide.



Fonte: Adaptado de Shalev-Shwartz e Ben-David (2014).

Os métodos de regressão para prever o desempenho dos algoritmos usam um conjunto finito de relações entre as variáveis dependentes e independentes, gerando uma função preditiva que modela essas associações. O método de regressão logística para prever a classe dos dados é normalmente usado para descrever as associações entre uma série de variáveis independentes que podem ser categorizadas como binárias, categóricas e contínuas (ZEINEDDINE *et al.*, 2020).

2.2.3 Máquinas de vetor de suporte

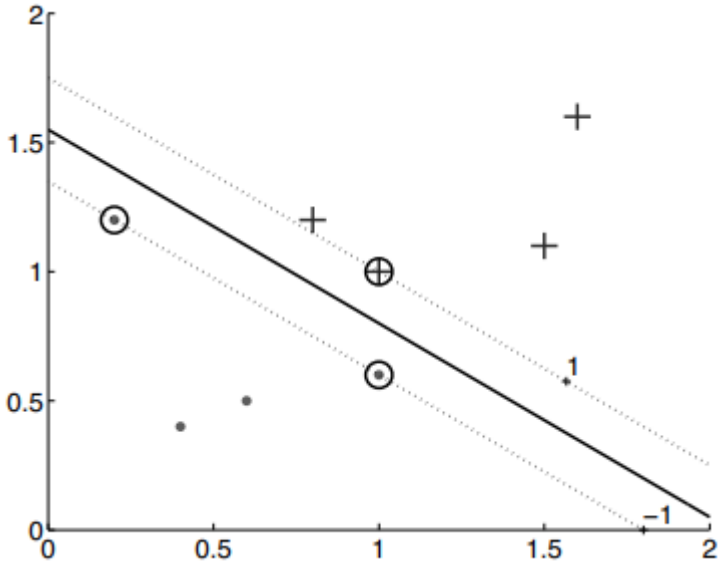
Máquinas de vetor de suporte (do inglês *Support Vector Machine* - SVM) são algoritmos de classificação que visam aproximar as margens de um dado a ser

classificado com os dados mais próximos. O algoritmo apresenta grande capacidade de generalização e robustez, possibilitando sua aplicação em vetores de grandes dimensões (AMARAL, 2016).

Zeineddine *et al*. (2020) definem que SVM é um método de aprendizado supervisionado que classifica os pontos de dados, segregando-os usando um hiperplano N-dimensional, onde N é o número de atributos que caracterizam um ponto de dados. A ideia geral é encontrar uma função com uma margem de erro definida que mapeie as variáveis de entrada para a variável de saída de modo que a saída prevista não se desvie da saída real mais do que a margem de erro definida (BHAVSAR *et al.*, 2017).

A Figura 7 representa um problema de duas classes em que as instâncias das classes são mostradas por sinais de mais e pontos, a linha espessa é o limite e as linhas tracejadas definem as margens em ambos os lados.

Figura 7 – Exemplo de SVM para Classificação.



Fonte: Alpaydin (2010).

Por ser um método baseado em discriminante, o SVM se preocupa apenas com as instâncias próximas ao limite e descarta aquelas que estão no interior (ALPAYDIN, 2010).

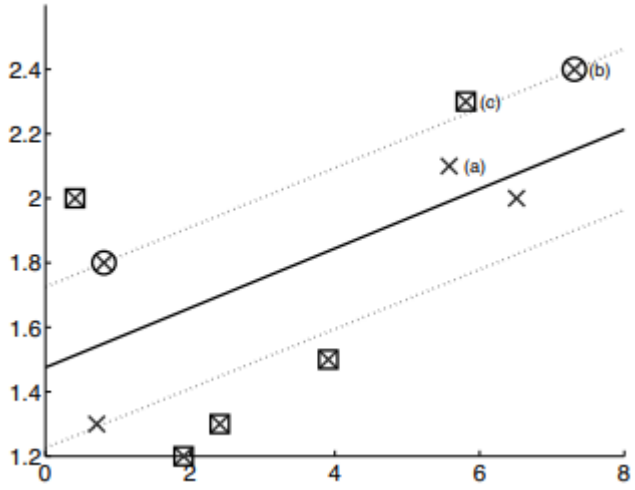
Alpaydin (2010) reforça que as SVMs podem ser generalizadas para regressão, que utiliza a função de perda sensível:

$$e_2\left(r^t, f(x^t)\right) = [r^t - f(x^t)]^2 \quad (16)$$

Em que $\left(r^t, f(x^t)\right)$ é uma nova observação que não foi usada para estimar e_2. Quanto menor o risco, melhor é a função de predição e_2. Isto significa que são tolerados erros e também que erros além tenham um efeito linear e não quadrático. Esta função de erro é, portanto, mais tolerante ao ruído e, consequentemente, mais robusta. Como na perda da dobradiça, existe uma região sem erro, o que causa dispersão.

A Figura 8 exemplifica SVM para regressão. A linha de regressão é ajustada aos pontos de dados, seus pontos são mostrados como cruzes, e para isso é mostrado com duas retas paralelas a área para a dispersão dos erros. Existem três casos: Em (a), a instância está na área de dispersão; em (b), a instância está no limite da área de dispersão (instâncias circuladas); em (c), está fora da área de dispersão tubo com uma folga positiva.

Figura 8 - Exemplo de SVR para Regressão



Fonte: Alpaydin (2010).

Uma das vantagens do SVM é a possibilidade de realizar classificações não lineares, pois a propriedade de convexidade forte garante mínimos globais e oferece uma solução esparsa porque podemos nos concentrar apenas nos 'vetores de suporte'. Enquanto as desvantagens são as maneiras mais fáceis de criar um classificador de relacionamento *n*-ário e criar n SVMs e treinar cada uma delas uma por uma, não é robusto para *outliers* no espaço de entrada e a escalabilidade computacional é limitada (SYAM; SHARMA, 2018).

Bhavsar *et al.* (2017) reitera que ao usar o SVM, é preciso ter muito cuidado

com os tipos de dados, algoritmo e implementação. Um dado desbalanceado é um dos problemas potenciais frequentemente observados para este método. A classificação correta de pequenos pontos de classe muitas vezes se torna mais importante do que a classificação de outros, nesses casos, devendo-se procurar um modelo SVM que trate desse problema.

2.2.4 K-vizinhos mais próximos (KNN)

Zeineddine *et al* (2020) descreve que K-vizinhos mais próximos (do inglês k-*Nearest Neighbors* - KNN) é um algoritmo simples que classifica um ponto de dados com base na classe predominante de seus K-vizinhos mais próximos. Os dados nesta técnica abrangem vários atributos multivariados que são usados para classificação.

Lopes (2018) reitera que o algoritmo de KNN é um método antigo e intuitivo de ML. Para este método, são utilizadas as observações dos dados de treino para encontrar os KNN para o dado que se deseja prever, seguindo uma métrica de proximidade.

A funcionalidade do algoritmo é estruturada para procurar os registros classificados mais próximos, e a partir dessa informação, classificar o novo registro, sendo o número de vizinhos mais próximos a serem comparados definido no valor "K" (FERRERO, 2009).

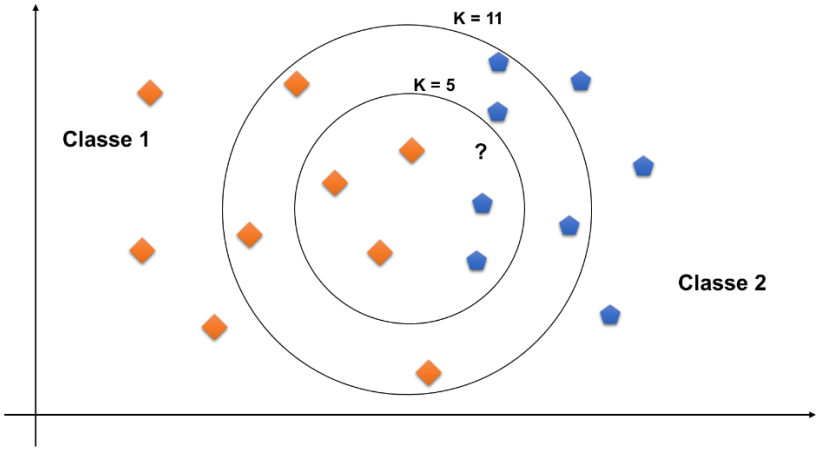
Amaral (2016) relata que a implementação do algoritmo KNN é realizado por meio da distância euclidiana, com base no teorema de Pitágoras. A parametrização da quantidade de vizinhos consideráveis na análise é representando por K, sendo muito menor a N, o tamanho da amostra (ALPAYDIN, 2010).

Para o cálculo da distância euclidiana, Roza (2016) escreve que a distância entre elementos x_i e x_j, é realizada pela equação 17:

$$D(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2} \tag{17}$$

Para ilustrar, foi criada a Figura 9, tendo duas classes, uma de pentágonos azuis e a outra de losango laranjas, na qual a chegada de uma nova variável é simbolizada pela interrogação, e a mesma precisa ser classificada.

Figura 9 - Exemplo de classificação por meio do KNN.



Fonte: Adaptado de Cover e Hart (2018)

Para o caso do exemplo de cassificação da Figura 9, com um k=5, a nova instância deve ser classificada como laranja (três vizinhos desta classe), permanecendo na classe laranja para um k=11, pois são seis vizinhos pertencentes à este rótulo, enquanto cinco são azuis.

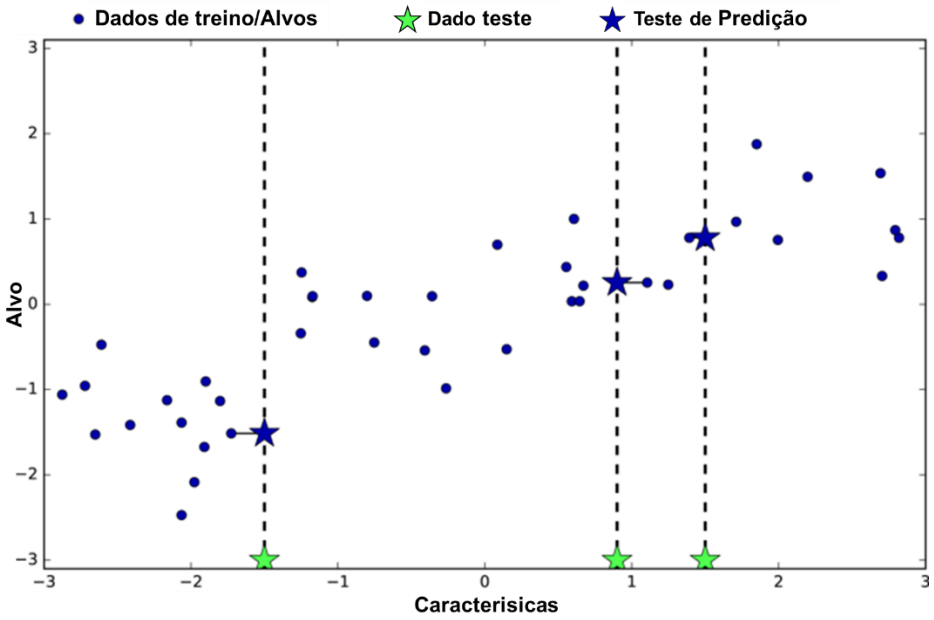
Rhys (2020) supõe que ao usar o algoritmo de KNN para classificação, os dados são para associação de classe, e o dado vencedor seleciona a classe que o modelo produz para os novos dados. O processo de votação ao usar K-vizinhos mais próximos para regressão é muito semelhante, exceto que consideramos a média desses k dados como o valor previsto para os novos dados.

Para regressão, Muhajir (2019) explica que as equações utilizadas são as mesmas que para classificação, podendo ser Euclidiana, Manhattan ou Minkowski. A equação utilizada foi a de Minkowsi, que é dada pela equação 18:

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^2\right)^{\frac{1}{q}} \quad (18)$$

Para ilustrar, foi criada a Figura 10 em que os dados são representados por círculos azuis, os dados testes por estrelas verdes e os dados de teste de predição com estrelas azuis:

Figura 10 - previsões feitas por regressão de KNN.



Fonte: Adaptado de Muhajir (2019).

Para o caso da Figura 10, a representação de novos dados testes com base em suas características, irá resultar em um valor alvo com base na distância para cálculo utilizada.

2.2.5 Árvore de Decisão

Muitos pesquisadores têm usado o método de previsão de árvore de decisão por sua clareza e facilidade em expor conjuntos de dados pequenos e grandes e prever o valor. A lógica na aplicação de técnicas de árvore de decisão equivale a uma série de declarações *IF-THEN*, que podem ajudar a simplificar o entendimento deste método (ZEINEDDINE *et al*. 2020).

Bhavsar *et al*. (2017) definem que árvore de decisão é um método não paramétrico que possui uma estrutura semelhante a uma árvore ou fluxograma, podendo ser utilizado para problemas de classificação e regressão.

A árvore de decisão começa com uma pergunta primária para um determinado problema que deve ser respondida para sua resolução. Esta questão é então dividida em possíveis soluções que podem ou não ter uma resposta definitiva. Cada solução possível é então examinada considerando o resultado que pode ou não exigir outras tomadas de decisão. Se o resultado de uma possível solução requer outra decisão, o

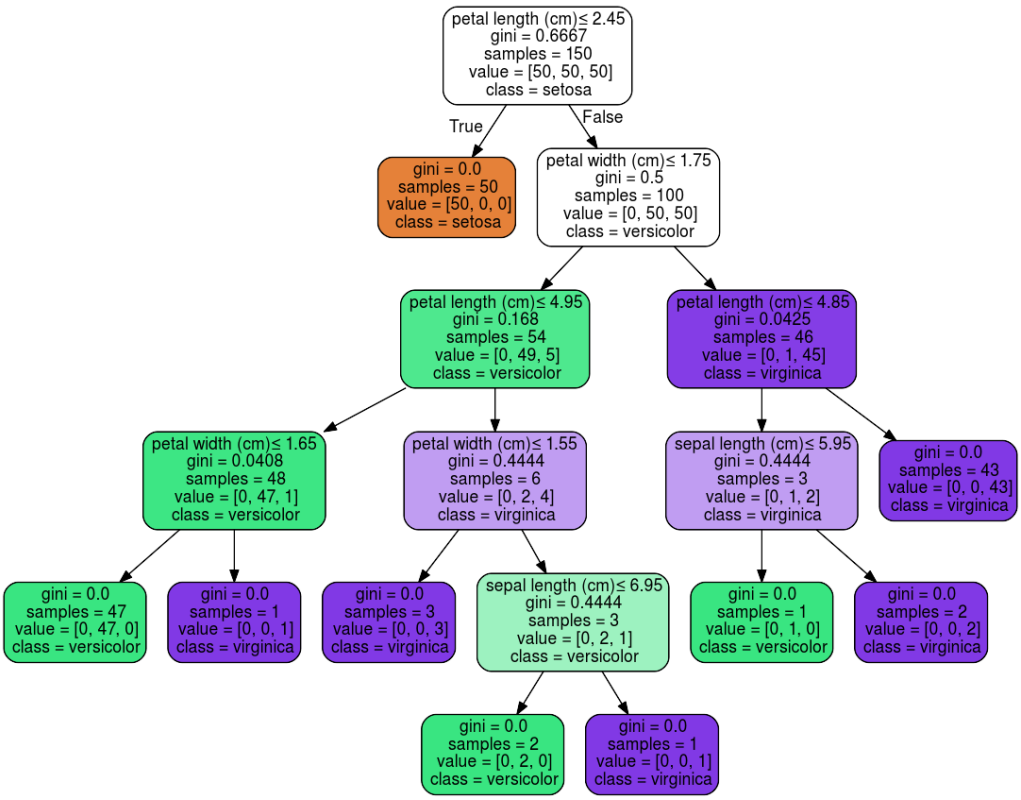
processo continua com a identificação de possíveis resultados para a nova decisão e considerando os resultados de cada um dos resultados. O processo termina quando todas as decisões e resultados possíveis são considerados, resultando em uma estrutura semelhante a uma árvore na qual a sequência lógica de decisões leva, em última análise, à decisão.

Roza (2016) afirma que a representação da árvore de decisão consiste em uma estrutura em que cada nó interno corresponde a um teste sobre um determinado atributo, e que cada ramo descendente representa uma possibilidade para esse teste, e cada folha contém a classe respectiva às instâncias anteriormente classificadas e a decisão obtida após testar os atributos de forma sequencial. O caminho percorrido para chegar à classe corresponde a uma regra de classificação.

Bhavsar *et al.* (2017) reitera que, para um problema de classificação, o processo segue a estrutura de uma árvore onde um atributo mais informativo para classificar os dados é dividido em ramos hierárquicos de forma que a próxima pergunta a ser feita dependa da resposta da pergunta atual. Para uma árvore de decisão de classificação, o nível de impureza é medido para avaliar o desempenho da árvore. Caso a árvore de decisão classifique todos os padrões de dados em classes às quais eles realmente pertencem, as divisões entre a classe e as ramificações são consideradas puras.

A Figura 11 ilustra o processo de classificação por árvores de decisão, utilizando uma base de aprendizagem chamada Iris (DUA; GRAFF, 2019):

Figura 11 – Representação de Árvore de Decisão para o conjunto "*Iris Dataset*"



Fonte: Adaptado de Scikit-Learn (2007).

Uma árvore de decisão para tarefas de regressão é construída quase da mesma maneira que uma árvore de classificação, exceto que a medida de impureza apropriada para classificação é substituída por uma medida apropriada para regressão (ALPAYDIN, 2010). Para isso, considera-se que para o nó *m*, x_m é um subconjunto de x atingindo o nó *m*, define-se que:

$$b_m(x) = \begin{cases} 1 & se\ x\ \in\ Xm;\ x\ atinge\ o\ nó\ m \\ 0 & caso\ contrário \end{cases} \tag{20}$$

Na regressão, uma divisão é medida pelo erro quadrático médio do valor estimado, na qual g_m é o valor estimado no nó *m*:

$$E_m = \frac{1}{Nm} \sum_t (r^t - g_m)^2\, b_m(x^t) \tag{21}$$

Onde, $N_m = |x_m| = \sum_t b_m(x^t)$;

E em um nó, usa-se a média das saídas necessárias de instâncias que alcançam o nó:

$$g_m = \frac{\sum_t b_m(x^t)\, r^t}{\sum_t b_m(x^t)} \tag{22}$$

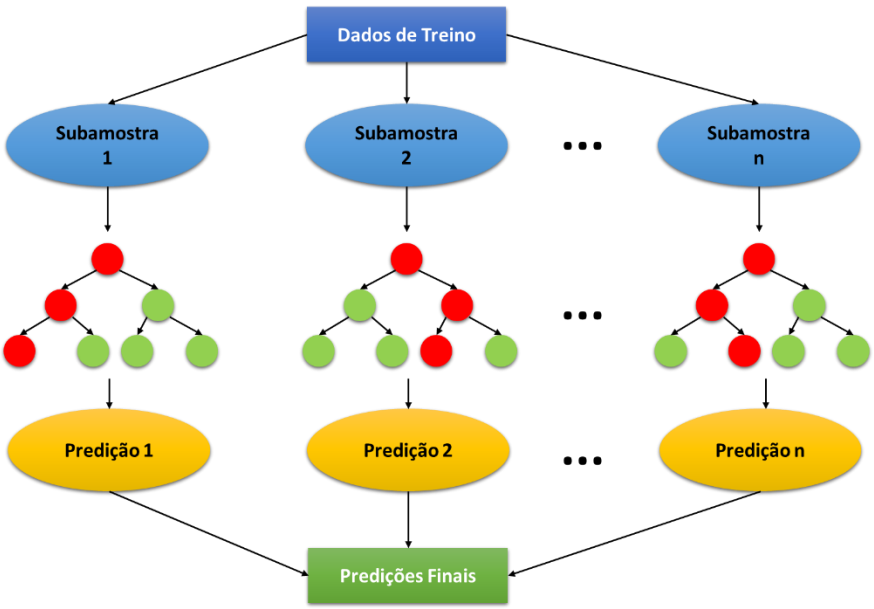
2.2.6 Florestas Aleatórias

O algoritmo Floresta Aleatória é um classificador que consiste em uma coleção de árvores de decisão, onde cada árvore é construída aplicando um algoritmo em um conjunto de treinamento e um vetor aleatório adicional. A previsão da floresta aleatória é obtida por maioria de votos sobre as previsões das árvores individuais para o método de classificação (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Para Cutler *et al*. (2012) *apud* Zhang e Yang (2020), o modelo de algoritmo de Floresta Aleatória, também chamado de modelo de floresta de decisão aleatória, é uma técnica de conjunto que pode ser realizada em problemas de regressão e classificação. Para a execução do algoritmo, deve-se selecionar aleatoriamente n subamostras, treinar a árvore de regressão para cada amostra e calcular a média de todos os resultados de predição de todas as árvores.

Radiya-Dixit, Zhu e Beck (2017) reforçam que os algoritmos de Floresta aleatória é um método *ensemble* de algoritmos de árvores de decisão, cada uma das quais recebe um subconjunto de n características totais, e embora as árvores de decisão por si mesmas sejam propensas a alta variância ou viés, muitos erros são contrabalançados quando compilados em um conjunto.

Figura 12 - Estrutura da Floresta Aleatória



Fonte: Adaptado de Zhang e Yang (2020).

Silveira (2019) afirma que o algoritmo da Floresta Aleatória mostra-se eficaz, capaz de resolver problemas práticos, pois ele fornece um treinamento de alta qualidade, com um alto número de aleatoriedade, introduzido durante o processo de construção do modelo. O algoritmo consegue ser eficaz em estimar quando há dados faltantes e mantém a sua precisão. Ainda, os erros em conjuntos de dados onde as classes são desequilibradas, são equilibrados pelo modelo.

Ainda Silveira (2019) reitera que outro benefício da floresta aleatória é o poder de lidar com dados em grandes volumes e com muitas dimensões. Este algoritmo pode trabalhar com milhares de variáveis de entrada e identificar as variáveis com maior significatividade, o que o torna um dos métodos de redução de dimensões.

2.2.7 Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs), conforme Bhavsar *et al*. (2017), são projetadas para imitar as funções e arquitetura do sistema nervoso. A sua primeira introdução foi por McCulloch e Pitts em 1943, ganhando popularidade significativa no domínio do aprendizado de máquina e da análise de dados. A unidade fundamental da RNA é um neurônio que utiliza uma função de transferência para calcular a saída de uma determinada entrada.

Bhavsar *et al.* (2017) reitera que esses neurônios são conectados para formar uma rede através da qual os dados fluem. As conexões são ponderadas e escalam o fluxo de dados por funções de transferência para que cada entrada corresponda à uma saída do mapa de neurônios.

Já Zeineddine *et al.* (2020) descrevem que uma RNA pode detectar todas as interações existentes entre as variáveis independentes. A capacidade da RNA de detectar associações complexas de alta confiança entre as variáveis independentes e dependentes a torna uma ferramenta poderosa.

Para Falqueto (2018) o algoritmo de RNA é um sistema físico que pode adquirir, armazenar e utilizar conhecimentos experimentais, e pode alcançar um ótimo desempenho por causa do grande número de conexões entre os neurônios da rede.

Zhang e Yang (2020) descrevem que, na literatura, os modelos de RNA mais implementados são do tipo *Perceptron* Multicamadas (*Multilayer Perceptron* - MLP), que pode ser expresso da seguinte forma, na equação 23:

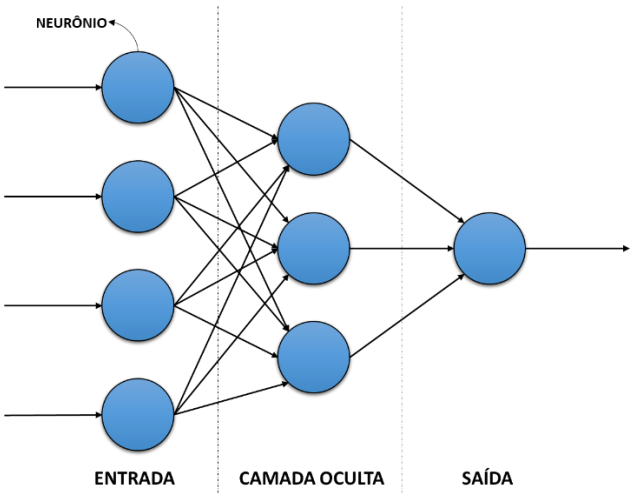
$$y = h\left(\varphi_0 + \sum_{j=1}^{N} \varphi_j g\left(\sum_{i=1}^{M} \theta_i x_i\right)\right) \tag{23}$$

Onde M e N denotam o número de neurônios na camada de entrada e na camada oculta, respectivamente; g e h representam as funções de transferência da camada de entrada e da camada oculta; e as matrizes vetoriais de θ e φ denotam os valores de peso dos neurônios nas camadas de entrada e ocultas, respectivamente.

O MLP é uma estrutura de rede neural artificial e é um estimador não paramétrico que pode ser usado para classificação e regressão (ALPAYDIN, 2010).

A Figura 13 ilustra o processo de aprendizagem do algoritmo RNA, que segundo Bhavsar *et al*. (2017) geralmente, a estrutura da RNA possui três camadas, sendo elas de entrada, oculta e saída. A camada oculta conecta as camadas de entrada e saída com um conjunto extra de neurônios:

Figura 13 - estrutura típica de RNA.



Fonte: Adaptado de Bhavsar *et al*. (2017).

Para Santos *et al*. (2005) há uma necessidade de cuidado para a definição dos parâmetros de implementação, pois estes podem interferir no desempenho das RNAs. Entre eles, estão o número de nós na camada de entrada, número de neurônios e camadas escondidas que serão criadas, e o número de neurônios na camada de saída, todos dependentes do número de variáveis de entrada e saída.

2.3 Aplicações de ML no ambiente Industrial

As indústrias estão em constante busca de inovações para obterem vantagens competitivas sobre seus concorrentes. Conforme explicam Mateus e Mendonça (2020), essa busca por inovações inclui a aplicação de modelos de ML, os quais têm ajudado as empresas a melhorar seu desempenho a e crescer apesar do ambiente fortemente competitivo.

Diversos autores têm estudado formas de aplicar ML ao contexto industrial. Como exemplo, Dogan e Birant (2021) realizaram uma revisão bibliográfica sobre as tendências atuais de aplicações de ML e DM desenvolvidas na indústria de manufatura. Foram identificados diferentes tipos de aprendizado de máquina, como supervisionado (classificação e regressão), não supervisionado (*clustering*, ARM, SPM, detecção de anomalias), aprendizado por conjunto e aprendizado profundo. O estudo apresentou as vantagens dos estudos baseados em ML no domínio industrial e também dá uma ideia clara sobre as dificuldades que os praticantes de aprendizado de máquina enfrentam ao estudar processos e equipamentos de manufatura.

Seguindo o mesmo cenário, na busca por otimização de trabalhos usando IA, o autor Ruberu (2020) estudou as perspectivas de aprendizado de máquina de acoplamento para otimizar a impressão de extrusão de biotintas para obter uma impressão 3D reproduzível, com boa fidelidade de formato utilizando as técnicas de otimização Bayesiana. Os resultados obtidos pelo estudo demonstram uma nova abordagem quantitativa para a otimização, e pode ser prontamente aplicada à otimização da capacidade de impressão de outros tipos de sistemas de tinta, onde a modalidade baseada em extrusão é empregada.

Pereira (2020) busca relatar a aplicação de redes neurais para previsão do desgaste de fresas de topo esférico. O autor ressalta que os resultados finais obtidos com ML cumprem o objetivo de seu estudo, mas recomenda a avaliação com outros modelos de aprendizado de máquina.

Já Silva (2020) propôs o estudo da inferência de vazão de um medidor por efeito térmico a partir do uso das técnicas de ML como KNN, Árvore de Decisão e Florestas Aleatórias, que demonstraram uma maior capacidade de manter a leitura estável ao longo das mudanças de vazão do medidor, mas tendo um baixo desempenho sob a métrica do erro de fundo de escala nas vazões mais baixas.

Com o mesmo intuito, o trabalho de Malakin *et al.* (2021) elaborou um estudo de aplicação de técnicas de classificação em lotes numa indústria farmacêutica para classificar itens como defeituosos ou não defeituosos, e ao avaliar os algoritmos

utilizados, o modelo de Florestas Aleatórias foi o que obteve um melhor desempenho, tanto em erros absolutos, quanto na faixa de erros de precisão e *recall*, para indústrias farmacêuticas.

Reforçado a argumentação para aplicação no segmento da indústria de embalagens, Lara (2020) descreve a utilização do método RNA na otimização da programação de máquinas, e com o treinamento dos dados, conseguiu correlacionar a maioria dos atributos de entrada com a saída esperada, em que dos 10 atributos selecionados, 9 tiveram influência direta no valor de saída do modelo, e com a análise dos resultados obtidos pelo algoritmo, os dados obtidos se aproximam muito dos dados reais.

Alhindawi e Altarazi (2018) descrevem, em seus estudos, que os algoritmos de aprendizado de máquina supervisionado foram usados na previsão da resistência à tração de filmes de extrusão de Polietileno de alta densidade, considerando as características do material e os parâmetros do processo. Foram utilizadas as técnicas de ANN, árvore de decisão e KNN, e descrevem que a vantagem de usar esses algoritmos de ML é que eles podem revelar relações complicadas entre os parâmetros e ingredientes do processo e a resistência à tração dos filmes de extrusão.

E buscando melhorar a determinação de informações dentro de empresas de embalagens, Borges (2020) relata que em seus treinamentos utilizando algoritmos de RNA, obteve resultados ótimos, com um erro médio baixo para a determinação de coeficiente de atrito dos materiais de extrusão. Descreveu-se que as RNA mostram um caminho interessante para a evolução do processo, por ter uma resposta concisa, e entrega com velocidade e de forma clara.

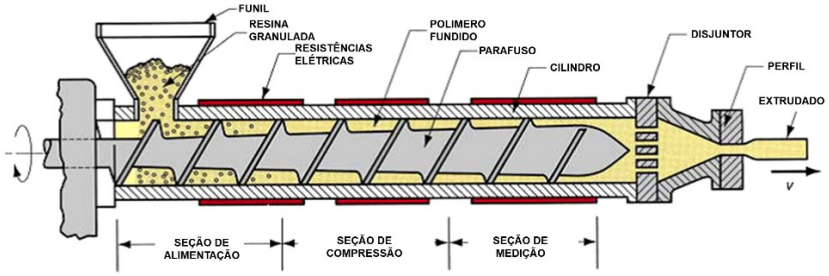
2.4 Extrusão de Embalagens

2.4.1 Processo de extrusão

Khayyami (2019) relata que para a produção de um filme de extrusão, a matéria-prima, grânulos de polímero, é alimentada na tremonha que então entra no funil de alimentação da extrusora. Quando o material entra na garganta de alimentação, na parte traseira do barril, é transportado para um parafuso rotativo, que é acionado por um motor elétrico. A tarefa do parafuso rotativo, independentemente de ser único ou duplo, é gerar pressão suficiente para empurrar o polímero para dentro

do cilindro. O barril é aquecido até a temperatura de fusão desejada e, juntamente com o cisalhamento, os grânulos de polímero são convertidos ao seu estado fundido. No final do parafuso, o polímero está completamente em seu estado fundido, conforme ilustrado na Figura 14:

Figura 14 - Representação esquemática de uma extrusora.



Fonte: Adaptado de Khayyami (2019).

Dooley e Rudolph (2003) descrevem que a Coextrusão de múltiplas camadas é um processo no qual dois ou mais polímeros são extrusados e unidos em um bloco de alimentação ou matriz para formar uma única estrutura com várias camadas. Esta técnica permite que o processador combine as propriedades desejáveis de vários polímeros em uma estrutura com características de desempenho aprimoradas. O processo de coextrusão tem sido amplamente utilizado para produzir folhas multicamadas, filme soprado, filme fundido, tubulação, revestimento de fio e perfis.

Khayyami (2019) continua a descrição do processo de extrusão, definindo que o processo pode ser dividido em três zonas de calor controladas que aumentam gradualmente a temperatura do cilindro com a temperatura mais baixa na parte traseira do cilindro. A divisão das zonas de calor é feita para permitir que o polímero derreta gradualmente à medida que é empurrado através do cilindro e para evitar a degradação do polímero. Depois de passar pelo cilindro, o plástico derretido passa pela matriz e é então transportado por fluxo de pressão para uso no processo de fabricação subsequente.

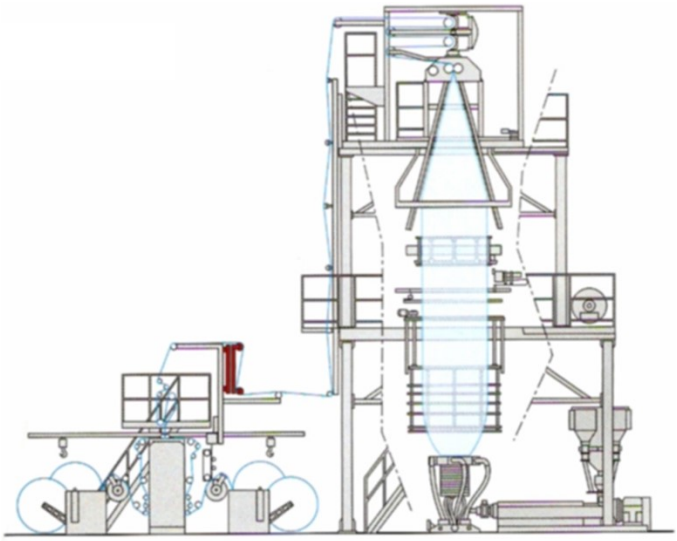
Borges (2020) relata que, após a geração do perfil do material extrusado com a espessura desejada, o material é levado para o balão de extrusão no intuito de reduzir sua temperatura ao longo de sua passagem pelo balão e, assim, se transformar em um filme plástico. A Figura 15 mostra o procedimento de balão extrusor:

Figura 15 - Procedimento de Balão extrusor.

Fonte: Carnevalli (2021)

Após o processo de inflar o material e o guiar até o topo, o material é conduzido para os rolos puxadores, as saias e os rolos guias. A torre é a estrutura metálica que constitui a unidade de arraste, e ela se faz necessária na obtenção de filmes com boa qualidade. Então, os rolos puxadores e as saias devem estar alinhados com a matriz de modo que o filme tubular corra verticalmente até o rolo (SOARES, 2012), conforme ilustra o processo na Figura 16:

Figura 16 - Procedimento de extrusão balão



Fonte: Plasfil (2021).

O embobinamento do filme extrusado só é efetuado quando os materiais saem dos rolos puxadores, sem dobras ou avarias. Quando a temperatura do material extrusado diminui, ele se torna rígido (CANTOR, 2006).

Durante o processo de criação do material Leme e Silva (2018) descrevem que a baixa velocidade de operação e produção atuam como fator preponderante para reduzir o desempenho do equipamento, demonstrando a real necessidade de a velocidade correta para produção não afetar a qualidade do material. Tal como o estudo de Chiaradia (2004) complementa descrevendo que a operação em velocidade fora do padrão recomendado prejudica a produtividade, por necessitar desprender uma maior quantidade de tempo na conclusão da produção planejada.

2.4.2 Extrusão na indústria de embalagens

Khayyami (2019) descreve que o processo de extrusão é uma técnica de revestimento amplamente utilizada em empresas de materiais de embalagem. No processo de revestimento por extrusão, uma resina plástica fundida é aplicada a um substrato e resfriada em seguida para formar uma camada fina e lisa de espessura uniforme. A aplicação de um plástico fundido a um substrato é útil quando o objetivo é laminar um filme plástico ou uma folha de metal, uma vez que o plástico fundido pode atuar como um adesivo.

Borges (2020) destaca que, sobre embalagens plásticas, deve-se levar em conta que elas são formuladas praticamente por filmes plásticos que são compostos por polímeros e que a produção destes filmes pode ser realizada de diversas formas, distinguindo-se entre si de acordo com a necessidade de cada produto.

Khayyami (2019) reitera que ao embalar produtos secos, materiais como papelão são úteis. Mas, como o papel cartão não tem propriedades de barreira e resistência mecânica limitada, ele não é o mais adequado para contato direto com alimentos úmidos e gordurosos. Para superar esse problema, o papel cartão é aprimorado e revestido em um laminador com diferentes camadas de polímeros, para obter um material de embalagem com aparência atraente, resistência à graxa, excelentes propriedades mecânicas e propriedades de selantes.

3 MÉTODO DE PESQUISA

Este capítulo apresenta os passos abordados de maneira sintetizada para o desenvolvimento desta pesquisa, como a estrutura dos dados que compõem as definições de velocidade e *setup*, metodologia de coleta de informações, assim como descrição dos meios utilizados para implementação das técnicas de aprendizado de máquina.

3.1 Descrição do Processo

Como já introduzido anteriormente, a empresa estudada é do ramo de embalagens. Para todo procedimento de produto novo elaborado na empresa, em que o material possui uma composição, espessura e rota de produção específicas, ao serem listados em um banco de dados, necessitam ser vinculados a uma velocidade de produção diária e um tempo de *setup* de sua configuração em máquina. Porém, estas informações não são fornecidas quando é realizado o cadastro destes novos materiais, e para tal, o especialista da função busca as características comuns deste novo material e define estas informações faltantes. Atualmente, essa tarefa é feita com base na experiência do especialista, porém os usos de técnicas de ML podem ajudar na absorção desse conhecimento especialista e ajudar na determinação de dados para futuros lotes de produção por outros funcionários.

3.2 Estrutura da Base de Dados

A estrutura dos dados foi obtida pela empresa por meio de uma planilha eletrônica preenchida pelo funcionário responsável pela atividade de cadastro. Os dados obtidos correspondem aos dados históricos.

O banco de dados é composto por 756 registros do procedimento de materiais extrusados, em que todos os dados já estão tratados e não necessitou realizar a remoção de campos vazios ou com informações incorretas.

A base de dados é constituída de oito colunas, cada uma correspondendo a uma característica distinta, sendo elas:

- **Número de camadas**: é a quantidade de camadas que o material

extrusado possui em sua confecção.

- **Tipo de material extrusado**: corresponde às características e propriedades que o material possui.
- **Espessura**: campo com a espessura do material em micrômetro (μm).
- **Item de Linha**: Diferenciação se o material extrusado pode ser comercializado ou é uma amostra de processos.
- **Cor**: Código da cor do material coextrusado produzido.
- **Máquina**: determina em qual máquina o material será processado e produzido.
- **Centro de Trabalho**: Segue as mesmas condições que as máquinas, e auxilia para dizer em qual máquina será executado.
- **Velocidade de produção**: corresponde à quantidade de material que será extrusado no período de 24 horas. Os valores são expressos em Kg/hora.
- ***Setup***: tempo necessário para configurar as máquinas para produção dos materiais extrusados. A representação é em horas decimais.
- **Perda**: quantidade de material que é perdido durante a produção do material, sendo definido em porcentagem.

A Tabela 1 mostra a relação de quantidade de categorias por atributo:

Tabela 1 - Quantidade de categorias por atributo

Categorias	Quantidade de atributos	Atributos
Nº Camadas	3	3, 7, 9
Filme	6	Material 01, Material 02, Material 03, Material 04, Material 05, Material 06,
Espessura (μm)	80	14 a 308
Item de Linha	2	Sim (1), Não (0)
Cor	8	Cor 01, Cor 02, Cor 03, Cor 04, Cor 05, Cor 06, Cor 07, Cor 08
Máquina	4	Máquina 01, Máquina 02, Máquina 03, Máquina 04
Centro de trabalho	4	Centro 01, Centro 02, Centro 03, Centro 04
VELOCIDADE (Kg/Hora)	16	290, 300, 320, 330, 340, 350, 360, 370, 380, 400, 420, 430, 440, 450, 460, 500

Setup (hora)	4	0,197, 0,251, 0,291, 0,362
Perda (%)	8	2,69, 2,78, 3,25, 3,51, 5,44, 6,71, 10,18, 13,08

Fonte: Elaborado pelo Autor (2021).

Quando é desenvolvido um produto novo para um cliente, a equipe que define a estrutura e composição do material da empresa repassa estas informações para a equipe que realizará o cadastro. Nas informações técnicas, contém os campos de número de camadas, tipo de material, espessura, máquina e centro de trabalho. Porém, os campos de velocidade de produção, *setup* e perda não são informados. Para os funcionários que executam a vinculação das informações, necessitam dos dados de velocidade e *setup* de início, desconsiderando o campo da perda.

Para determinar os dados faltantes, no processo atual o funcionário busca um material de referência que tenha propriedades mais próximas a do novo material que será produzido em todas as colunas que já foram preenchidas. Na Tabela 2, consta uma representação dos dados fornecidos para a produção dos materiais.

Tabela 2 - Representação prévia dos dados

N° CAMADAS	FILME	µm	Item de Linha	Cor	Máquina	Centro de Trabalho	VELOCIDADE (Kg/Hora)	Setup (hora)	Perda (%)
7	Material 01	180	1	Cor 01	Máquina 01	Centro 01	500	0,362	3,25
7	Material 01	200	1	Cor 01	Máquina 01	Centro 01	500	0,362	3,25
7	Material 01	220	1	Cor 01	Máquina 01	Centro 01	500	0,362	3,25
7	Material 01	250	1	Cor 01	Máquina 01	Centro 01	500	0,362	3,25
7	Material 01	100	1	Cor 01	Máquina 01	Centro 01	500	0,362	3,25
7	Material 01	130	1	Cor 01	Máquina 01	Centro 01	500	0,362	3,25

Fonte: Elaborado pelo Autor (2021).

3.3 Descrição dos modelos de análise

Como todos os dados obtidos estão prontos para análise, isto permite definir que as variáveis de saída serão as colunas de velocidade de produção e *setup*. Os

campos estão previstos da seguinte maneira:

- **Velocidade de produção por velocidade**: as velocidades cadastradas correspondem a 16 (desessei) velocidades, que podem ser associadas no momento de produção.
- **Velocidade de produção por faixa de velocidade**: as velocidades cadastradas correspondem a três faixas de valores, sendo elas alta (para velocidades de 440 à 500 kg/hr), média (de 360 à 430 kg/hr) e baixa (de 290 à 350 kg/hr), que podem ser associadas no momento de produção.

Para a elaboração dos testes de aprendizado de máquina, foram realizados 12 testes, sendo sete utilizando métodos de classificação de ML e cinco métodos utilizando as técnicas de regressão, conforme Tabela 3 apresentada no subcapítulo 3.4.

3.4 Implementação de técnicas de *Machine Learning*

A aplicação das técnicas de análise dos dados foi executada por meio de plataformas computacionais, que utilizam a linguagem de programação *Python*. As plataformas foram o *Jupyter Notebook,* e o *Google Colab*, em suas versões *open source*. Para o desenvolvimento de algoritmos e implementação das técnicas de aprendizado de máquina, a biblioteca *Scikit-learn* foi a principal utilizada, por ser a mais difundida no cenário de *Data Science*, seu fácil aprendizado, ampla acessibilidade, simplicidade, eficiência e com uma grande comunidade de desenvolvedores.

As duas plataformas citadas são capazes de ler diversas extensões de arquivo, como “csv”, “txt”, “json”, entre outras e realizar integrações com bancos de dados relacionais e não relacionais. Para o levantamento de informações de máquinas, o formato do arquivo utilizado está em “xlsx”, uma vez que a empresa utiliza a ferramenta *Excel* para seu armazenamento de informações, e maior facilidade de manipulação dos dados.

A partir do referencial teórico foram identificadas as técnicas de ML que serão aplicadas, sendo selecionadas técnicas de classificação e regressão, entre elas, *Naïve Bayes*, Regressão Logística, K-vizinhos mais próximos, SVM, Árvore de Decisão, Florestas Aleatórias e RNA com auxílio da Análise de Componentes Principais (*Principal Component Analysis* - PCA) para serem implementadas.

O computador usado nos testes possui sistema operacional *Windows* 11 *Home Single Language* 64bits, processador Intel(R) *Core*(TM) i7-10510U CPU de 1.80GHz até 2.30 GHz, com 24,00 GB de memória RAM, e placa de vídeo *GeForce* MX110 com arquitetura Maxwell de 28 nanômetros, com 2GB de memória dedicada DDR3.

Os dados apresentados para definição de velocidade de máquina e *setup* já estão padronizados para aplicar os modelos de aprendizado de máquina, mas para que o aproveitamento das ferramentas de ML seja completo, os campos identificados como *string*, foram submetidos à categorização binária (variável *Dummy*), e em seguida, as colunas com dados numéricos submetidos à normalização dos valores, para obtenção de dados mais concisos.

Antes da implementação das técnicas de classificação e regressão, foi realizada a normalização dos dados, pois em sua estrutura, há intervalos de velocidades de máquinas relativamente amplo, quando comparado com os demais campos. A normalização dos dados foi realizada com auxílio da biblioteca *MinMaxScaler*, padrão do *Scikit Learn*, que possui como base a equação 24, executando a normalização:

$$x_{novo} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{24}$$

Definimos os elementos desta equação como:

- x_{novo} – valor resultante normalizado;
- x_i – valor a ser normalizado;
- x_{min} – menor valor da base de dados;
- x_{max} – maior valor da base de dados.

O resultado será o dado normalizado, sendo distribuído no intervalo de 0 a 1.

O método de separação de dados utilizados foi o *Holdout* Repetido, com 100 repetições, alterando o *Random state* com um laço *for* variando de 1 à 100 para todas as técnicas e hiper parâmetros. A divisão dos dados foi estabelecida em 70% para treino (529 instâncias) e 30% para testes (227 instâncias).

A Tabela 3 mostra a relação das técnicas utilizadas para a avaliação e aprendizagem de máquina para a definição de velocidade máquina.

Tabela 3 - Técnicas utilizadas para realização da Aprendizagem de máquina.

Técnicas	Classificação	Regressão

Naïve Bayes	X	
Regressão Logística	X	
SVM	X	X
K-Vizinhos mais próximos	X	X
Árvore de Decisão	X	X
Florestas Aleatórias	X	X
Redes Neurais	X	X

Fonte: Elaborado pelo Autor (2021).

Para a aplicação das técnicas de ML da Tabela 3, foram utilizados os seguintes parâmetros de cada algoritmo, conforme a relação da Tabela 4.

Tabela 4 - Relação dos testes utilizados para cada técnica

Técnica	Parâmetros	Testes
NB	Padrão	Padrão
LR	Padrão	Padrão
SVM	*Kernel* (função)	*RBF*
	C	10^{-1} a 10^{3}
	Gamma	10^{0} a 10^{-4}
KNN	Nº de vizinhos	1 a 30
DT	Profundidade máxima	1 a 30
RF	Nº de estimadores	10 a 1000
ANN	Nº máximo de iterações	2500 a 20000

Fonte: Elaborado pelo Autor (2021).

Começando pela técnica de *Naïve Bayes* foi implementada utilizando a biblioteca *BernoulliNB*, sem necessidade de alterar os parâmetros e os atributos.

Para a sua execução utilizou-se a biblioteca *LogisticRegression*, da *Scikit Learn* sendo alterado o padrão do *solver* para “lbfgs” e do parâmetro *multi_class* de “*auto*” para “*multinomial*”, pois a perda minimizada é o ajuste de perda multinomial em toda a distribuição de probabilidade, mesmo quando os dados são binários, uma vez que esse método permite a generalização da regressão logística.

A utilização e implementação da técnica de máquina de vetores de suporte se ocorreu a partir da biblioteca SVC da *Scikit Learn*, no entanto com o intuito de definir o ótimo local, foi definido a Função de Base Radial (*RBF*, do inglês Radial Basis Function) como o *kernel* para treinar os dados.

A biblioteca utilizada para execução dos testes foi a *GridSearchCV*, permitindo

testar os valores do parâmetro "*C*", que variam de 10^{-1} a 10^{3}, que consiste em modificar as margens da função de decisão, em que, quanto maior o valor do parâmetro, menor será a distâncias entre as margens.

Outro parâmetro alterado foi o "*Gamma*", na qual os seus valores variavam de 10^{0} a 10^{-4}, definindo assim o nível de influência de um único exemplo de treinamento, quanto maior o valor do parâmetro "*Gamma*" maior será a influência dos pontos que estiverem mais perto dele.

Para a execução dos KNNs, o primeiro ponto importante é a escolha do K número vizinhos, assim sendo, em um laço de repetição variando de 1 a 60 para o parâmetro "*n_neighbors*", e utilizando a biblioteca *KNeighborsClassifier* para classificação e *KNeighborsRegression* para regressão, foi realizado a plotagem de um gráfico que apresentava as taxas de erros referente a estes intervalos, sendo escolhido o número de vizinhos que possui a menor média de erros, buscando dar preferência para os números ímpares, como meio de evitar empate no momento das classificações.

Para a execução dos testes de Árvore de decisão, a estrutura foi semelhante a utilizada em KNN, em que foi estabelecido um laço de repetição variando de 1 a 60 para o parâmetro *"max_depth"*, e utilizando a biblioteca *DecisionTreeClassifier* para classificação e *DecisionTreeRegression* para regressão, e plotado um gráfico que apresentava as taxas de erros referente a estes intervalos, sendo escolhido o número de ramificações que possui a menor média de erros.

Mantendo a estrutura dos algoritmos aplicados para KNN e árvores de decisão, a técnica de florestas aleatórias também possui um laço de repetição que varia de 10 a 1000, com intervalos de 10, para o parâmetro *"n_estimator"*.

Foi utilizada a biblioteca *RandomForestClassifier* para classificação e *RandomForestRegression* para regressão, e elaborado um gráfico que apresentava as taxas de erros referente aos intervalos de aleatoriedade, sendo escolhido o número de estimações que possui a menor média de erros.

Para a técnica de redes neurais, usou-se as bibliotecas *MLPClassifier* (*Multi-layer Perceptron classifier*) e Keras, na qual este modelo possui uma capacidade de aprender com modelos não lineares. Também houve a necessidade de definir o número de camadas dos módulos.

Para a avaliação das técnicas de regressão, além das métricas escritas no tópico 2.1.1.2, foi utilizado o algoritmo *Score*, fornecido pelo *Scikit Learn* para todas

as técnicas de ML, que retorna o percentual de acertos e eficiencia média baseado nos dados de teste da base de dados.

4 RESULTADOS PRELIMINARES

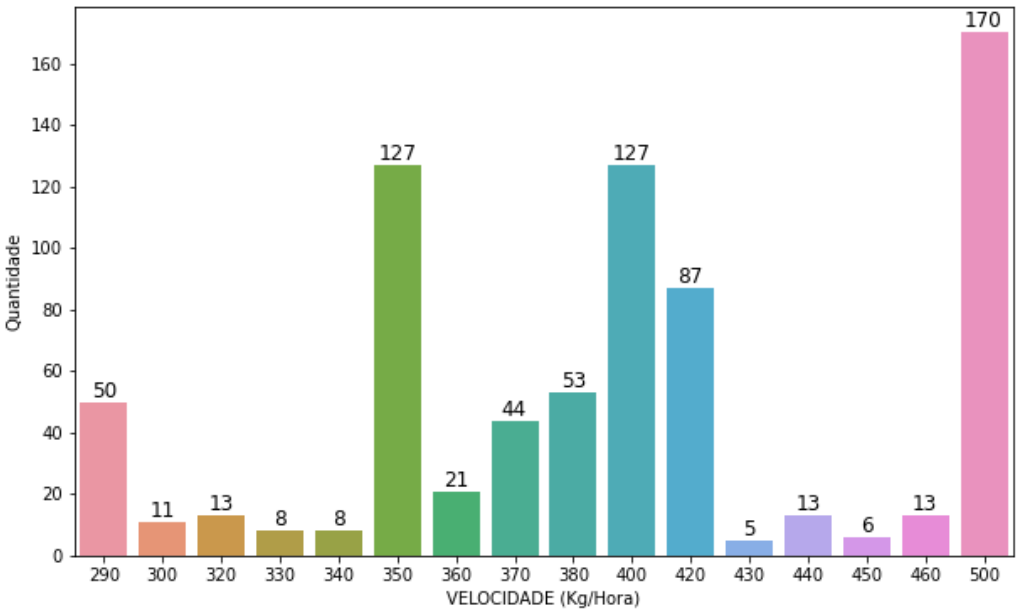
Este capítulo apresenta os resultados obtidos nesse trabalho.

4.1 Análise Descritiva da Base de Dados

Primeiramente foi realizada uma análise exploratória dos dados, buscando interpretar e compreender seu comportamento e tendências, afim de extrair informações relevantes para a criação dos modelos de implementação. Para isso, foi utilizado principalmente técnicas gráficas, como mapa de calor, histogramas e gráfico de pizza.

A base de dados possui ao todo 756 registros, a primeira análise realizada é verificar quais são as velocidades que a base de dados possui, e sua distribuição para cada velocidade. A Figura 17 mostra a quantidade de instâncias para cada velocidade.

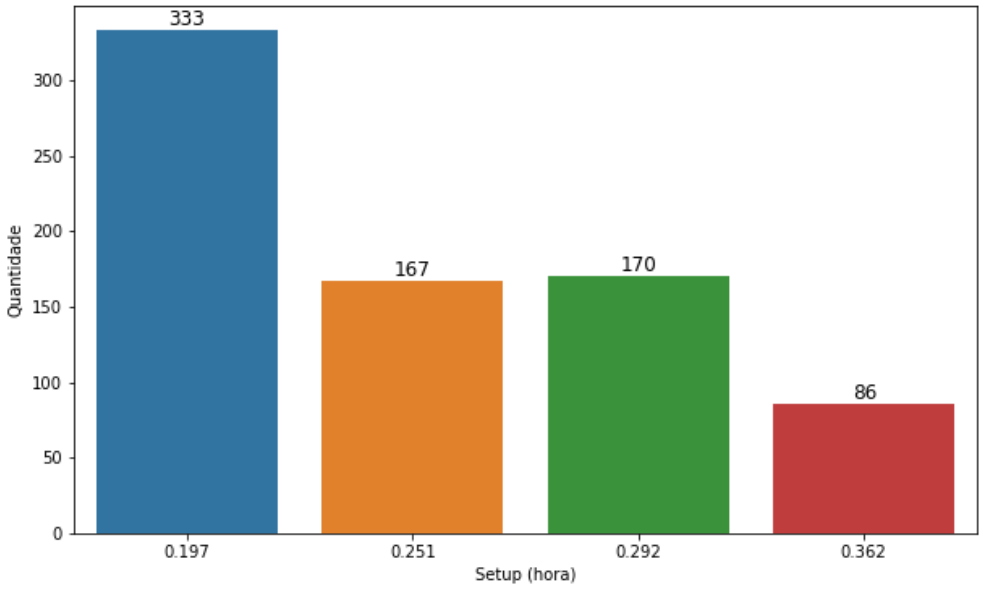
Figura 17 – Distribuição das velocidades de produção.



Fonte: Elaborado pelo Autor (2021).

Conforme visto na Figura 17 acima, a maioria das instancias correspondem as velocidades 500, 400, 350 e 420 Kg/Hora, ou seja 67,60% da base, enquanto as outras velocidades possuem 5 a 13 instâncias. Seguindo a mesma ideia, foi elaborado o histograma da distribuição das velocidades, e ilustrado na Figura 18:

Figura 18 - Distribuição dos tempos de Setup de máquina.

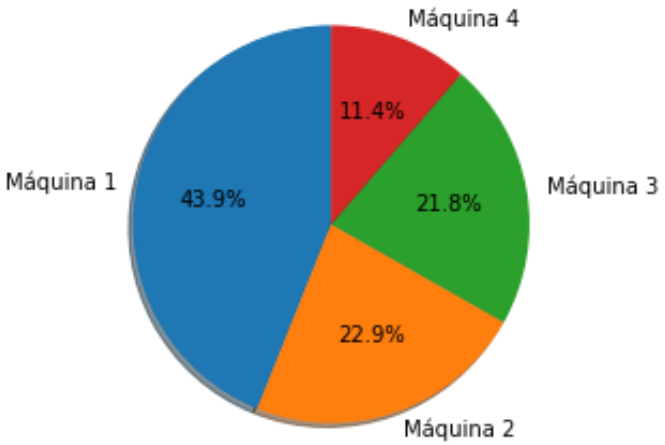


Fonte: Elaborado pelo Autor (2021).

Na distribuição das velocidades, o tempo de setup 0,197 correponde à 44% das instancias, enquanto o tempo de setup 0,362 corresponde a 12%. Quando comparado às informações de velocidade de produção, os dados de setup estão melhor distribuídas.

As instancias estão distribuídos entre quatro diferentes máquinas para produção, a Figura 19 mostra a frequência de distribuição dos registros entre as máquinas:

Figura 19 - Percentual de Filmes por Máquinas



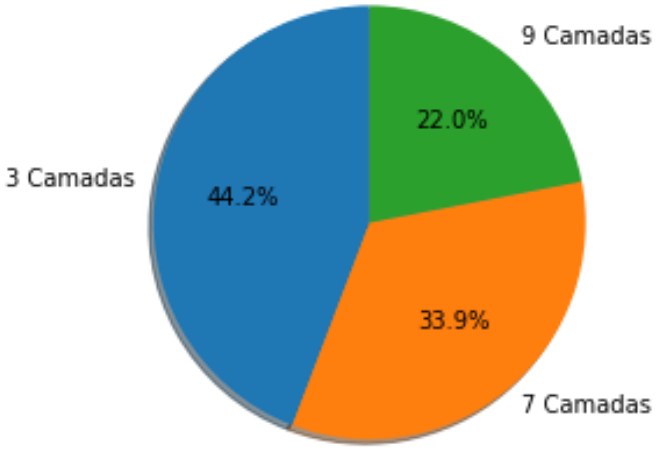
Fonte: Elaborado pelo Autor (2021).

Ao analisar o gráfico da Figura 19, foi possível verificar que a Máquina 1 é a

que possui maior número de materiais extrusados vinculados para produção, dominando quase que a metade de todos os dados. A Máquina 4 é a responsável pela menor fatia de materiais vinculados para a produção. Uma das razões que caracterizam este fato é de ter um número maior de produtos que utilizam os filmes produzidos por esta máquina.

Também foi analisada a quantidade de materiais extrusados por quantidade de camadas, conforme a Figura 20, buscando compreender como está a distribuição da confecção dos materiais.

Figura 20 - Percentual de Filmes por Camadas

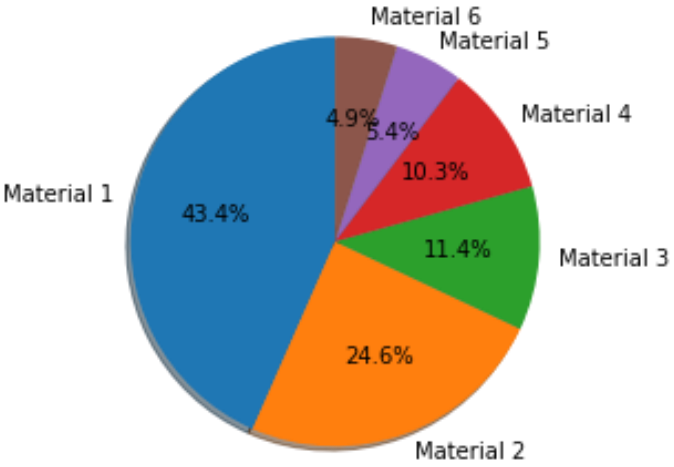


Fonte: Elaborado pelo Autor (2021).

Quando levado em conta a divisão por camadas, o motivo da produção de um material extrusado ser maior que outro se baseia na variedade de mercados que os filmes podem atender.

Quanto à variabilidade de materiais, a Figura 21 mostra que há uma predominância na produção de filmes extrusados do tipo 1 e 2, com 68%, enquanto os materiais do tipo 5 e 6 correspondem a 10,3% de todos os cadastros.

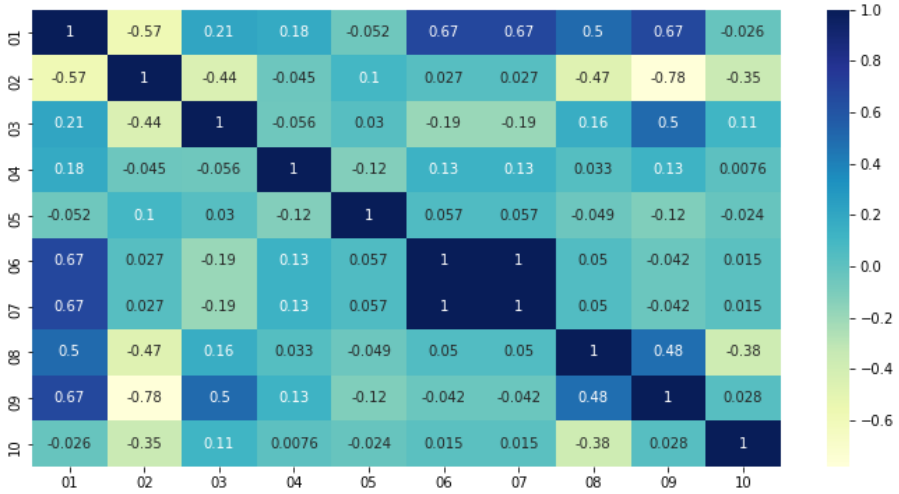
Figura 21 - Percentual de Filmes por Material



Fonte: Elaborado pelo Autor (2021).

Para identificar as relações entre as variáveis, foi calculada uma matriz de correlação dos dados amostrais e posteriormente foi plotado um mapa de calor divididos por *clusters*, com o intuito de criar uma representação visual. Cada célula é exibida de uma cor distinta, sendo essa proporcional à sua posição ao longo de um gradiente de cores. Cabe ressaltar que esta análise retornou apenas o mapa de calor com variáveis do tipo inteiro e *float*. A ordem das linhas se dá através das análises de *clusters* hierárquicos, conforme mostrado na Figura 22.

Figura 22 – Mapa de calor das variáveis



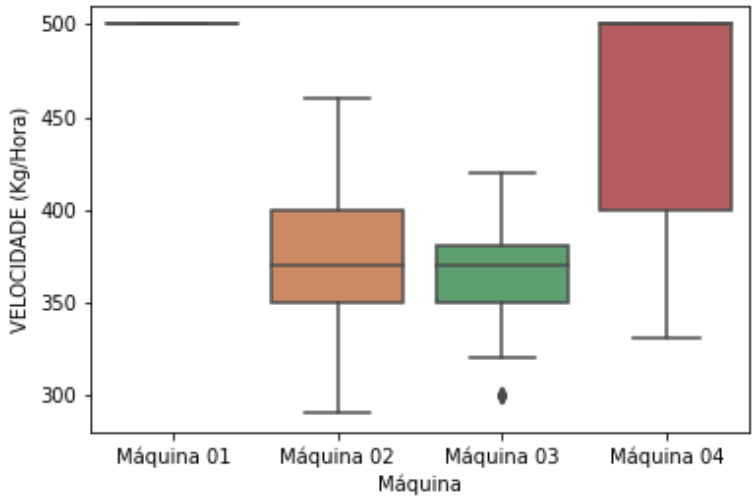
Legenda: (01) Nº Camadas; (02) Filme; (03) Espessura (µm); (04) Item de Linha; (05) Cor; (06) Máquina; (07) Centro de Trabalho; (08) Velocidade (Kg/Hora); (09) Setup (Hora); (10) Perda(%).

Fonte: Elaborado pelo Autor (2021).

Ao observar o mapa de calor na Figura 22 foi possível notar que há uma fraca correlação entre as variáveis listadas e a taxa de perdas, enquanto o número de camadas é o que tem melhor correlação com os outros atributos, em destaque especial com os atributos Máquina, Centro de trabalho, Velocidade e *Setup*. Tornando a percepção de que o número de camadas pode vir a inferir a velocidade de produção e *setup*. Destaca-se que, a correlação perfeita entre máquina e centro de trabalho dá-se ao fato de as informações serem semelhantes, porém necessárias na construção dos dados de extrusão para a empresa.

A partir do mapeamento inicial, é importante ter o conhecimento de quais as velocidades de produção são encontradas, e em quais máquinas são encontradas. A Figura 23 estabelece a distribuição das velocidades dos dados históricos, com as máquinas disponíveis.

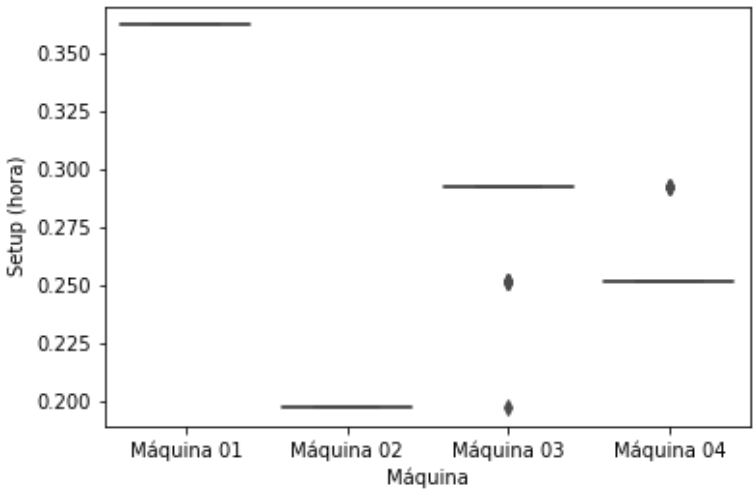
Figura 23 - Distribuição da velocidade de máquina por máquina



Fonte: Elaborado pelo Autor (2021).

Percebe-se que a Máquina 01 tem sua velocidade definida em 500 kg/hora, por ter configurações especificas para produção, enquanto as outras 3 são bem distribuídas entre as velocidades disponíveis. Seguindo a mesma ideia, a Figura 24 define a distribuição do tempo de *setup* com as máquinas disponíveis.

Figura 24 - Distribuição da tempo de *setup* por máquina

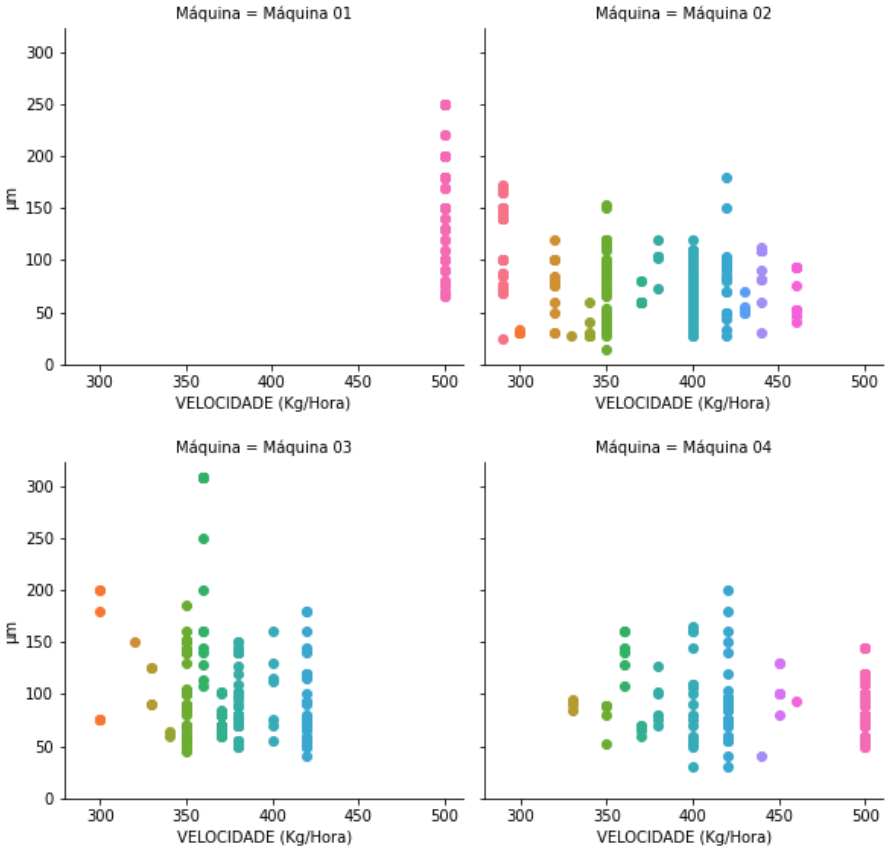


Fonte: Elaborado pelo Autor (2021).

Nesta correlação, a definição do tempo de *setup* mostra que cada máquina tem o seu próprio tempo, e para as Máquinas 03 e 04, há pontos anômalos, determinando que em alguns cadastros há um tempo diferente em suas configurações.

Tornando o mapeamento dos dados mais completos, a Figura 25 retoma ao vínculo das espessuras dos materiais com as máquinas, e as velocidades de sua produção.

Figura 25 - Correlação entre máquina, velocidade e espessura

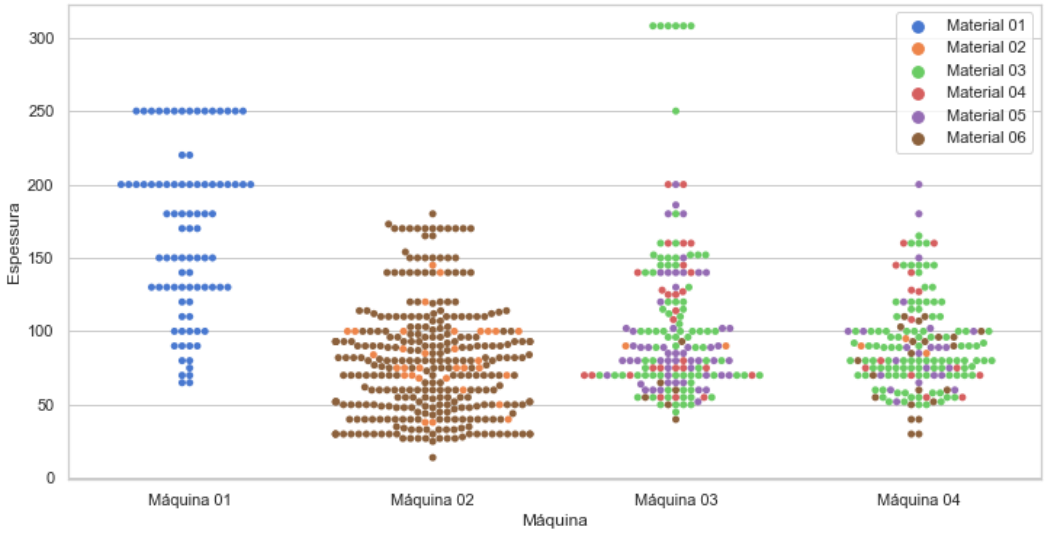


Fonte: Elaborado pelo Autor (2021).

Consegue-se definir pela Figura 25 que as configurações para a Máquina 01, são específicas, tanto de velocidade, quanto de *setup*, porém a espessura varia muito. Enquanto a fabricação e cadastro dos outros extrusados nas outras três máquinas, tem maior variabilidade.

O mapeamento realizado na figura 26 identifica em quais máquinas e quais as espessuras mais comuns para os materiais.

Figura 26 - Correlação entre máquina, espessura e material

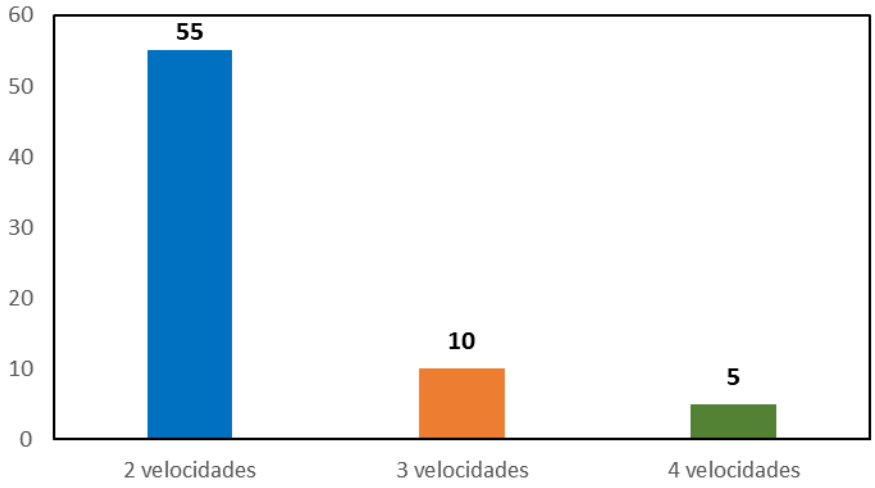


Fonte: Elaborado pelo Autor (2021).

Conforme escrito anteriormente, todas as configurações da Máquina 01 são específicas, mas isto não se aplica às espessuras. Assim como o Material 01 roda exclusivamente na Máquina 01, os demais materiais podem ser produzidos nas outras duas máquinas.

Com as correlações acima escritas, outro ponto a ser verificado é o número de ocorrências em que os atributos iguais resultam em valores alvos diferentes. Ao todo há 70 ocorrências de combinações de atributos que resultam em, ao menos, duas velocidades diferentes conforme mostrado na Figura 27:

Figura 27 – Quantidade de velocidades alvos por combinação de características

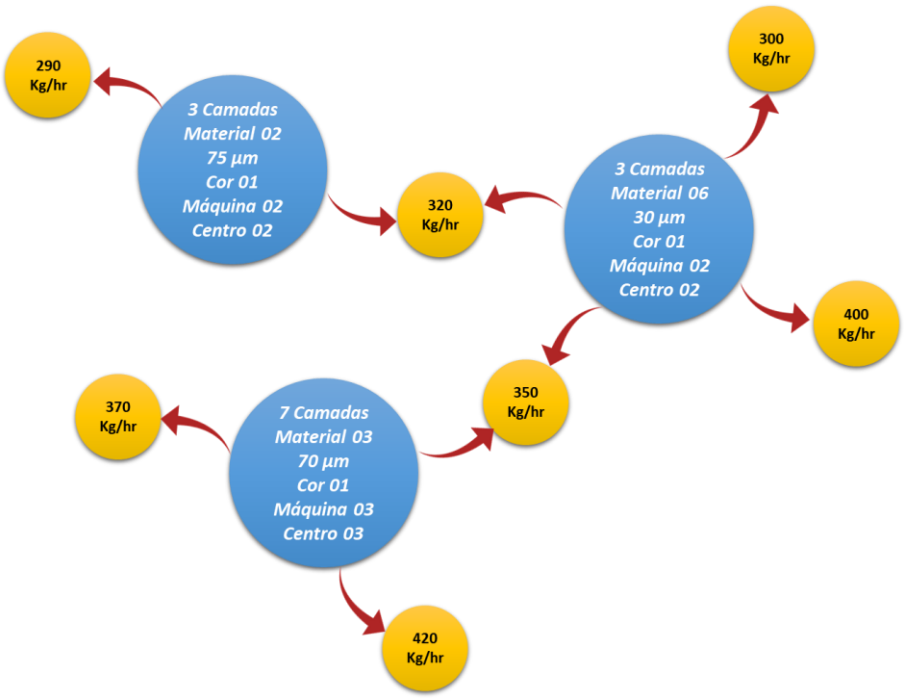


Fonte: Elaborado pelo Autor (2021).

A representatividade da atribuição de duas ou mais velocidades de produção

dos materiais coextrusados conforme suas características, é melhor representado na Figura 28:

Figura 28 – Atribuição de velocidades aos materiais coextrusados



Fonte: Elaborado pelo Autor (2021).

4.2 Comparação e Análise dos Resultados

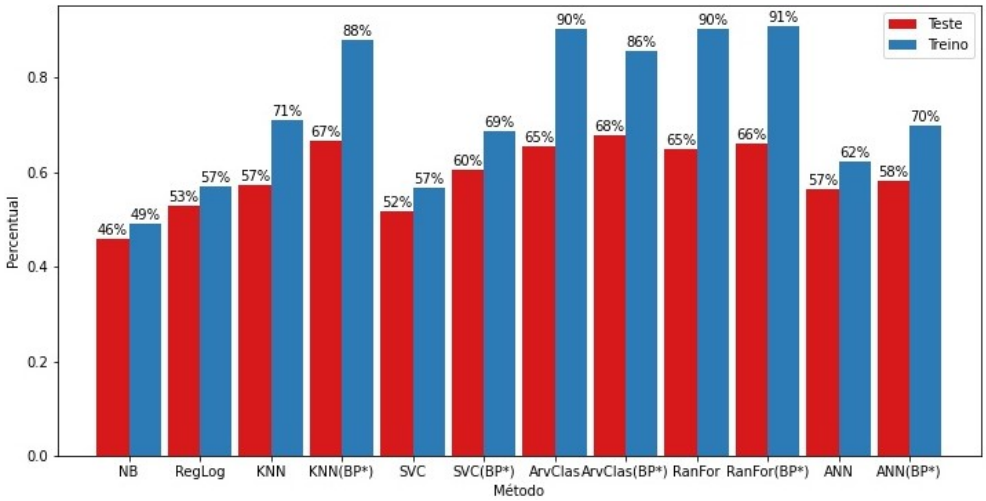
Após realizada a implementação das técnicas de aprendizado de máquina, os resultados tanto da fase de teste como a de treinamento foram analisados utilizando as métricas descritas pelas equações 1, 2 e 5 para técnicas de Classificação e as equações 6, 8, 9 e 10 para técnicas de Regressão. As avaliações foram realizadas de acordo como os modelos de análises e as suas respectivas configurações, conforme referidas na seção 3.4.

4.2.1 Técnicas de Classificação Com Velocidades Definidas

Foram iniciadas as análises pelo método de classificação com as 16 velocidades distintas, mantendo os atributos conforme descrito na seção 3.4. Entretanto, como foi trabalhado com um banco de dados relativamente pequeno, essa abordagem se demonstrou inadequada para a classificação. A Figura 29 mostra os

percentuais de acurácia desse modelo, tanto para os dados de teste como para os dados de treinamento.

Figura 29 – Média da Acurácia dos treinamentos nas bases de Teste e Treino



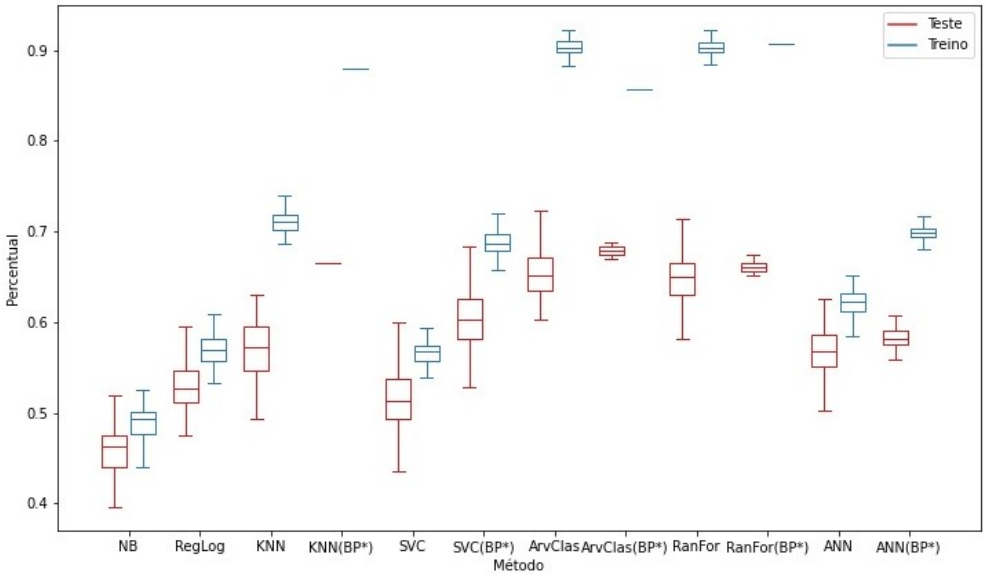
Legenda: (NB) Naïve Bayes; (RegLog) Regressão Logística; (KNN) K-Vizinhos mais próximos; (SVC) SMV; (ArvClas) Árvore de Classificação; (RanFor) Florestas Aleatórias; (ANN) Redes Neurais; (BP) Melhores Locais (do inglês *Best Places*).

Fonte: Elaborado pelo Autor (2021).

Conforme mostrado na Figura 29, as técnicas de Árvore de decisão (*Default* e com 12 nós) e *Random Forest* (*Default* e com 12 nós) foram as que apresentaram melhor desempenho neste modelo com todas as velocidades, obtendo uma acurácia máxima de 68% e 66%, consecutivamente. Enquanto a técnica de *Naïve Bayes* foi a que obteve o pior desempenho.

Por este motivo, a Figura 30 ilustra a diferença entre valores máximos e mínimos a partir de *boxplots* e reforça que as técnicas acima citadas tiveram os melhores desempenhos em todos os testes em geral.

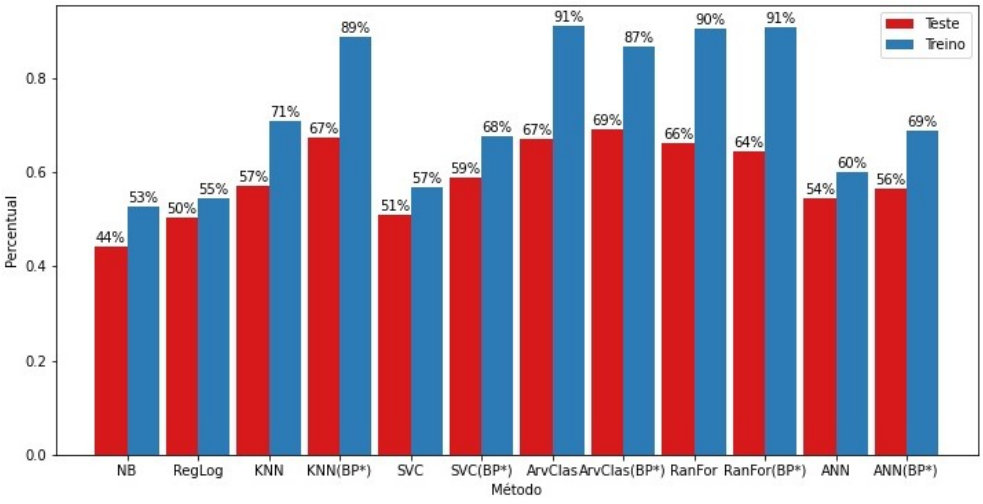
Figura 30 - *Boxplot* da acurácia dos treinamentos nas bases de Teste e Treino



Fonte: Elaborado pelo Autor (2021).

Também foram plotadas as precisões das técnicas para os dados de treinamento, tendo como objetivo verificar a diferença de percentuais entre a fase de treinamento e a de teste, conforme observado pela Figura 31:

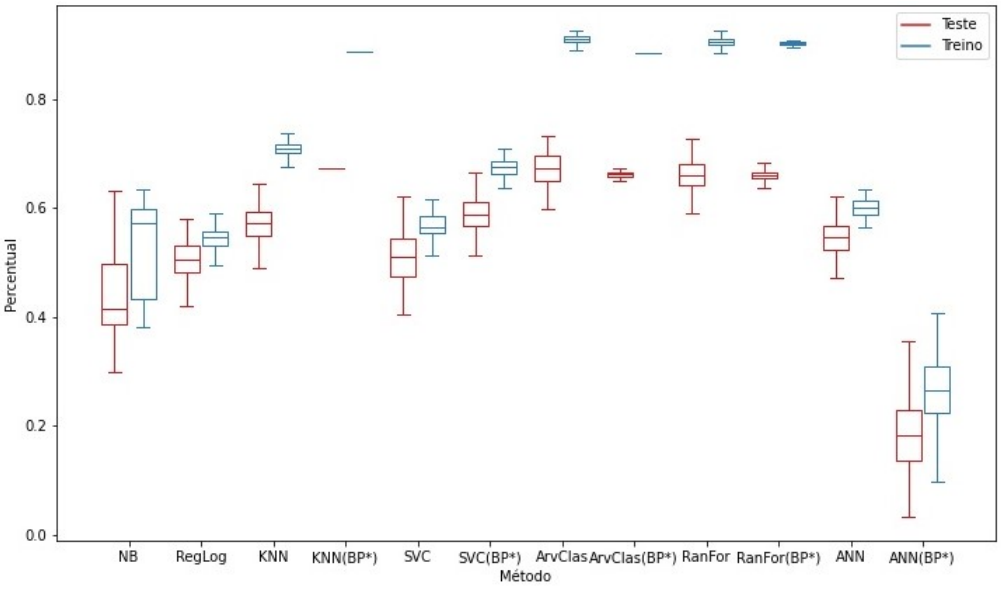
Figura 31 - Média da Precisão dos treinamentos nas bases de Teste e Treino



Fonte: Elaborado pelo Autor (2021).

E a Figura 32 ilustra a diferença entre valores máximos e mínimos a partir de *boxplots* para a precisão das técnicas de ML, e as técnicas de Árvore de Decisão e *Random Forest* foram as que obtiveram os melhores desempenhos em todos os testes em geral.

Figura 32 - *Boxplot* da precisão dos treinamentos nas bases de Teste e Treino



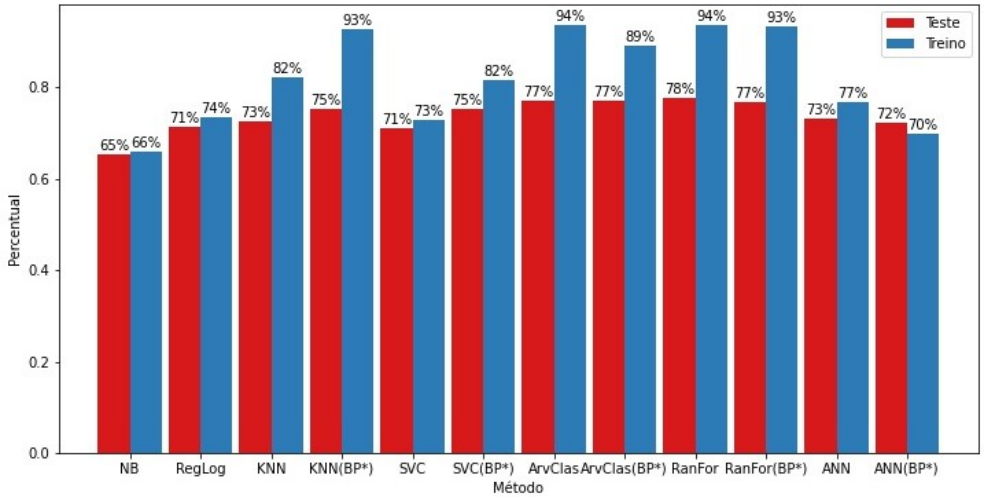
Fonte: Elaborado pelo Autor (2021).

Porém, um ponto a se avaliar, os resultados da acurácia e da precisão demonstra que as técnicas de KNN (com 1 vizinho), Árvore de Decisão (*default* e com 12 ramificações) e *Random Forest* (*default* e com 29 árvores), tiveram bom desempenho com a base de treino, mas muito inferior na base de teste, resultando em *overfitting*.

4.2.2 Técnicas de Classificação por Faixa de Velocidades

As técnicas de classificação por faixa de velocidades buscaram prever entre três classes, conforme escrito no tópico 3.3, contudo, com o objetivo de minimizar a complexidade dos atributos de previsão. Os resultados da acurácia desse método foram obtidos utilizando a média aritmética e podem ser observados na Figura 33:

Figura 33 - Média da Acurácia dos treinamentos nas bases de Teste e Treino



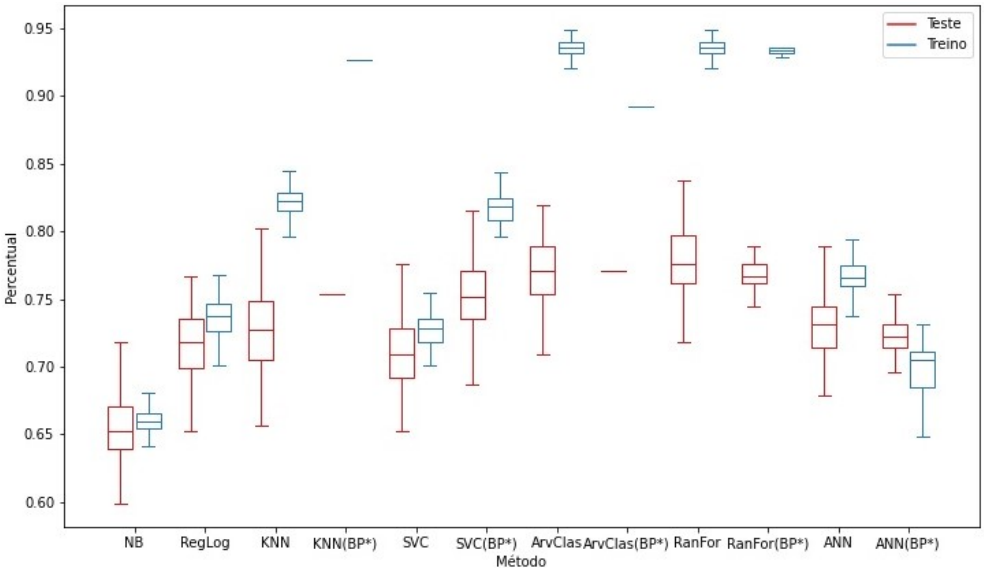
Legenda: (NB) Naïve Bayes; (RegLog) Regressão Logística; (KNN) K-Vizinhos mais próximos; (SVC) SVM; (ArvClas) Árvore de Classificação; (RanFor) Florestas Aleatórias; (ANN) Redes Neurais; (BP) Melhores Locais (do inglês *Best Places*).

Fonte: Elaborado pelo Autor (2021).

Conforme mostrado na Figura 33, a técnica *Random Forest* na base de teste foi a que apresentou melhor desempenho, obtendo uma acurácia média em relação aos testes de 78%. As demais técnicas obtiveram valores médios dos testes de predição próximos à *Random Forest*. Enquanto a técnica *Naïve Bayes* foi a que obteve a pior performance.

Por este motivo, a Figura 34 ilustra a diferença entre valores máximos e mínimos a partir de *boxplots* e reforça que as técnicas de *Random Forest*, tiveram os melhores desempenhos em todos os testes em geral.

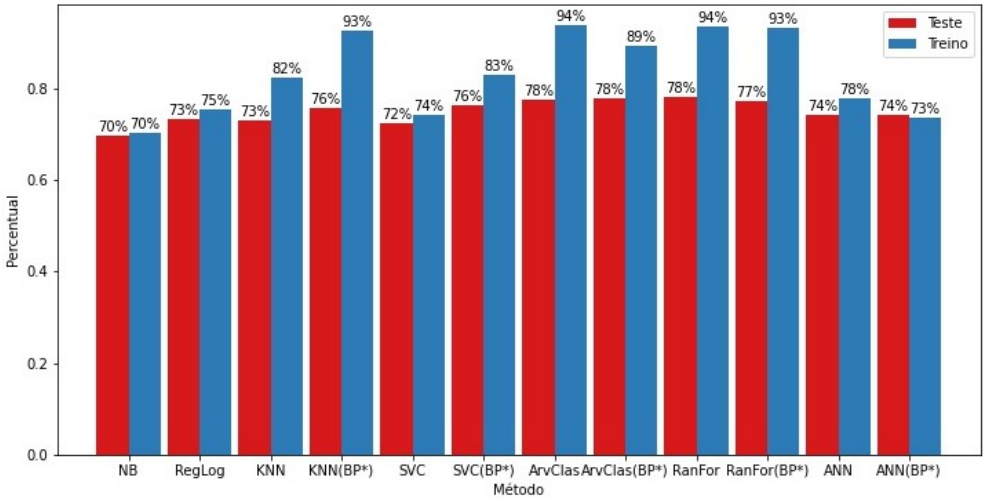
Figura 34 - *Boxplot* da acurácia dos treinamentos nas bases de Teste e Treino



Fonte: Elaborado pelo Autor (2021).

Também foram plotadas as precisões das técnicas para os dados de treinamento, tendo como objetivo verificar a diferença de percentuais entre a fase de treinamento e a de teste, conforme observado pela Figura 35:

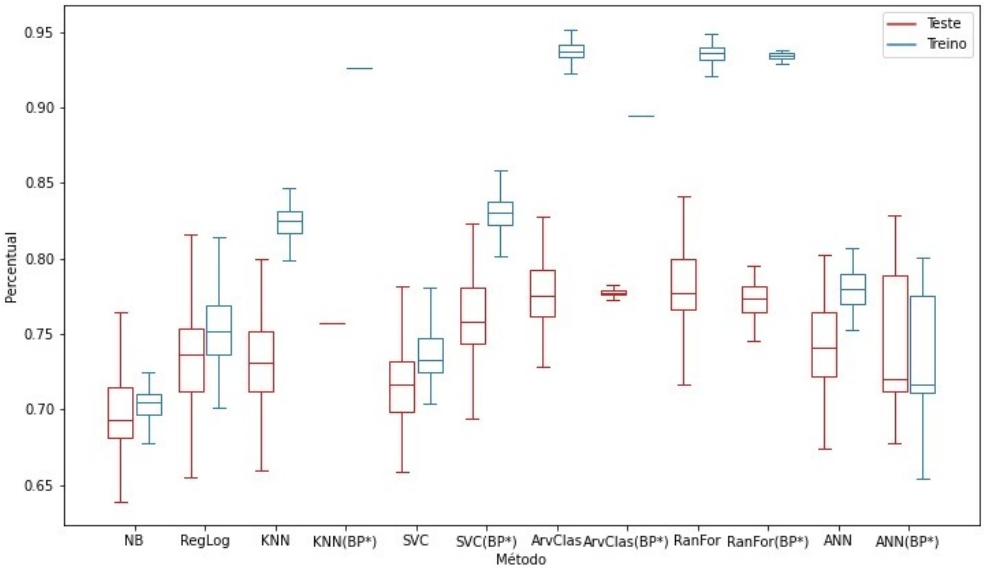
Figura 35 - Média da Precisão dos treinamentos nas bases de Teste e Treino



Fonte: Elaborado pelo Autor (2021).

E a Figura 36 ilustra a diferença entre valores máximos e mínimos a partir de *boxplots* para a precisão das técnicas de ML, e as técnicas de Árvore de Decisão e *Random Forest* foram as que obtiveram os melhores desempenhos em todos os testes em geral.

Figura 36 - *Boxplot* da precisão dos treinamentos nas bases de Teste e Treino



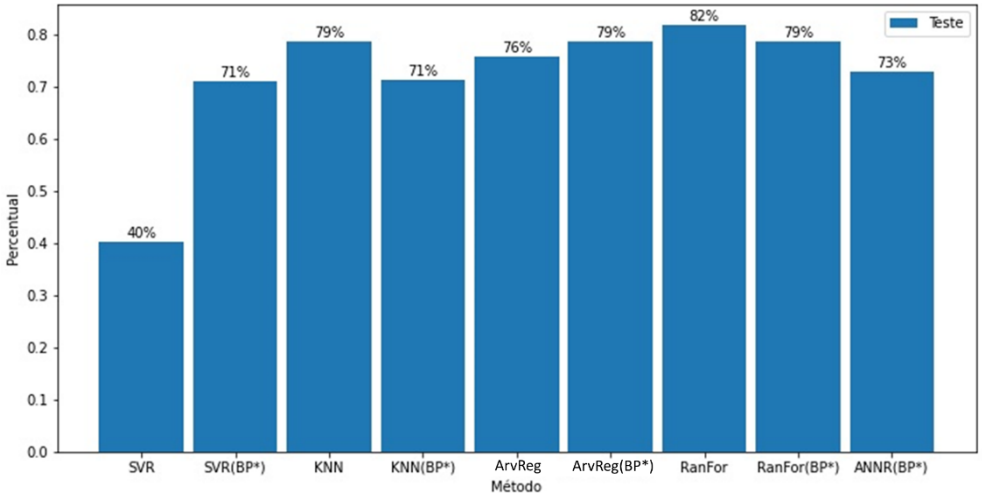
Fonte: Elaborado pelo Autor (2021).

Porém, um ponto a se avaliar, os resultados da acurácia e da precisão mostra que as técnicas de KNN (com 1 vizinho), Árvore de Decisão (*default* e com 13 ramificações) e *Random Forest* (*default* e com 25 árvores), tiveram bom desempenho com a base de treino, mas muito inferior na base de teste, resultando em *overfitting*.

4.2.3 Técnicas de Regressão

As técnicas de regressão buscaram prever as velocidades exatas para novos materiais extrusados. Os resultados dos *scores* dos métodos foram obtidos utilizando a média aritmética e podem ser observados na Figura 37:

Figura 37 - Média do Score dos testes com técnicas de Regressão



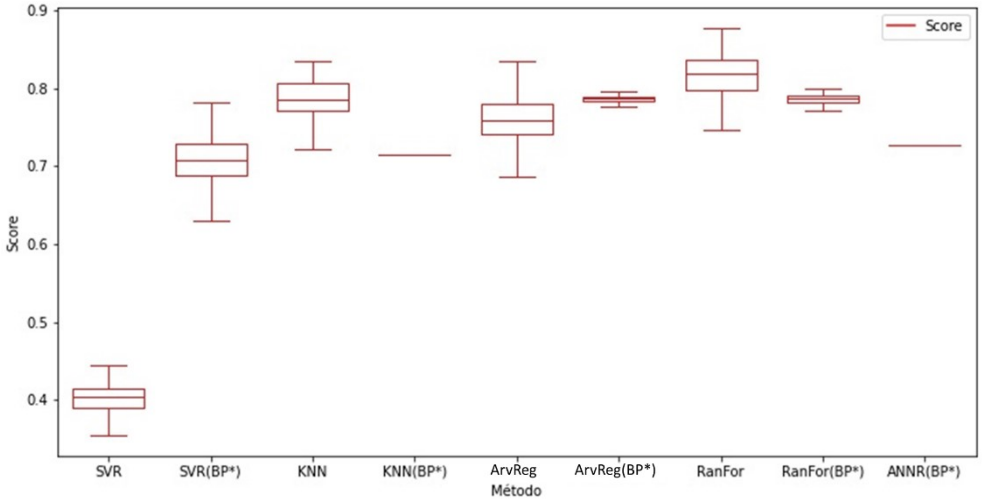
Legenda: (SVR) SVM; (KNN) K-Vizinhos mais próximos; (ArvReg) Árvore de Regressão; (RanFor) Florestas Aleatórias; (ANNR) Redes Neurais de Regressão; (BP) Melhores Locais (do inglês *Best Places*).

Fonte: Elaborado pelo Autor (2021).

Conforme mostrado na Figura 37, com exceção da técnica SVM com configurações *default*, o *Score*, que é a média da acurácia das técnicas de ML, apresentou bom desempenho, com percentual acima de 70%. A técnica de *Random Forest* com configuração *Default* é a que obteve o melhor desempenho, com 82%.

Por este motivo, a Figura 38 ilustra a diferença entre valores máximos e mínimos a partir de *boxplots* e reforça que as técnicas acima citadas tiveram os melhores desempenhos em todos os testes em geral. Percebe-se que não houve *outliers* nos resultados.

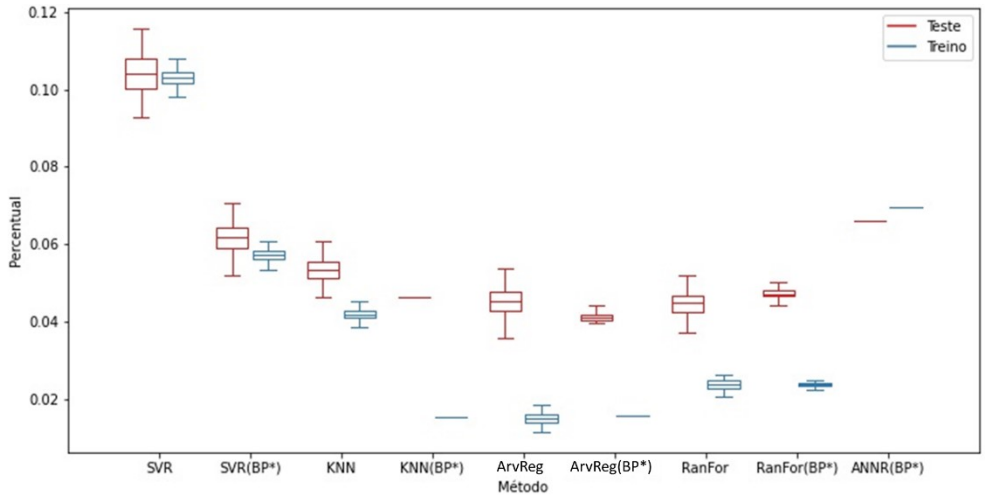
Figura 38 - *Boxplot* do Score dos testes com técnicas de Regressão



Fonte: Elaborado pelo Autor (2021).

Também foi plotado o *Mean Absolute Percentual Error* (MAPE), e para esta métrica, quanto menor o percentual obtido, melhor é a técnica de ML. A análise principal tem como objetivo verificar a diferença de percentuais entre a fase de treinamento e a de teste, conforme observado pela Figura 39:

Figura 39 - *Boxplot* do MAPE dos treinamentos e testes com técnicas de Regressão



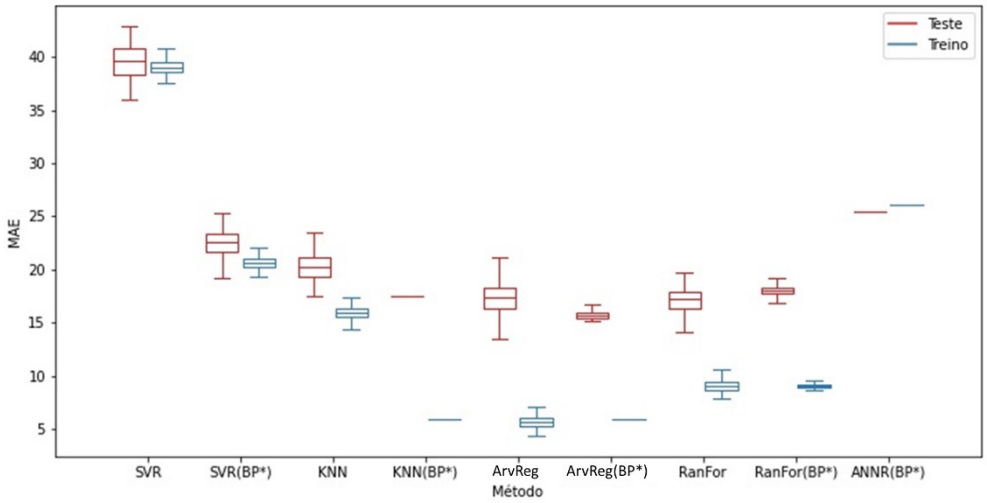
Fonte: Elaborado pelo Autor (2021).

Avaliando os percentuais obtidos, percebe-se que as chances de os valores obtidos por técnicas de regressão estarem incorretos é baixa, ou seja, a diferença entre os valores reais e os valores obtidos, em percentual, correspondem à valores entre 4% e 12%.

E a Figura 40 ilustra a diferença entre valores máximos e mínimos a partir de *boxplots* para a *Mean Absolute Error* (MAE) das técnicas de ML, e monstram que as

técnicas de regressão tiveram os melhores desempenhos em todos os testes em geral.

Figura 40 - *Boxplot* do MAE dos treinamentos e testes com técnicas de Regressão



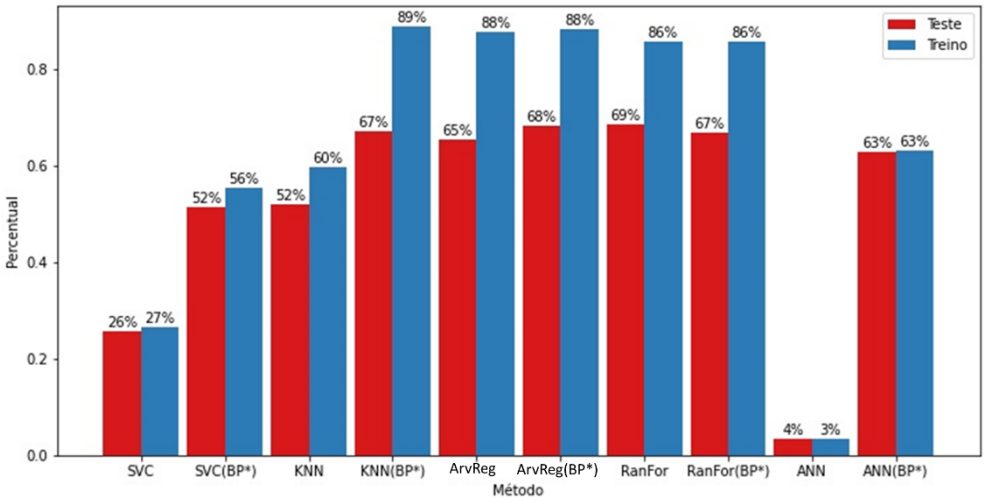
Fonte: Elaborado pelo Autor (2021).

Avaliando os resultados, as técnicas de Regressão em geral obtiveram um bom desempenho, com exceção da técnica de SVM com configuração *Default*, que teve *score* menor que 50%, e seus erros percentuais absolutos foram altos, gerando *underfitting.*

4.2.4 Classificando as Técnicas de Regressão

Com a geração das previsões para as bases de teste e treino para as técnicas de Regressão, foi realizado um comparativo com as velocidades originais da base de dados, buscando avaliar os acertos das técnicas, conforme ilustra a Figura 41:

Figura 41 - Média dos acertos das técnicas de Regressão



Fonte: Elaborado pelo Autor (2021).

O desempenho das técnicas de regressão se assemelha às primeiras técnicas de classificação utilizadas para previsão de velocidades, em que o algoritmo de *Random Forest* é que possui melhor desempenho, com 68%. Em compensação, SVR e ANN com configurações *default* tiveram os piores desempenhos no geral.

4.2.5 Comparativo de técnicas

Com o propósito comparar as técnicas de aprendizado de máquina foi estruturado a Tabela 05, que mostra o comportamento da acurácia das mesmas de acordo com os modelos.

Tabela 5 - Comparativo dos percentuais de acurácia das Técnicas de ML

	Classificação por Velocidade Definida		Classificação por Faixa de Velocidade		Regressão	
Técnicas	Teste	Treino	Teste	Treino	Teste	Treino
Naïve Bayes	46%	49%	65%	66%	-	-
Regressão Logistica	53%	57%	71%	74%	-	-
KNN	57%	71%	73%	82%	79%	60%
KNN parametrizado	67%	88%	75%	93%	71%	89%
SVC/SVR	52%	57%	71%	73%	40%	27%
SVC/SVR (BP*)	60%	69%	75%	82%	71%	56%
Árvore de Decisão	65%	90%	77%	94%	76%	88%
Árvore de Decisão (BP*)	68%	86%	77%	89%	79%	88%
Random Forest	65%	90%	78%	94%	82%	86%
Random Forest (BP*)	66%	91%	77%	93%	79%	86%

Redes Neurais	57%	62%	73%	77%	4%	3%
Redes Neurais (BP*)	58%	70%	72%	70%	73%	63%

Fonte: Elaborado pelo Autor (2021).

Observando os percentuais da tabela pode-se notar que, de maneira geral, as técnicas apresentam melhor resultado se considerado um menor número de classificadores, em que possui uma maior acurácia. Em todos os testes realizados, o algoritmo de *Random Forest* (Classificação e Regressão), foi o que apresentou melhor desempenho.

5 CONSIDERAÇÕES FINAIS

As técnicas de aprendizado de máquina propostas buscaram classificar e definir a velocidade de produção dos materiais coextrusados, apresentando duas formas de avalição com utilização de técnicas de classificação e regressão. As técnicas de classificação foram aplicadas de duas formas. A primeira delas buscou classificar entre as 16 velocidades da base de dados, e a segunda entre faixa de velocidades.

Entre as técnicas apresentadas, a técnica de *Random Forest* foi a que atingiu o melhor índice de acurácia e precisão nas técnicas de classificação, e os melhores índices de *score*, MAPE e MAE para técnicas de regressão. Sendo superada a expectativa, devido à complexidade da problemática em estudo, e a baixa correlação dos dados obtidos. Ressaltando assim a relevância do estudo, uma vez que as previsões possam vir a auxiliar na programação de velocidades de diferentes materiais.

Por meio da utilização da RNA para Regressão, os resultados obtidos com as configurações *default* foram insuficientes, a ponto de não poder ser utilizado para predição das velocidades de máquina, pois obteve *score* negativo e a média dos erros absolutos (MAPE e MAE) foram valores extremamente altos se comparados a outras técnicas.

Um dos problemas identificados durante a execução desta atividade, dá-se ao fato de não ter maiores informações que permitam identificar e diferenciar todas as instancias para reduzir o impacto das instâncias iguais com mais de uma velocidade final. Outro ponto, é que a empresa antes do presente trabalho, não dava a real importância do armazenamento das informações para o procedimento de coextrusados, dificultado a veracidade de todos os dados.

Acredita-se que as utilizações das técnicas de aprendizado de máquina sejam úteis para encontrar boas previsões para a problemática em industrias, bem como presumiu-se que essas técnicas possam ser aplicadas em outras etapas do processo produtivo da empresa, que também utilizem velocidades de produção e consigam automatizar as suas execuções.

A partir dos aspectos que levam a gerar resultados não desejáveis, destaca-se as configurações de produção de coextrusados semelhantes, mas que resultam

em velocidades diferentes, gerando ruídos para a aprendizagem de máquina, e o método atual de definição de velocidade, cujo procedimento não possui documentação de sua execução para averiguar se as velocidades cadastradas estão corretas.

Por mais que haja a utilização do método de divisão das faixas de velocidade, não é uma forma válida de trabalho para a empresa em seu dia a dia durante a programação dos materiais coextrusados, servindo apenas como referência de informação de maior assertividade para quem realiza a etapa de cadastro dos dados.

Das principais contribuições deste trabalho, é a permissão de utilização de ML e IA dentro da indústria, mostrando uma aceleração da Indústria 4.0 por meio da automação dos serviços e aprendizagem de rotinas morosas, ou que faltem procedimentos para sua execução.

Sugere-se que em pesquisa futuras sejam utilizados outras técnicas e outros ajustes de parametrização de aprendizagem para a predição da velocidade de produção durante a extrusão. Também se sugere a utilização de métodos como Ganho de Informação (tradução livre do termo *Information Gain*) para a formação das Árvores de Decisão, Análise de Correspondência Múltipla (MCA) e Análise Multivariada para a análise de dados e das técnicas de separação de conjuntos como *K-fold cross-validation* e *Leave-one-out*.

REFERÊNCIAS

ALBUQUERQUE, A. **Como a inteligência artificial pode ajudar sua empresa de engenharia**. ONYZ, 15 Agosto 2021. Disponivel em: <https://www.onyz.com.br/page/pt-br/posts/InteligenciaArtificial.php>. Acesso em: 16 de Outubro de 2021.

ALHINDAWI, F.; ALTARAZI, S. **Predicting the tensile strength of extrusion-blown high density polyethylene film using machine learning algorithms**. 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2018. 715-719.

ALMEIDA, A.; CARVALHO, F.; MENINO, F. **Introdução ao machine learning**. [S.l.]: Grupo DataAt, 2017. Disponivel em: <https://dataat.github.io/introducao-ao-machine-learning/introdu%C3%A7%C3%A3o.html#aprendizado-supervisionado>. Acesso em: 22 Julho 2021.

ALPAYDIN, E. **Introduction to machine learning**. 2ª. ed. Londres: [s.n.], 2010.

AMARAL, F. **Aprenda mineração de dados:** Teoria e Pratica. São Paulo: Alta Books, 2016.

AMBROISE, J.; GIARD, J.; GALA, J. L.; MACQ, B. **Identification of relevant properties for epitopes detection using a regression model**. IEEE/ACM Transactions on Computational Biology and Bioinformatics, v. 8, n. 6, p. 1700–1707, 2011. IEEE.

AMIDI, A.; AMIDI, S. **Dicas e truques de aprendizado de máquina**. Stanford University CS 229, p. 2–4, 2018. Disponivel em: <https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-truques-aprendizado-maquina>. Acesso em: 01 Agosto 2021.

BHAVSAR, P. et al. **Machine learning in transportation data analytics**. In: CHOWDHURY, M.; APON, A.; DEY, K. Data Analytics for Intelligent Transportation Systems. [S.l.]: Elsevier, 2017. p. 283-307.

BISHOP, C. M. **Pattern recognition and machine learning**. Cambrigde: Springer, 2006.

BORGES, P. P. **Aplicação de redes neurais para a determinação do coeficiente de atrito de filmes plásticos flexíveis**. Londrina: [s.n.], 2020.

CANTOR, K. **Blown film extrusion** – an introduction. [S.l.]: Hanser Publishers, 2006.

CARNEVALLI. **Extrusión – Linea Polaris Plus**. Disponível em: <http://carnevalli.com/es/produtos/extrusion-2/>.

CARVALHO, H. M. Aprendizado de máquina voltado para mineração de dados: Arvores de Decisão. Brasilia: [s.n.], 2014.

CHIARADIA, Á. J. P. **Utilização do indicador de eficiência global de equipamentos na gestão e melhoria contínua dos equipamentos:** Um estudo de caso na indústria automobilística. [S.l.]: [s.n.], 2004.

COSSETTI, M. C. **O que é inteligência artificial?** TECNOBLOG, 21 Agosto 2019. Disponivel em: <https://tecnoblog.net/263808/o-que-e-inteligencia-artificial/>.

DOGAN, A.; BIRANT, D. **Machine learning and data mining in manufacturing**. Expert Systems With Applications, 166, 15 Março 2021.

DOOLEY, J.; RUDOLPH, L. **Viscous and elastic effects in polymer coextrusion**. JOURNAL OF PLASTIC FILM & SHEETING, p. 111-122, Abril 2003.

Dua, D.; Graff, C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. 2019. Disponível em: <http://archive.ics.uci.edu/ml>.

FERRERO, C. A. **Algoritimo kNN para previsão de dados temporais:** funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia. São Carlos: [s.n.], 2009.

GOMES, D. D. S. **Inteligência artificial**. Olhar Científico, p. 234-246, 2010.

JAMES, G. et al. **An introduction to statistical learning:** With Applications in R. Nova Iorque: Springer Science + Business Media, 2013.

JONSSON, L. et al. **Automated bug assignment:** Ensemble-base machine learning in large scale industrial contexts. Empirical Software Engineering, 21, n. 4, 10 Setembro 2015. 1533–1578.

KHAYYAMI, S. **Predicting mechanical properties of polymer films after extrusion coating using supervised machine learning algorithms**. [S.l.]: [s.n.], 2019.

LARA, L. D. **Otimização da programação da produção em uma indústria de embalagens utilizando redes neurais artificiais**. Caçador: [s.n.], 2020.

LEAL, I. H. D. S. **O uso de aprendizagem de máquina para identificação e classificação de fake news no twitter referentes a eleição presidencial de 2018**. FACULDADES DOCTUM DE CARATINGA. CARATINGA. 2018.

LEME, R. L. D. A.; SILVA, J. E. A. R. D. **Manutenção produtiva total:** Um estudo de caso sobre a implementação de um modelo de gestão da manutenção. Encontro Nacional de Engenharia de Produção, Maceió, 2018.

LOPES, H. F. **Aprendizado de máquina aplicado a previsão de desempenho de jogadoresde futebol**. São Carlos: [s.n.], 2018.

LUDERMIR, T. B. **Inteligência artificial e aprendizado de máquina**: estado atual e tendências. SciELO Brasil, 19 Abril 2021.

MALAKIN, L. A.; AL, E. **Classificação de defeito em lotes numa indústria farmacêutica**: Uma abordagem prática com aprendizado de máquina em processos de qualidade. I Workshop de Matemática, Estatística e Computação Aplicadas à Indústria, 2021.

MATEUS, F. M. Q.; MENDONÇA, M. D. C. **Machine learning na melhoria de processos internos:** estudos de caso na indústria de varejo brasileira. Rio de Janeiro: [s.n.], 2020.

MUHAJIR, I. **K-neighbors regression analysis in python.** Analytics Vidhya - Medium, 20 Abril 2019. Disponivel em: <https://medium.com/analytics-vidhya/k-neighbors-regression-analysis-in-python-61532d56d8e4>. Acesso em: 01 Agosto 2021.

NEVES, S. A. D. **Técnicas de aprendizado de máquina aplicadas a classificação da qualidade de pavimentos asfálticos utilizando smartphones**. João Monlevade: TCC, 2018.

PEREIRA, E. L. **Aplicação de um modelo de aprendizado de máquina para previsão do desgaste de fresas de topo esférico**. Florianópolis: [s.n.], 2020.

PLASTICS TECHNOLOGY. **The Extrusion Process**. Plastics Technology.

POÇAS, M. F. F.; MOREIRA, R. **Segurança alimentar e embalagem**. Porto: [s.n.], 2003.

RADIYA-DIXIT, E.; ZHU, D.; BECK, A. H. **Automated classifcation of benign and malignant proliferative breast lesions**. Scientiflc Reports, 2017.

RHYS, H. I. **Machine learning with r, the tidyverse, and mlr**. [S.l.]: Manning, 2020. p. 536.

ROZA, F. S. **Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas**, Florianópolis, 2016. Originalmente apresentado como monografia de graduação do Curso de Engenharia de Controle e Automação para Universidade Federal de Santa Catarina.

RUBERU, K. E. A. **Coupling machine learning with 3D bioprinting to fast track optimisation of extrusion printing**. Applied Materials Today, 05 Dezembro 2020.

RUSSEL, S. J.; NORVIG, P. **Artificial intelligence:** A modern approach. 3º. ed. Prentice-Hall: GEN LTC, 2013. ISBN-10 : 8535237011.

SCHÜSSLER, P. J.; BASTIANI, E.; BUSSLER, N. R. C. **Inteligência artificial e aprendizado de máquina**: Utilizando o entendimento da inteligência humana para reprodução na computação. Salão do Conhecimento: Ciência para a redução das desigualdades - XXVI Seminário de Iniciação Científica, 04 Outubro 2018.

SCIKIT LEARN. Logistic regression. **Scikit Learn**, 2007. Disponivel em: <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>. Acesso em: 31 Julho 2021.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning:** From theory to algorithms. Nova iorque: Cambridge University Press, 2014.

SILVA, J. R. D. C. **Algoritmos de aprendizado de máquinas aplicados na inferência de vazão de um medidor de vazão por efeito térmico**. Rio de Janeiro: [s.n.], 2020.

SILVEIRA, I. V. **Modelo de previsão de demanda com o uso de aprendizado supervisionado de máquina:** Um estudo de caso em uma empresa de varejo. Florianópolis: [s.n.], 2019. 49-50 p.

SOARES, A. L. F. **Estudo da Permeabilidade em FIlmes de Polietileno Verde**, 2012.

SOTO, T. **Regression analysis**. In: VOLKMAR, F. R. Encyclopedia of Autism Spectrum Disorders. Nova Iorque: Springer New York, 2013.

SOUZA, W. B. D.; ALMEIDA, G. S. G. D. **Processamento de polímeros por extrusão e injeção:** Conceitos, Equipamentos e Aplicações. São Paulo: Érica, 2015.

STYLIANOU, N. et al. **Mortality risk prediction in burn injury**: Comparison of logistic regression with machine learning approaches. Burns, p. 925–934, 28 Março 2015.

SYAM, N.; SHARMA, A. **Waiting for a sales renaissance in the fourth industrial revolution**: Machine learning and artificial inteligence in sales research and practice. Industrial Marketing Management, p. 135-146, Dezembro 2018.

ZEINEDDINE, H.; BRAENDLE, U.; FARAH, A. **Enhancing prediction of student success**: Automated machine learning approach. Computers and Electrical Engineering, 24 Novembro 2020.

ZHANG, H. **The optimality of naive bayes**. FLAIRS2004 conference, Fredericton, 2014.

ZHANG, Z.; YANG, X. **Freeway Traffic Speed Estimation by Regression Machine-Learning Techniques Using Probe Vehicle and Sensor Detector Data**, 2020.