

EXPERIÊNCIAS DE UM ESTATÍSTICO

Guilherme Parreira



AGENDA

- FORMAÇÃO ACADÊMICA
- EXPERIÊNCIA PROFISSIONAL
- PROFISSÕES DO FUTURO
- SIMILARIDADES COM CIENTISTA DE DADOS
- CASOS
- CONTATO



FORMAÇÃO

Graduação em Estatística em 2015 pela UFPR
com período sanduíche na University of
Western Australia

Mestrado em Métodos Numéricos aplicados a
estatística pelo PPGMNE/UFPR em 2021



Experiência Profissional



PROFISSÕES DO FUTURO

Considerando seu potencial para criação de vagas entre 2023 e 2027, o Fórum Econômico Mundial listou 50 profissões emergentes para os próximos anos. “Espera-se que o emprego de analistas e cientistas de dados, especialistas em big data, **especialistas em aprendizado de máquina de IA** e profissionais de segurança cibernética cresça, em média, 30% até 2027”, diz o relatório.

Veja as primeiras colocadas na lista de profissões emergentes do Fórum Econômico Mundial:

- Especialistas em Inteligência Artificial (AI) e Machine Learning
- Especialistas em sustentabilidade
- Analistas de business intelligence (BI)
- Analistas de segurança da informação
- Engenheiro de fintech
- Analistas e cientistas de dados
- Engenheiros robóticos
- Especialistas em Big Data
- Operadores de equipamentos agrícolas

Fonte: Fórum econômico mundial
2023

No que aprimorar

ENGENHARIA DE SOFTWARE

Ciclo de Vida de Desenvolvimento de Software, Design Pattern, Clean Code, Fácil manutenção, Unit Tests, Modular Programming

ARQUITETURA DE HARDWARE

CPU, GPU, TPU, RAM, I/O

PORTABILIDADE DE SOFTWARE

Ambientes virtuais e Containers

NUVEM

Azure, AWS, GCP

Pontos fortes

ANALISAR DADOS DE R&D

Planejamento e otimizar experimentos, ensaios clínicos (cohort, ...);

ANALISAR DADOS DE OPINIÃO PÚBLICA

Elaborar Perguntas de Pesquisa

DESENHAR PLANOS AMOSTRAIS

PENSAMENTO CRÍTICO
PARA ANALISAR
DIFERENTES TIPOS DE
DADOS

EXPERIÊNCIA PROFISSIONAL

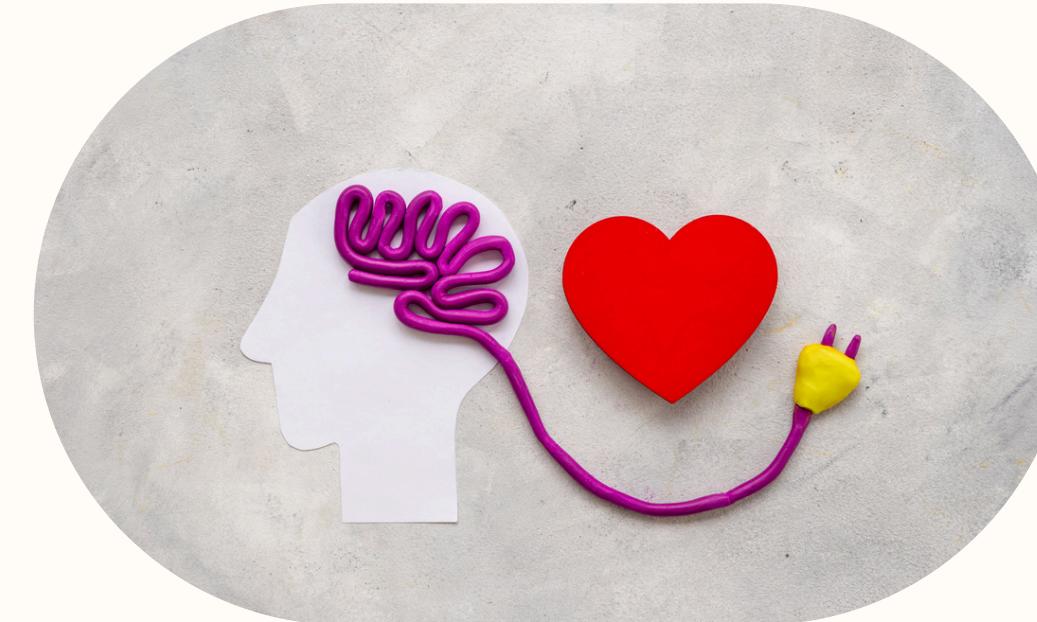
- Estatístico
- Cientista de Dados



SESSÕES DE TERAPIA DE CASAL



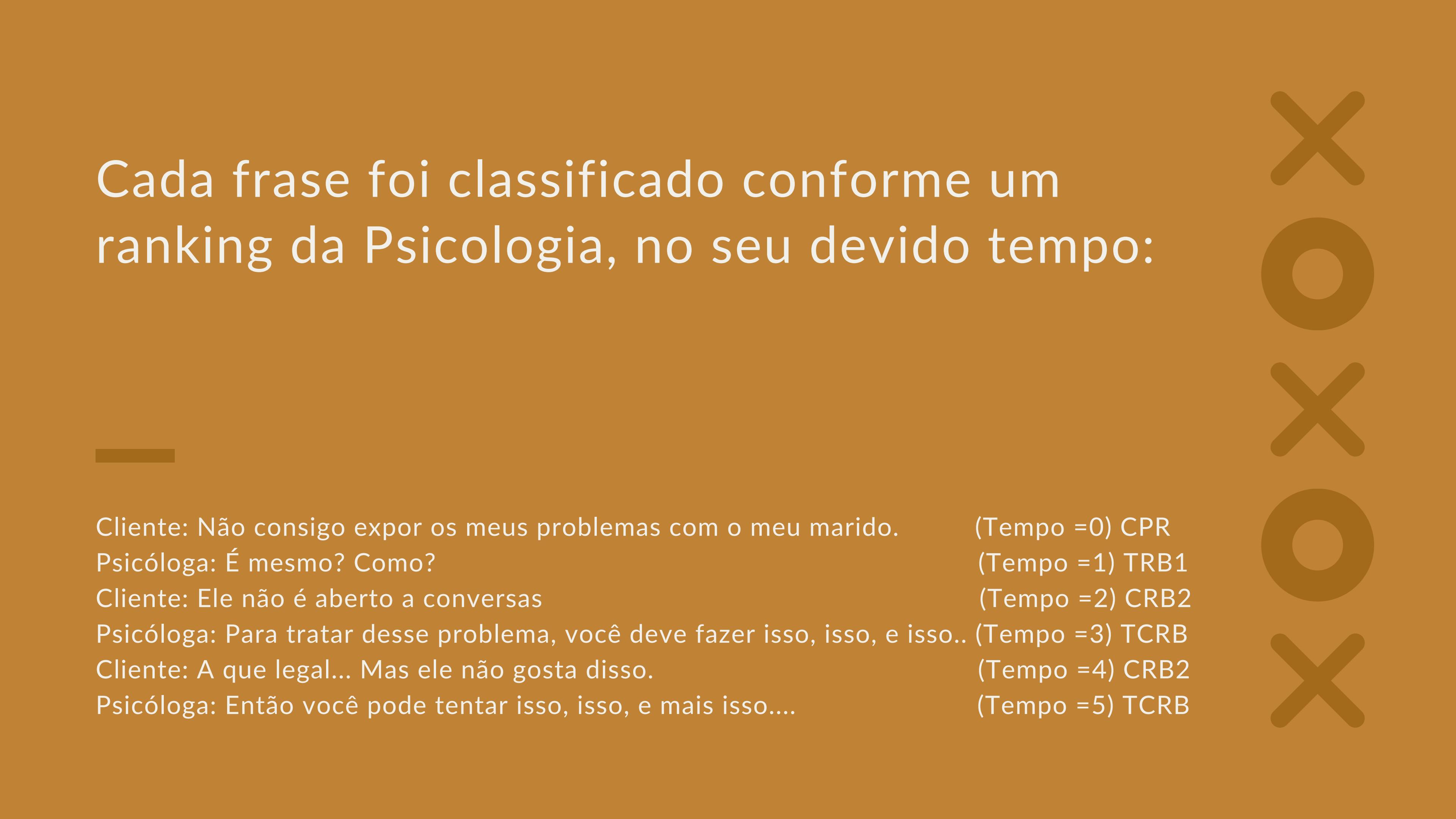
Pesquisadora: Antoniela Yara
Programa de Pós-Graduação em Psicologia da
UFPR (PPGPs/UFPR)



O objetivo era analisar o uso da Psicoterapia
Analítica Funcional, que foca no aqui/agora para
avaliar se existia uma melhora na **intimidade do
casal**.

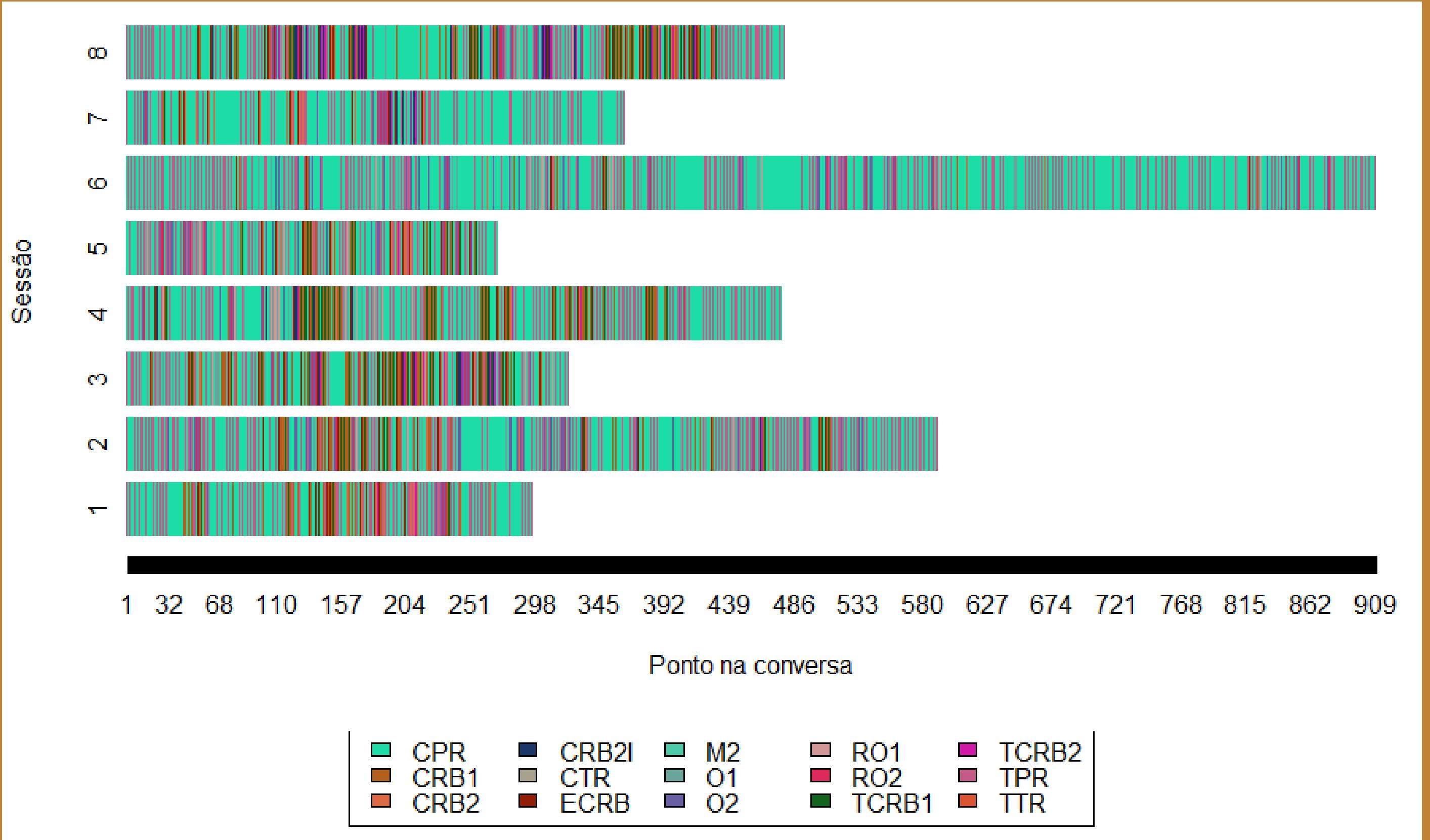


N = 1 casal
8 sessões



Cada frase foi classificado conforme um ranking da Psicologia, no seu devido tempo:

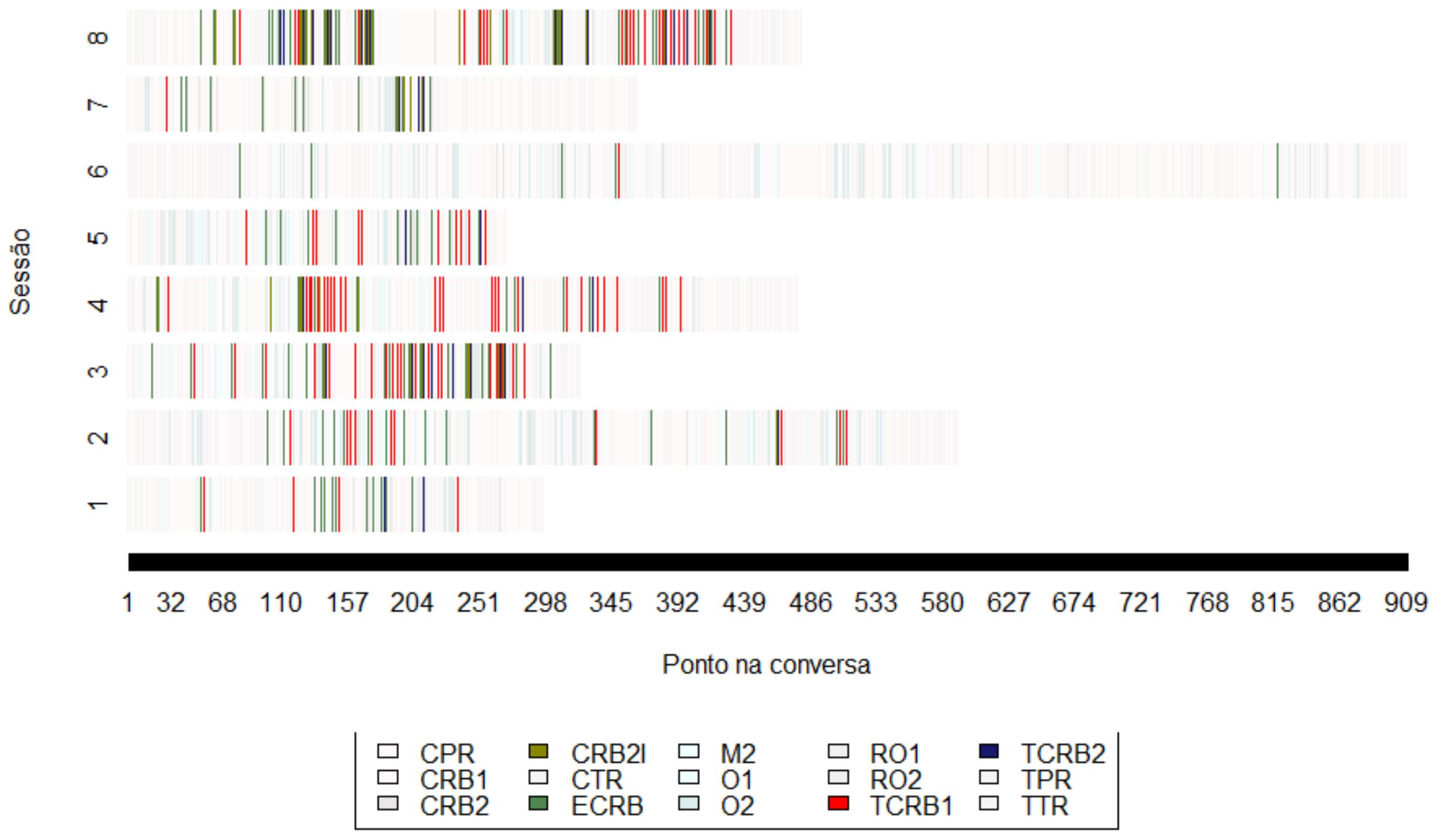
-
- Cliente: Não consigo expor os meus problemas com o meu marido. (Tempo =0) CPR
 - Psicóloga: É mesmo? Como? (Tempo =1) TRB1
 - Cliente: Ele não é aberto a conversas (Tempo =2) CRB2
 - Psicóloga: Para tratar desse problema, você deve fazer isso, isso, e isso.. (Tempo =3) TCRB
 - Cliente: A que legal... Mas ele não gosta disso. (Tempo =4) CRB2
 - Psicóloga: Então você pode tentar isso, isso, e mais isso.... (Tempo =5) TCRB



Sessões ímpares eram de intervenção padrão de terapia de casal foco na resolução de problema

Seções pares que focavam em provocar o aumento da intimidade do casal

Sequência da conversa por Sessão



Matriz de Transição Geral com Tempo = 3													
	[CPR]	[CRB1]	[CRB2]	[CRB2I]	[ECRB]	[O1]	[O2]	[Outras]	[RO1]	[RO2]	[TCRB1]	[TCRB2]	[TPR]
[CPR ->]	0.53	0.03	0.01	0.00	0.03	0.03	0.03	0.00	0.01	0.01	0.02	0.01	0.29
[CRB1 ->]	0.29	0.07	0.00	0.00	0.08	0.04	0.02	0.00	0.01	0.01	0.19	0.06	0.22
[CRB2 ->]	0.40	0.05	0.05	0.00	0.09	0.05	0.02	0.02	0.00	0.00	0.04	0.04	0.25
[CRB2I ->]	0.17	0.07	0.00	0.02	0.10	0.00	0.00	0.00	0.05	0.02	0.22	0.17	0.17
[ECRB ->]	0.42	0.28	0.05	0.06	0.04	0.03	0.03	0.00	0.00	0.00	0.03	0.01	0.04
[O1 ->]	0.38	0.05	0.01	0.01	0.07	0.08	0.03	0.00	0.10	0.01	0.05	0.01	0.18
[O2 ->]	0.39	0.01	0.01	0.00	0.08	0.05	0.05	0.00	0.01	0.10	0.05	0.01	0.24
[Outras ->]	0.25	0.25	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.00
[RO1 ->]	0.67	0.00	0.00	0.00	0.00	0.19	0.12	0.00	0.00	0.00	0.02	0.00	0.00
[RO2 ->]	0.52	0.06	0.00	0.06	0.00	0.06	0.20	0.00	0.02	0.02	0.00	0.00	0.06
[TCRB1 ->]	0.33	0.32	0.06	0.06	0.00	0.02	0.04	0.01	0.01	0.01	0.04	0.00	0.11
[TCRB2 ->]	0.29	0.29	0.00	0.15	0.02	0.02	0.10	0.00	0.00	0.00	0.02	0.02	0.07
[TPR ->]	0.66	0.05	0.02	0.01	0.01	0.05	0.06	0.00	0.00	0.01	0.00	0.00	0.12

Analisa-se a conversa de origem no tempo 0 com a conversa de destino no Tempo 3;

Matriz de Transições -> Processos Markovianos



ENGORDA DE FRANGO



Objetivo era analisar o tratamento que resultasse no maior ganho de peso corporal de frangos da granja da empresa Impextraco



Ajuste de Modelo Linear Misto

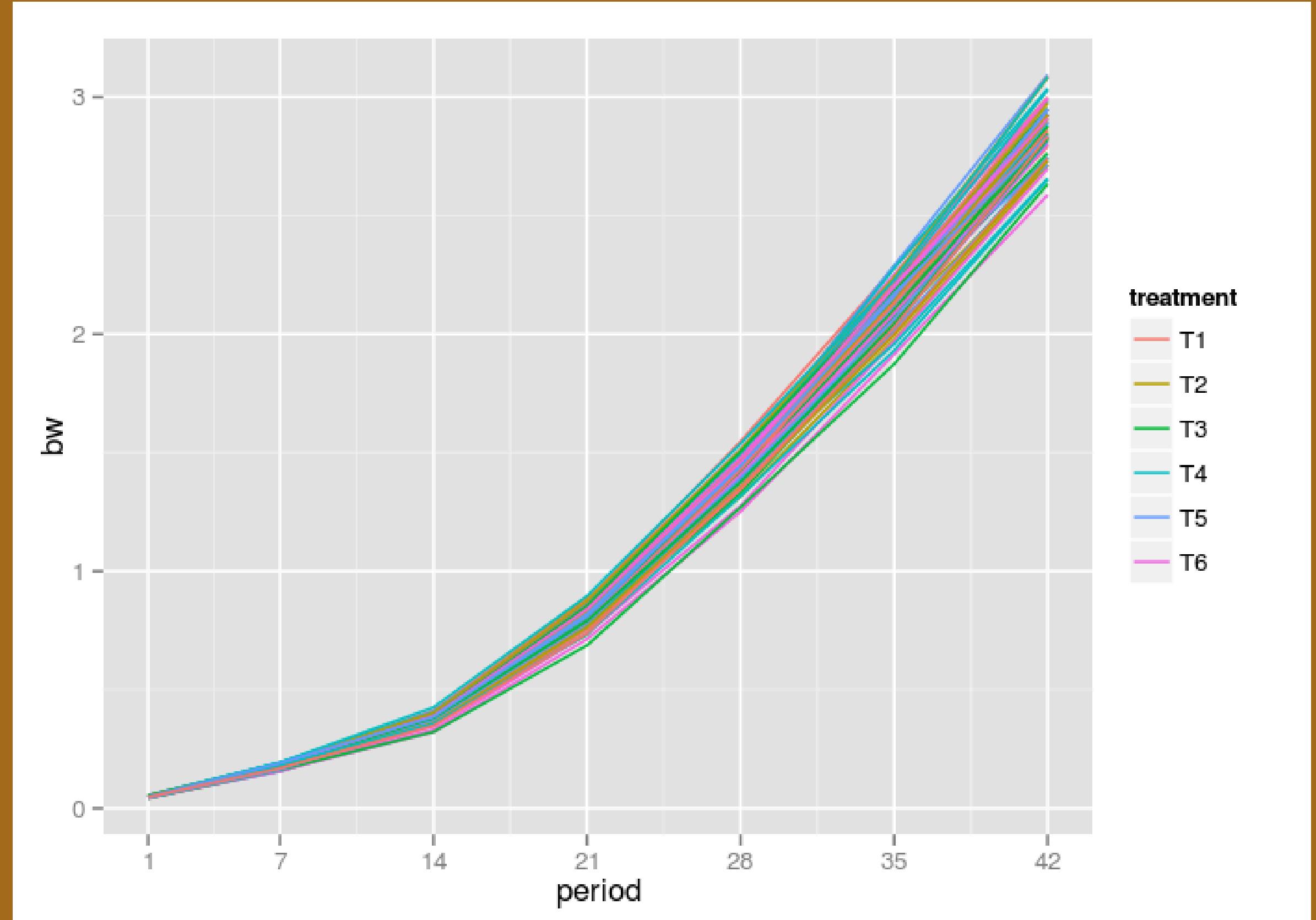
Efeito aleatório:

- Inclinação para período ao longo do tempo
- Intercepto Para cada frango

Estrutura heterocedástica do dado:

- Função de variância exponencial para o período

Extensões de modelos de regressão



GRUPO MARISTA

PREVISÃO DE DEMANDA DE MATRÍCULAS



Regras de Negócio

- + O período de matrículas ocorre entre Agosto e Março
- + As rematrículas ocorriam no período de 2 semanas no mês de Novembro

Contexto

- 24 escolas da rede da Educação Básica do Marista Centro Sul
- 4 segmentos (EI, EFAI, EFAF, EM)
- Alunos Novos/Rematrículas
- 4 anos de histórico

Variável modelada

- + Modelamos o saldo semanal de matrículas = Número de matrículas total - Número de matrículas desistentes

Regras

- + Quantidade de rematrícula não pode ser maior que o número de alunos que terminaram o último ano - alunos formados; Ex.: EM1 e EM2
- + Vaga de novas matrículas limitadas ao tamanho do colégio;
- + Total de matrículas não poderia ser maior que a capacidade do colégio e segmento

Erros de base no Sistema

- + Para o primeiro ano que o sistema foi implantado, algumas matrículas foram feitas em lote;
- + Em outros casos, as matrículas começaram no meio da campanha da matrícula, sendo que esse comportamento é característico no início

Definições de previsão

- Modelamos o saldo semanal de matrículas
- Novas previsões disponibilizadas diariamente
- Período fixo entre Agosto e Março
- Previsões até o final da campanha - 34 semanas

Variáveis utilizadas

- Meta
- Pandemia
- Período da Rematrícula
- Número de potenciais estudantes
- % de Fidelização histórica dos estudantes

Modelos Avaliados:

- Thyme Boost
- Prophet
- Holt Winters
- AutoArima
- AutoML

Dificuldades na modelagem:

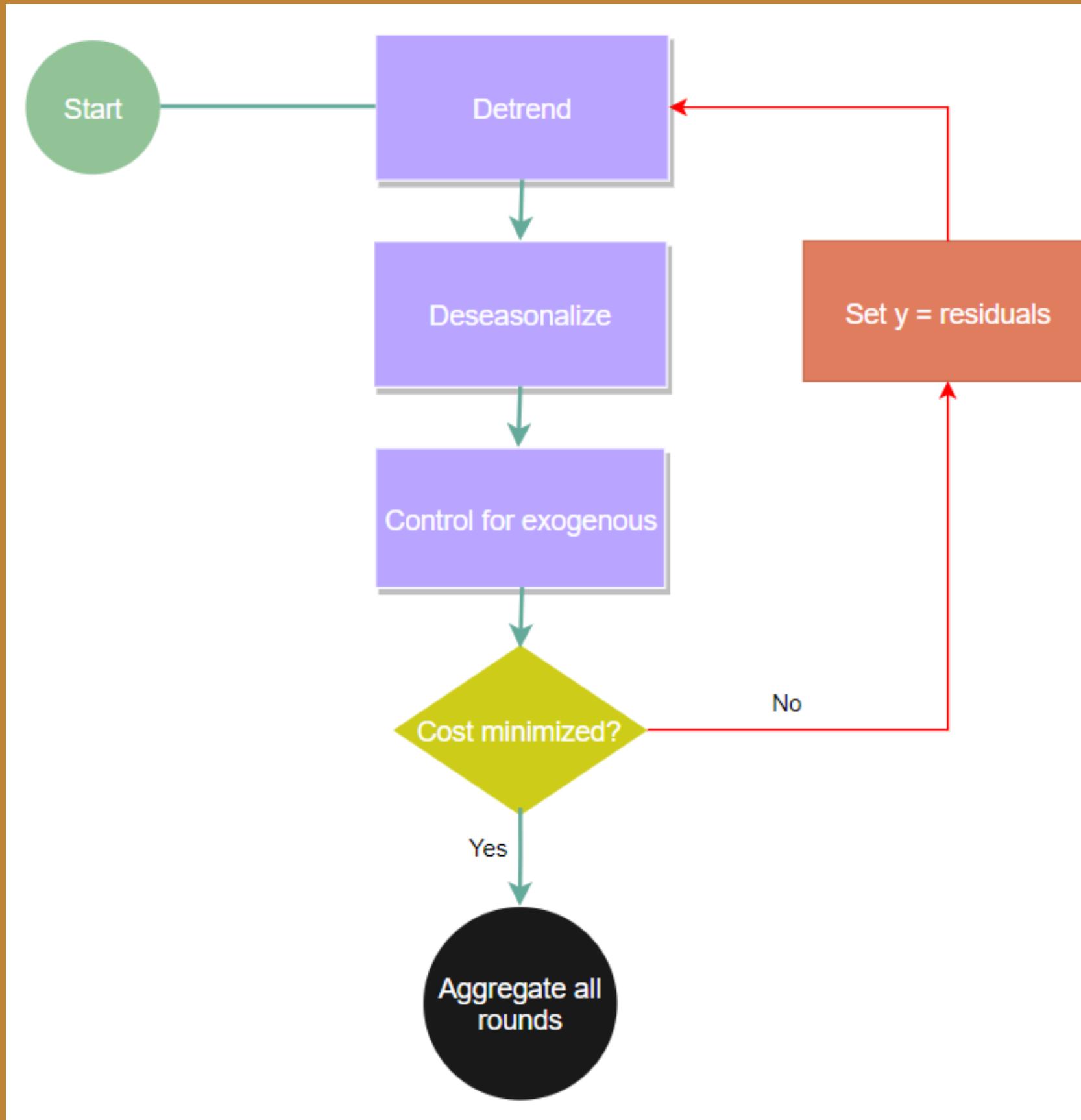
- Rateio da previsão semanal para diário;
- Dados de entrada poderiam ser negativo

Stack:

- Python - Jupyter Notebook
- Databricks
- MI Flow
- Azure (Blob storage)

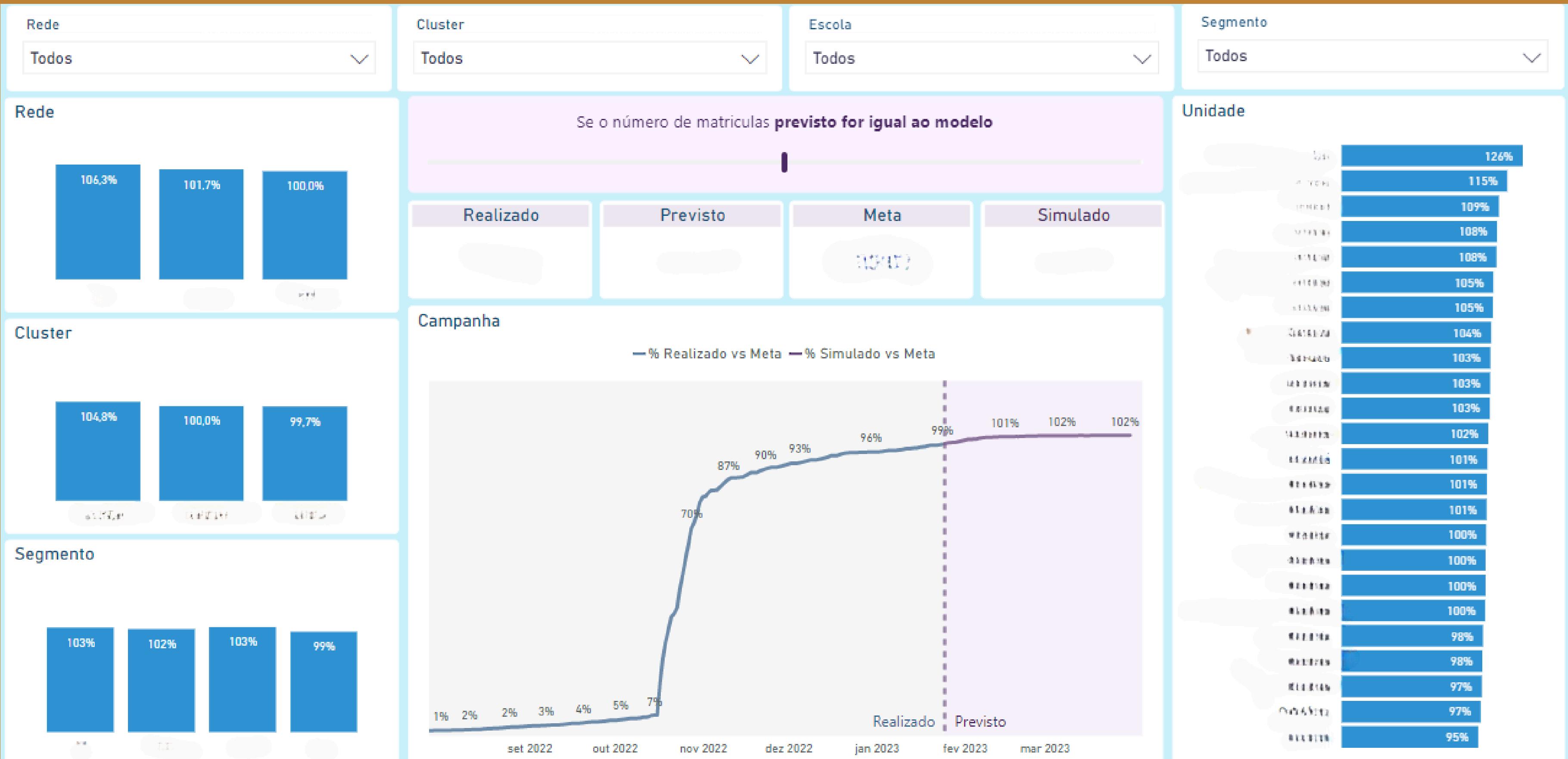
- Consultoria semanal dos professores da pós-graduação da PUC-PR

Modelo ThymeBoost



Decomposição de Séries Temporais +
Gradient boosting

Define-se estimadores de tendência
(média, linear, arima) e sazonalidade
(fourier)



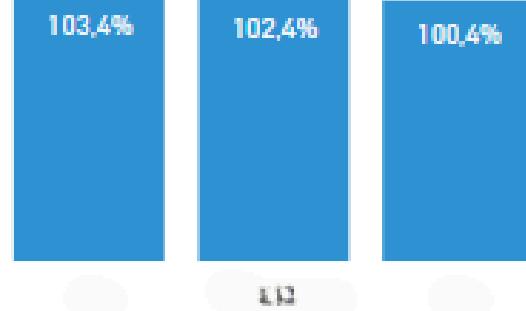
Simulador

Tabela

Simulador Receita

Rede

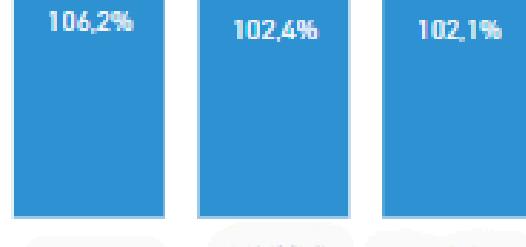
Todos



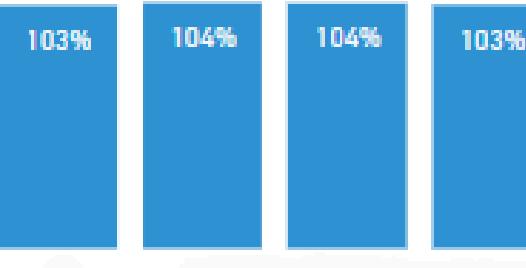
Cluster

Todos

Cluster



Segmento

Se as matrículas **for igual ao modelo**Se o desconto **previsto for igual ao modelo**

Receita Bruta

Desconto

Receita Líquida

Orçado

Simulado

% Orç

103,8%

Orçado

Simulado

% Orç

105,7%

Orçado

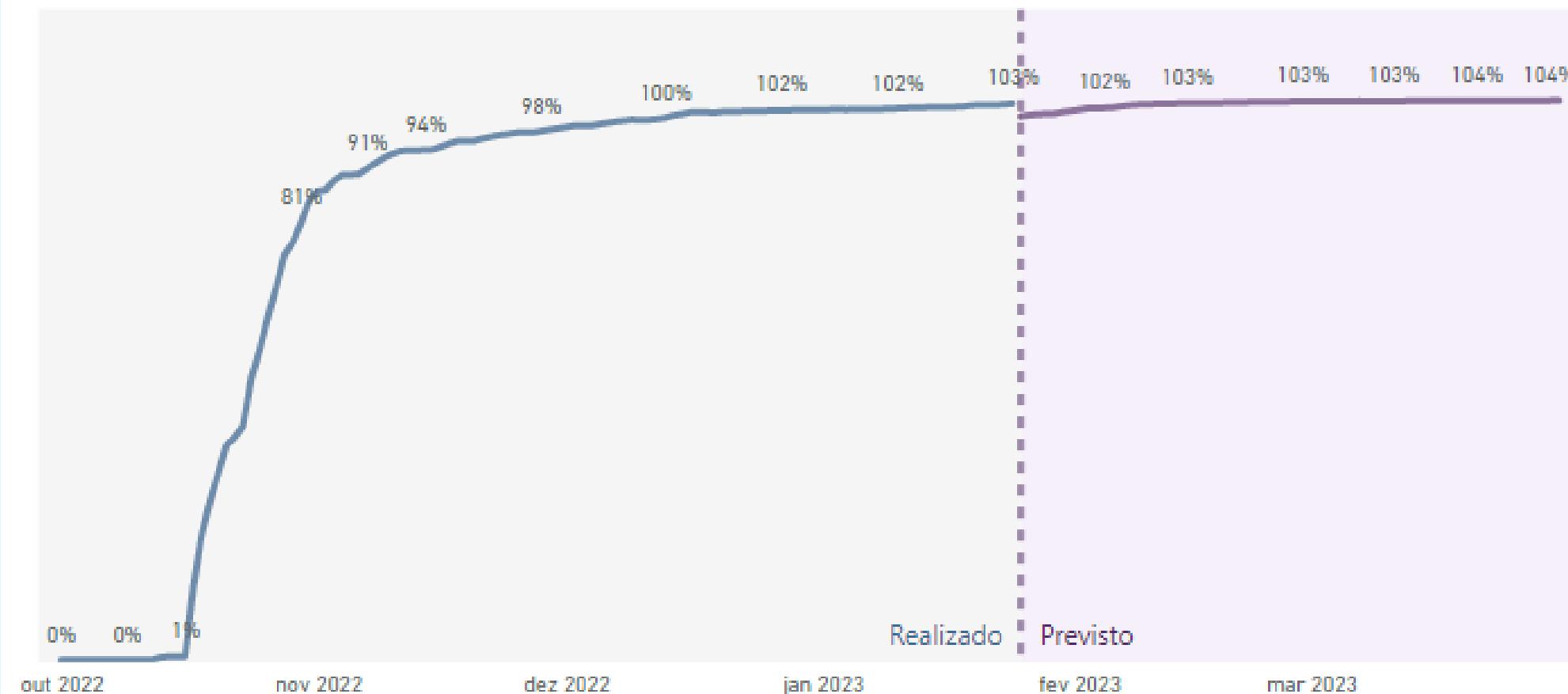
Simulado

% Orç

103,6%

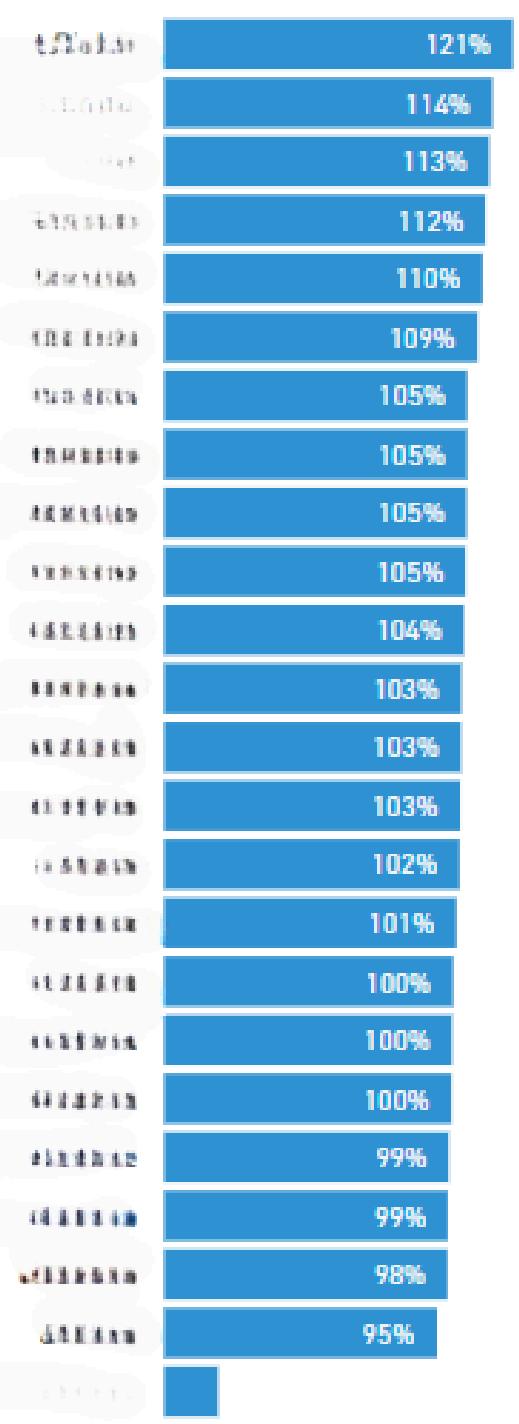
Campanha

— % Receita Líquida Realizada vs Orçamento — % Receita Líquida Simulada vs Orçamento



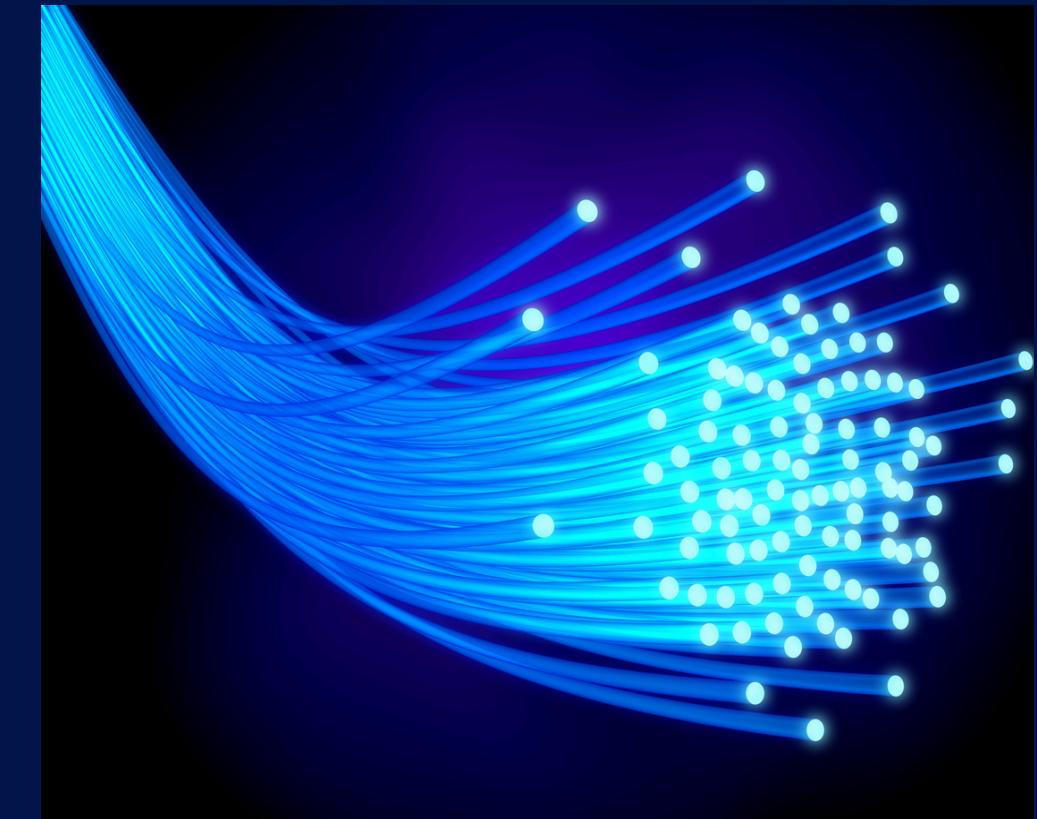
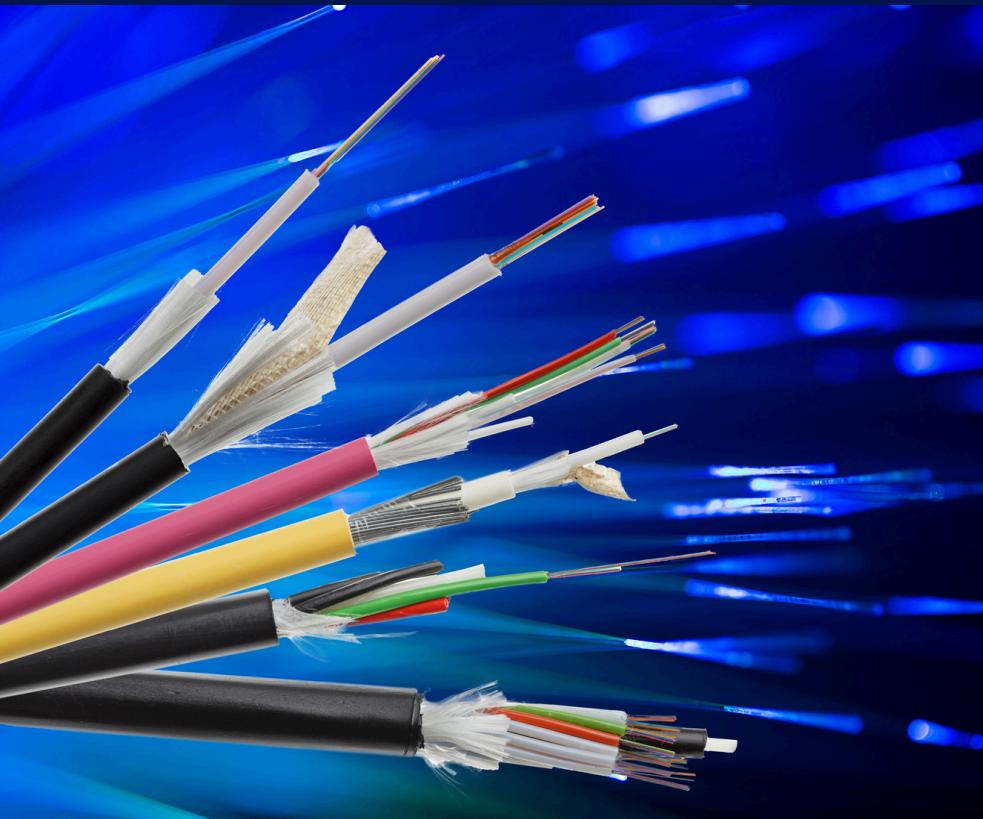
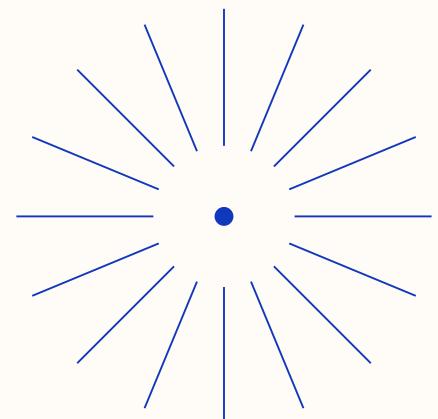
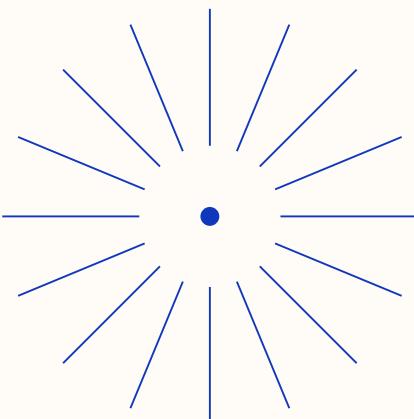
Segmento

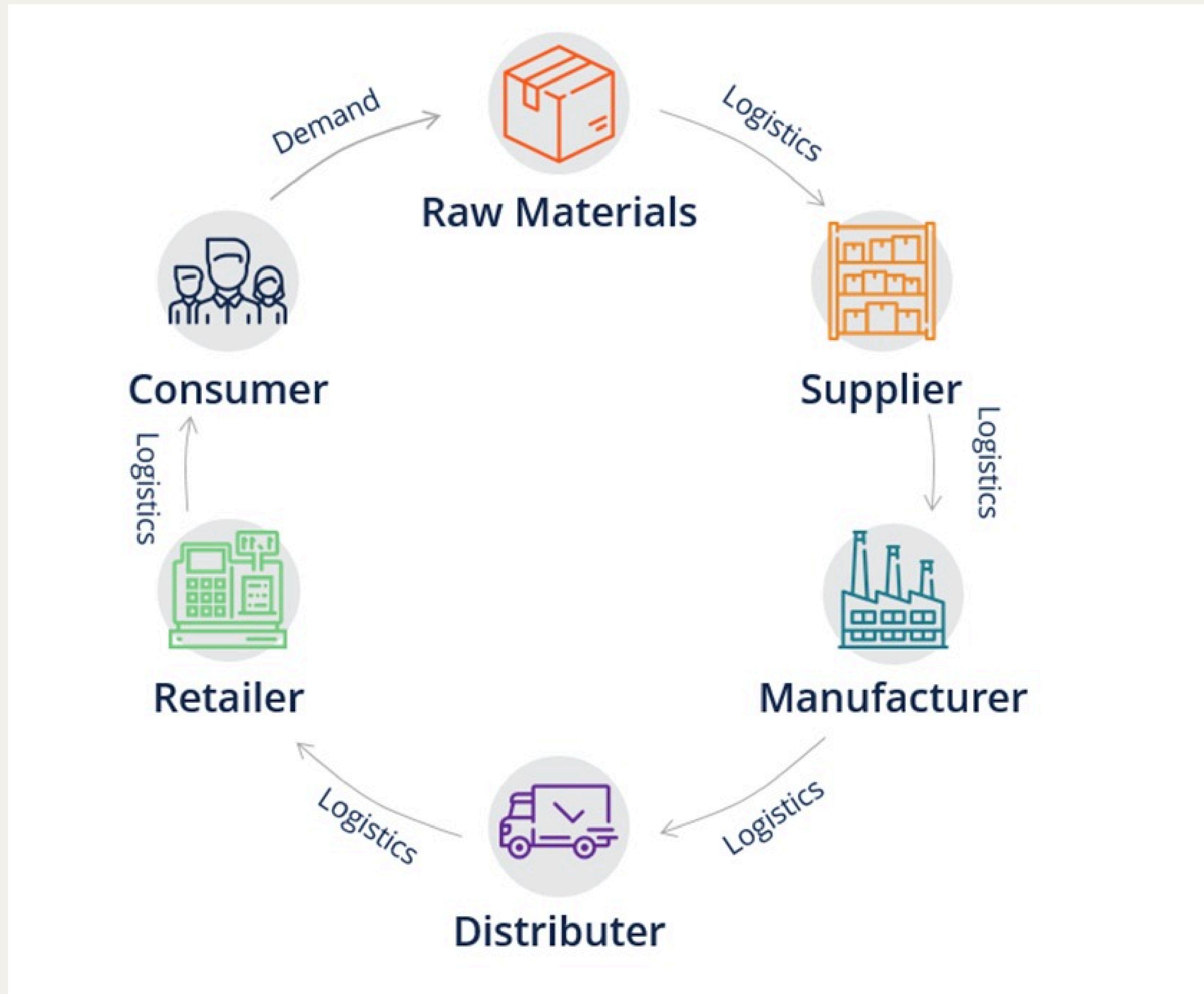
Todos

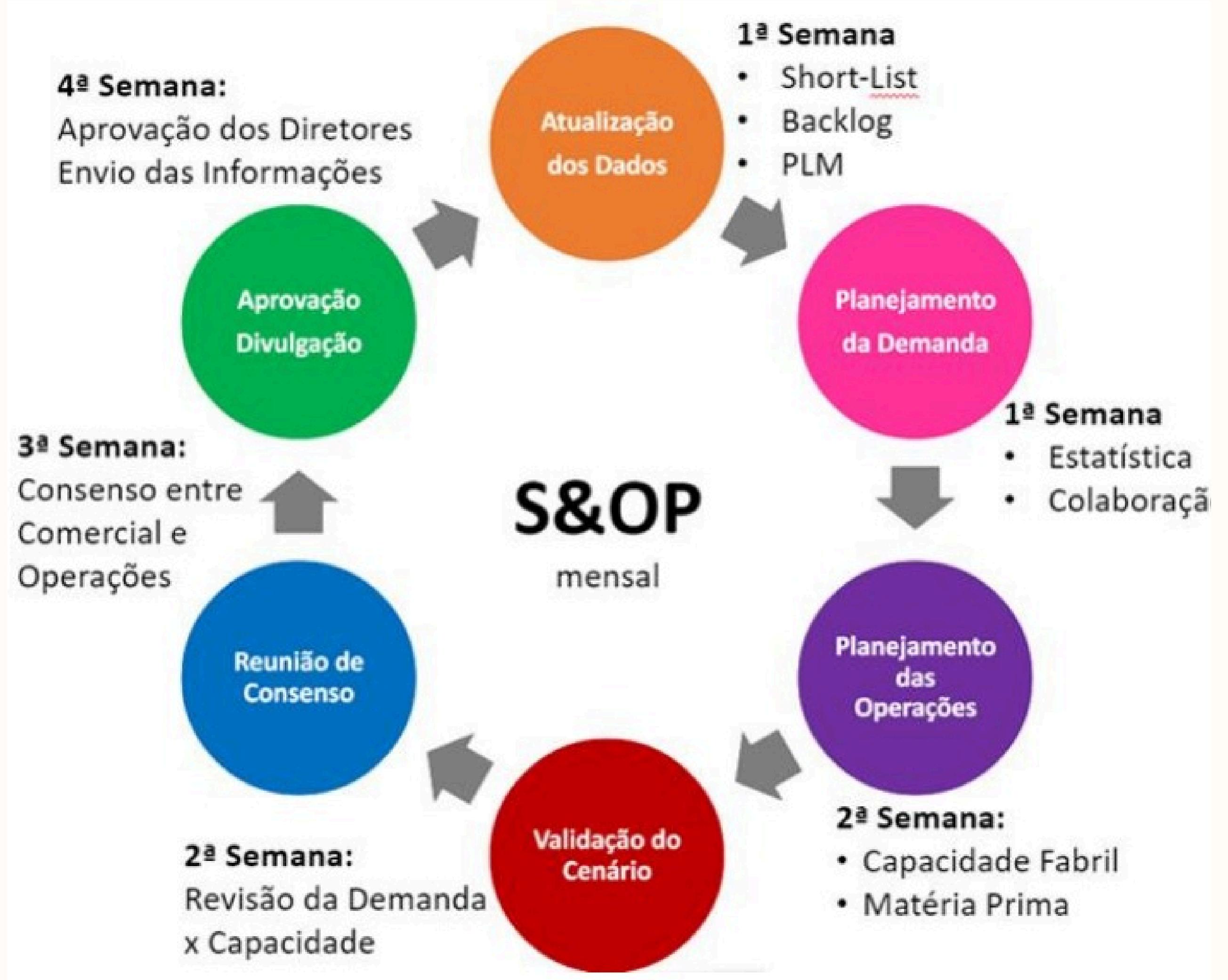


PREVISÃO DE DEMANDA

Furukawa Electric Latam









Escopo

Dados de 15 “hierarquias” modeladas individualmente

Previsões mensais do volume de venda

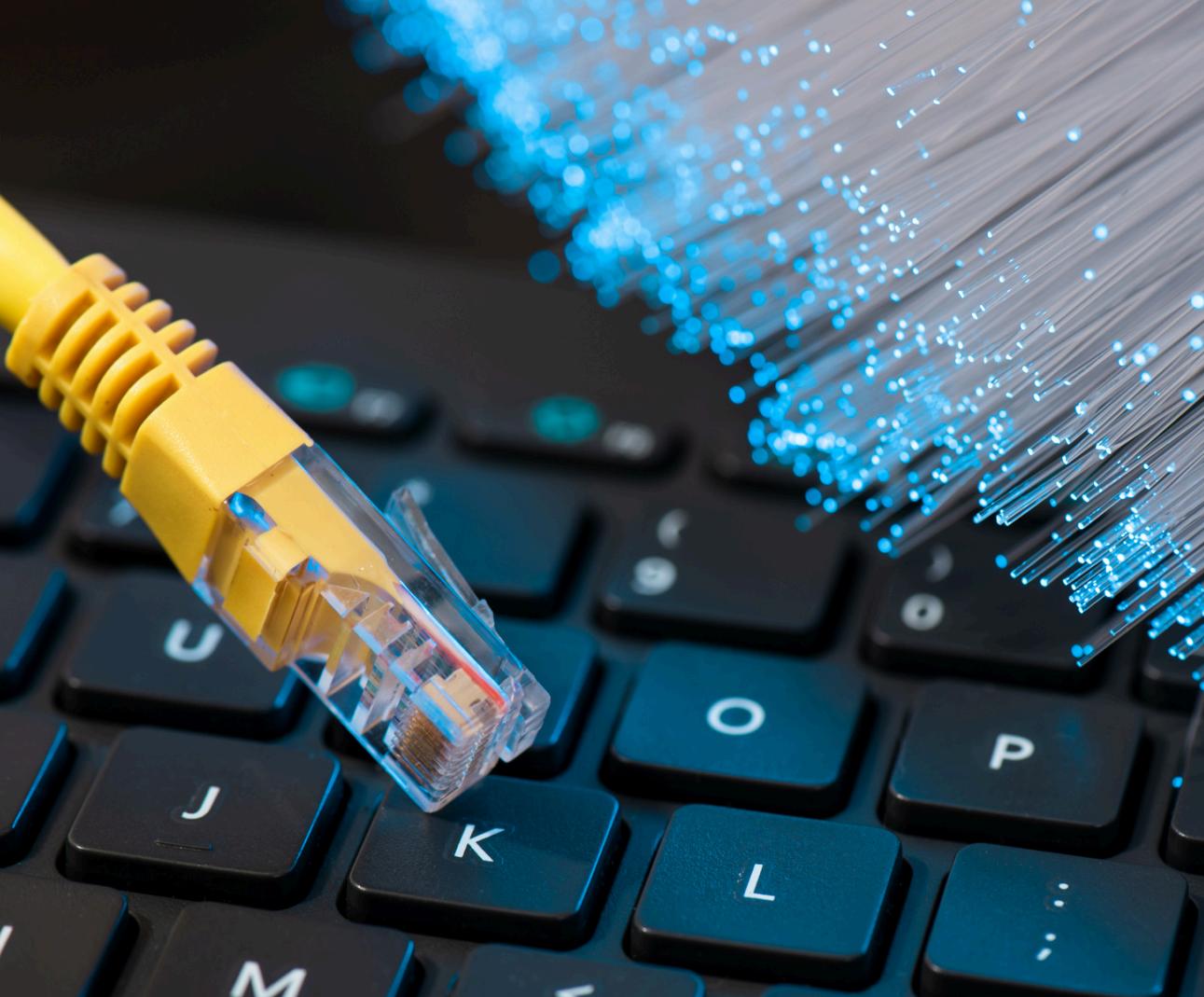
Unificação de Unidade de Medida

Covariáveis Externas:

Dolar | IPCA | SELIC | BSRIA | IBOVESPA | PIB | PIB Industrial | PIB Serviços | Cobre

Covariáveis Internas:

Faturamento | Valor líquido unitário | Volume de Estoque



Middle-out-approach

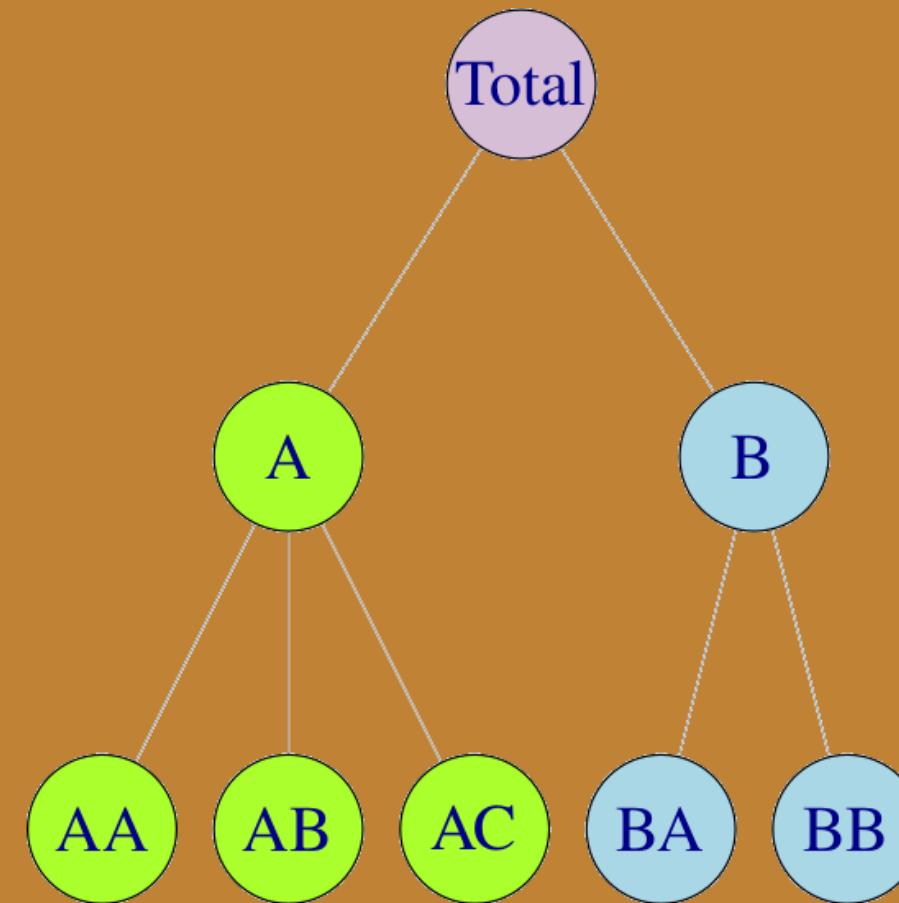
Você nem define 1 única série e depois distribui a previsão para todas abaixo
(Top down)

Nem prevê produto a produto e depois concilia tudo (Bottom up)

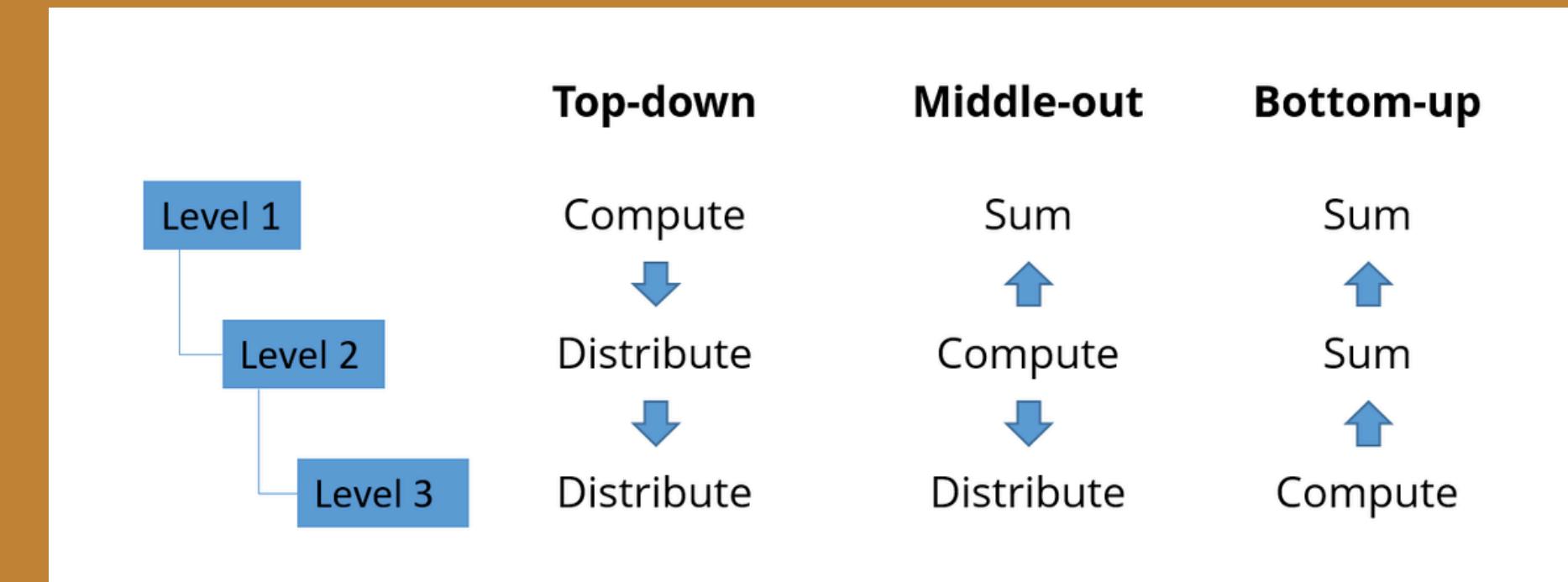
Ela fica no meio entre as duas abordagens: Foram escolhidos os produtos que representam maior % de venda de acordo com os respectivos mercados

As hierarquias são um agrupamento de produtos.

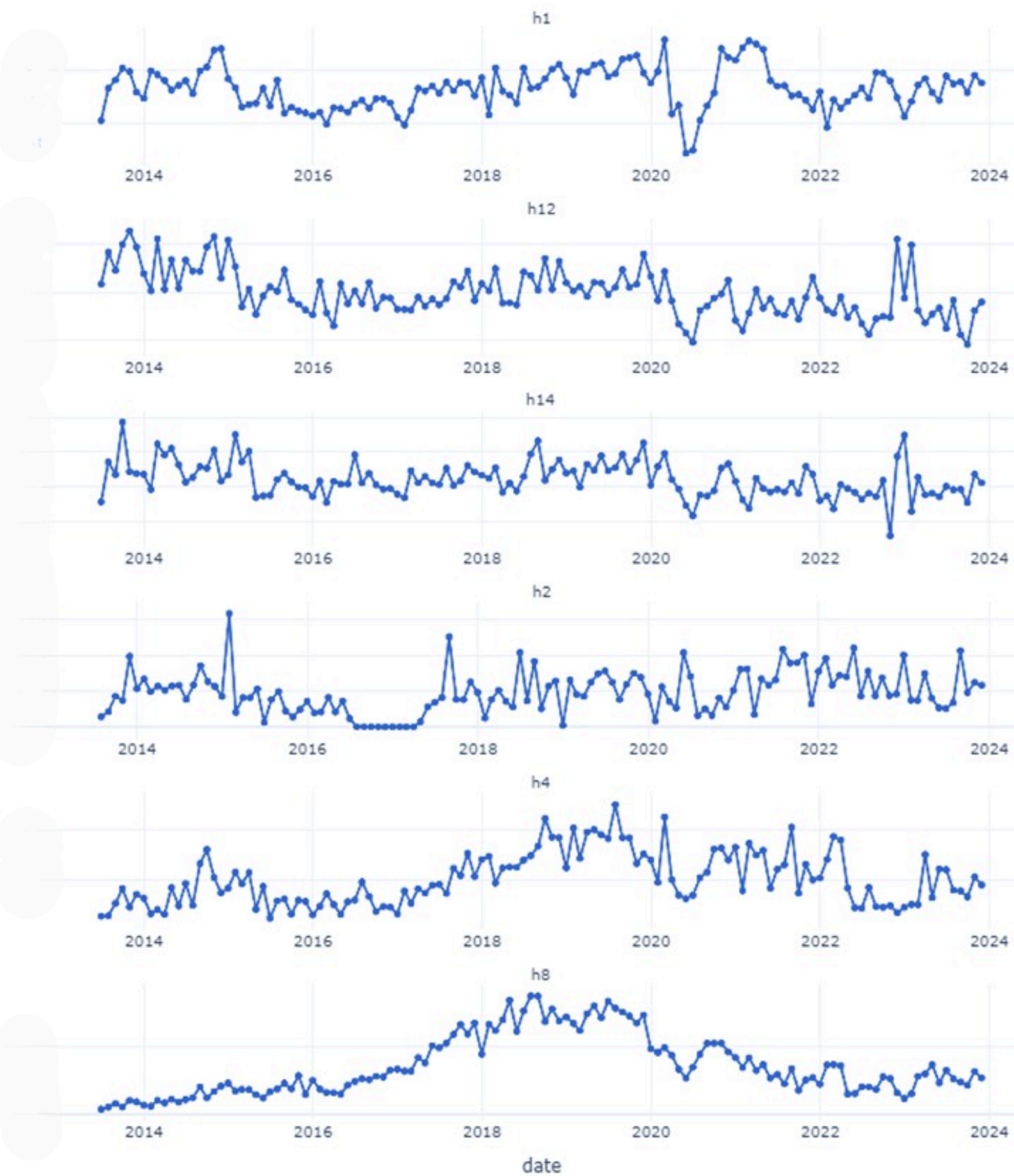
Ao final da previsão, faz-se se o rateio da previsão para a previsão de volume chegar a nível de produto.



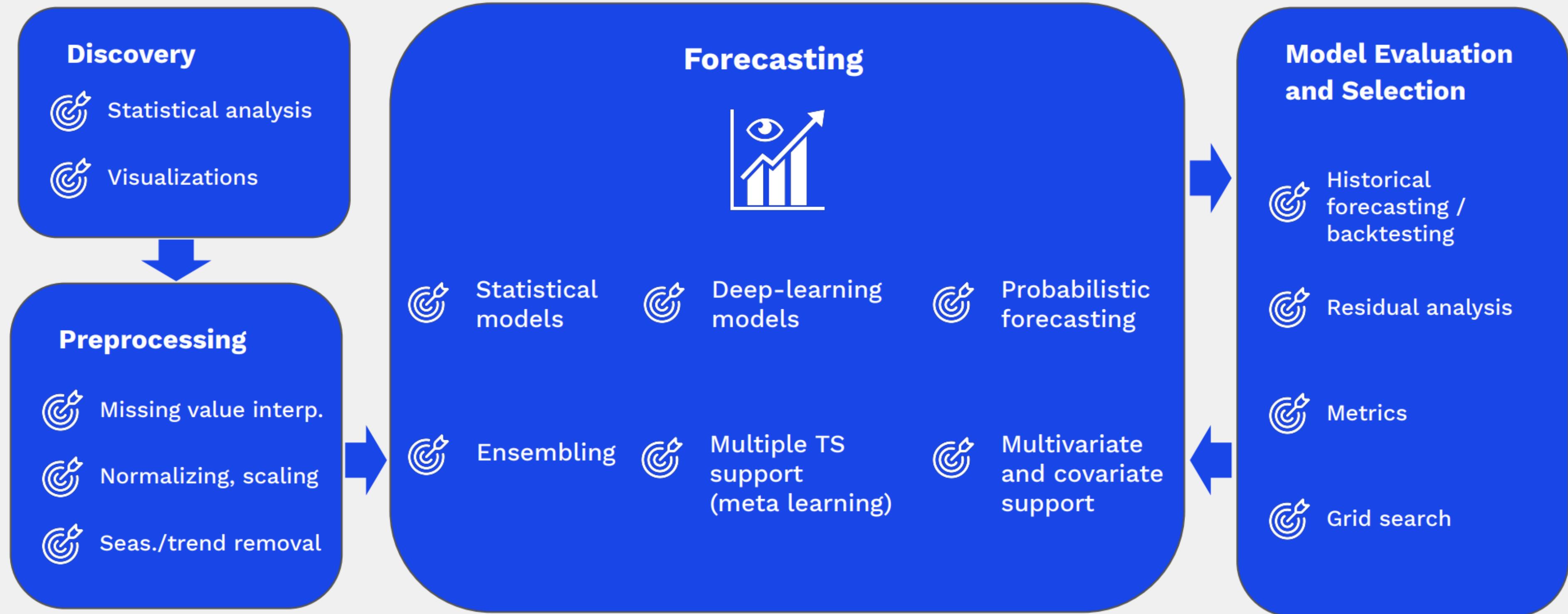
Foi escolhida a Middle-out-approach pois permite agrupar produtos que possuem a mesma matéria prima; diminuindo a complexidade de modelar a nível de produto, situação em que a volatilidade é ainda maior.



volume



Darts



PROCESSO DE TUNING:

15 HIERARQUIAS X
6 MODELOS X
HYPERPARÂMETROS X
11 COVARIÁVEIS X
6 LAGS

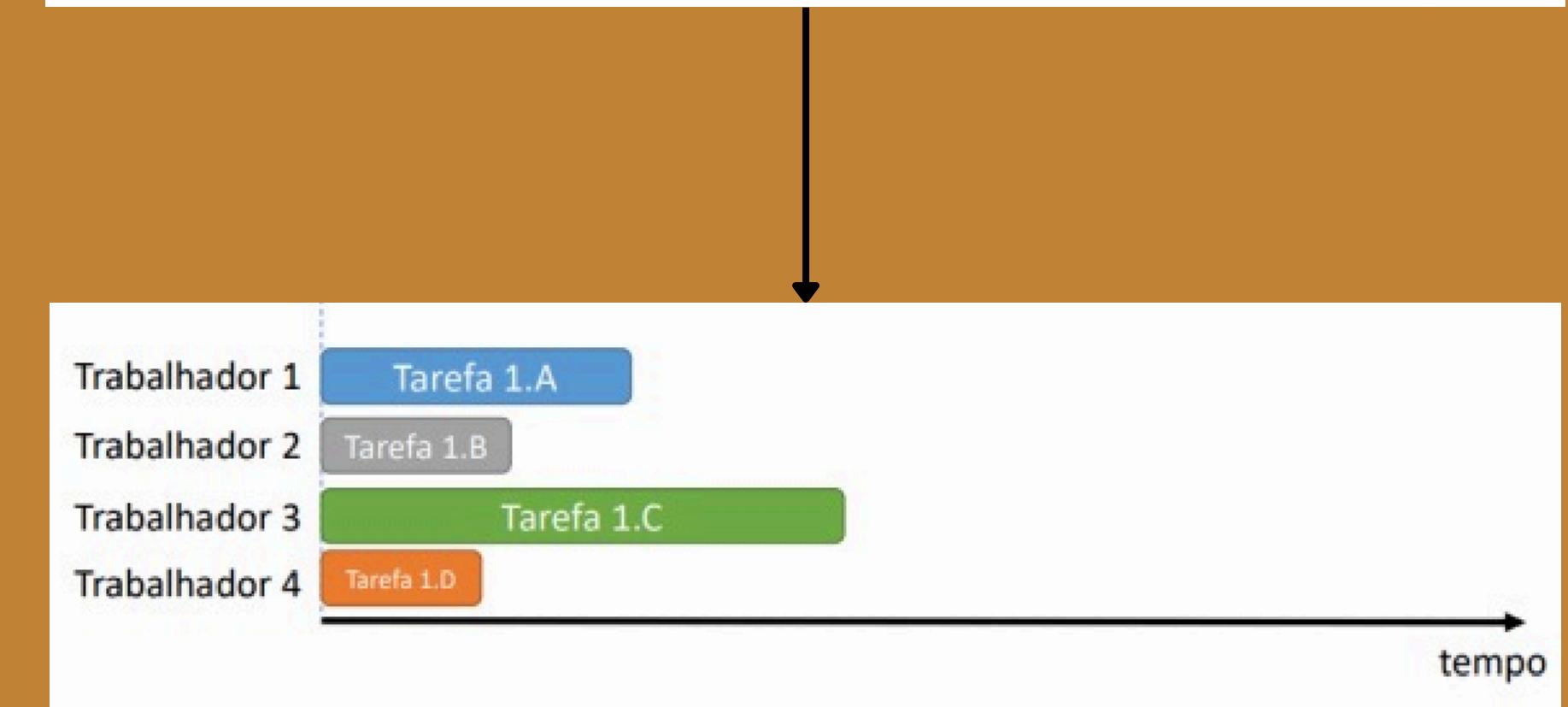
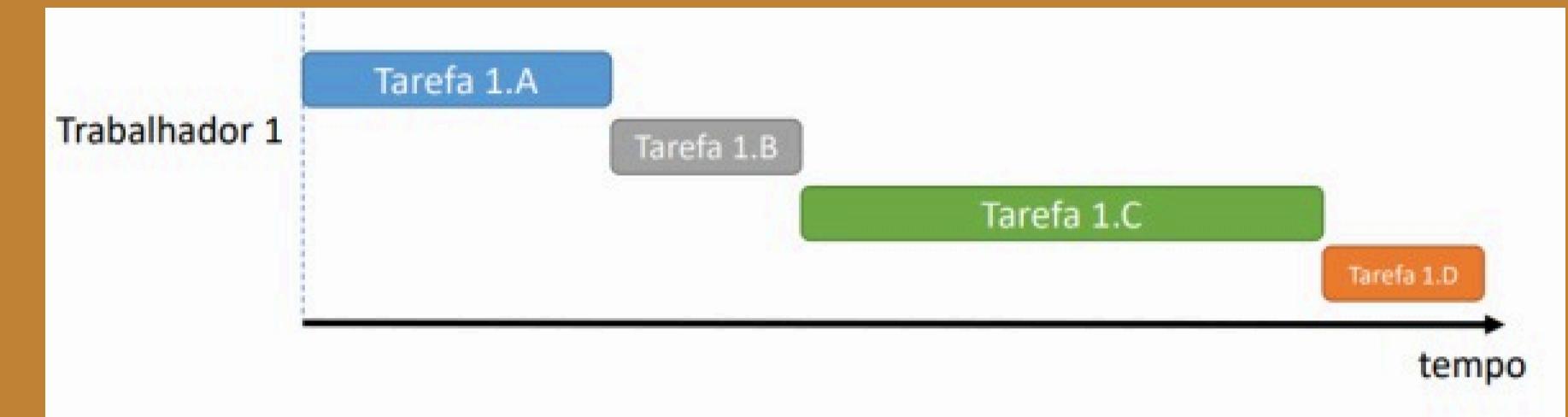
LAGS:

TARGET [3, 6, 12, 18, 24]
PASSADA [3, 6, 12, 18, 24]
FUTURA [6]

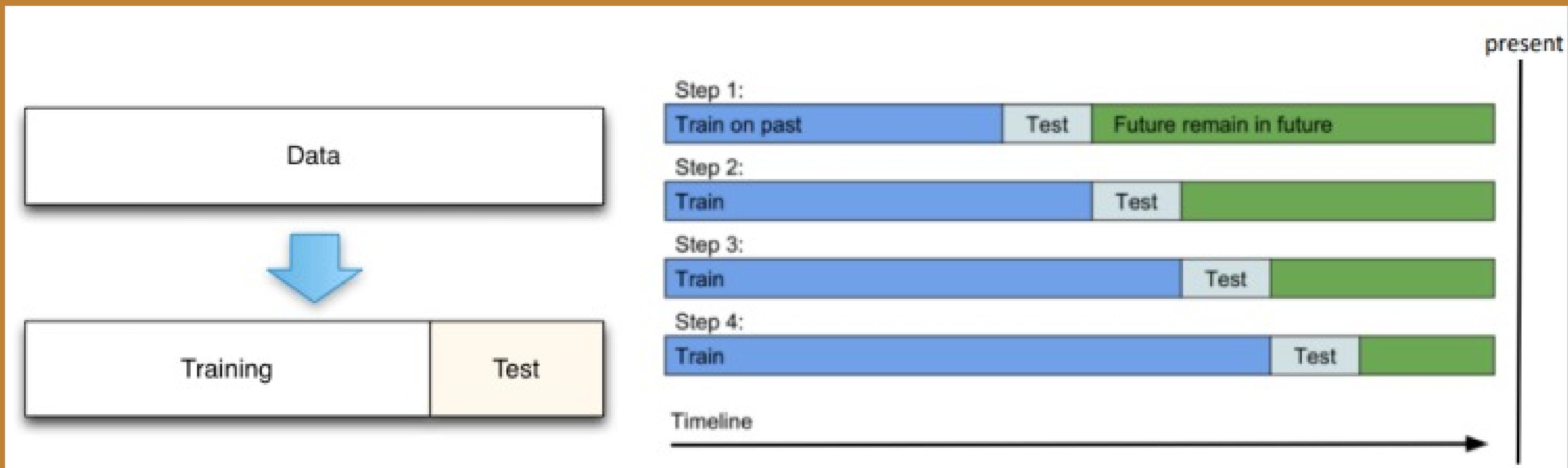
MODELOS:

KNEIGHBORS REGRESSOR
KERNEL RIDGE
SVR
RIDGE CV
LINEAR SVR
LINEAR REGRESSION

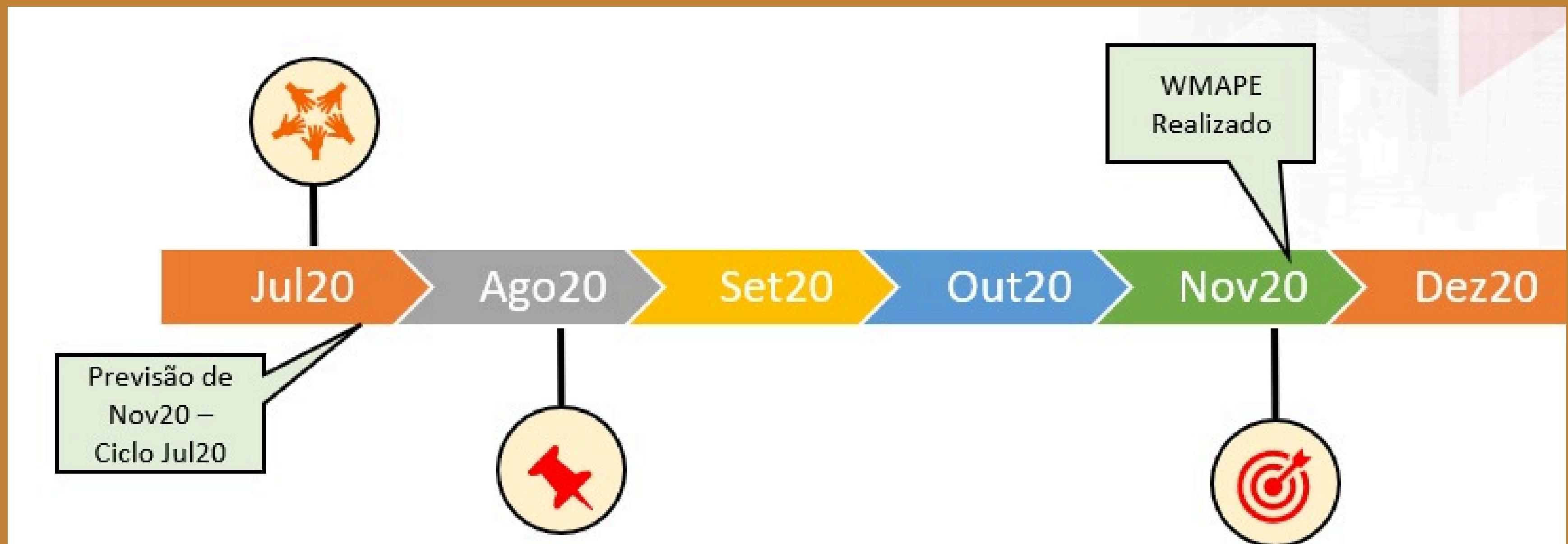
PARALELISMO



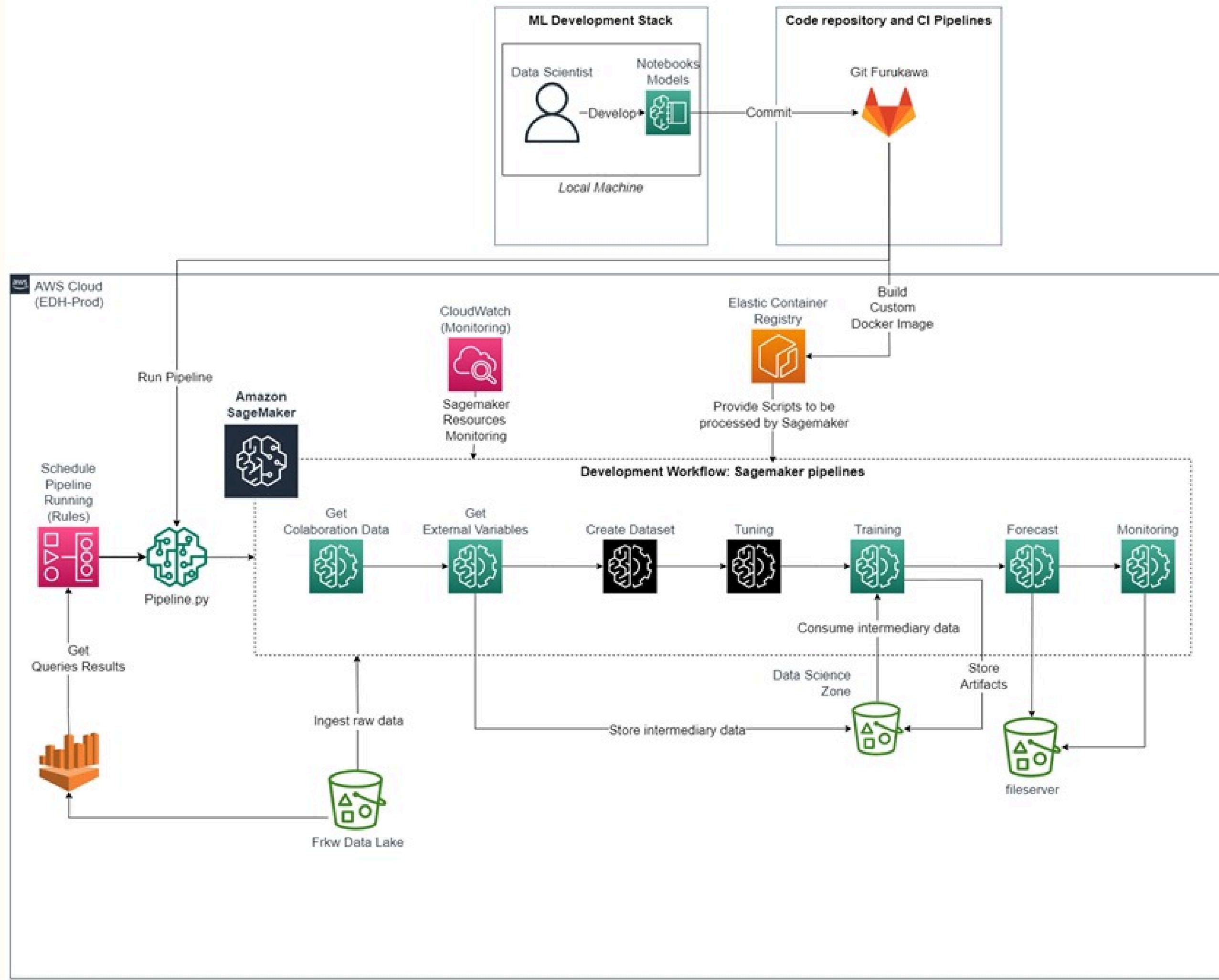
Time Series Cross Validation (TSCV)



Avaliação do Erro



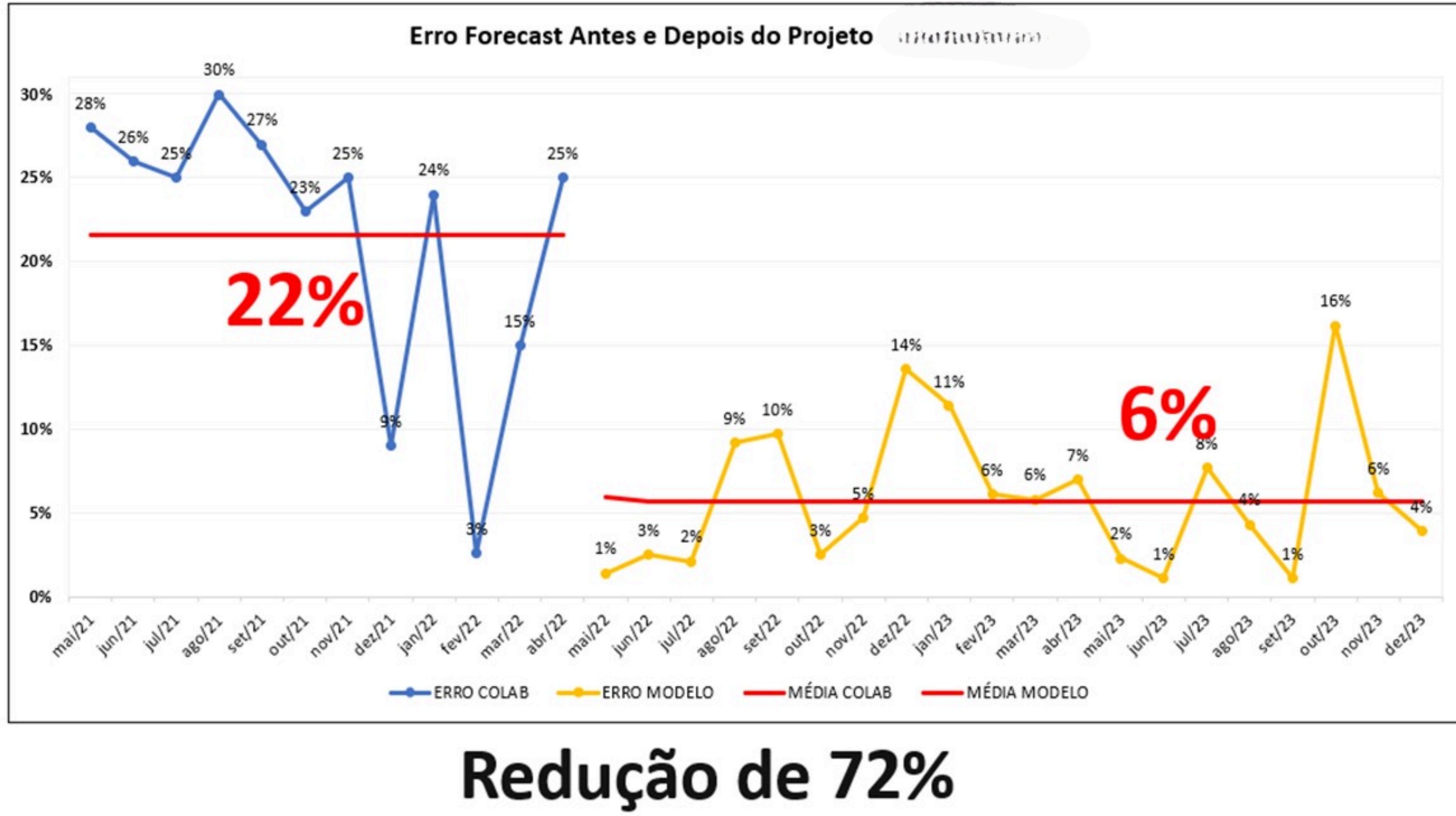
Infraestrutura na AWS

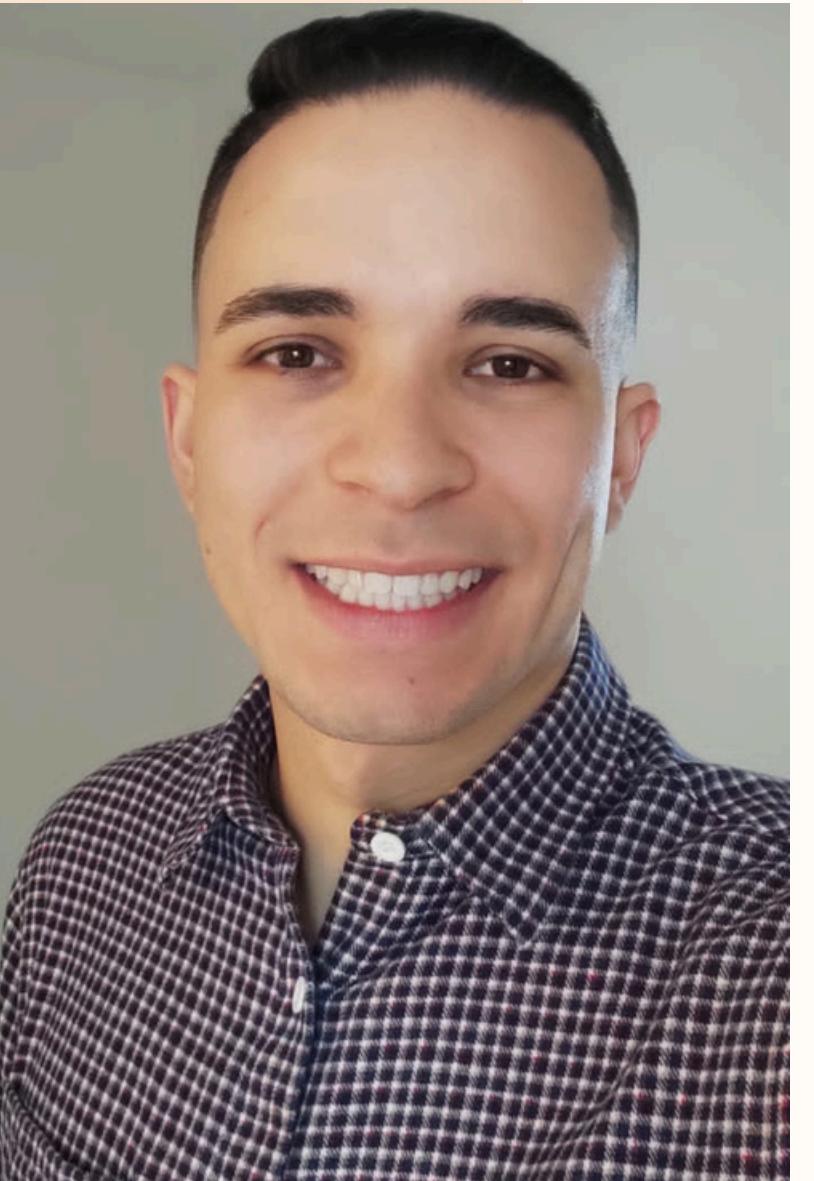


Erro Forecast Antes e Depois do Projeto -



Redução de 63%





OBRIGADO

Guilherme Parreira

 /guilhermeparreira
 /guilherme-parreira

