

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA

TRABALHO DE GRADUAÇÃO

**ALGORITMOS DE APRENDIZADO DE MÁQUINAS APLICADOS
NO DIMENSIONAMENTO E CONTROLE DE ESTOQUE NA
INDÚSTRIA DE BEBIDAS**

Alan Motta Ganem

Trabalho de Graduação apresentando ao
Departamento de Engenharia Química da
Universidade Federal de São Carlos

Orientador: Prof. Dr. **Antonio Carlos Luperni Horta**

São Carlos – SP

2020

BANCA EXAMINADORA

Trabalho de Graduação apresentado no dia 13 de Outubro de 2020 perante a seguinte banca examinadora:

Orientador: Antonio Carlos Luperni Horta, DEQ/UFSCar

Convidado: Thiago Mesquita, DEQ/UFSCar

Professor da Disciplina: Alice Medeiros de Lima, DEQ/UFSCar

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, pela vida, pelo amor e afeto, pelas oportunidades que me deram e pela confiança depositada, mesmo a mais de 2000km de distância.

A meus familiares e amigos de Natal, que mesmo com a distância, sempre fizeram o máximo para não deixar-me senti-la.

À família que São Carlos me deu, pelos diversos momentos de aprendizado e diversão. Família de verdade é aquela que a gente escolhe e eu tenho a sorte de poder ter escolhido mais de uma.

Aos meus irmãos e irmãs do PET-EQ UFSCar, pelos diversos ensinamentos, momentos de alegria e tristeza compartilhados e por terem me feito sentir na pele que nada grande se alcança sozinho e que certas coisas – como a alegria - ao serem divididas se multiplicam.

Ao meu orientador, professor Dr. Antonio Carlos Luperni Horta, pela paciência, ensinamentos e a sensibilidade de me dar a liberdade e tranquilidade necessária para concluir este trabalho.

Por fim, agradeço a cada brasileira e brasileiro que com parte do suor do seu trabalho me proporcionou a possibilidade de frequentar uma universidade pública, gratuita e de qualidade. Espero um dia poder devolver parte daquilo que me foi por eles possibilitado.

RESUMO

O planejamento e controle de estoque é uma tarefa extremamente importante dentro de qualquer indústria. Essa importância se torna ainda mais notável quando a indústria tem que lidar com uma complexa e intrincada cadeia de suprimentos, com insumos variados, tempos de ressuprimento distintos e sazonais e diversos fornecedores. Nesse trabalho, foram utilizadas técnicas de ciência de dados e aprendizado de máquinas para simular a previsão e a simulação do controle de estoques de diversos insumos em diversas unidades fabris, em uma indústria de bebidas. Para a previsão de consumo, foram utilizados os modelos de Holt-Winters, Gradient Boosting (implementação LGBM), Redes Neurais Densas e um Modelo simples de persistência (prever o futuro como sendo exatamente os valores passados). Os modelos preditivos gerados foram validados utilizando a métrica de erro absoluto médio (MAE) e os resíduos foram testados para normalidade (Shapiro-Wilk), média zero (Teste t) e auto correlação (Ljung-Box). Foi desenvolvido um programa na linguagem Python para simular um sistema de controle de estoque utilizando a previsão de cada um desses modelos em conjunto com uma política de estoque heurística, utilizada pela companhia em que esse estudo de caso foi realizado. Foi testado também o redimensionamento dessa política de estoque (tamanho do estoque mínimo) levando em conta a performance preditiva dos modelos para cada série temporal. Por fim, utilizando métricas de estoque e a técnica de fronteira de Pareto para otimização multiobjetivo, foram selecionados os modelos Holt-Winters e de redes neurais densas como candidatos para passarem por uma etapa futura de validação em produção.

Palavras-chave: Planejamento. Controle. Estoque. Aprendizado de máquinas. Indústria. Cadeia de suprimentos.

ABSTRACT

Inventory planning is an extremely important task within any industry. This importance becomes even more remarkable for industries that deals with complex and intricate supply chains, with plants spread around many regions of the country, varying lead times and different suppliers. In this work, we will use data science and machine learning techniques to carry out forecasting and inventory control of various supplies in different plants, in a large beer industry. For the consumption forecast, Holt-Winters, Gradient Boosting (LGBM implementation), Dense Neural Networks and a simple persistence model (predicting the future as being exactly the past values) were used. The predictive models generated were validated using the mean absolute error (MAE) metric and the residuals were tested for normality (Shapiro-Wilk), zero mean (t-test) and autocorrelation (Ljung-Box). A Python software was developed in order to simulate a predictive inventory control system using the prediction of each of these models alongside with a heuristic inventory policy provided by the company. The resizing of the inventory policy was also tested (lower bound threshold), taking into account the predictive performance of the models for each time series. Finally, using inventory metrics from simulation and the Pareto front technique for multiobjective optimization, the best candidates were selected for further validation in production stage.

Key Words: Planning. Control. Inventory. Machine Learning. Industry. Supply Chain.

*“Os fabricantes de ferramentas foram
recriados por suas próprias ferramentas”*

(Arthur C. Clark)

SUMÁRIO

Sumário

RESUMO	II
ABSTRACT	II
SUMÁRIO	V
LISTA DE FIGURAS	VIII
LISTA DE TABELAS	X
NOMENCLATURA	XI
1. INTRODUÇÃO.....	1
1.1. Objetivos.....	2
1.1.1. Objetivos gerais	2
1.1.2. Objetivos específicos.....	3
2. REVISÃO BIBLIOGRÁFICA.....	5
2.1. Sistemas de controle automático.....	5
2.1.1. Controle preditivo.....	6
2.2. Ciência e análise de dados	6
2.3. Modelagem matemática.....	7
2.4. Correlação vs Causalidade em modelos empíricos.....	8
2.5. Ciência de dados na indústria	10
2.5.1. Ciência de dados descritiva	10
2.5.2. Ciência de dados preditiva.....	11
2.5.3. Ciência de dados prescritiva.....	11
2.5.4. Aprendizado de máquinas	12
2.5.5. Tipos de variáveis.....	12
2.5.6. Aprendizado supervisionado	13
2.5.6.1. Regressão	13
2.5.6.2. Classificação.....	14
2.5.7. Aprendizado não supervisionado.....	15
2.6. Series temporais	16
2.7. Suavização exponencial.....	17
2.7.1. Modelo Holt-Winters.....	18
2.8. Árvores de decisão.....	19

2.8.1.	Boosting.....	20
2.9.	Redes Neurais Artificiais	21
2.9.1.	Redes Neurais Densas	22
2.9.2.	Gradiente Descendente Estocástico.....	23
2.10.	Gestão de estoque	24
2.10.1.	Aprendizado de máquinas e previsão de estoque	25
3.	MATERIAIS E MÉTODOS.....	26
3.1.	Descrição do problema	26
3.1.1.	Planejamento de demanda atual	26
3.1.2.	Gestão de estoque	28
3.2.	Modelagem do problema	28
3.3.	Preparação dos dados.....	31
3.3.1.	Preprocessamento Modelo baseline (Naive)	32
3.3.2.	Preprocessamento Holt-Winters.....	33
3.3.3.	Gradient Boosting e Redes Neurais.....	34
3.4.	Treinamento dos modelos.....	35
3.4.1.	Holt-Winters:.....	35
3.4.2.	Gradient Boosting:.....	36
3.4.3.	Rede Neural	37
3.5.	Simulação de estoque.....	39
4.	Resultados.....	41
4.1.	Validação dos modelos	41
4.1.1.	Análise visual	42
4.1.2.	Testes estatísticos dos resíduos	43
	Teste T para viés.....	43
	Teste Shapiro-Wilk para normalidade.....	44
	Teste Ljung-Box para autocorrelação.....	45
4.1.3.	Erro absoluto médio (MAE).....	46
4.2.	Validação das simulações	48
4.2.1.	Estoque mínimo atual	49
4.2.2.	Estoque mínimo sugerido	50
4.3.	Análise desagregada	53
4.3.1.	Pontos estritamente dominantes	55

5.	Conclusão	58
5.1.	Modelos candidatos	58
5.2.	Nova política de estoque	58
5.3.	Próximos passos.....	59
6.	Referências	60

LISTA DE FIGURAS

Figura 1: Sistematização de modelos mecanísticos, empíricos e validação experimental.	8
Figura 2: Imagem de uma correlação espúria entre consumo de margarina e divórcios no estado do Maine.....	10
Figura 3: Exemplo de regressão linear.	14
Figura 4: Classificador multi-classe (3 classes).	15
Figura 5: Clusterização (3 clusters). Note a semelhança entre a classificação multi-classe. A diferença é que os dados não possuem rótulos, diferente do problema de classificação.....	16
Figura 6: A venda de remédios para diabetes na australia por ano, um exemplo de série temporal.	17
Figura 7: Aplicação do método de Holt-Winters com efeito de nível e tendência.....	19
Figura 8: Representação gráfica de um modelo de árvore de decisão com dois estimadores (x_1 e x_2).....	20
Figura 9: Esquematização do perceptron.	21
Figura 10: Propagação e processamento de informação em uma Rede de perceptrons em multicamadas.	22
Figura 11: Exemplo de minimização com gradiente descendente estocástico em uma superfície não convexa.	24
Figura 12: Processo atual de previsão de consumo de insumos.....	27
Figura 13: Analogia de estoques do processo produtivo com um sistema de tanques...	29
Figura 14: Diagrama de blocos do controle de reposição	30
Figura 15: Exemplo de previsão Naive.	33
Figura 16: Arquitetura da modelagem Holt-Winters.....	36
Figura 17: Arquitetura da modelagem LGBM.	37
Figura 18: Arquitetura da rede neural utilizada.....	38
Figura 19: Arquitetura da modelagem DNN.	39
Figura 20: Exemplo do funcionamento do simulador com dados fictícios para leadtime de 20 dias.	41
Figura 21: Previsão de consumo semanal do modelo DNN para um insumo em uma fábrica específica. leadtime do insumo é de uma semana.	42
Figura 22: Previsão de consumo para horizonte de 10 semanas do modelo DNN para um insumo em uma fábrica específica. leadtime do insumo é de dez semanas.	43
Figura 23: Mapa de calor para percentual de modelos com menor MAE para cada fábrica.	47
Figura 24: Perfil de consumo real e resíduos para melhores modelos.	48
Figura 25: Análise de tamanhos de estoque mínimos padrão e sugerido.....	51
Figura 26: Análise de delta de estoque mínimos padrão e sugerido.	52
Figura 27: Simulação com estoque mínimo padrão.	53
Figura 28: Simulação com estoque mínimo reduzido.	53

Figura 29: Distribuição de número de pontos na fronteira de Pareto para todas as séries temporais.	55
Figura 30: distribuição de modelos estritamente dominantes.	55
Figura 31: Ganho relativo ao caso base (PCP) para modelos estritamente dominantes.	56

LISTA DE TABELAS

Tabela 1: Resumo dos 3 principais tipos de "Analytics"	11
Tabela 2: Bibliotecas utilizadas e suas referências.....	30
Tabela 3: Exemplo da estrutura dos dados históricos com valores fictícios.	31
Tabela 4: Exemplo fictício de estrutura de input para modelo Holt-Winters.....	33
Tabela 5: Estrutura de input para modelos multivariados.	34
Tabela 6: Estrutura de output para modelos multivariados.	35
Tabela 7: Estrutura padrão de output de todos os modelos para validação.	35
Tabela 8: Percentual de séries temporais com $p > 0.05$ para teste de viés dos resíduos para modelos em horizontes de previsão específicos.	44
Tabela 9: Resultados dos normalidade dos resíduos para modelos em horizontes de previsão específicos.	44
Tabela 10: Resultados dos testes de autocorrelação dos resíduos para modelos em horizontes de previsão específicos.	45
Tabela 11: Percentual agregado de modelos com menor MAE.	47
Tabela 12: Métricas de estoque agregadas para simulação levando em conta política de estoque padrão. caso base é a reposição pela previsão do PCP levando em conta o estoque mínimo padrão.....	49
Tabela 13: Métricas de estoque agregadas para simulação levando em conta política de estoque sugerida. caso base é a reposição pela previsão do PCP levando em conta o estoque mínimo padrão.....	50
Tabela 14: Tabela resumo de tamanhos de estoque mínimos padrão e sugerido.	52
Tabela 15: Ganho relativo em métricas de estoque por modelo estritamente dominante.	56

NOMENCLATURA

E_k^i	Estoque do insumo i no instante k
\hat{r}_k^i	Reposição prevista do insumo i no instante k
\hat{R}^i	Vetor com previsões de reposição para o insumo i
c_k^i	Consumo real do insumo i no instante k
C^i	Vetor de consumos do insumo i
$\hat{C}_{k,h}^i$	Vetor de previsão de consumos do insumo i no instante k com horizonte h
$\hat{C}_{k:k+h}^i$	Soma cumulativa dos consumos previstos para o insumo i para o horizonte h a partir do instante k
$Cov_{k-l,k}^i$	Vetor de variáveis covariantes para previsão do insumo i dos últimos l períodos a partir do instante k
$C_{k-l,k}^i$	Vetor de consumos do insumo i dos últimos l períodos a partir do instante k
$\hat{c}_{k,j}^i$	Previsão de consumo feita no instante k do insumo i , para j períodos após o instante k
$\hat{R}_{k:k+h}^i$	Soma cumulativa das previsões de reposição para o insumo i para o horizonte h a partir do instante k
$\hat{r}_{k,j}^i$	Previsão de reposição feita no instante k do insumo i , para j períodos após o instante k
$\hat{E}_{k,h}^i$	Vetor de previsão do volume de estoque do insumo i , feita no instante k para o horizonte h
E_{min}^i	Estoque mínimo do insumo i
E_{max}^i	Estoque máximo do insumo i
\hat{r}_{k+h+1}^i	Reposição gerada no instante k com previsão de entrega $h + 1$ períodos após o instante k
$\hat{y}_{t+h t}$	Previsão no horizonte h a partir de t , dado t
ℓ_t	Nível (média) suavizado em t
b_t	Tendência suavizada no tempo t

s_t	Fator sazonal
m	Período de sazonalidade
α	Fator de suavização para nível
β^*	Fator de suavização para tendência
γ	Fator de suavização para efeito sazonal

1. INTRODUÇÃO

A gestão de cadeia de suprimentos é uma atividade fundamental em todos os setores da economia, desde o setor primário, passando pela indústria, até por fim, o comércio. Quanto maior o tamanho e a diversidade dos produtos fornecidos por uma indústria, mais complexa se torna essa tarefa, mostrando-se hoje um grande diferencial competitivo (MAQBOOL; RAFIQ, 2014)

Dentro da intrincada tarefa de gerir complexas redes de suprimento de materiais, pode-se destacar, do ponto de vista operacional, três grandes desafios:

- i) Planejamento de demanda;
- ii) políticas de ressuprimento;
- iii) dimensionamento de estoque.

Todos os desafios estão intimamente ligados e fazem parte de uma grande área de estudo chamada de planejamento e controle da produção (PCP) (TUBINO, 2007). A capacidade de planejar a demanda em um horizonte futuro com precisão é fundamental para tomar decisões acertadas a respeito de estoque de insumos produtivos, infraestrutura logística e produtiva (MULLER, 2003).

O tamanho do horizonte de previsão, comumente tratados (por conveniência) como previsão de curto, médio e longo prazo, define a utilidade da tarefa de previsão. A ampliação e criação de unidades produtivas está geralmente relacionada com previsões de longo prazo, enquanto a decisão de ressuprimento do estoque de um insumo está associada à previsão de curto prazo (MULLER, 2003).

Dada a aplicação operacional do estudo e seu conseqüente foco na previsão de curto prazo, vê-se conveniente também abordar o problema sob a ótica do controle de processos preditivo, ainda que a dinâmica do processo seja trivial (sistema puramente capacitivo com tempo morto). A analogia será útil principalmente no desenvolvimento do simulador de estoque (STEPHANOPOULOS, 1983).

A necessidade de criar estoques de insumos nasce do fato de que estes possuem um tempo de ressuprimento maior que zero, de modo que, caso haja uma necessidade no tempo $t = t_0$, essa necessidade só poderá ser suprida, não havendo estoque, em um tempo $t = t_0 + h$, em que h é o tempo de ressuprimento desse insumo. Ainda que necessário, estoques devem ser projetados e geridos de forma a minimizar uma determinada função custo, que geralmente está associada aos custos de ressuprimento, custo de capital imobilizado e risco de furo de estoque. O estudo de otimização de estoque não é escopo desse trabalho.

Quanto à modelagem da demanda, existem diversas maneiras de realizá-la, desde métodos simples, como média histórica, modelo de persistência de valor mais recente até modelos mais complexos, como algoritmos de aprendizado de máquinas (Machine Learning) e aprendizado profundo (Deep Learning). Alguns deles serão detalhados na seção de revisão bibliográfica.

1.1. Objetivos

1.1.1. Objetivos gerais

O presente estudo foi desenvolvido com o objetivo de avaliar a performance de diferentes modelos de previsão de séries temporais aplicadas na previsão de estoque, política de reabastecimento e dimensionamento do estoque de insumos produtivos em plantas espalhadas pelo Brasil.

O foco será em previsões de curto prazo, com horizontes de previsão variáveis, dado que será modelada a demanda de todos os insumos produtivos de uma grande indústria de bebidas, que por sua vez possuem tempos de ressuprimento diferentes.

Os dados foram compilados, tratados, preprocessados (tratar valores inconsistentes, agregar consumo por semana, preencher valores faltantes), e

separados em um subconjunto de treino (para treinar os modelos) e teste (para validar os modelos).

Foram utilizados para a regressão das séries temporais os modelos de Holt-Winters, Gradient Boosting (GBM), Redes Neurais Artificiais Densas (DNN) e modelo de persistência (prever o futuro exatamente igual ao passado) como a base de comparação, utilizando bibliotecas de código aberto em Python e algumas extensões customizadas pelo autor.

A performance dos modelos será avaliada utilizando testes estatísticos para os resíduos (auto correlação, normalidade e viés) e o valor do erro absoluto médio (MAE) para cada uma das séries temporais (previsões de consumo) do conjunto de teste.

Será realizada uma simulação (malha fechada) utilizando a política de estoque (heurística) da empresa em questão em conjunto com as previsões de consumo geradas pelos modelos e o consumo real de insumos no intervalo de tempo simulado, dentro do conjunto de teste. Todo código Python será rodado numa instância de um notebook Jupyter.

Uma segunda simulação será realizada utilizando uma política de estoque sugerida em função da variância histórica do resíduo (erro) entre a previsão do modelo e o consumo real.

A política de estoque define qual o estoque mínimo, máximo e o nível desejado após o reabastecimento. Existem custos associados à geração de pedidos e o processo de reabastecimento (transporte, armazenagem) de modo que a otimização total de estoque justificaria um outro estudo.

1.1.2. Objetivos específicos

1. Avaliar a performance de modelos preditivos de aprendizado de máquinas na previsão de consumo de insumos produtivos.
2. Verificar o impacto da utilização desses modelos em métricas de estoque através de simulação de reposição, utilizando a política de estoque atual da companhia (que compreende nível mínimo, máximo e desejado de estoque)

3. Verificar o impacto do redimensionamento de estoque levando em conta o poder preditivo de cada modelo em cada série temporal específico (dimensionar o estoque mínimo como 1 desvio padrão do erro de previsão)

2. REVISÃO BIBLIOGRÁFICA

Nessa seção serão revisados temas pertinentes ao trabalho. Alguns mais genéricos e recentes (com definições relativamente difusas, mesmo na academia) como a ciência de dados e *Analytics*. Para esses tópicos mais abrangentes será dada uma profundidade menor. Já outros tópicos mais consolidados e com relação mais direta com o trabalho, como aprendizado de máquinas, controle automático, gestão de estoque e séries temporais, serão explorados um pouco mais a fundo, ainda que detalhes mais específicos podem ser encontrados nas referências citadas.

2.1. Sistemas de controle automático

De acordo com Morari et al. (1988), controle automático de um processo pode ser definido como a manipulação dos graus de liberdade do processo para que sejam satisfeitos determinados critérios operacionais.

De modo genérico, pode-se resumir a tarefa de um controlador como aquela de garantir que determinados estados de um sistema sejam atingidos e/ou mantidos, manipulando o fluxo de quantidades (matéria, energia, informação...) que interferem no sistema.

Stephanopoulos (1989) elenca cinco requerimentos principais que os sistemas de controle têm de endereçar:

1. Segurança:
 - a. É inegociável, em uma planta química, o bem estar e segurança dos operadores. Desse modo, é imprescindível que temperaturas, pressões, concentrações de componentes e demais parâmetros de processo estejam sempre dentro das faixas permissíveis e seguras.
2. Especificações de produto:
 - a. Uma planta deve atender as especificações de produção, tanto em quantidade quanto em qualidade do produto final
3. Regulamentação ambiental:
 - a. Diversas regulamentações ambientais determinam faixas aceitáveis de temperatura, concentração e quantidade de emissões de efluentes, impondo mais um tipo de restrição nos processos químicos.
4. Restrições operacionais:

- a. Equipamentos devem sempre operar em faixas predeterminadas pelo seu projeto. Por exemplo: tanques devem operar abaixo do volume máximo e também não podem esvaziar.

5. Economia:

- a. A operação e produção deve ocorrer em conformidade com os aspectos econômicos de oferta de matérias primas e demanda de produto acabado, de modo a minimizar os custos.

2.1.1. Controle preditivo

O controle preditivo utiliza-se não apenas dos estados passados do sistema como inputs para o controlador, mas também representações do sistema a ser controlado, de modo a estimar o efeito futuro de uma atuação realizada. Essas representações matemáticas são modelos dinâmicos do sistema com suas incertezas associadas. A importância da contabilização das incertezas do modelo tem se mostrado cada vez mais importantes, como pontuado por Morari et al. (1988).

A experiência demonstrou que o modelo preditivo (MPC) fornece uma metodologia eficaz para lidar com as restrições do sistema durante o projeto e a implementação do controlador (MORARI et al. 1988). Além disso, o MPC não é restrito em termos de qual modelo pode ser utilizado na modelagem do fenômeno, função objetivo, funcionalidade de restrição tornando-o uma opção extremamente flexível no projeto de controladores em suas diversas instancias.

Com o MPC os estados futuros das variáveis manipuladas são determinados por otimização, minimizando alguma função custo associada, dado restrições operacionais.

A otimização é repetida a cada tempo de amostragem, com base em informações atualizadas (medições) do sistema.

2.2. Ciência e análise de dados

Ciência de dados é uma área emergente que ainda não possui uma única definição consensual (SCHUTT, 2013), ainda assim, é possível delinear contornos acerca de algumas características desta (ZHU; XIONG, 2015). Ciência de dados é uma área de conhecimento que intersecciona estatística, habilidades de programação e conhecimento de um domínio específico. Desse modo, torna-se conveniente definir ciência de dados em função do que faz o cientista de dados: ele faz uso de ferramentas matemáticas,

computacionais e estatísticas para poder realizar análises, visualizações e modelos preditivos que geram valor a partir de um conjunto de dados. Essas técnicas envolvem em geral, análises descritivas, preditivas e prescritivas (SHARDA et al. 2013).

De forma bastante abrangente, pode-se dizer que um cientista de dados é responsável por criar modelos matemáticos, em geral puramente empíricos ou semiempíricos visando uma aplicação específica.

2.3. Modelagem matemática

Pode-se definir como modelo matemático de um sistema, uma representação deste através de uma formulação matemática ou algorítmica, de modo que os estados do sistema e suas alterações devem ser representadas no modelo (SODRÉ, 2007).

Um ponto extremamente importante quando se trata de modelagem é entender o tipo de modelo que está sendo criado. Segundo Sodr  (2007), podemos dividir modelos em 3 categorias:

- Fenomenol gicos
- Emp ricos
- Semi emp ricos

Essas categoriza es dizem respeito a forma que o fen meno/processo est  sendo modelado.

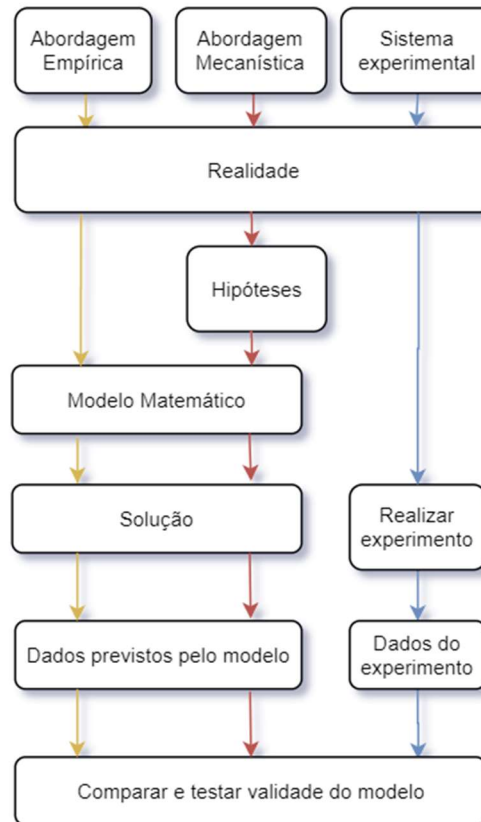
O primeiro (tamb m chamado de te rico ou mecan stico)   criado a partir de leis tidas como universais, e a partir da , as subseqentes deriva es s o feitas (exemplo: Equa es de navier-stokes, utilizando segunda lei de newton e leis de conserva o). Esses modelos podem ser atrelados   ideia de causalidade.

O segundo, por sua vez, leva em conta puramente dados experimentais e aproxima es de fun es desses dados. Em geral   utilizado quando o fen meno em quest o   muito complexo e/ou n o possui leis universais que o descrevem. Essa fam lia de modelos se sustenta na ideia de correla o.

Por último, modelos semi-empíricos são uma espécie de híbrido entre os dois primeiros, em que há uma composição entre leis universais e leis empíricas específicas.

A metodologia para a modelagem matemática de sistemas utilizando as abordagens empírica, mecânica e experimental podem ser sintetizadas nas etapas da Figura 1.

Figura 1: Sistematização de modelos mecânicos, empíricos e validação experimental.



Fonte: Acervo pessoal.

2.4. Correlação vs Causalidade em modelos empíricos

Quando tratamos de modelagens empíricas, é extremamente importante levantar o ponto da dicotomia Correlação-Causalidade.

A causalidade é entendida como uma descrição fenomenológica (perceba a referência circular) da conexão necessária, universal e uniforme entre dois eventos temporalmente

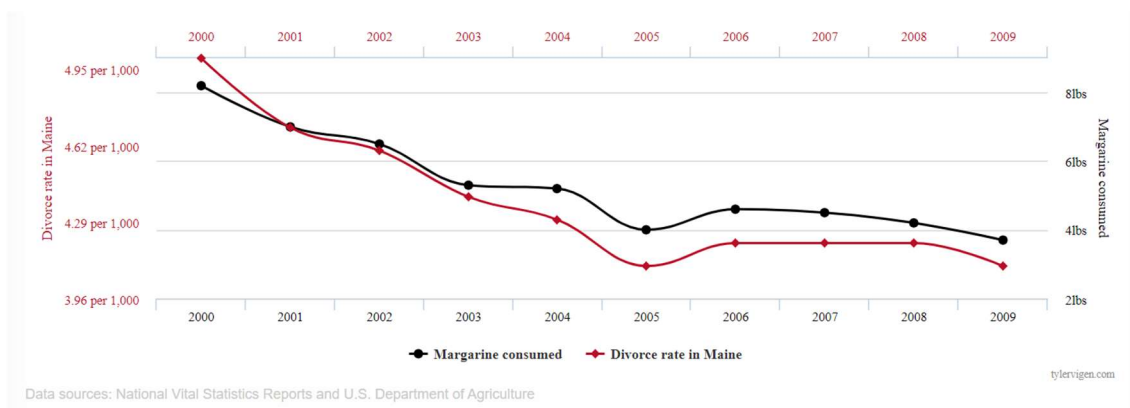
simultâneos ou sucessivos, em que um é entendido como a causa e o outro o efeito. Essas condições permitem a formulação de leis, uma vez que se presume que os dois eventos sejam inevitavelmente sucedidos, um ao outro, em uma certa ordem (SASSOWER, 2017). Todavia, tecer uma rede de causalidades em sistemas complexos, em que eventos são a soma de sinais de diversos outros eventos (e.g. a cadeia de suprimentos de uma indústria de bebidas), a abordagem causal (mecanística, fenomenológica, teórica...) vai se tornando gradativamente inviável.

Por outro lado, a correlação, em contraste com a causalidade, é comumente discutida em termos estatísticos e descreve o grau ou nível de associação entre duas variáveis aleatórias e não relacionadas. Mesmo quando existe uma correlação estatisticamente significativa entre dois eventos (ou conjuntos de eventos) observáveis, o grau em que há de fato uma relação causal entre esses eventos permanece incerto.

Uma implicação muito importante dessas duas definições é que causalidade implica correlação, mas o contrário não é sempre válido, o que torna indispensável em qualquer trabalho que leve em conta a correlação entre variáveis, uma interpretação causal desse resultado, ainda que hipotético, para que seja mais profundamente investigado.

Essa análise causal pode evitar a modelagem através de correlações espúrias, isso é, quando duas variáveis que não possuem relação causal possuem correlação estatística, como ilustrado na Figura 2. Isso pode ocorrer devida à coincidências em função de um tamanho de amostra pequeno, ou devido a confusão ou viés de seleção (HANEUSE, 2016).

Figura 2: Imagem de uma correlação espúria entre consumo de margarina e divórcios no estado do Maine.



Fonte: <https://www.tylervigen.com/spurious-correlations>. Acessado em: 20/09/2020

2.5. Ciência de dados na indústria

Na indústria, em geral, a aplicação de ciência de dados se dá por um subconjunto do espaço total de suas possibilidades, comumente chamado de *Analytics*. pode-se entender *Analytics* como o processo de desenvolvimento de decisões acionáveis ou recomendação de ações com base em informações geradas a partir de dados históricos (POWER et al. 2018). Sua aplicação aparece majoritariamente de três formas diferentes, porém diretamente conectadas em uma cadeia hierárquica.

2.5.1. Ciência de dados descritiva

Segundo Sharda et al. (2013), a Análise descritiva é a mais simples e mais empregada forma de ciência de dados. Ela consiste na aquisição, padronização e análise estatística de bases de dados, gerando relatórios que guiarão a tomada de decisões em uma questão específica. Essa forma de ciência de dados se suporta fortemente em criação de *dashboards* e visualizações de dados em gráficos, medidas estatísticas como média, mediana, desvio padrão, matriz de correlação entre outros. Essa forma ainda exige a inferência de um humano antes da tomada de decisão. De forma resumida, a análise

descritiva é a organização e condensação dos dados em uma estrutura que facilite sua interpretação e guie a interpretação destes por parte de um humano.

2.5.2. Ciência de dados preditiva

A ciência de dados preditiva, por sua vez, pressupõe modelos de aprendizado estatístico (aprendizado de máquinas) que irão utilizar os dados históricos para prever, a partir de determinadas variáveis, estados futuros (SHARDA et al. 2013). Ferramentas comumente usadas estão na família de algoritmos de classificação, regressão e segmentação (*clustering*). Pode-se usar técnicas de classificação, como modelos de árvore de decisão e redes neurais, para prever o desempenho de um rótulo novo. Também podemos usar algoritmos de segmentação para segmentar unidades fabris em função do seu histórico de furos de estoque.

2.5.3. Ciência de dados prescritiva

Uma vez que se tem uma boa descrição de um sistema específico, através da análise preditiva, é possível simular decisões e seus efeitos futuros, viabilizando assim a exploração de um espaço de possibilidades de decisão antes de tomá-las de fato. É possível também utilizar técnicas para otimizar alguma função objetivo que possa ser calculada utilizando o nosso modelo preditivo (SHARDA et al. 2013). Por exemplo: utilizar dados de processo e ensaios sensoriais e químicos de várias cervejas para criar um modelo para prever as notas sensoriais do produto final. Usar então esse modelo para explorar o espaço de possíveis matérias primas e parâmetros de processo que retornam um conjunto de notas sensoriais o mais próximo possível do desejado.

Um resumo de cada uma das formas de análise descritas acima pode ser encontrado na Tabela 1.

Tabela 1: Resumo dos 3 principais tipos de "Analytics".

Análise descritiva	Análise preditiva	Análise prescritiva

Descrever os dados usando estatísticas (média, desvio padrão, etc.) e visualizações gráficas.	Criar modelos preditivos usando dados históricos.	Usar a descrição de um sistema (modelo) para simular resultado de diferentes tomadas de decisão. Otimizar algum objetivo
---	---	--

Fonte: Acervo pessoal.

2.5.4. Aprendizado de máquinas

O aprendizado de máquinas surgiu da área de inteligência artificial, que tem como objetivo tornar máquinas inteligentes, ou reproduzir habilidades da inteligência humana (RÄTSCH, 2014). Hoje há uma grande intersecção entre as áreas compreendidas como ciência de dados e aprendizado de máquinas, uma vez que o aprendizado de máquinas faz uso da teoria de modelagem matemática e estatística.

O Aprendizado de máquina não se preocupa apenas em criar máquinas (modelos matemáticos) inteligentes, mas máquinas que aprendem (inferência indutiva) a abstrair padrões de um conjunto de dados e usar essa abstração para fazer inferência estatística em exemplos não vistos antes.

2.5.5. Tipos de variáveis

Ao modelar um problema, uma tarefa fundamental é identificar as variáveis que são informativas para gerar a previsão desejada e codificá-la de um modo que está seja compreensível pelo modelo matemático. Tendo em vista essa tarefa, é conveniente classificar variáveis como contínuas ou categóricas.

Em geral, os tipos de variáveis que surgem na modelagem com algoritmos de aprendizado de máquinas são as **contínuas** (ou numéricas) e as **categóricas** (ou discretas). As variáveis numéricas podem possuir qualquer valor real ou inteiro e é possível atribuir a seus valores uma relação de ordinalidade (ex.: 23.7 °C é mais quente que -12.3°C, um site com 235 visitas é menos popular que outro com 1276). Tratando-se de variáveis

categóricas, essa relação de ordinalidade não é mais possível. Por exemplo: é possível comparar o tamanho da população, a extensão territorial e o número de copas do mundo vencidas por países como Brasil e Argentina, mas não é possível determinar uma relação de ordinalidade do tipo Brasil > Argentina.

Vale salientar que existem situações limítrofes em que é difícil definir com que tipo de variável estamos lidando, como é o caso, por exemplo, de tamanhos de roupa (P, M ou G) ou meses do ano. Nessas situações, é importante a interpretação da variável no contexto do problema específico, por parte do cientista de dados. A partir daí então define-se como ela deve ser tratada.

Os algoritmos de aprendizado de máquinas seguem uma taxonomia em função do tipo de variáveis que se dispõe como input e o que se espera como output (OLADIPUPO, 2010). As partições mais comuns dessas subdivisões encontram-se abaixo:

2.5.6. Aprendizado supervisionado

O aprendizado supervisionado é talvez o caso mais intuitivo para quem possui um conhecimento consolidado sobre funções matemáticas, pois se dá a partir da criação de uma função que mapeia um vetor de variáveis aleatórias (presunção de independência entre si) a um vetor de variáveis resposta. É desejável que essas funções possuam a capacidade de generalizar bem para problemas diversos, uma vez que sua formulação não é fenomenológica e deve modelar problemas de diferentes naturezas (OLADIPUPO, 2010).

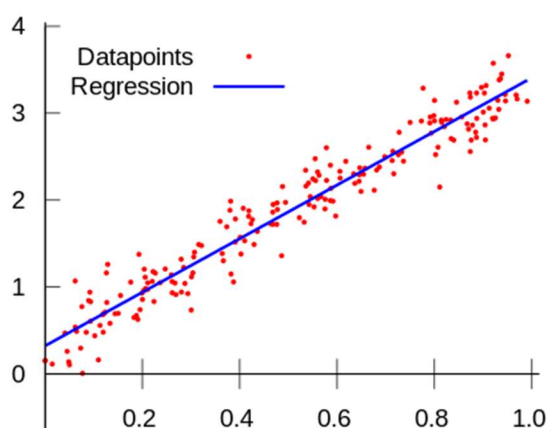
O aprendizado supervisionado se divide em duas grandes famílias: Regressão e Classificação. Em ambos os casos, o determinante é o tipo de variável de saída.

2.5.6.1. Regressão

Regressão é uma modalidade de aprendizado supervisionado em que se mapeia um vetor de variáveis quaisquer (numéricas, categóricas ou mistas) para um vetor de variáveis contínuas. Uma das variedades mais conhecidas de regressão, é a regressão linear através do método dos mínimos quadrados, como ilustrado na Figura 3.

Uma aplicação muito comum é em regressões lineares (JAMES et al. 2017) e análise de séries temporais, mais especificamente em previsão de consumo de energia elétrica (BONETTO; ROSSI, 2017).

Figura 3: Exemplo de regressão linear.



Fonte: https://pt.wikipedia.org/wiki/Regress%C3%A3o_linear. Acessado em: 20/09/2020

2.5.6.2. Classificação

A classificação difere da regressão em um único aspecto: o seu vetor resposta deve ser composto por variáveis categóricas. Em classificação é possível encontrar problemas em que as possíveis "classes" (variáveis resposta) a que pertencem uma leitura podem ser mutuamente excludentes (classificação multi-classe ou binária, dependendo do número de classes) ou não mutuamente excludentes (classificação multi-rótulo).

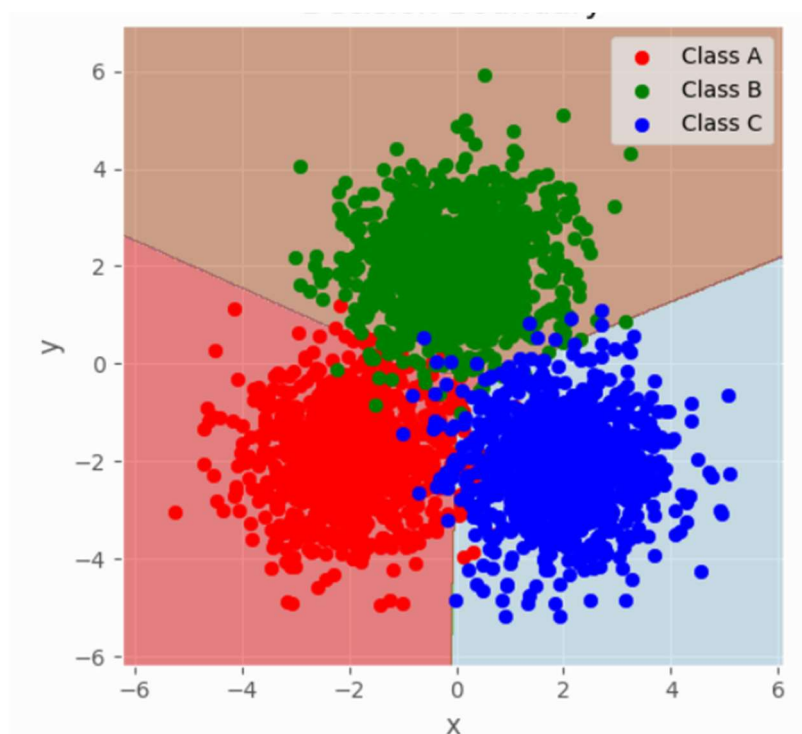
Um exemplo de separação de 3 classes em um espaço de duas dimensões pode ser observado na Figura 4.

Um exemplo de classificação binária é um algoritmo de floresta aleatória (de árvores de decisão) treinado para determinar se um paciente possui ou não uma doença, baseado em sintomas apresentados, idade, sexo, etc (ZRIQAT et al., 2016).

Uma aplicação de classificação multi-classe, por Roth et al. (2015) é a utilização de redes neurais convolucionais para classificação de imagens médica.

Por fim, uma aplicação comum de classificação multi-rótulo é em classificação de imagens de satélite, como visto em Gardner e Nichols (2017).

Figura 4: Classificador multi-classe (3 classes).



Fonte: http://scikit-hep.org/root_numpy/auto_examples/tmva/plot_multiclass.html.
Acessado em: 20/09/2020.

2.5.7. Aprendizado não supervisionado

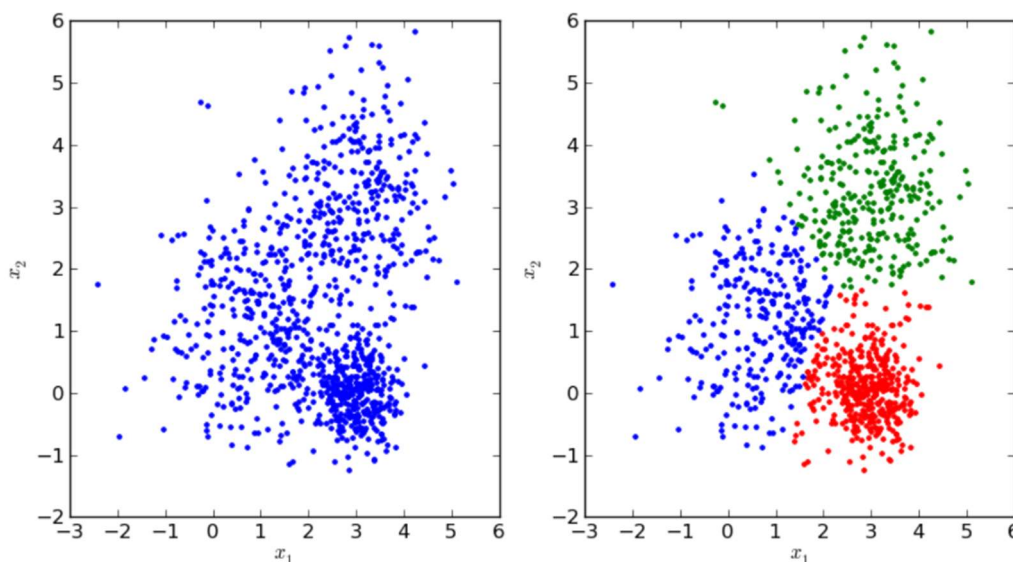
Ao contrário do aprendizado supervisionado, em que são fornecidos os valores das variáveis de entrada e os valores das variáveis resposta, e espera-se que o algoritmo mapeie os inputs aos outputs, sem o fornecimento dos valores históricos de nenhuma variável resposta. Existem duas abordagens principais de aprendizado não supervisionado: aprendizado por reforço e clusterização (DEY, 2016).

No primeiro, um algoritmo toma alguma decisão e atua em um sistema específico. Dependendo do retorno dessa ação, a decisão tomada é "reforçada" ou não. No fim, os

parâmetros do algoritmo são selecionados de modo a otimizar uma função objetivo (por isso por reforço).

Quanto à clusterização, trata-se de uma maneira de estimar, a partir de critérios de similaridade (ou distância), quais entidades podem ser colocadas em um mesmo grupo. Diferentemente de algoritmos de classificação, em que 'mostramos' para o algoritmo a classe a qual pertencem as entidades utilizadas durante o treino, na clusterização essa 'classe' é desconhecida, como pode-se observar na Figura 5.

Figura 5: Clusterização (3 clusters). Note a semelhança entre a classificação multi-classe. A diferença é que os dados não possuem rótulos, diferente do problema de classificação.



Fonte: <https://i.stack.imgur.com/cIDB3.png>. Acessado em: 20/09/2020.

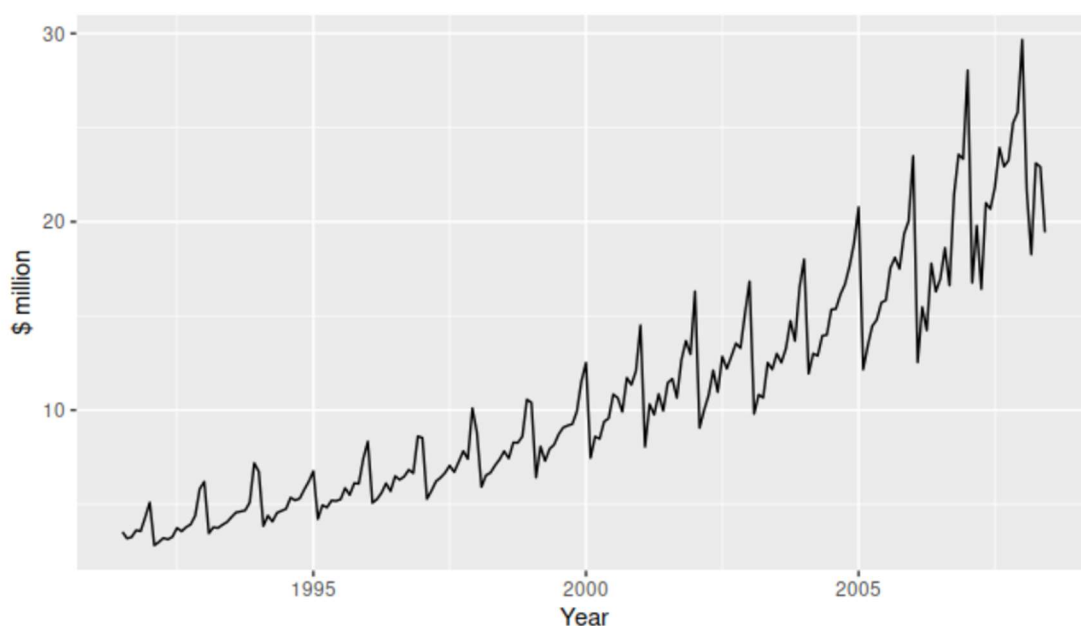
2.6. Series temporais

Series temporais são uma aplicação muito comum de algoritmos de aprendizado de máquinas. Uma série temporal é uma sequência de valores de uma quantidade, ordenados por ocorrência no tempo (ADHIKARI; AGRAWAL, 2013).

Pode-se observar na Figura 6 o histórico de vendas de fármacos para tratamento de diabetes na Austrália, um exemplo de série temporal.

O aprendizado de máquinas pode ser utilizado na modelagem de séries temporais. Problemas de séries temporais podem ser de análise de apenas uma variável (os valores da própria série no passado) ou multivariada (quando séries temporais de mais variável é usada). Abaixo estão detalhados três algoritmos comumente utilizados para realizar a previsão de valores futuros de uma série temporal.

Figura 6: A venda de remédios para diabetes na australia por ano, um exemplo de série temporal.



Fonte: Hyndman & Athanasopoulos (2018).

2.7. Suavização exponencial

Modelos de suavização exponencial utilizam médias móveis de quantidades (tendência, nível, sazonalidade) históricas com pesos que decaem exponencialmente de acordo com a distância dos valores em relação ao valor presente (valores mais distantes do passado tem pesos menores). Uma grande vantagem dessa família de métodos é que eles não exigem que a série seja estacionária, dando a esta uma grande relevância por parte da indústria (HYNDMAN; ATHANASOPOULOS, 2018).

De acordo com Hyndman e Athanasopoulos (2018), uma série temporal estacionária possui propriedades que não dependem do instante em que a série é observada. Assim, as

séries temporais com tendências ou sazonalidade não são estacionárias. A tendência e a sazonalidade afetarão o valor da série temporal em tempos diferentes. Por outro lado, uma série de ruídos brancos (distribuição normal com média zero e sem auto correlação) é estacionária. Não importa quando você a observa, ela deve ter a mesma aparência a qualquer momento.

2.7.1. Modelo Holt-Winters

Segundo Hyndman e Athanasopoulos (2018), o método de Holt-Winters é definido por uma equação de previsão e três equações de suavização, uma para o nível, uma para a tendência e uma para o componente sazonal. Para cada equação há um fator de amortecimento associado. Um último fator (fator de amortecimento de tendência) é definido para a equação de previsão. É necessário também definir m , que é a frequência da sazonalidade ($m=3$ para dados trimestrais, $m = 12$ para dados anuais).

Existem duas ramificações desse método. Essa separação se dá em função da natureza do componente sazonal. No modelo de sazonalidade aditiva a componente sazonal é representada em termos absolutos das séries temporais e na equação de nível. A série é ajustada com a subtração do componente sazonal, por isso, o modelo de sazonalidade aditiva é preferível quando as oscilações sazonais são aproximadamente constantes. Já o método multiplicativo, expressa-se o componente sazonal em termos relativos (proporções) e o ajuste sazonal é feito dividindo o nível pelo componente sazonal.

A equação do modelo sazonal de Holt-Winters é dada por (eq. de 1.1 a 1.4):

$$\hat{y}_{t+h|t} = \ell_{t_h} b_t + s_{t+h-m(k+1)} \quad (1.1)$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1}) \quad (1.2)$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1} \quad (1.3)$$

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m} \quad (1.4)$$

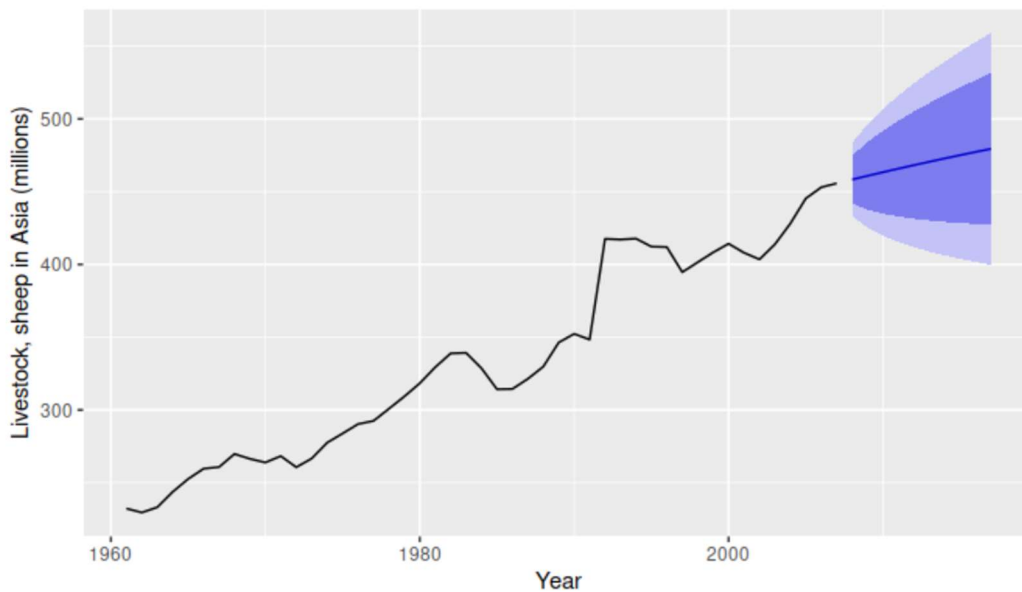
em que $\hat{y}_{t+h|t}$ é a previsão h passos a frente de t, ℓ_t é o nível (média) suavizado em t, b_t é a tendência suavizada no tempo t e s_t é o fator sazonal. Todos os parâmetros devem estar no intervalo (0,1).

Os parâmetros da equação, juntamente com seus valores iniciais (necessários para completar os graus de liberdade do sistema) são estimados de modo a minimizar o erro quadrado médio entre as previsões e os valores reais da série temporal:

$$MSE = \frac{1}{n} \sum_{t_0}^{t_f} (\hat{y}_{t+h|t} - y_{t+h})^2 \quad (1.5)$$

De modo que geramos previsões como a da Figura 7.

Figura 7: Aplicação do método de Holt-Winters com efeito de nível e tendência.

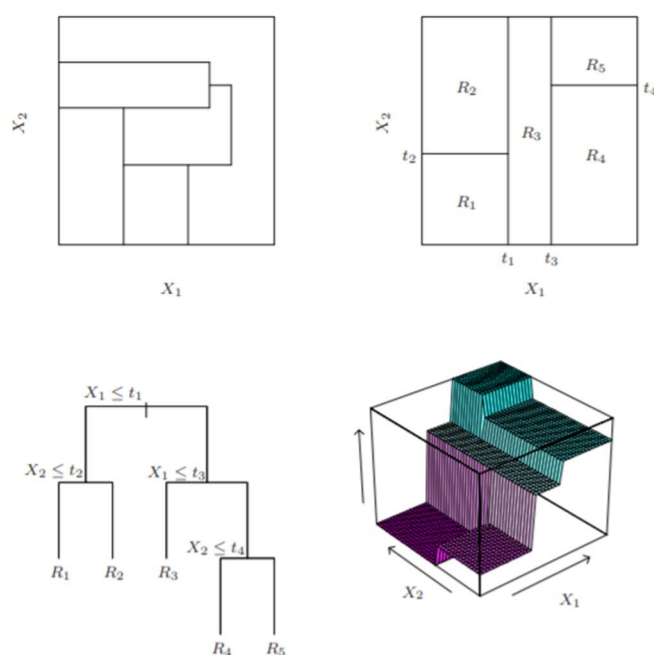


Fonte: Hyndman & Athanasopoulos (2018)

2.8. Árvores de decisão

Algoritmos de árvores de decisão dividem o espaço dos preditores (variáveis usadas como entrada dos modelos) em um conjunto de retângulos e depois ajusta um modelo simples (como uma $f = \text{constante}$ dentro daquele retângulo). Ainda que conceitualmente simples, esses modelos são extremamente poderosos. O processo de otimização consiste em achar as partições que minimizem alguma função custo (HASTIE et al, 2008). Uma representação esquemática dessas partições pode ser encontrada na Figura 8.

Figura 8: Representação gráfica de um modelo de árvore de decisão com dois estimadores (x_1 e x_2).



Fonte: Hastie et al. (2008)

2.8.1. Boosting

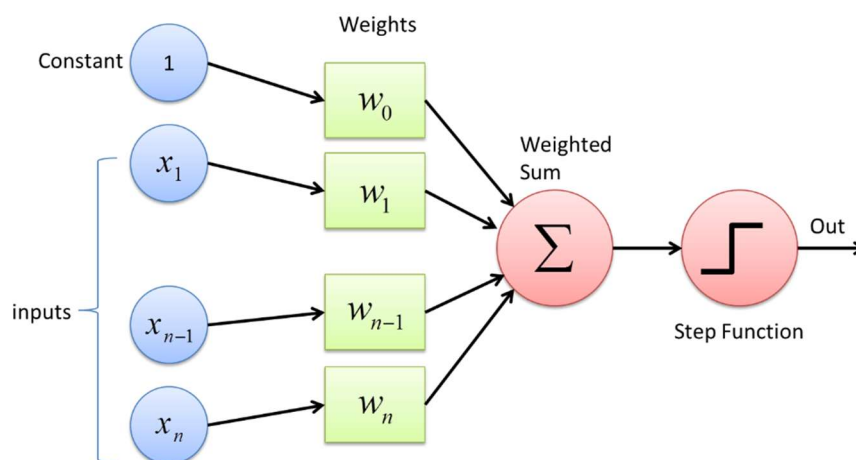
O algoritmo de boosting é um modelo aditivo de árvores de decisão. A ideia é que sejam criadas diversas árvores simples (com um número relativamente baixo de partições) e que o resultado final do modelo seja uma média ponderada do resultado de cada um desses modelos simples. De acordo com Hastie et al. (2008), Esse tipo de estratégia previne o problema de sobre ajuste encontrado em modelos simples de árvores de decisão com muitas ramificações (alta variância), uma vez que são criadas várias árvores com poucas ramificações (alto viés).

2.9. Redes Neurais Artificiais

Redes neurais artificiais são modelos capazes de representar padrões não lineares em um conjunto de dados. Trata-se de modelos inspirados no processamento de informação de neurônios. A analogia que é feita é de que os neurônios recebem informação de outros neurônios, integram essa informação, aplicam a ela uma função de ativação e passam a informação para os demais neurônios (LECUN et al. 2015)_como observado na Figura 9.

Os parâmetros aprendidos pelo modelo são os pesos da conexão entre os neurônios. Esses neurônios podem se agrupar em blocos e/ou camadas. Arquiteturas mais complexas possuem também células especializadas, como as redes neurais convolucionais e as rede neurais recorrentes (LECUN et al. 2015). A idealização das redes neurais e seu bloco fundamental - o perceptron - não é recente e remonta a Rosenblatt (1958), contudo o crescimento de aplicações utilizando estes algoritmos tem se dado pelo grande volume de dados produzidos (necessários para que o modelo generalize bem o tipo de problema), a grande capacidade atual de processamento, principalmente em unidades de computação gráfica (GPUs) e o desenvolvimento de algoritmos de otimização mais robustos. Essa grande capacidade computacional é necessária para que os parâmetros do modelo (por vezes milhões) sejam aprendidos através de uma otimização da função custo da rede, utilizando algoritmos de gradiente descendente estocásticos (LECUN et al. 2015).

Figura 9: Esquematização do perceptron.

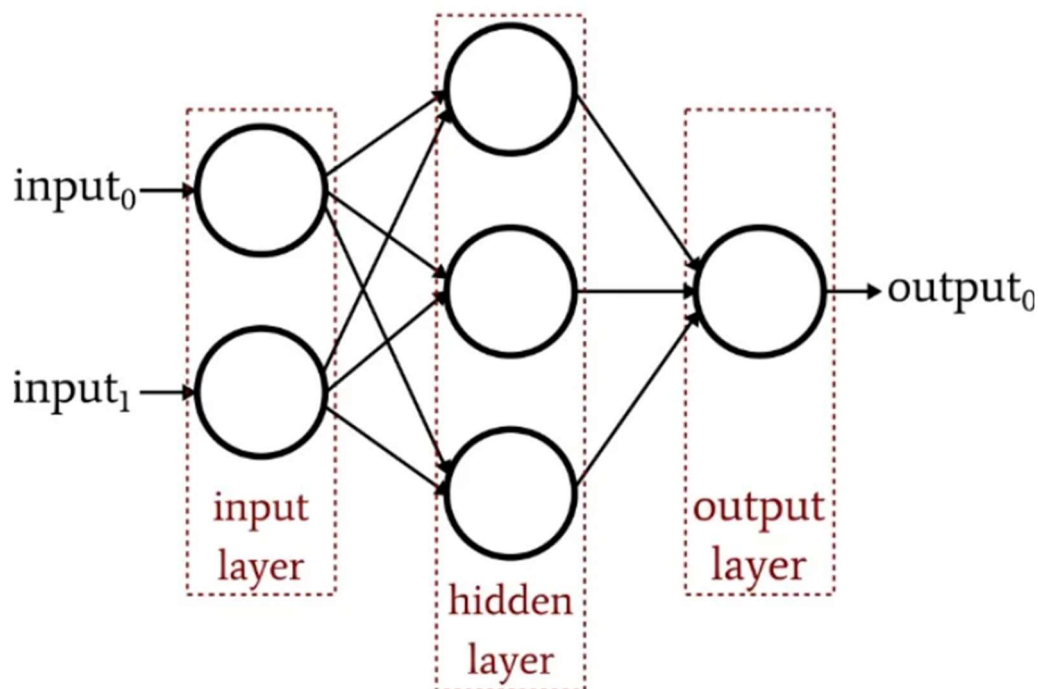


Fonte: <https://www.gratispng.com/png-rev2hl/>

2.9.1. Redes Neurais Densas

As redes neurais densas (ou perceptrons em multicamadas) são a arquitetura mais comuns para as redes neurais. Nela, os neurônios (perceptrons) se agrupam em camadas e essas camadas se conectam com as camadas seguintes (HASTIE et al. 2008) como pode ser visto na Figura 10. Em geral, o número de camadas e o número de neurônios em cada camada são hiperparâmetros (um parâmetro do modelo que não é aprendido durante o processo de treino) definidos pelo usuário. Outro hiperparâmetro importante é a função de ativação utilizada nos neurônios. É importante ressaltar que Neurônios de uma mesma camada não se conectam entre si e conexões entre neurônios não possuem loops de feedback (o que aumentaria a complexidade do processo de otimização) uma vez que nesta configuração de RN, a passagem de informação é sempre para a camada seguinte. A última camada da rede é a camada de saída, e seu número de neurônios, bem como sua função de ativação, são escolhidas levando em conta o tipo de problema que se quer tratar (classificação ou regressão).

Figura 10: Propagação e processamento de informação em uma Rede de perceptrons em multicamadas.



Fonte: <https://www.allaboutcircuits.com/technical-articles/how-to-train-a-multilayer-perceptron-neural-network/> . Acesso em 20/09/2020.

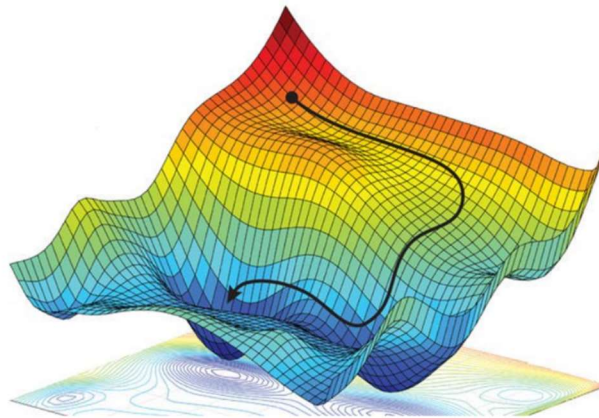
2.9.2. Gradiente Descendente Estocástico

Algoritmos de gradiente descendente estocástico consistem em utilizar vetores de input contendo alguns exemplos (batch) (não todos de uma vez, por isso estocástico), calcular o valor de saída dados esses inputs, calcular os erros, os seus gradientes (em relação aos pesos) e atualizar o valor dos pesos em conforme (HASTIE et al. 2008). Isso é repetido até que todos os *batches* (um subconjunto amostrado aleatoriamente do conjunto total e dados) do conjunto de treino são passados pelo modelo, completando uma *epoch* (uma rodada de treino em que todos os *batches* do conjunto de dados são alimentados pela rede). Esse processo se repete de forma iterativa, até que a média da função custo do subconjunto de validação pare de diminuir.

A componente estocástica da família de algoritmos se deve ao fato de que cada 'batch' fornece uma estimativa ruidosa do gradiente médio em todos os exemplos. Esse procedimento simples geralmente encontra um bom conjunto de pesos de forma rápida quando comparado com técnicas de otimização muito mais elaboradas. Além disso, o ruído gerado pela componente estocástica, em conjunto com uma escolha adequada do parâmetro de taxa de aprendizado (learning rate, representado como η), pode ajudar o otimizador a sair de pontos de sela e mínimos locais em superfícies não convexas (BOTTOU, 1991), como a da Figura 11.

Passada a etapa de treino, o desempenho do modelo é medido em um conjunto diferente do que foi utilizado no treinamento. Assim, é possível testar a capacidade de generalização do sistema, isso é, sua capacidade de produzir respostas coerentes para entradas as quais ele não foi exposto durante a fase de treinamento.

Figura 11: Exemplo de minimização com gradiente descendente estocástico em uma superfície não convexa.



Fonte: <https://towardsdatascience.com/coding-deep-learning-for-beginners-linear-regression-gradient-descent-fcd5e0fc077d>. Acessado em: 20/08/2020.

2.10. Gestão de estoque

Planejamento e controle de estoque de insumos é uma tarefa vital para uma empresa. Altos níveis de estoque significam custo de capital imobilizado, enquanto baixo nível de estoque pode criar o risco de falta de insumos, o que pode ter um custo ainda maior. Tendo isso em vista, é fundamental que o planejamento e controle de estoque se dê de maneira a mitigar os riscos associados à falta de estoque sem que isso gere um grande custo de capital imobilizado (MULLER, 2003).

Soma-se a complexidade do planejamento e controle de estoque o fato de que insumos podem ter tempos de reabastecimento (“leadtime”) muito altos, tornando necessária uma forma de previsão de consumo em um horizonte próximo a esse “leadtime”.

Além disso, é também importante minimizar o custo associado à compra de insumos, que não consiste apenas no valor do material e do transporte, mas também despesas associadas ao reabastecimento e despesas relacionadas à equipe de compras.

Essas componentes do problema de gestão de estoque citadas torna inevitável a comparação de um sistema de gestão de estoque com um problema de controle preditivo (MPC), como proposto por Hai et al. (2011), diferindo apenas no fato de que não haverá

uma função objetivo que será otimizada nem um conjunto de restrições explícitas as serem respeitadas (ainda que estejam implícitas na política de estoque).

Desse modo, Muller (2003) pontua como fundamental a capacidade de prever o consumo futuro com exatidão e faixas de confiabilidade bem definidas (importantes para análise de risco) como crucial para um bom planejamento e controle de estoque, além de necessária para um planejamento de estoque dependente de demanda.

2.10.1. Aprendizado de máquinas e previsão de estoque

Uma aplicação natural de algoritmos de previsão de séries temporais é no planejamento e gestão de estoque, dada uma previsão de demanda.

Bandara et al. (2019) utilizou redes neurais recorrentes para prever demanda de uma vastidão de produtos diferentes na área de e-commerce.

Abolghasemi et al. (2019) verificou, utilizando vários modelos estatísticos (incluindo modelos de suavização exponencial) e de aprendizado de máquinas, o impacto das promoções na volatilidade de demanda de itens em uma cadeia de suprimentos.

Singh et al. (2019) modelou a demanda por itens novos (sem histórico de compras) no mercado de vestimentas, utilizando variáveis latentes dos novos itens, para que fosse possível utilizar informações de itens que possuíam histórico, dado um critério de similaridade entre os conjuntos de dados.

Esses trabalhos nos mostram a versatilidade de tais modelos e sua aplicabilidade em diversos contextos relacionados à previsão de demanda e gestão de estoque. Ainda assim, não foi encontrado nenhum trabalho que realizasse um estudo de caso específico ao controle de estoque na indústria de bebidas.

3. MATERIAIS E MÉTODOS

3.1. Descrição do problema

O objeto de estudo desse trabalho é a previsão e simulação de estoque em uma companhia da indústria de bebidas, que possui unidades por todo o Brasil e possui um portfólio extremamente diversificado.

Antes de modelar o problema, é importante possuir uma noção macro dos processos relevantes neste trabalho: Produção, planejamento de demanda e gestão de estoque.

3.1.1. Planejamento de demanda atual

Para adereçar o problema de gestão de estoque, a empresa possui processos internos de planejamento de demanda.

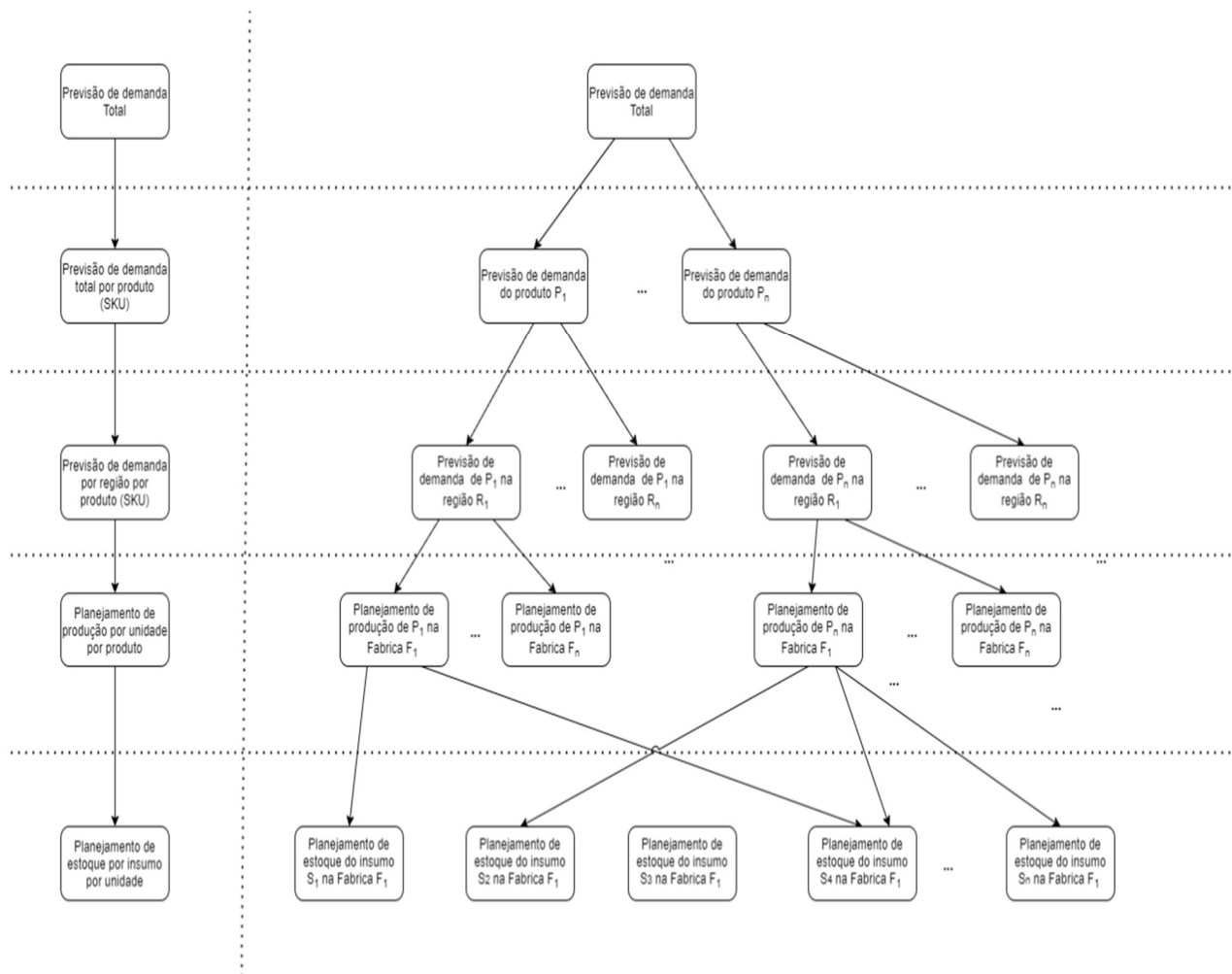
O processo de previsão se dá primeiramente de modo agregado e vai sendo recursivamente refinado, até o nível de demanda por produto específico por região econômica.

O próximo passo então é definir as metas produtivas de todas as fabricas, de modo a otimizar os custos de aquisição de matéria prima, produção, distribuição e até aspectos fiscais, levando em conta restrições como a capacidade produtiva máxima das fábricas.

Por fim, cada produto possui uma receita, que leva em conta a quantidade específica de cada insumo para cada unidade de produto acabado. Desse modo, é possível estimar a demanda e planejar o estoque de um insumo específico dentro de cada fábrica. Vale ressaltar que por produto acabado (SKU), nos referimos ao produto do modo que será obtido pelo consumidor, o que não leva em conta apenas o conteúdo (líquido), mas também a embalagem e os demais aspectos relacionados ao produto acabado como um todo.

O processo como um todo pode ser descrito através do fluxograma da Figura 12:

Figura 12: Processo atual de previsão de consumo de insumos



Fonte: Acervo pessoal.

É importante observar no fluxograma que pode haver cruzamento de insumos, isso é, o mesmo insumo é utilizado para produzir produtos (SKUs) diferentes. Alguns exemplos desses insumos são:

- Produtos químicos:
 - Agentes reguladores de pH, produtos de limpeza para os tanques. Esses insumos são utilizados independentemente do produto que será produzido.
- Malte:
 - Algumas receitas levam o mesmo tipo de malte, ainda que concentrações diferentes. Além disso, dois produtos (SKU) podem conter o mesmo líquido (“Cerveja X” em lata e “Cerveja X” em garrafa).
- Lúpulo:
 - Mesmos motivos citados acima para o malte.

- Garrafa:
 - Algumas garrafas são comuns em produtos diferentes, uma vez que o que diferencia as marcas são os filmes (rótulos).

Essa abordagem de refinamento de previsões de um nível mais agregado (volume total) para um nível mais granular (volume de SKU por região) é conhecida como método de previsão “top-down”. Ainda que muito utilizado na indústria, Hyndman et al. (2011) provou que é impossível gerar uma previsão não enviesada usando o método “top-down”.

Assim, os métodos de previsão proposto nesse trabalho são feitos unicamente no nível de granularidade Insumo/Unidade, ao invés de prever a demanda agregada e cascatear os resultados para um nível de resolução mais baixo.

3.1.2. Gestão de estoque

Como já pontuado, a gestão de estoque é feita de modo otimizar os custos de estoque (capital imobilizado, risco de falta de insumos, risco de vencimento...). Esse processo de otimização pode ser extremamente complexo, dada a natureza diversa dos insumos presentes no processo de produção e o grande número de restrições e parametrizações necessárias para definir o problema.

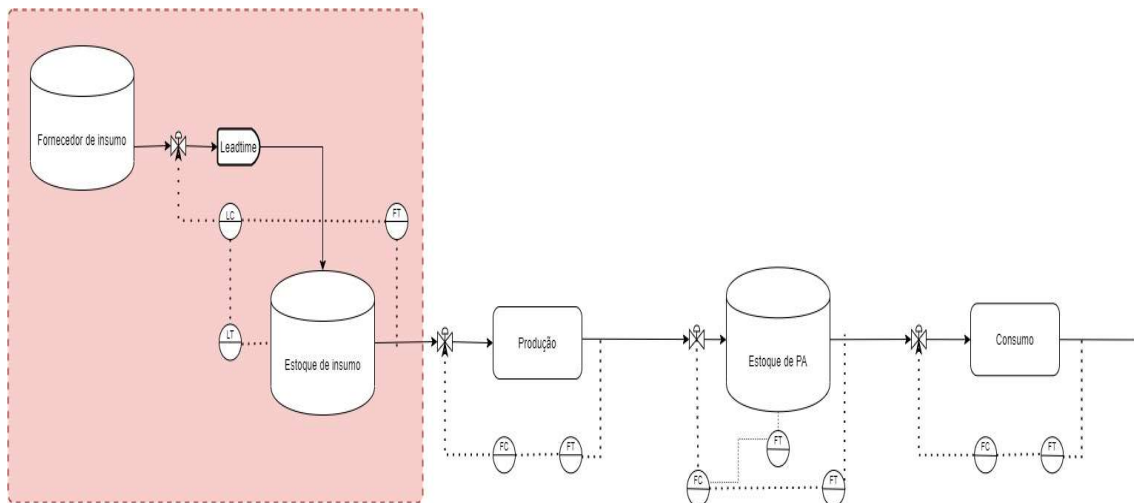
Dito isso, a empresa em que esse trabalho foi realizado possui diretrizes (heurísticas) que guiam a gestão de estoque de insumos específicos. Essa política de estoque define um estoque mínimo e um máximo, de modo que o nível de estoque deve se manter em um ponto intermediário entre os dois extremos. Além disso, a política de reposição é tal que toda reposição deve elevar o nível de estoque até o ponto intermediário entre o estoque máximo e mínimo

Uma vez que a política de estoque é dada, o escopo desse trabalho se restringe a sugerir um novo limite inferior de estoque (estoque mínimo) e um novo modelo de previsão de consumo.

3.2. Modelagem do problema

O problema de controle de estoque pode ser representado como um sistema de tanques em série, onde os fluxos entre esses tanques possuem tempos de atraso específicos e são controlados por válvulas, de acordo com a Figura 13:

Figura 13: Analogia de estoques do processo produtivo com um sistema de tanques



Fonte: Acervo pessoal.

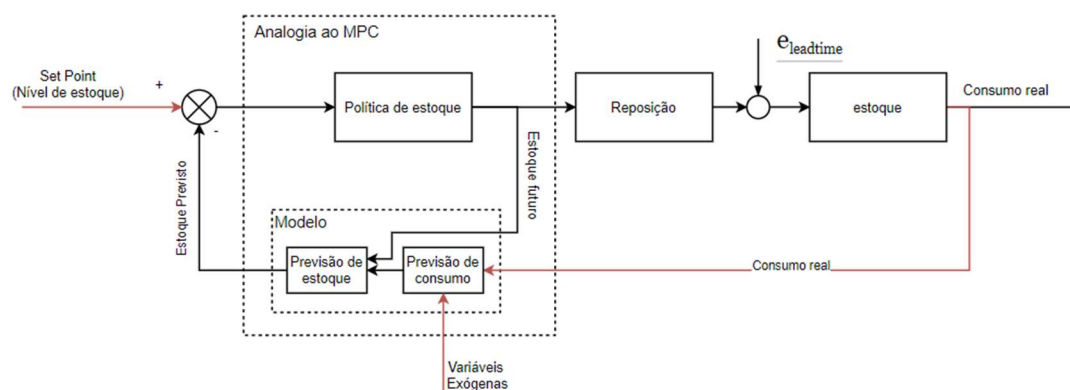
O diagrama acima representa um sistema “puxado”, em que o controle de fluxo de saída de uma etapa anterior do processo é sempre regido pelo fluxo posterior. Nesse caso:

- O consumo de estoque de produto acabado é puxado pelo consumo dos clientes finais;
- A produção é puxada, ou seja, é determinada pelo nível de estoque de produto acabado;
- O consumo de estoque de insumos é puxado pela produção;
- O fornecimento de insumos é determinado pela previsão do nível de estoque futuro (uma vez que existe um atraso entre o sinal de reposição e a reposição).

A área vermelha na Figura 13 representa os subprocessos que são objeto do estudo atual, em especial o controlador LC.

Essa área hachurada pode ser representada em um diagrama de blocos de acordo com a Figura 14:

Figura 14: Diagrama de blocos do controle de reposição



Fonte: acervo pessoal.

Nesse diagrama, pode-se perceber que o nível de estoque é controlado pelo sinal da política de estoque para o processo de reposição. A política de estoque é alimentada com o set point desejado de nível de estoque, restrições, e a previsão de estoque futuro (horizonte do “leadtime”) do modelo.

Além do tempo de reposição padrão, existe uma perturbação ($e_{leadtime}$) que interfere no tempo de reposição real (esse efeito não será modelado no presente trabalho).

Assim, para definirmos a reposição de estoque através do sistema projetado, é necessário conhecermos três inputs:

- O nível de estoque desejado (Set point);
- O histórico de consumo do insumo para alimentar o modelo preditivo;
- As variáveis exógenas do modelo (outras variáveis importantes além do histórico de consumo).

É necessário também que conheçamos a política de estoque para cada insumo (estoque máximo e mínimo).

As ferramentas utilizadas para implementar as etapas da modelagem estão disponíveis em código aberto para a linguagem python. A Tabela 2 mostra um esquema dessas ferramentas e a sua etapa de utilização.

Tabela 2: Bibliotecas utilizadas e suas referências.

Etapa	Biblioteca	Referência
Pré-processamento	Pandas	(MCKINNEY, 2011)

Treino de modelos	Statsmodels Fast.ai LightGBM	(SEABOLD; PERKTOLD, 2010) (HOWARD; GUGGER, 2020) (KE et al. 2017)
Visualização	Seaborn Matplotlib	https://github.com/mwaskom/seaborn https://github.com/matplotlib/matplotlib
Interface de programação	Jupyter Notebooks	https://github.com/jupyter/notebook
Simulação	Código próprio	-

Fonte: Acervo pessoal.

3.3. Preparação dos dados

A Base de dados utilizada possui, para cada combinação insumo-fábrica, o histórico dos últimos dois anos (aproximadamente) de consumo de diversos insumos produtivos, os seus respectivos “leadtimes” teóricos, e as previsões de consumo para os próximos 75 dias utilizando o modelo de previsão descrito na seção 3.1.1.

Uma amostra dos dados pode ser vista na Tabela 3:

Tabela 3: Exemplo da estrutura dos dados históricos com valores fictícios.

Data	Fabrica	Cod.Insumo	“leadtime”	Consumo D0	Previsão D1	Previsão D2	...	Previsão D75
01/07/2018	F1	INSUMO1	15	450	475	480	...	433
01/07/2018	F2	INSUMO1	20	0	0	0	...	0
02/07/2018	F1	INSUMO1	15	-	414	398	...	-
02/07/2018	F3	INSUMO2	7	15	15	15	...	14
02/07/2018	F2	INSUMO3	60	1850	-	1880	...	1741
02/07/2018	F2	INSUMO3	60	1850	-	1880	...	1741

Fonte: Acervo Pessoal

É possível identificar que existem valores faltantes e colunas duplicadas, além do fato de que os dados não estão no formato necessário para utilizarmos como input para nossos modelos

As etapas necessárias para pré-processamento dos dados independentemente do modelo:

- Preenchimento de valores faltantes (interpolação);
- Remoção de valores duplicados;
- Redução da resolução de horizonte de previsão (de dias para cumulativo da semana);
- Separação em conjunto de treino e teste (18 meses para treino, 6 meses para teste).

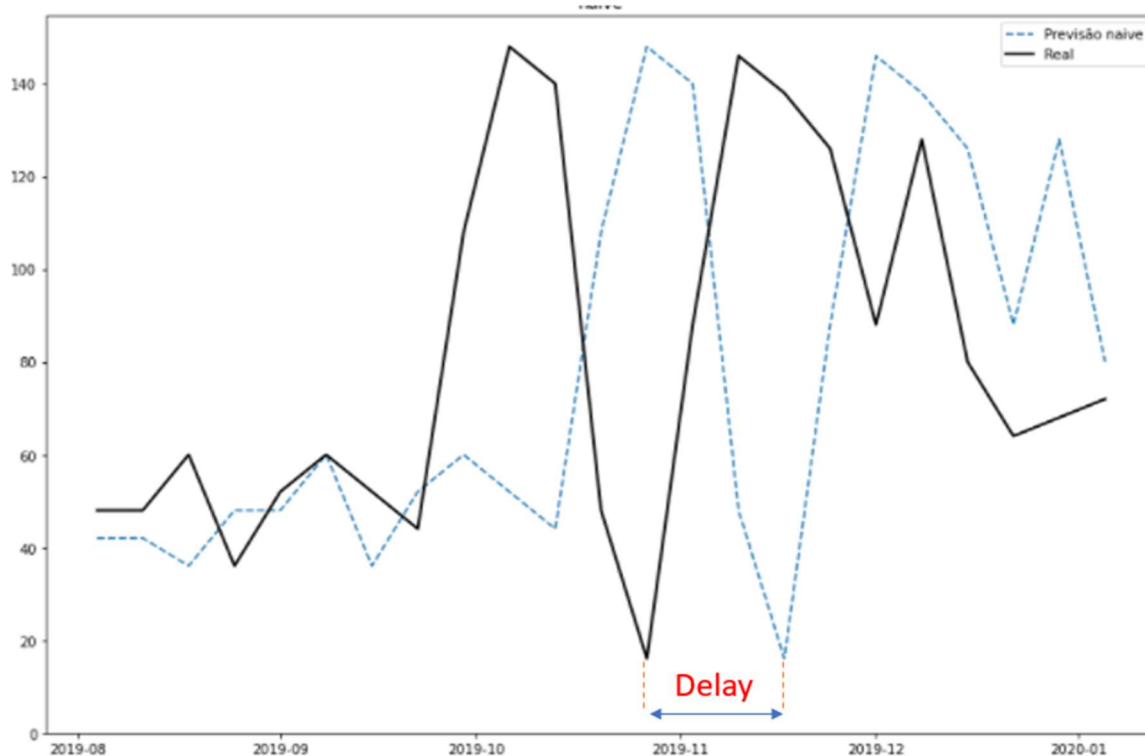
Além dessas três etapas, existe um pré-processamento específico para o modelo Holt-Winters e outro para os modelos de Redes Neurais e Gradient Boosting.

3.3.1. Pré-processamento Modelo baseline (*Naive*)

Segundo Hyndman e Athanasopoulos (2018), é recomendável que ao validar o resultado de previsões, utilize-se um modelo como base de comparação (ou baseline). Esse modelo em geral é um modelo muito simples de ser implementado e que seu resultado seja interpretável.

Para esse trabalho, o modelo baseline escolhido foi o modelo de persistência (ou simplesmente *Naive*), isso é, os valores do passado persistem no futuro, porém atrasados por um período igual ao horizonte de predição, como pode ser visto na Figura 15.

A preparação dos dados é relativamente simples e consiste em usar em conjunto os métodos “groupby()” e “shift()” da biblioteca Pandas (MCKINNEY, 2011) para processamento de dados.

Figura 15: Exemplo de previsão *Naive*.

Fonte: acervo pessoal.

3.3.2. Pré-processamento Holt-Winters

Para o modelo Holt-Winters, por se tratar de um modelo que leva em conta apenas variáveis endógenas (histórico de consumo apenas), as etapas de pré-processamento restante são apenas as genéricas citadas acima.

O formato do input requerido pelo modelo de Holt-Winters é ilustrado na Tabela 4:

Tabela 4: Exemplo fictício de estrutura de input para modelo Holt-Winters.

Fabrica	Cod.Insumo	Data	Consumo S0
F1	INSUMO_1	01/07/2018	2747.551724
		08/07/2018	2615.297796
		15/07/2018	2742.969918
	...		
	INSUMO_N	01/07/2018	2693.096624
08/07/2018		2559.405784	
15/07/2018		2538.632578	

...

FN		01/07/2018	2444.436988
	INSUMO_1	08/07/2018	2831.069158
		15/07/2018	2737.266002
	...		
		01/07/2018	2532.931794
	INSUMO_N	08/07/2018	2515.079385
		15/07/2018	2980.890101

Fonte: Acervo pessoal.

Para cada série temporal (combinação entre fábrica e insumo) será gerado um modelo, isso é, valores para α , β^* e γ que minimizam o erro entre as previsões feitas pelo modelo e os valores reais da série temporal.

3.3.3. Gradient Boosting e Redes Neurais

Esses dois modelos utilizarão tanto variáveis endógenas (histórico recente de consumo) quanto exógenas (mês, derivada do consumo), da seguinte forma:

- Criação de variáveis categóricas:
 - Mês;
 - Centro;
 - Código do Insumo;
- Criação de variáveis numéricas:
 - Derivada do consumo ($c_i - c_{i-1}$);
- Criar o vetor passado (vetor contendo o consumo das últimas 11 semanas);
- Criar o vetor futuro (vetor contendo o consumo das próximas 11 semanas);
- Estandarização das variáveis e targets numéricos.

O formato do input e do output (target) para esses modelos podem ser vistos nas Tabelas 5 e 6, respectivamente

Tabela 5: Estrutura de input para modelos multivariados.

Mês	Fabrica	Cod.Insumo	Consumo S-11	...	Consumo S-1	Consumo S0	Derivada S-11	...	Derivada S0
7	F1	INSUMO1	2589		2856	2690	-268		-167
7	F2	INSUMO1	695		625	611	69		-14
7	F1	INSUMO1	1349		1179	1260	170		81
8	F3	INSUMO2	291		339	304	-49		-36
8	F2	INSUMO3	3781		4002	3451	-222		-551
8	F1	INSUMO1	201		184	173	17		-11
9	F7	INSUMO5	17		19	20	-2		1

Fonte: Acervo pessoal.

Tabela 6: Estrutura de output para modelos multivariados.

Consumo S+1	Consumo ... S+11
2435	2639
617	676
1306	1220
321	322
4166	3462
182	190
18	18

Fonte: Acervo pessoal.

É importante ressaltar que para esses dois modelos, o tipo de problema resolvido pelo algoritmo é uma regressão multi-output (o target possui 11 dimensões).

3.4. Treinamento dos modelos

Nessa sessão serão explicados alguns detalhes das implementações dos modelos e os formatos de input e output.

Para que os modelos sejam validados da mesma maneira, todos os outputs, independentemente do modelo, devem possuir o mesmo formato da Tabela 7.

Tabela 7: Estrutura padrão de output de todos os modelos para validação.

Data	Fabrica	Cod.Insumo	“leadtime”	Previsão S1	...	Previsão S11
01/07/2018	F1	INSUMO1	15	2475	...	27105
01/07/2018	F1	INSUMO1	20	2430	...	32462
02/07/2018	F1	INSUMO1	15	2978	...	31643
02/07/2018	F1	INSUMO1	7	2461	...	32842
02/07/2018	F1	INSUMO1	60	2713	...	31028
03/07/2018	F1	INSUMO1	15	2946	...	32962

Fonte: Acervo pessoal.

3.4.1. Holt-Winters:

A implementação usada pelo modelo Holt – Winters foi a disponível na biblioteca de Python Statsmodels (SEABOLD; PERKTOLD, 2010), com algumas adaptações.

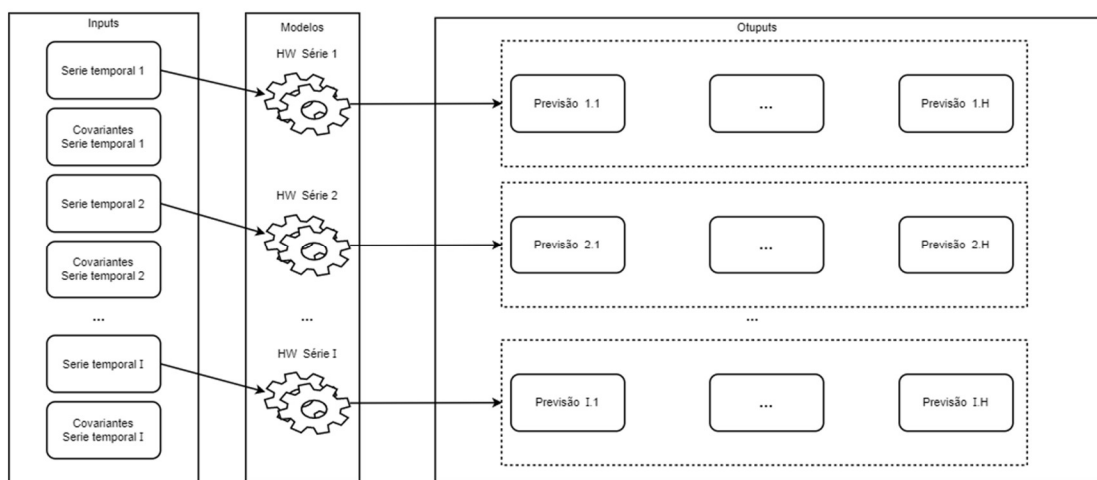
A primeira delas é que foi criada uma extensão da classe *tsa.holtwinters.ExponentialSmoothing* chamada *GroupWiseHoltWinters*, para suportar o input de séries temporais distintas (uma para cada combinação Fabrica-Insumo) e fazer internamente a separação dessas séries em séries individuais. Além disso, a classe *GroupWiseHoltWinters* possui um atributo que é um dicionário contendo uma instância da classe original para cada combinação Fabrica-Insumo. Assim, uma vez criada uma instância de *GroupWiseHoltWinters*, é possível treinar os modelos para cada série através do método *fit*, em que é passado como input os dados históricos e alguns hiper parâmetros do modelo (como método de inicialização de parâmetros, otimizador, etc...).

Após treinamento do modelo, é possível acessar as previsões realizadas para cada série temporal, bem como incluir novas leituras de input para gerar previsões levando em conta um histórico atualizado.

A previsão com o modelo Holt-Winters pode ser realizada para o horizonte de semanas que for desejado pelo usuário. Entretanto, utilizaremos o horizonte de 11 semanas para que seja possível alinhar os outputs do modelo Holt-Winters com os modelos Gradient Boosting (LGBM) e Rede Neural.

A representação esquemática dessa modelagem pode ser vista na Figura 16.

Figura 16: Arquitetura da modelagem Holt-Winters.



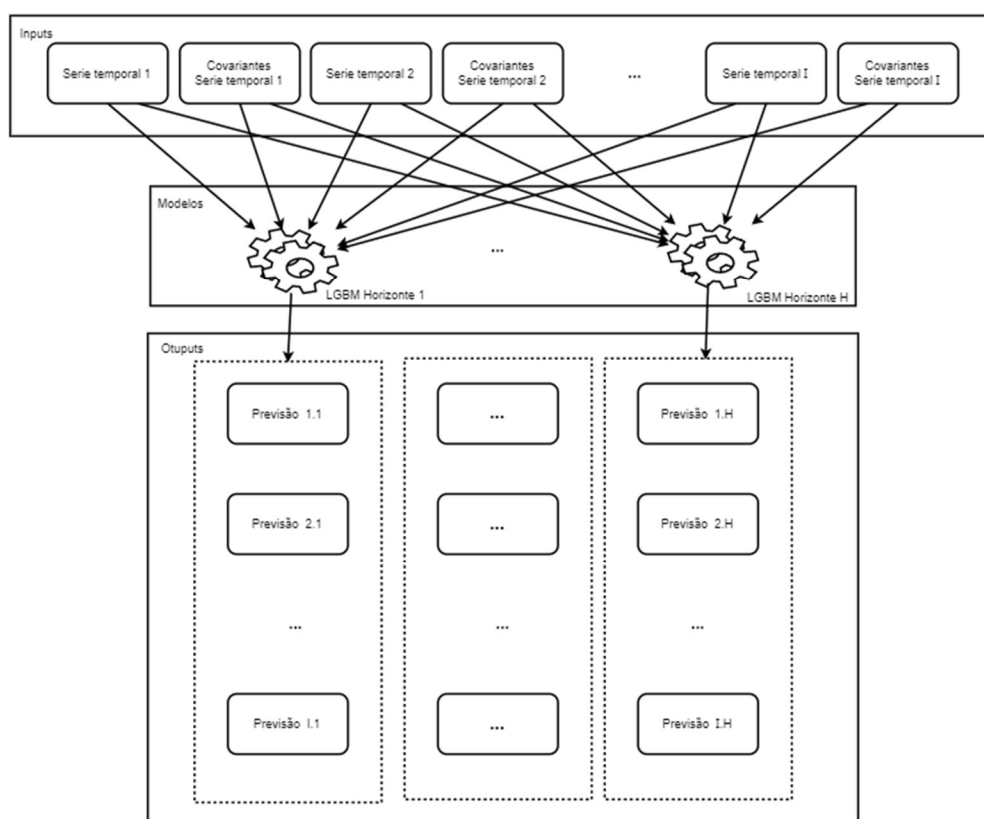
Fonte: acervo pessoal.

É importante notar que existe um modelo para cada série temporal e que esses modelos não utilizam variáveis exógenas na previsão

3.4.2. Gradient Boosting:

A implementação utilizada desse modelo foi da biblioteca de python LightGBM (KE et al. 2017) da Microsoft. Como o módulo de regressão dessa biblioteca suporta apenas outputs unidimensionais, foi necessária utilizar a classe MultiOutputRegressor do scikit-learn (BUTINCK et al, 2013) para treinar múltiplos regressores single-output com o mesmo input, um para cada horizonte de previsão, totalizando 11 modelos (e por consequência uma função objetivo por modelo), como mostrado na figura 17.

Figura 17: Arquitetura da modelagem LGBM.



Fonte: acervo pessoal.

É importante notar que existe um modelo para cada horizonte de previsão, mas cada modelo recebe como input informações de todas as séries temporais, além de covariantes. Isso permite que o modelo consiga identificar certos padrões globais compartilhados entre diferentes séries temporais (caso eles existam).

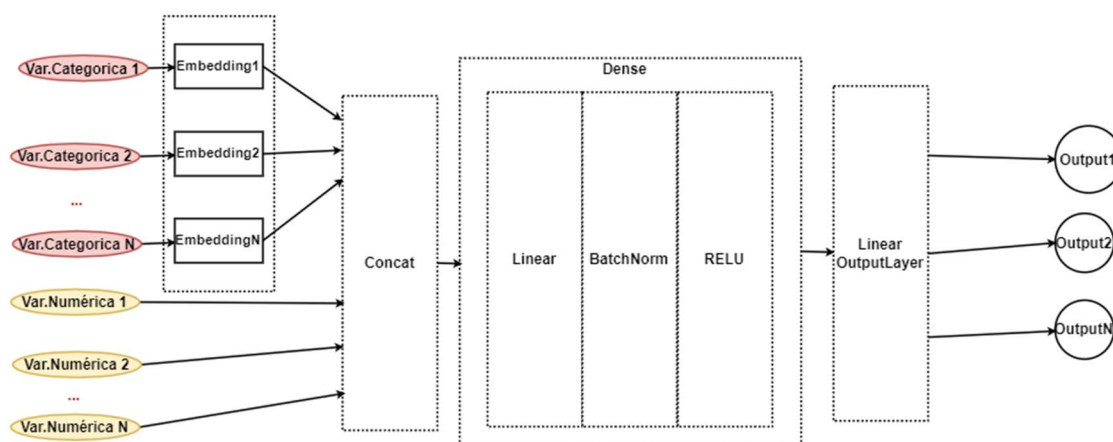
3.4.3. Rede Neural

A implementação e arquitetura de rede neural utilizada foi a do framework Fast.ai (HOWARD; GUGGER, 2020). A arquitetura da rede possui uma camada de embeddings, que são representações densas de vetores esparsos, utilizado para as variáveis categóricas como proposto por Guo e Berkhahn (2016), uma camada para concatenar as embeddings contínuas com o vetor de variáveis numéricas, uma camada densa e uma camada de output (Linear + BatchNorm + ReLu) com 20 neurônios e uma camada linear de output. Um esquema da arquitetura do modelo pode ser visto na Figura 18.

Para a rede neural, utilizou-se um único modelo (uma única função objetivo) responsável pela previsão de todos os horizontes para todos os insumos. Essa abordagem (ilustrada na Figura 19) permite que o modelo aprenda padrões globais compartilhados entre séries temporais de diferentes insumos para diferentes horizontes de previsão (caso existam).

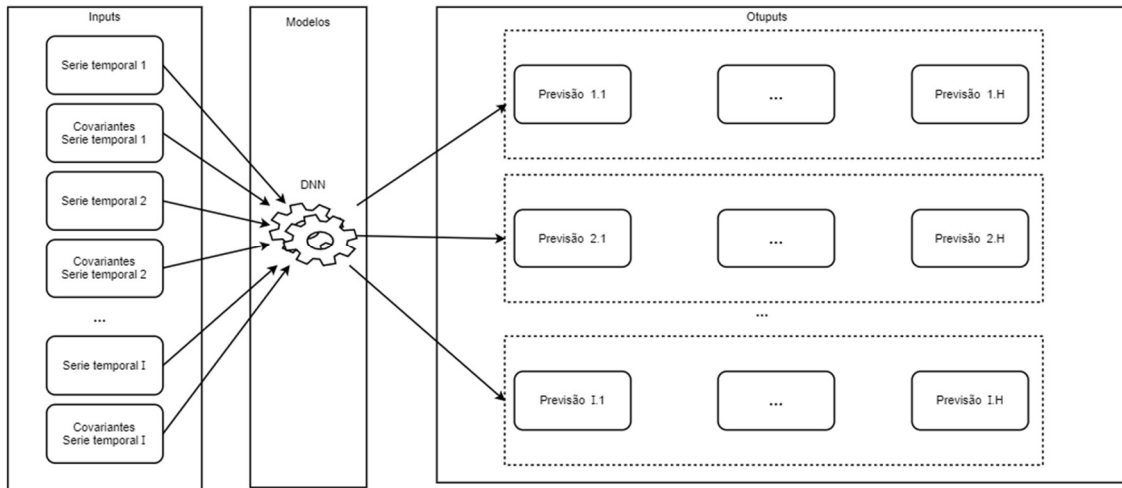
Para evitar viés no sentido da função objetivo dar mais importância a séries temporais com valores numéricos maiores, os consumos reais (targets) são padronizados (representados em termos do número de desvios padrão).

Figura 18: Arquitetura da rede neural utilizada.



Fonte: acervo pessoal.

Figura 19: Arquitetura da modelagem DNN.



Fonte: acervo pessoal.

3.5. Simulação de estoque

A simulação foi realizada através de um programa escrito em Python, que simula o comportamento do sistema. A classe criada possui métodos para consumo, reposição e aplicação da política de estoque. As etapas realizadas pelo algoritmo são as seguintes:

1. Adiciona-se ao estoque atual a reposição prevista para o dia

$$E_k^i = E_k^i + \hat{r}_k^i, \quad \hat{r}_k^i \in \hat{R} \quad (3.1)$$

2. Subtrair consumo real do estoque, mesmo que o resultado seja negativo. Isso simula transferência de insumos entre fábricas (o alto custo de falta de insumo justifica essa prática)

$$E_k^i = E_k^i - c_k^i, \quad c_k^i \in C^i \quad (3.2)$$

3. Prever consumo cumulativo para o horizonte do *leadtime*. O vetor de previsão é um vetor contendo as previsões para os próximos dias a partir do dia i e é função do histórico de consumo e das variáveis covariantes:

$$\hat{C}_{k+h}^i = f(C_{k-l:k}^i, Cov_{k-l:k}^i) \quad (3.3)$$

$$\hat{C}_{k:k+h}^i = \sum_{j=0}^h \hat{c}_{k,j}^i, \quad \hat{c}_{k,j}^i \in \hat{C}_k^i \quad (3.4)$$

4. Checar reposições previstas no horizonte do *leadtime*:

$$\hat{R}_{k:k+h}^i = \sum_{j=0}^h \hat{r}_{k,j}^i, \quad \hat{r}_{k,j}^i \in \hat{R}_k^i \quad (3.5)$$

5. Subtrair do estoque atual o valor de consumo cumulativo previsto E_i é o estoque real no dia i :

$$\hat{E}_{k+h}^i = E_k^i - (\hat{C}_{k:k+h}^i - \hat{R}_{k:k+h}^i) \quad (3.6)$$

6. Checar se o estoque previsto está abaixo do estoque mínimo:

$$\hat{E}_{k+h}^i < E_{min}^i \quad (3.7)$$

7. Se a condição 4 for satisfeita, realizar um pedido (que será repostado teoricamente em h dias) de modo a ter um estoque teórico igual ao ponto médio entre o estoque máximo e mínimo:

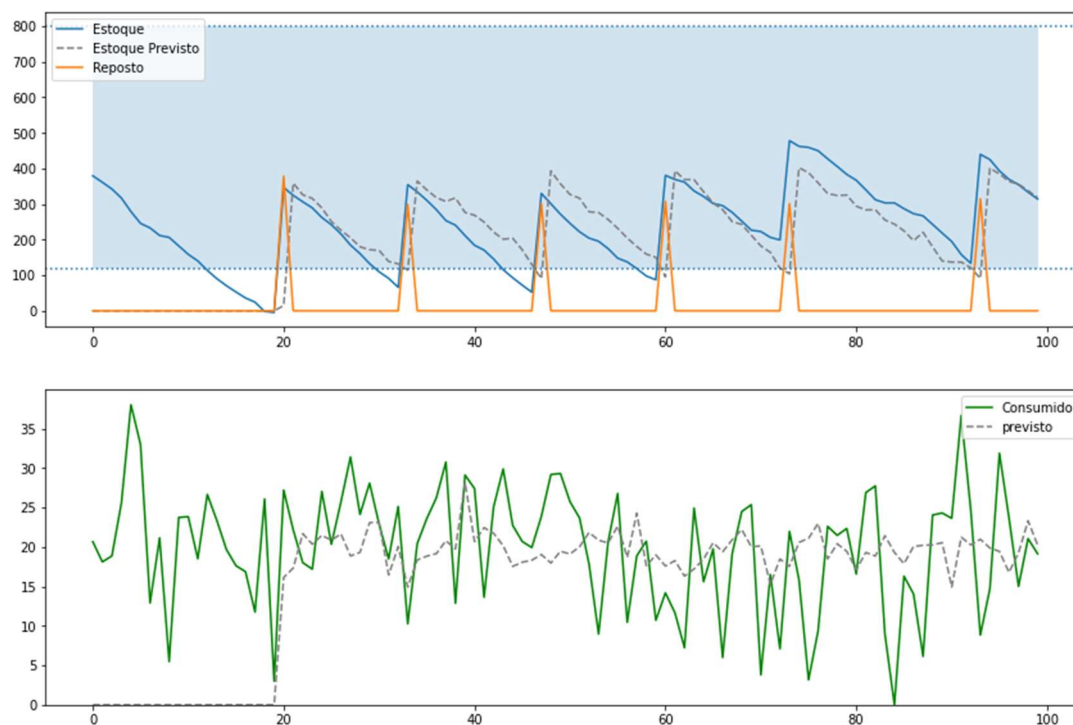
$$\hat{r}_{k+h+1}^i = (E_{min}^i - \hat{E}_{k+h}^i) + \frac{(E_{max}^i - E_{min}^i)}{2}, \quad \hat{r}_{k+h+1}^i \in \hat{R}^i \quad (3.8)$$

8. Próxima iteração

$$k = k + 1 \quad (3.9)$$

Esse processo é repetido para todas as séries temporais (i). No final, é possível obter a matriz de estoque “E” para todos os insumos em todas as unidades, dado a política de estoque, o estoque inicial e as previsões de consumo futuras. Uma ilustração do resultado final pode ser vista na Figura 20.

Figura 20: Exemplo do funcionamento do simulador com dados fictícios para *leadtime* de 20 dias.



Fonte: Acervo pessoal.

Esse processo de simulação é então repetido para todos os modelos preditivos usados, gerando diferentes matrizes de estoque de insumo simulados E^i , a partir desses estoques simulados é possível calcular métricas como dias de furo, capital imobilizado total e dias de sobre estoque.

4. Resultados

4.1. Validação dos modelos

As previsões geradas foram validadas utilizando a métrica do erro absoluto médio, bem como testes estatísticos para os resíduos. Os erros foram testados para viés (t-test), autocorrelação (Ljung-Box) e normalidade (Shapiro-Wilk).

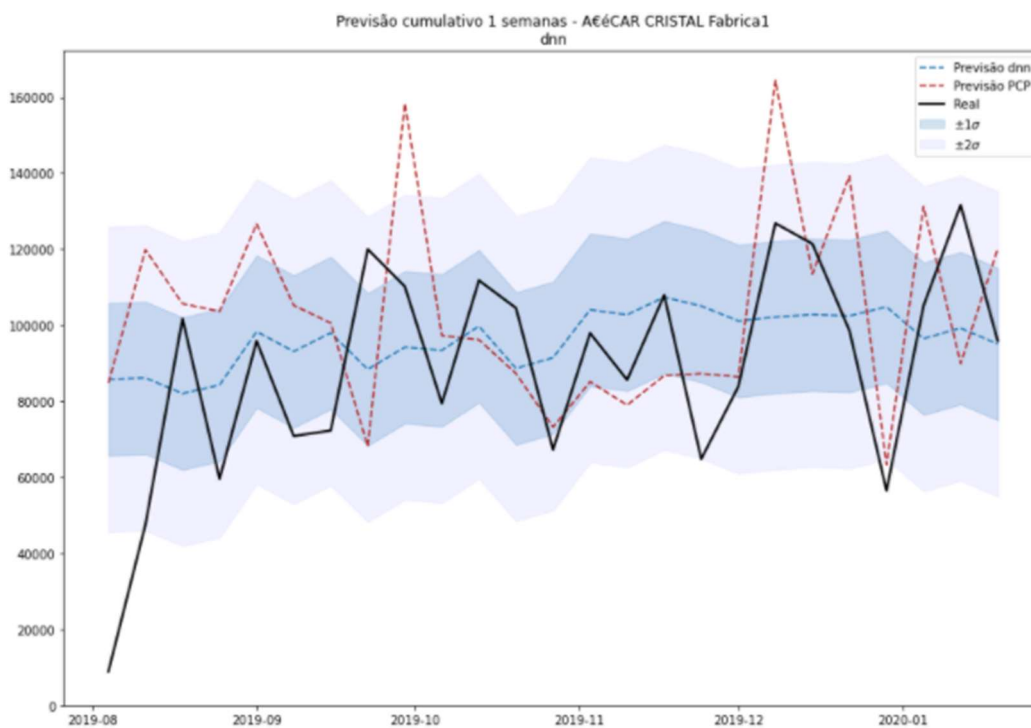
Todas as validações foram realizadas para a soma cumulativa dos 11 horizontes de previsão, juntamente com o horizonte de previsão do *leadtime*, específico para cada série temporal.

4.1.1. Análise visual

Mesmo sendo um método relativamente simples, a análise visual nos permite entender, criticar e validar previsões realizadas, além de avaliar alguns critérios qualitativos das previsões.

Na Figura 21, por exemplo, podemos observar uma série temporal aproximadamente estacionária, oscilando em torno de uma média. Além disso, é possível observar que o modelo consegue se adequar rapidamente a mudanças recentes no padrão de consumo. Isso se dá porque o *leadtime* do produto é de apenas uma semana, e a previsão de consumo cumulativo é realizada nesse curto horizonte. A previsão do PCP é a previsão de consumo realizada pelas unidades produtivas.

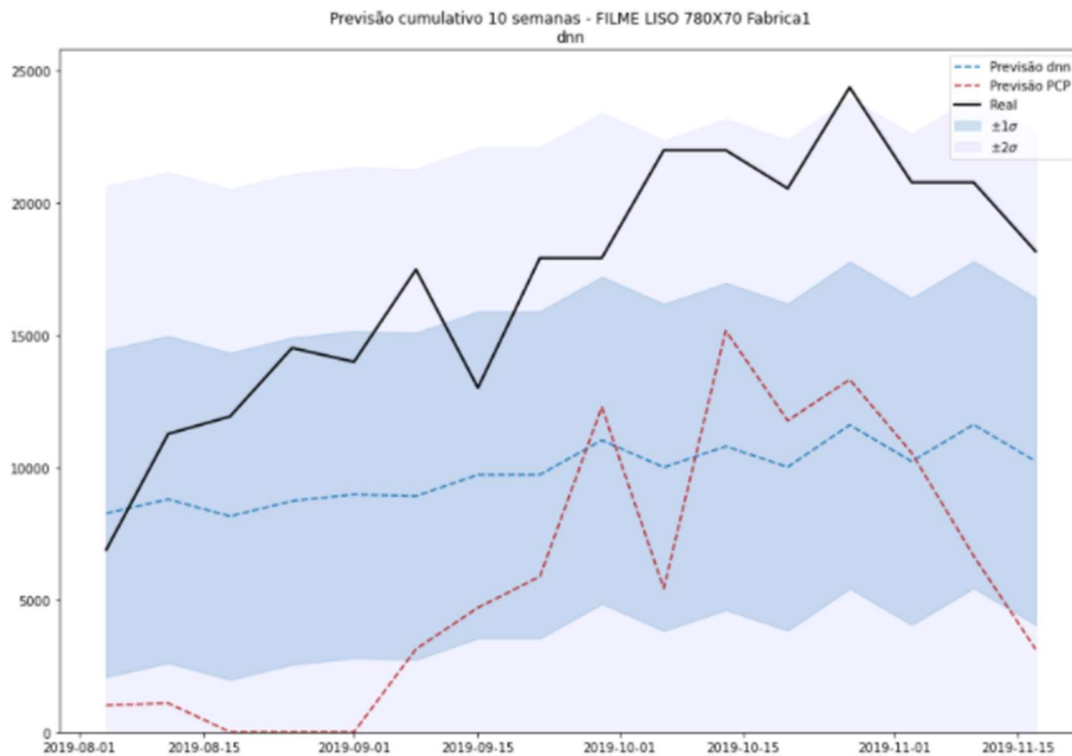
Figura 21: Previsão de consumo semanal do modelo DNN para um insumo em uma fábrica específica. *leadtime* do insumo é de uma semana.



Fonte: acervo pessoal.

Já na Figura 22, pode-se observar o contrário. Temos um *leadtime* de 10 semanas, o que torna o consumo altamente imprevisível, em função do horizonte de previsão mais alongado.

Figura 22: Previsão de consumo para horizonte de 10 semanas do modelo DNN para um insumo em uma fábrica específica. *leadtime* do insumo é de dez semanas.



Fonte: acervo pessoal.

Entretanto, por se tratar de aproximadamente 10000 séries temporais (2000 para cada modelo), torna-se inviável fazer a análise visual para todas as séries, tornando assim imprescindível outros tipos de validação, como testes estatísticos para resíduos e simulações de estoque.

4.1.2. Testes estatísticos dos resíduos

Os testes estatísticos de resíduos servem para verificar certas propriedades estatísticas destes. Calcula-se o erro entre o previsto e o real (cumulativo) para cada horizonte e para cada uma dessa série de erros. Com essa série temporal de resíduos, calculam-se os testes estatísticos e gera-se um resumo das séries que passaram nos testes de hipótese, para cada horizonte e modelo.

Teste T para viés

O teste T checka a validade de rejeitar a hipótese nula de que a média dos resíduos é diferente de zero. Nesse caso, rejeitaremos a hipótese nula para todos os casos em que o p-valor é maior que 0,05.

Na Tabela 8 é possível observar que os modelos multivariados (LGBM e DNN) retornaram um alto viés nos erros em comparação com os demais modelos, principalmente olhando a previsão cumulativa no horizonte do *leadtime*. Isso pode ser explicado pela presença das covariantes. Uma vez que o modelo não utiliza apenas o histórico da série temporal, é possível que as covariantes guiem as previsões para um valor distante da média do histórico recente, gerando viés nas previsões.

Tabela 8: Percentual de séries temporais com $p > 0,05$ para teste de viés dos resíduos para modelos em horizontes de previsão específicos.

Modelo	dnn	holt-winters	lgbm	naive	pcp
Horizonte					
h-1	31%	70%	53%	84%	62%
h-2	22%	63%	43%	84%	52%
h-3	19%	59%	38%	82%	48%
h-4	18%	56%	35%	79%	46%
h-5	18%	53%	32%	74%	44%
h-6	18%	51%	29%	68%	43%
h-7	17%	47%	28%	62%	41%
h-8	16%	45%	26%	57%	39%
h-9	16%	43%	25%	52%	38%
h-10	16%	41%	23%	49%	37%
h-11	16%	38%	21%	45%	39%
h-leadtime	15%	40%	22%	49%	39%

Fonte: Acervo pessoal.

Teste Shapiro-Wilk para normalidade

O teste Shapiro-Wilk é realizado para checar se os erros são distribuídos de acordo com uma distribuição normal. A hipótese nula é de que uma amostra é distribuída normalmente, desse modo, estamos buscando as séries em que o p-valor são maiores que 0,05

Tabela 9: Resultados dos normalidade dos resíduos para modelos em horizontes de previsão específicos.

Modelo	dnn	holt-winters	lgbm	naive	pcp
Horizonte					
h-1	22%	35%	29%	64%	46%

h-2	31%	45%	40%	66%	56%
h-3	35%	47%	42%	65%	58%
h-4	35%	46%	43%	61%	58%
h-5	33%	44%	41%	58%	56%
h-6	31%	42%	36%	54%	54%
h-7	28%	40%	34%	49%	52%
h-8	25%	38%	35%	44%	49%
h-9	23%	35%	34%	39%	47%
h-10	22%	35%	36%	36%	46%
h-11	21%	34%	38%	34%	46%
h-leadtime	24%	38%	40%	45%	47%

Fonte: Acervo pessoal.

Teste Ljung-Box para autocorrelação

O teste Ljung-Box serve para verificar se há autocorrelação entre os resíduos do modelo. A autocorrelação em uma série temporal é o grau em que os valores de uma série temporal estão correlacionados com o valor dessa mesma série, só que em instantes passados (HYNDMAN; ATHANASOPOULOS, 2018).

A autocorrelação entre resíduos indica que o conteúdo de informação dos inputs não foi “esgotado”, isso é, ainda existe informação que poderia ser utilizada pelo modelo.

Tabela 10: Resultados dos testes de autocorrelação dos resíduos para modelos em horizontes de previsão específicos.

Modelo	dnn	holt-winters	lgbm	naive	pcp
Horizonte					
h-1	71%	75%	74%	45%	79%
h-2	50%	56%	60%	56%	70%
h-3	23%	28%	43%	30%	54%
h-4	12%	15%	30%	15%	41%
h-5	6%	9%	24%	9%	35%
h-6	5%	8%	21%	7%	29%
h-7	4%	6%	19%	5%	24%
h-8	3%	5%	17%	4%	20%
h-9	3%	5%	16%	3%	18%
h-10	2%	5%	16%	3%	14%

h-11	2%	5%	16%	3%	13%
h-leadtime	4%	6%	18%	6%	18%

Fonte: Acervo pessoal.

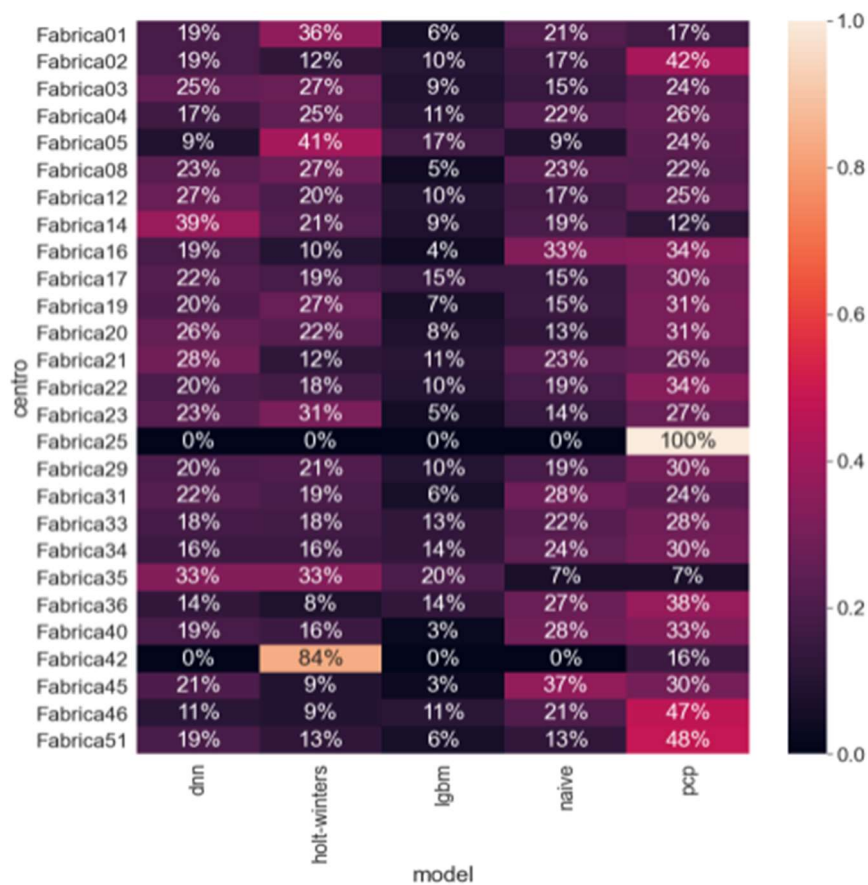
4.1.3. Erro absoluto médio (MAE)

O Desvio absoluto médio nos diz quanto, em média, uma série temporal errou em termos absolutos. Essa métrica nos permite quantificar qual o valor esperado (valor médio) para o erro absoluto em cada previsão. O seu cálculo é feito de acordo com a relação de equivalência da equação 4.1

$$MAE = E[|e|] = \frac{\sum |\hat{y}_i - y_i|}{n} \quad (4.1)$$

Assim, é possível definir, para cada série temporal, de forma desagregada, o modelo que obteve menor MAE. Desse modo, determina-se o percentual de vezes em que cada modelo foi a melhor opção dentre as demais por unidade. Um resumo dessa análise pode ser encontrado no mapa de calor da Figura 23.

Figura 23: Mapa de calor para percentual de modelos com menor MAE para cada fábrica.



Fonte: acervo pessoal.

É possível perceber que de fato a previsão do PCP (atual) é a que em geral, performa melhor em relação aos demais modelos. Contudo, existe uma grande oportunidade de melhoria levando em conta que no total acumulado, apenas 28,8% dos casos o PCP estima a série temporal com o menor MAE dentre os demais, como observado na Tabela 11.

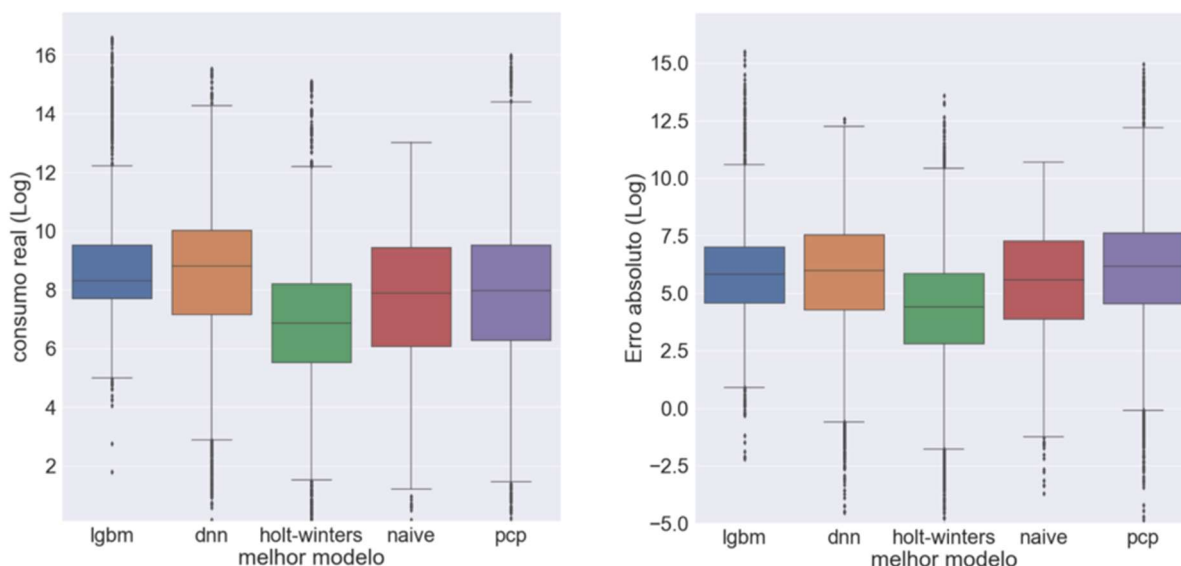
Tabela 11: Percentual agregado de modelos com menor MAE.

	pcp	naive	holt-winters	dnn	lgbm
Total	28.8%	21.3%	21.1%	20.9%	8.0%

Fonte: Acervo pessoal.

Além disso, ao plotar gráficos de caixas para o valor do consumo real e dos erros gerados pela previsão feita pelos melhores modelos (Figura 24), percebe-se que alguns modelos possuem menor MAE para séries temporais com valores mais elevados (DNN) enquanto outros performam melhor para valores de consumo menores (Holt-Winters).

Figura 24: Perfil de consumo real e resíduos para melhores modelos.



Fonte: Acervo pessoal.

Essas observações mostram que uma abordagem híbrida pode trazer resultados mais promissores no que diz respeito à minimização do MAE.

4.2. Validação das simulações

Ao ponto em que foram validadas as métricas relacionadas ao desvio das previsões com relação ao valor de consumo real, é interessante simular o sistema representado na Figura 14, a fim de estimar os impactos reais no processo para cada escolha de modelo específica.

Esses impactos foram estimados a partir das seguintes métricas:

- Número de dias de furo:
 - Número de dias em que houve falta do insumo.
- Número de dias críticos:

- Número de dias em que o nível de estoque ficou abaixo do estoque mínimo.
- Número de dias de sobre estoque:
 - Número de dias em que o estoque ficou acima do estoque máximo.
- Volume de estoque acumulado total:
 - A soma do volume total de estoque em cada dia, para todos os dias da simulação.
- Volume de capital imobilizado total:
 - A soma do custo total de estoque em cada dia, para todos os dias da simulação.
- Número de reposições:
 - Número total de reposições de estoque realizadas.

4.2.1. Estoque mínimo atual

O primeiro resultado diz respeito à simulação utilizando a política de estoque padrão (estoque mínimo e máximo predefinidos). Os resultados obtidos podem ser encontrados na Tabela 12.

Tabela 12: Métricas de estoque agregadas para simulação levando em conta política de estoque padrão. Caso base é a reposição pela previsão do PCP levando em conta o estoque mínimo padrão.

Modelo	best	dnn	holt-winters	lgbm	naive	pcp
Métrica						
N_DIAS_FURO	-24%	-22%	-9%	-38%	-25%	0%
N_DIAS_CRITICO	-15%	-17%	-2%	-31%	-17%	0%
N_DIAS_SOBREESTOQUE	-53%	54%	-54%	180%	-36%	0%
ESTOQUE_ACUMULADO_TOTAL	-35%	-24%	-35%	-19%	-32%	0%
CAPITAL_IMOBILIZADO_TOTAL	-27%	-19%	-31%	62%	-27%	0%
ESTOQUE_NEGATIVO_ACUMULADO_TOTAL	-53%	-40%	-56%	-58%	-52%	0%
CAPITAL_IMOBILIZADO_NEGATIVO_TOTAL	-26%	6%	-11%	-31%	-17%	0%
N_REPOSICOES	8%	12%	10%	6%	11%	0%

Fonte: Acervo pessoal.

É importante ressaltar que esse resultado é o agregado de um único modelo, isso é, as métricas representam um cenário em que um único modelo foi aplicado para todas as séries temporais.

A coluna *best* se refere à escolha do modelo com o menor MAE para cada uma das séries temporais, isso é, uma previsão usando uma composição de modelos diferentes, porém um único modelo por série temporal.

Pelo fato de haver mais de uma métrica de estoque definida, em alinhamento com especialistas da área de negócio da empresa responsável pelo controle de estoque, foram escolhidas duas delas como mais importantes: número de dias de furo e capital imobilizado total. Isso é, deseja-se minimizar o número de furo de estoque ao passo em que também se minimiza o capital imobilizado total.

É possível observar que os modelos que causaram um impacto mais desejado nessas duas métricas são: *best*, DNN e *Naive*. Sobretudo, como será tratado mais a frente, isso não é indicativo suficiente de escolha desses modelos para piloto pois é necessário olhar de forma desagregada, para cada série temporal, qual modelo melhor performou levando em conta as métricas de estoque.

4.2.2. Estoque mínimo sugerido

O Segundo resultado das simulações nos traz as métricas de estoque para uma nova política de estoque mínimo, que será definida como o equivalente a um desvio padrão do erro histórico entre o consumo previsto pelo modelo e o consumo real. O caso base para esses resultados é o mesmo caso base para a simulação anterior, isso é: previsão do PCP e estoque mínimo padrão. Os resultados encontram-se na Tabela 13.

Tabela 13: Métricas de estoque agregadas para simulação levando em conta política de estoque sugerida. caso base é a reposição pela previsão do PCP levando em conta o estoque mínimo padrão.

Modelo	<i>best</i>	<i>dnn</i>	<i>holt-winters</i>	<i>lgbm</i>	<i>naive</i>	<i>pcp</i>
Métrica						
N_DIAS_FURO	-38%	-32%	-27%	-36%	-41%	-9%
N_DIAS_CRITICO	4%	4%	12%	-13%	-6%	16%
N_DIAS_SOBREESTOQUE	-33%	68%	-34%	217%	1%	40%
ESTOQUE_ACUMULADO_TOTAL	-32%	-28%	-24%	-6%	-24%	1%

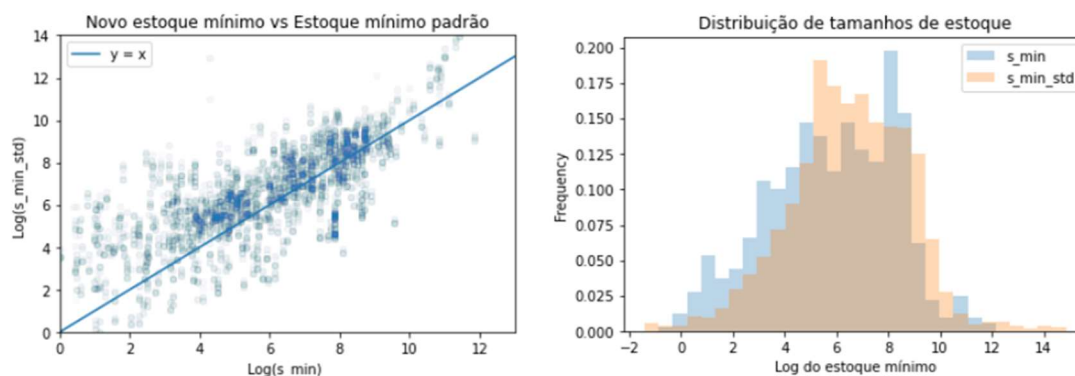
CAPITAL_IMOBILIZADO_TOTAL	-19%	-12%	-19%	84%	-10%	21%
ESTOQUE_NEGATIVO_ACUMULADO_TOTAL	-52%	-46%	-49%	-51%	-48%	-6%
CAPITAL_IMOBILIZADO_NEGATIVO_TOTAL	-23%	-9%	-9%	-11%	-13%	9%
N_REPOSICOES	17%	21%	22%	17%	27%	9%

Fonte: Acervo pessoal.

Uma tendência geral que pode ser observada é que o número de dias de furo diminuiu significativamente, ao custo de um maior capital imobilizado (ainda que negativo em comparação com o caso base) e um número maior de reposições.

Isso nos mostra que a definição de estoque mínimo feita pela indústria em questão acaba sendo menos conservadora (diminui o volume de estoque ao custo de aumentar o risco de falta). Isso se torna evidente ao analisar as distribuições e estoque mínimo padrão da companhia (s_{min}) versus o novo estoque mínimo sugerido (s_{min_std}), que será definido como um desvio padrão do erro histórico do modelo preditivo (*standard deviation* do inglês), de acordo com a Figura 25:

Figura 25: Análise de tamanhos de estoque mínimos padrão e sugerido.



Fonte: Acervo pessoal.

É possível observar que a maioria dos pontos no gráfico de dispersão estão acima da reta identidade, mostrando que para um mesmo insumo, a tendência é que o estoque mínimo sugerido seja maior que o estoque mínimo padrão. Já ao observar as distribuições dos dois estoques, percebe-se que a distribuição dos estoques mínimos sugeridos é mais deslocada para a direita, o que evidencia um volume maior de estoques mínimos com valor alto. Um painel resumo das estatísticas descritivas (contagem, média e percentis) dessas duas distribuições pode ser encontrado na Tabela 14:

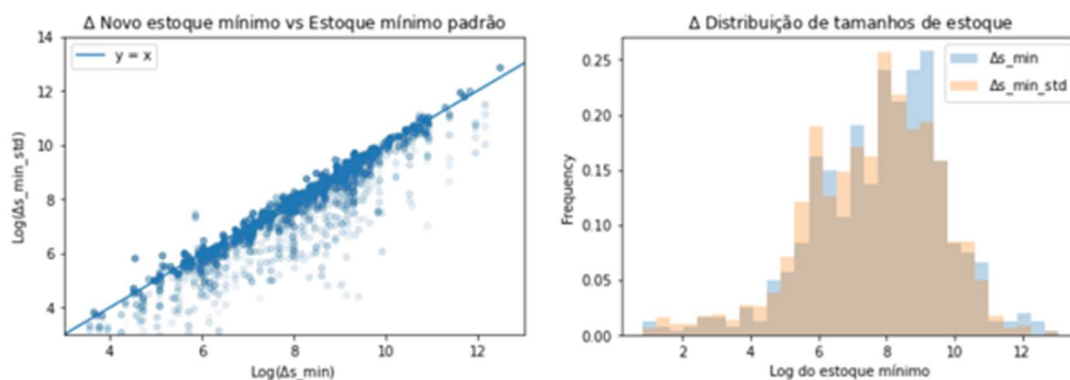
Tabela 14: Tabela resumo de tamanhos de estoque mínimos padrão e sugerido.

	count	mean	std	min	25%	50%	75%	max
s_min	95348	3278	11355	0	57	407	2574	138397
s_min_std	95348	13635	121750	0	166	654	3349	2817623

Fonte: Acervo pessoal.

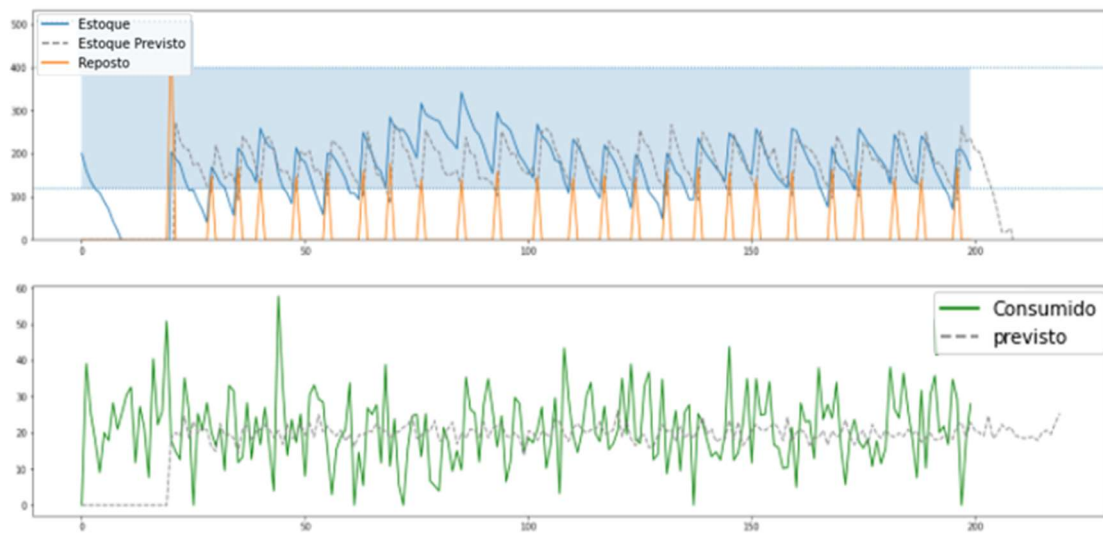
Já no que diz respeito ao aumento significativo no número de reposições, isso se deve ao fato de que, ao ponto em que o estoque mínimo aumentou, o estoque máximo permaneceu constante, de modo que o intervalo em que o estoque pode flutuar se tornou menor (Figura 26), aumentando o número de reposições (Figura 27 e Figura 28).

Figura 26: Análise de delta de estoque mínimos padrão e sugerido.



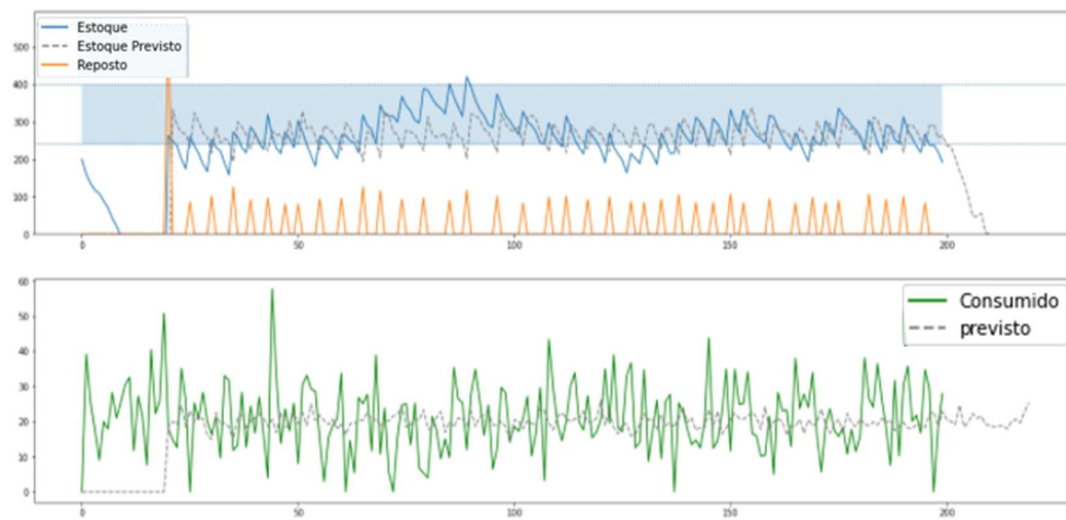
Fonte: Acervo pessoal.

Figura 27: Simulação com estoque mínimo padrão.



Fonte: Acervo pessoal.

Figura 28: Simulação com estoque mínimo reduzido.



Fonte: Acervo pessoal.

4.3. Análise desagregada

Dado que a sessão anterior se sustentou na análise agregada da performance dos modelos, a fim de encontrar um conjunto de séries temporais candidatas a uma bateria de testes em produção, deve-se realizar a análise desagregada de performance de estoque para cada série temporal.

Para essa análise, não foi levado em conta as simulações com a política de estoque sugerida, uma vez que é difícil de estimar o custo do crescimento do número de reposições.

Visto que a estratégia para a escolha de bons candidatos passa por um a otimização multiobjetivo (duas variáveis: número de dias de furo e capital imobilizado total) e que definir uma única função objetivo escalar que leve em conta essas duas variáveis é um esforço que vai além do escopo desse trabalho, foi utilizada uma estratégia que leva em conta a fronteira de Pareto do problema de otimização.

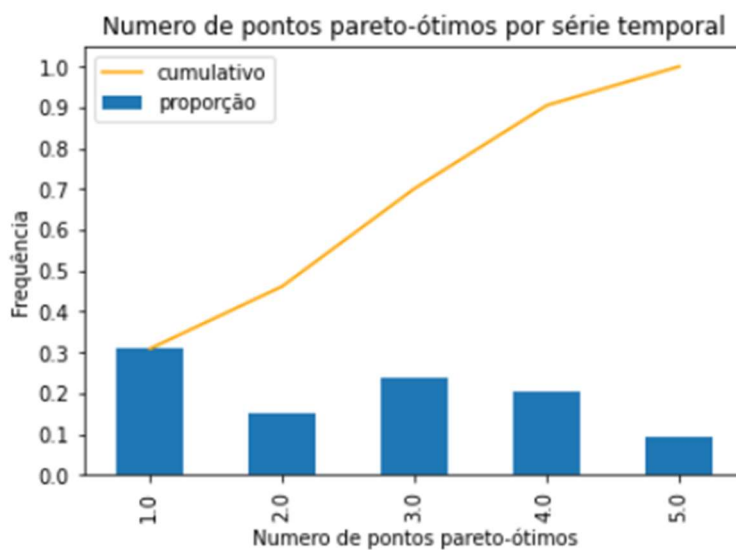
A fronteira de Pareto é o conjunto de pontos não dominados pertencentes ao espaço de solução de um problema de otimização multiobjetivo (GIAGKIOZIS; FLEMING, 2014). No problema de seleção de modelos candidatos, cada série temporal representa uma superfície em que os pontos são as coordenadas (número de dias de estoque e capital imobilizado total) de cada modelo, para aquela série temporal específica.

Desse modo, para cada série temporal, foi buscada a fronteira de Pareto, cujos pontos se referem aos modelos preditivos e as coordenadas representam as métricas de estoque (dias de furo e capital imobilizado).

Chamaremos de ponto estritamente dominante aquele em que não há nenhuma vantagem comparativa em escolher qualquer um dos demais pontos no espaço de solução (ele é um ponto ótimo para ambas as métricas individualmente).

Assim, foi criada uma função em Python para definir os pontos da fronteira de Pareto para o conjunto de métricas de estoque para cada série temporal. A distribuição do Número de pontos na fronteira de Pareto para as séries temporais encontra-se na Figura 29.

Figura 29: Distribuição de número de pontos na fronteira de Pareto para todas as séries temporais.

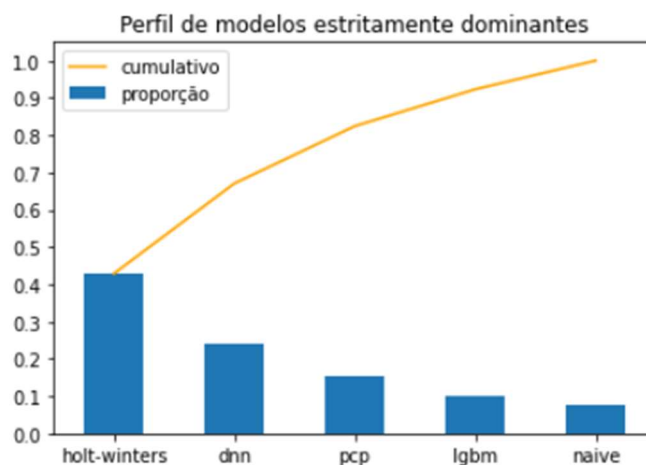


Fonte: Acervo pessoal

4.3.1. Pontos estritamente dominantes

Será dado enfoque aos pontos estritamente dominantes, uma vez que para as duas métricas de estoque avaliadas, estes pontos correspondem ao melhor cenário possível, isso evita indecisões na escolha dos modelos dada a natureza multiobjetivo do problema de escolha. O volume desses pontos corresponde a cerca de 30% de todos os pontos. Desses 30%, é possível observar sua composição na Figura 30.

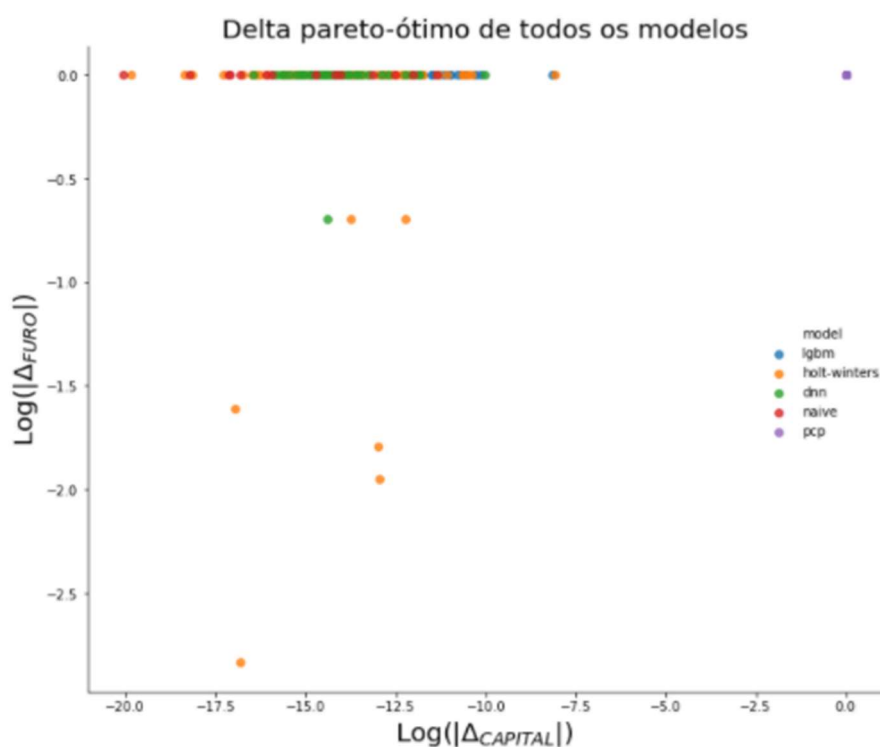
Figura 30: distribuição de modelos estritamente dominantes.



Fonte: Acervo pessoal.

Foi comparado então, para os modelos estritamente dominantes, qual seria o impacto percentual nas métricas de estoque, em relação ao caso base (previsão do PCP) caso esses modelos tivessem sido usados como modelo preditivo na simulação. Os resultados se encontram na Tabela 15. É possível também observar a dispersão desses pontos no gráfico da Figura 31. Nesse gráfico, os valores negativos implicam que houve uma diminuição nas métricas em relação ao caso base.

Figura 31: Ganho relativo ao caso base (PCP) para modelos estritamente dominantes.



Fonte: Acervo pessoal.

Tabela 15: Ganho relativo em métricas de estoque por modelo estritamente dominante.

Métrica	N_DIAS_FURO	CAPITAL_IMOBILIZADO_TOTAL	N_REPOSICOES
Modelo			
DNN	-11%	-18%	-3%
Holt-Winters	-24%	-22%	5%
LGBM	0%	-3%	1%

Naive	0%	-25%	10%
PCP	0%	0%	0%

Fonte: Acervo pessoal.

Tanto pelo gráfico de dispersão da Figura 31 quanto pela tabela 15, é possível observar que em metade dos modelos, o impacto maior é no capital imobilizado. Já nos modelos DNN e Holt-Winters, o impacto é em ambas as métricas, sem um grande impacto no número de reposições.

5. Conclusão

Dado os objetivos de definir bons modelos candidatos para testes em produção e definir uma nova política de estoque mínimo levando em conta o erro esperado dos modelos preditivos, em conjunto com os resultados obtidos, é possível gerar um conjunto de conclusões que irão direcionar ações futuras para a implementação da metodologia de controle de estoque sugerida nesse trabalho.

5.1. Modelos candidatos

Dentre as análises de performance realizadas, será dado ênfase à análise das simulações (em detrimento da análise estatística dos resíduos), visto que esta leva em conta critérios objetivos de negócio, como as métricas de estoque.

Dito isso, dentro da análise da simulação, será dado maior enfoque na análise desagregada, dado que cada família de modelos pode performar bem em um subconjunto das séries temporais, e pior em outros, como já formulado nos famosos teoremas do almoço grátis para otimização (WILLIAM; MACREADY, 1997).

Assim, visando simplificar a etapa de testes em produção e visando um maior impacto com menor esforço, O modelo Holt-Winters foi selecionado para realização dos testes, dado que cerca de 43% de todos os pontos estritamente dominantes são resultados de simulação usando o modelo Holt-Winters. Esses pontos constituem cerca de 13% de todas as séries temporais estudadas (43% vezes 30% de pontos estritamente dominantes), gerando um impacto estimado de menos 23,5% no número de dias de furo e diminuição de 22,2% no capital imobilizado total, para esse subconjunto de 13% das séries temporais.

5.2. Nova política de estoque

Com relação à nova política de estoque sugerida, pode-se concluir que o dimensionamento de estoque leva em conta fatores mais complexos do que simplesmente o erro histórico do modelo de previsão. Tais fatores como custo de ressuprimento, lote mínimo, validade e espaço físico necessário são outros fatores que tem um impacto relevante no dimensionamento de política de estoque.

Um outro ponto que se observou é a dificuldade de gerar uma função objetivo escalar para definir pontos ótimos de estoque e ressuprimento, dado que existem muitas interações entre diversas variáveis que justificariam um trabalho exclusivamente para a sua definição.

5.3. Próximos passos

Como ações futuras complementares a esse trabalho, podemos citar:

- Testar e validar o modelo Holt-Winters em produção para as séries temporais mais adequadas.
- Testar e validar o modelo DNN em produção para séries temporais mais adequadas.
- Definir uma função objetivo escalar a ser otimizada ao invés de uma política de estoque heurística, possibilitando uma adequação mais próxima ao conceito de MPC.

6. Referências

- ABOLGHASEMI, Mahdi; GERLACH, Richard; TARR, Garth; BEH, Eric; TRAN, Quoc; SEAMAN, Brian. **Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion.** Disponível em: <https://arxiv.org/abs/1909.13084>. Acesso em: 07 set. 2020.
- BANDARA, Kasun; SHI, Peibei; BERGMEIR, Christoph; HEWAMALAGE, Hansika; TRAN, Quoc; SEAMAN, Brian. **Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology.** Disponível em: <https://arxiv.org/abs/1901.04028>. Acesso em: 7 set. 2020.
- BUITINCK, Lars. **API design for machine learning software: experiences from the scikit-learn project.** 2013. Disponível em: <https://arxiv.org/pdf/1309.0238.pdf>. Acesso em: 07 set. 2020.
- CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2018, Long Beach. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree.** Long Beach: Neural Information Processing Systems Foundation, Inc., 2017.
- GIAGKIOZIS, Ioannis; FLEMING, Peter J.. Pareto Front Estimation for Decision Making. **Evolutionary Computation**, [S.L.], v. 22, n. 4, p. 651-678, dez. 2014. MIT Press - Journals. http://dx.doi.org/10.1162/evco_a_00128.
- GUO, Cheng; BERKHAHN, Felix. **Entity Embeddings of Categorical Variables.** 2016. Disponível em: <https://arxiv.org/abs/1604.06737>. Acesso em: 07 set. 2020.
- HAI, Dong; HAO, Zheng; PING, Li Yan. Model Predictive Control for inventory Management in Supply Chain Planning. **Procedia Engineering**, [S.L.], v. 15, p. 1154-1159, 2011. Elsevier BV. <http://dx.doi.org/10.1016/j.proeng.2011.08.213>.
- HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning.** 2. ed. Stanford: Springer, 2008.

- HOWARD, Jeremy; GUGGER, Sylvain. **Fastai: A Layered API for Deep Learning**. 2020. Disponível em: <https://arxiv.org/abs/2002.04688>. Acesso em: 07 set. 2020.
- HYNDMAN, Rob J.; AHMED, Roman A.; ATHANASOPOULOS, George; SHANG, Han Lin. Optimal combination forecasts for hierarchical time series. **Computational Statistics & Data Analysis**, [S.L.], v. 55, n. 9, p. 2579-2589, set. 2011. Elsevier BV. <http://dx.doi.org/10.1016/j.csda.2011.03.006>.
- HYNDMAN, R J; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. Melbourne: Otexts, 2018. Disponível em: <https://otexts.com/fpp2/index.html>. Acesso em: 07 set. 2020.
- LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, [S.L.], v. 521, n. 7553, p. 436-444, maio 2015. Springer Science and Business Media LLC. <http://dx.doi.org/10.1038/nature14539>.
- MAQBOOL, Saif; RAFIQ, Muhammad. CREATING COMPETITIVE ADVANTAGE THROUGH SUPPLY CHAIN MANAGEMENT. **International Journal Of Research In Commerce, It And Management**. p. 47-52. abr. 2014. Disponível em: https://www.researchgate.net/publication/261793707_CREATING_COMPETITIVE_ADVANTAGE_THROUGH_SUPPLY_CHAIN_MANAGEMENT. Acesso em: 07 set. 2020.
- MCKINNEY, Wes. **Pandas: a Foundational Python Library for Data Analysis and Statistics**. Disponível em: https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics. Acesso em: 07 set. 2020.
- MULLER, Max. **ESSENTIALS OF INVENTORY MANAGEMENT**. New York: Amacom, 2003.
- SEABOLD, Skipper; PERKTOLD, Josef. **Statsmodels: Econometric and Statistical Modeling with Python**. Disponível em: <https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf>. Acesso em: 07 set. 2020.

- SINGH, Pawan Kumar; GUPTA, Yadunath; JHA, Nilpa; RAJAN, Aruna. **Fashion Retail: Forecasting Demand for New Items**. Disponível em: <https://arxiv.org/abs/1907.01960>. Acesso em: 07 set. 2020.
- WOLPERT, David H.; MACREADY, William G.. No Free Lunch Theorems for Optimization. **Ieee Transactions On Evolutionary Computation**. p. 67-82. abr. 1997