

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Rogério Souza Mascarenhas

**A importância de estimar o Auxílio Emergencial no contexto econômico dos
Municípios Fluminenses**

Rio de Janeiro

2021

Rogério Souza Mascarenhas

**A importância de estimar o Auxílio Emergencial no contexto econômico dos
Municípios Fluminenses**

Trabalho de Conclusão de Curso apresentado
no Curso de Especialização em Ciência de
Dados e Big Data como requisito parcial à
obtenção do título de especialista.

Rio de Janeiro
2021

Sumário

1. Introdução.....	4
1.1. Contextualização.....	4
1.2. O Problema Proposto	9
2. Coleta de Dados	10
3. Processamento/Tratamento de Dados.....	14
4. Análise e Exploração dos Dados.....	22
5. Criação de Modelos de Machine Learning.....	34
6. Apresentação dos Resultados.....	53
7. Links.....	64
REFERÊNCIAS.....	65

1. Introdução

O trabalho visa prever qual deveria ser o novo valor de Auxílio Emergencial (AE) a ser repassado às populações dos municípios do Estado do Rio de Janeiro, para combater os efeitos econômicos da pandemia de Covid-19, com base nos valores pagos pelos principais benefícios assistenciais: Bolsa Família (BF) e Benefício de Prestação Continuada (BPC), e considerando os repasses do Fundo de Participação dos Municípios (FPM). Neste sentido, o estudo foi estruturado em seis partes. Nesta Introdução (Capítulo 1), será apresentada a contextualização do assunto tratado e o problema proposto. Na sequência, serão apresentadas as informações acerca da coleta dos dados (Capítulo 2) e os passos realizados para o tratamento inicial, que consiste em sua limpeza (ou correção) e estruturação (Capítulo 3). No passo seguinte, serão efetuadas a análise e a exploração desses dados, com vistas à obtenção de informações estatísticas relevantes (Capítulo 4). Como o trabalho gira em torno de uma predição, o Capítulo 5 trata da aplicação de Modelos de Regressão Linear (técnica de *Machine Learning*), bem como das métricas para validação dos resultados. Por fim, no Capítulo 6 serão apresentados os resultados obtidos.

1.1. Contextualização

Após mais de um ano da chegada da pandemia de Covid-19 no Brasil é consenso a necessidade de continuar amparando o segmento da população mais vulnerável aos impactos econômicos da pandemia, considerando os cenários de incerteza em nosso horizonte, potencializados pelas novas variantes do Coronavírus, morosidade da vacina (em função da escassez do imunizante) e a baixa oferta de empregos.

Segundo um estudo da Fundação Getúlio Vargas (FGV) do Rio de Janeiro (GONZALES et al, 2020), o Governo Brasileiro, injetou um montante significativo de recursos, por meio do pagamento do **Auxílio Emergencial (AE)**, estabelecido pela Lei nº 13.982, de 2 de abril de 2020 (BRASIL, 2020), cujos efeitos não se limitaram ao apoio às famílias que receberam diretamente o benefício. Ao irrigar o consumo, os efeitos se estenderam também às empresas e ao emprego, contribuindo para uma queda menor do PIB nacional em 2020, que ficou em -4,1%, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE). Em rápida retrospectiva, o programa

durou de abril a dezembro de 2020 e foi desenhado para repassar 3 parcelas de R\$ 600 ou R\$ 1200 (para a mulher provedora ou chefe de família monoparental). Posteriormente, foram anunciadas mais 2 parcelas de mesmo valor e, após isso, mais 4 parcelas adicionais, cujo valor foi cortado pela metade (R\$ 300 ou R\$ 600).

Ainda – de acordo com o estudo, o AE foi pago a mais de 67 milhões de brasileiros, sendo mais de 5 milhões de pessoas atendidas no Rio de Janeiro, num total de 35% da população. Se em âmbito nacional a pobreza no Brasil diminuiu 3,74% (de julho e agosto de 2020), curiosamente, no Rio de Janeiro a faixa da população mais pobre no estado (considerada aquela que possui renda familiar de menos de $\frac{1}{2}$ salário mínimo) cresceu 1,55%, sendo o único estado em que esta situação se registrou. A despeito desta excepcionalidade, é inegável a essencialidade desta renda no cenário pandêmico, pois a situação econômica das famílias e municípios seria desalentadora. Cabe ressaltar que, segundo um relatório da Federação das Indústrias do estado do Rio de Janeiro (FIRJAN) (FIRJAN, 2021), em 2020, a atividade econômica fluminense recuou 3,8%, registrando uma retração menor que o no PIB nacional.

A importância econômica do AE para os municípios fluminenses pode ser percebida quando observamos a linha de arrecadação do principal consumo do país: o ICMS (Imposto sobre a Circulação de Mercadorias e Serviços de Transporte Interestadual e Intermunicipal e de Comunicação, de competência dos Estados-Membros e do Distrito Federal), que demonstra a relevância do repasse no fomento ao consumo em 2020. A queda da arrecadação no 1º quadrimestre é revertida a partir da entrada das primeiras parcelas do AE.

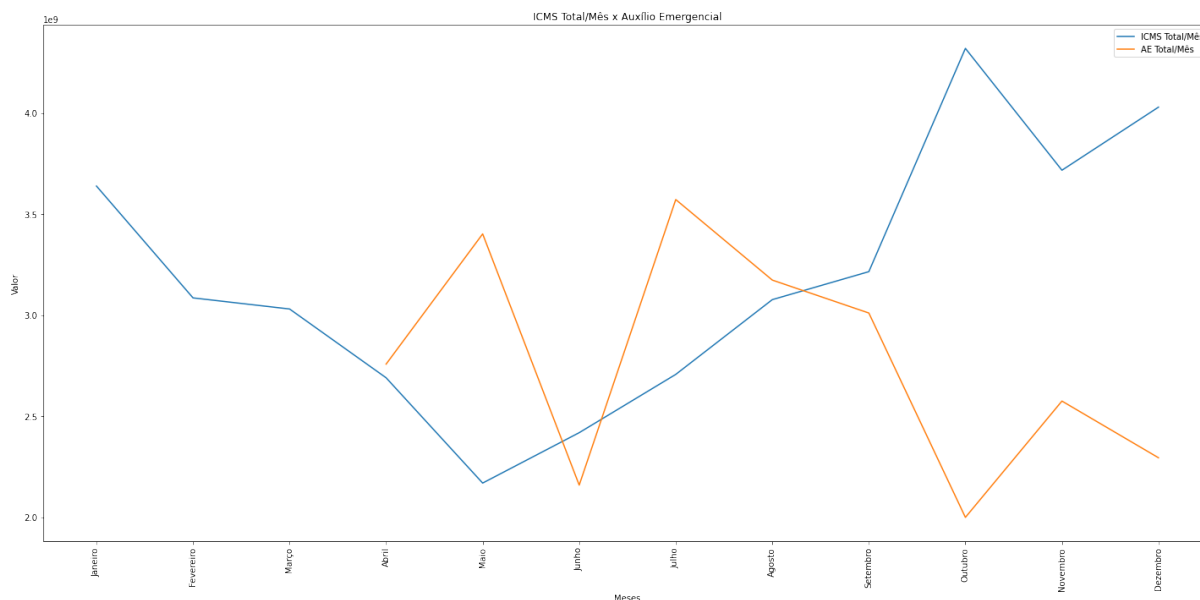


Figura 1.0 – Gráfico de Linha comparativo do pagamento do AE com a arrecadação de ICMS no Estado do Rio de Janeiro (em bilhões (R\$))

Também é possível perceber que, no último quadrimestre do ano, mesmo com a redução (em 50%) do benefício, há uma influência positiva na recuperação da arrecadação do tributo estadual, mantendo o perfil nos níveis de 2019. A arrecadação apresentou uma variação positiva de 4,76%, contrariando as previsões do estudo do Instituto Brasileiro de Planejamento e Tributação (IBPT) (IBPT, 2020), que previa uma queda da ordem de até 39%, considerando os cenários de isolamento social.

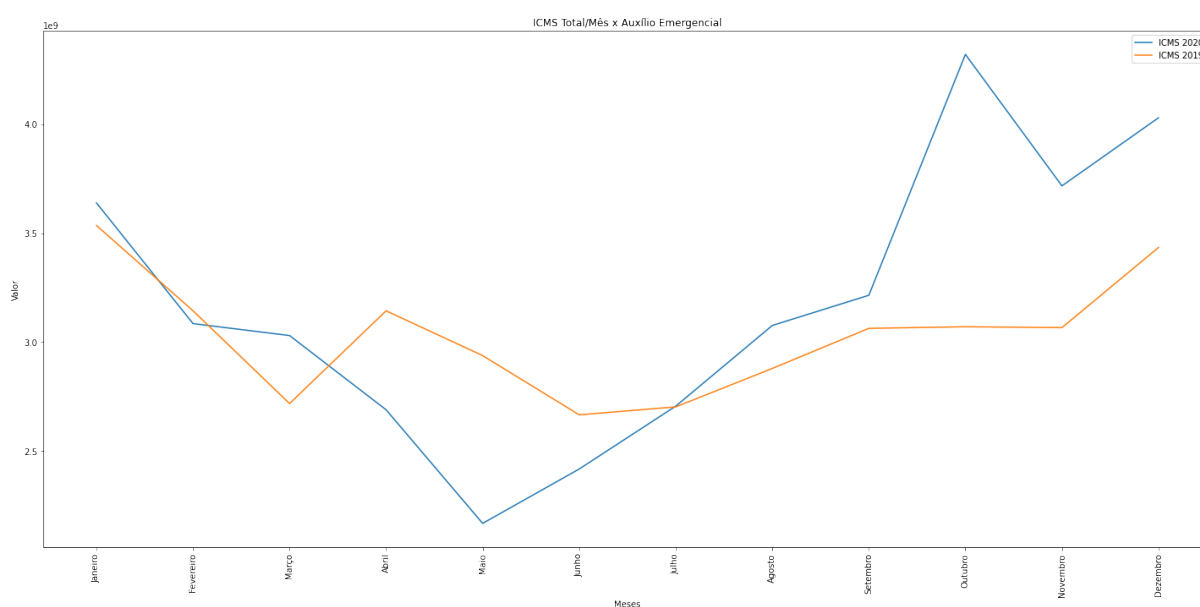
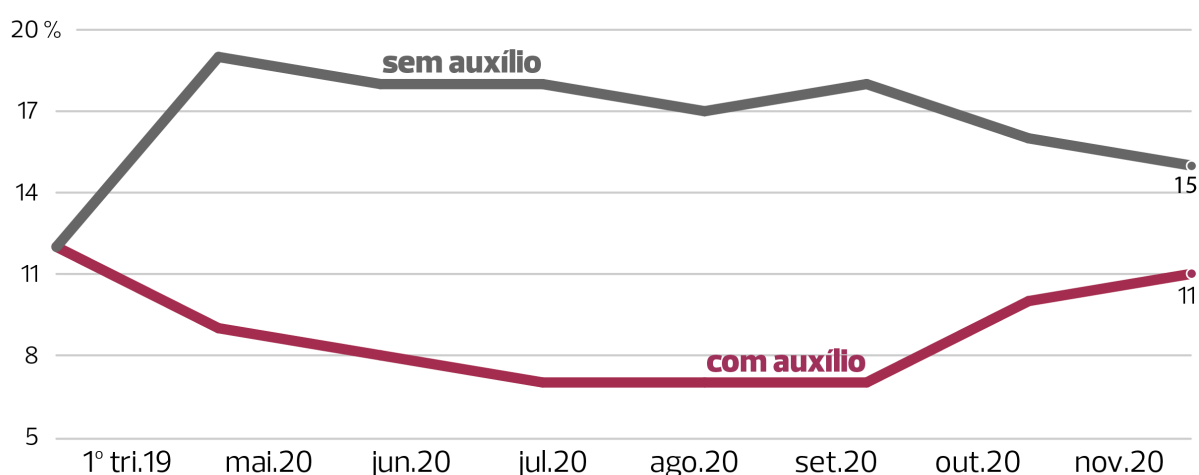


Figura 1.1 – Gráfico comparativo demonstrando o perfil da arrecadação de ICMS no Estado do Rio de Janeiro que totalizou R\$ 36.362.006.949,40 em 2019 e R\$ 38.093.118.748,75 em 2020 (em bilhões (R\$)).

No aspecto da renda das famílias no cenário nacional, segundo um documento do Instituto de Ensino e Pesquisa (INSPER) (MENEZES et al, 2021), estima-se que, durante a sua vigência, esta renda emergencial evitou que até 10% da população (cerca de 21 milhões de pessoas) passasse a viver na pobreza (sem renda suficiente para suprir suas necessidades básicas) e chegou perto de erradicar a miséria, sem falar no impacto econômico positivo para os Estados e Municípios.

Fatia da população brasileira em situação de pobreza com e sem auxílio emergencial



Fonte: Reducing Poverty and Inequality during the Coronavirus Outbreak: The Emergency Aid Transfers in Brazil (2020).

Insper

Figura 1.2 – Gráfico de Linha comparativo da fatia da população brasileira com e sem o AE (fonte: Insper)

É importante destacar que, no bojo da criação do AE, muito se discutiu se, ao invés de criar um outro benefício, o Governo não deveria fazer apenas fazer uma revisão do Bolsa Família (BF) que, pode se dizer, ao lado do Benefício de Prestação Continuada (BPC), são os principais benefícios assistenciais (sem contraprestação) existentes no Brasil. Por fim, o Governo optou por uma nova renda emergencial, popularmente chamada de “*coronavoucher*”, com regras de concessão mais flexíveis (e amplas) que o BF, entretanto não cumulativos (a família registrada no BF recebeu o mais vantajoso). O **Auxílio Emergencial (AE)** foi destinado aos cidadãos maiores de idade sem emprego formal, mas que estavam na condição de trabalhadores

informais, microempreendedores individuais (MEI) ou contribuintes da Previdência Social. Também era necessário ter renda familiar mensal inferior a $\frac{1}{2}$ salário mínimo per capita ou três salários mínimos no total e não ser beneficiário de outros programas sociais ou do seguro-desemprego.

Ainda sobre os benefícios de cunho assistencial, é importante contextualizar:

- o **Bolsa Família (BF)**, que segundo a CAIXA (agente operador do programa), atende a mais de 13,9 milhões de famílias cadastradas no Cadastro Único (CadÚnico). O programa (criado pela Lei nº 10.836, de 9 de janeiro de 2004) consiste na ajuda financeira às famílias extremamente pobres (definidas como aquelas que possuem renda per capita inferior a R\$ 89,00) ou pobres (definidas como aquelas que possuem renda per capita de R\$ 89,00 a R\$ 178,00) que tenham em sua composição gestantes e crianças ou adolescentes entre 0 e 17 anos e extremamente pobres (com renda per capita até R\$ 89,00). A contrapartida é que as famílias beneficiárias mantenham as crianças e os adolescentes entre 6 e 17 anos com frequência na escola, façam o acompanhamento de saúde das gestantes e mantenham a vacinação em dia dos menores e lactantes; e
- o **Benefício de Prestação Continuada (BPC)**, da Lei Orgânica da Assistência Social (LOAS), estabelecido pela Lei nº 8.742, de 7 de dezembro de 1993, também baseado no CadÚnico, é a garantia de um salário mínimo de benefício mensal à pessoa com deficiência e ao idoso que, comprovadamente, não possuem meios de prover à própria manutenção ou de tê-la provida por sua família. Para ter direito, é necessário que a renda por pessoa do grupo familiar seja menor que $\frac{1}{4}$ do salário mínimo.

Neste espaço de definição dos valores repassados diretamente à população ou aos municípios é importante citar um fundamental repasse da União feito às prefeituras: o **Fundo de Participação dos Municípios (FPM)**, composto de 22,5% da arrecadação do Imposto de Renda (IR) e do Imposto sobre Produtos Industrializados (IPI), que é a maneira como o Governo Federal repassa verbas para os municípios brasileiros. O percentual, dentre outros fatores, é determinado principalmente pela proporção do número de habitantes estimado anualmente pelo IBGE. Cada faixa de população determina os coeficientes de distribuição desse Fundo: que são variáveis em 16 faixas até o limite de 156.216 habitantes, mas - acima deste número - o

coeficiente máximo torna-se fixo. Estima-se que o FPM seja responsável por cerca de 60% das receitas disponíveis das prefeituras de municípios pequenos com até 5 mil habitantes, sendo, inclusive, utilizado como um importante “*benchmarking*” de avaliação da situação econômica dos municípios. Para se ter um exemplo, segundo um estudo da Associação Nacional dos Auditores Fiscais da Receita Federal do Brasil (ANFIP) (FRANÇA, 2019), o pagamento de benefícios previdenciários (aposentadorias, pensões, auxílios e salário-maternidade) efetuados pelo Instituto Nacional do Seguro Social (INSS) supera em cerca de 73% dos municípios o FPM. No Rio de Janeiro, este percentual é de 100% dos municípios fluminenses.

Neste contexto, o objetivo desta investigação (ou abordagem) é, a partir dos dados dos valores pagos à população por meio dos principais benefícios assistenciais: Bolsa Família (BF) e Benefício de Prestação Continuada (BPC), e considerando os repasses do Fundo de Participação dos Municípios (FPM), desenvolver um modelo capaz de prever (ou estimar) os valores de um novo Auxílio Emergencial (AE) para todos os municípios do Estado do Rio de Janeiro.

Para tal, aplicaremos técnicas de *Machine Learning*, mais detidamente de Aprendizado de Máquina Supervisionado: a Regressão Linear, para prever (ou estimar) o valor da nova renda emergencial recebida pela população dos municípios do Estado do Rio de Janeiro. Cabe ressaltar que a análise em questão não sofre influência do tempo, o que, dentre outras características, não nos levou a considerar os modelos de Séries Temporais.

1.2. O Problema Proposto

Com a finalidade de entender melhor a investigação, delimitando de forma mais precisa o escopo da análise, utilizaremos a técnica dos 5W para sistematizar melhor o estudo:

Why?

Neste cenário pandêmico, a predição contribui para os estudos do valor do AE a ser pago nos municípios do Estado do Rio de Janeiro, considerando a necessidade de manutenção do equilíbrio econômico destes entes e a sustentabilidade social.

Who?

Os dados objetos da análise são do Governo Federal do Brasil e foram disponibilizados como dados abertos no Portal da Transparência (da CGU), no Portal Brasileiro de Dados Abertos (do Ministério da Economia, ex-Planejamento), além do site do Instituto Brasileiro de Geografia e Estatística (IBGE) e da Secretaria de Estado de Fazenda do Rio de Janeiro (SEFAZ-RJ).

What?

A abordagem adotada será uma Análise Exploratória de Dados (AED) nas bases de dados dos benefícios assistenciais (BF e BPC) e dos repasses (FPM) do Governo Federal, buscando a construção de um modelo capaz de estimar os valores de um novo Auxílio Emergencial (AE) para os municípios do Estado do Rio de Janeiro. Para tal predição será utilizada a abordagem de Regressão Linear, no âmbito das técnicas de *Machine Learning*.

Where?

Os aspectos geográficos da análise são restritos aos municípios (92) do Estado do Rio de Janeiro.

When?

O período analisado contempla os benefícios pagos e repasses da União efetuados de abril a dezembro de 2020, período de duração da primeira fase do Auxílio Emergencial (AE).

2. Coleta de Dados

Os dados dos benefícios assistenciais: Auxílio Emergencial (AE), Bolsa Família (BF) e Benefício de Prestação Continuada (BPC); e dos repasses do Fundo de Participação dos Municípios (FPM) utilizados no estudo, estão disponíveis respectivamente nos sites do Governo Federal do Brasil, nos endereços/sites:

- Portal da Transparência (que é um site de acesso livre, gestado pela CGU, no qual o cidadão pode encontrar informações sobre como o dinheiro público é utilizado, além de se informar sobre assuntos relacionados à gestão pública do país):

<https://www.portaltransparencia.gov.br/beneficios>;

- Portal de Dados Abertos (que é um site, gestado pelo Ministério da Economia, onde todos possam encontrar e utilizar os dados e informações públicas. Um dos principais objetivos é promover a interlocução entre atores da sociedade com o governo para pensar a melhor utilização dos dados, promovendo impactos positivos sob os pontos de vista social e econômico): <https://dados.gov.br/dataset/transferencias-constitucionais-para-municipios>;

Além destes Portais, foram obtidos no site do Instituto Brasileiro de Geografia e Estatística - IBGE (<https://www.ibge.gov.br/cidades-e-estados/rj.html>), por meio da opção <EXPORTAR> os dados gerais (geográficos, demográficos, econômicos, etc) dos municípios fluminenses, no formato CSV, para enriquecimento dos estudos. Também foi utilizado o Sistema IBGE de Recuperação Automática – SIDRA para obtenção dos dados do PIB dos Municípios (base 2018), pela seleção da UF=33.RJ na tabela número 5938 (<https://sidra.ibge.gov.br/tabela/5938#resultado>).

Outra base complementar utilizada no trabalho foi a de “Dados sobre a Arrecadação” do ICMS (base 2019 e 2020) dos municípios fluminenses no Portal da Secretaria de Estado de Fazenda do Rio de Janeiro (http://www.fazenda.rj.gov.br/sefaz/faces/menu_structure/servicos).

A tabela abaixo apresenta características dos arquivos utilizados:

ORIGEM/ARQUIVO	ÓRGÃO/ENTIDADE RESPONSÁVEL PELO ENVIO	PERIODICIDADE	DADOS ATUALIZADOS ATÉ
Auxílio Emergencial (Dados disponíveis desde 04/2020)	Ministério da Cidadania	Mensal	12/2020
Caixa Econômica Federal – Bolsa Família (Dados disponíveis desde 01/2013)	Ministério da Cidadania	Mensal	12/2020

Folha de Pagamentos do Instituto Nacional do Seguro Social – BPC (Dados disponíveis desde 01/2019)	Ministério da Cidadania	Mensal	01/2021
Transferências Constitucionais para Municípios (Dados disponíveis desde 01/1997)	Ministério da Economia	Mensal (segmentado por decêndios)	01/2021
Dados Gerais do Portal das Cidades@ (Histórico de disponibilidade não informada)	IBGE	Anual	01/2021
PIB (Dados disponíveis desde 2002)	IBGE (SIDRA)	Anual	12/2020
Arrecadação Geral do ICMS por município (Dados disponíveis desde 2011)	Secretaria de Estado de Fazenda do Rio de Janeiro	Anual	2020

Figura 2.0 – Quadro de disponibilização dos dados no Portal da Transparência, Portal de Dados Abertos, IBGE e SEFAZ-RJ.

Destacamos também os campos/colunas dos principais arquivos de trabalho (insumos para a predição) utilizados:

1. Portal da Transparência:

(<http://www.portaltransparencia.gov.br/beneficios/consulta?paginacaoSimples=true&tamanhoPagina=&offset=&direcaoOrdenacao=asc&colunasSelecionadas=linkDetalhamento%2ClinguagemCidada%2CmesAno%2Cuf%2Cmunicipio%2Cvalor&ordenarPor=mesAno&direcao=desc>)

Auxílio Emergencial (AE)

UF	Município	Mês/Ano	Valor Disponibilizado (R\$)	Programa Social
= RJ	***	04 – 12/2020	***	= Auxílio Emergencial

A extração foi baixada/carregada como *auxilio_emergencial.csv* (no Google Colab)

Bolsa Família (BF)

UF	Município	Mês/Ano	Valor Disponibilizado (R\$)	Programa Social
= RJ	***	04 – 12/2020	***	= Bolsa Família

A extração foi baixada/carregada como *bolsa_familia.csv* (no Google Colab)

Benefício de Prestação Continuada (BPC)

UF	Município	Mês/Ano	Valor Disponibilizado (R\$)	Programa Social
= RJ	***	04 – 12/2020	***	= BPC

A extração foi baixada/carregada como *bpc.csv* (no Google Colab)

Obs: É importante destacar que não foram considerados, no estudo, os valores dos demais benefícios (Garantia-Safra, PETI e Seguro-Defeso) disponibilizados, no Portal da Transparência, pela irrelevância dos valores ou pelo seu caráter sazonal.

2. Portal Brasileiro de Dados Abertos:

(<https://dados.gov.br/dataset/transferencias-constitucionais-para-municipios>)

Fundo de Participação dos Municípios (FPM)

Município	UF	Ano	Mês	1º Decêndio	2º Decêndio	3º Decêndio	Item transferência	Transferência
***	= RJ	2020	4-12	***	***	***	= FPM	= FPM

Os arquivos foram carregados como *TransferenciaMensalMunicipios2020XX.csv*, onde XX varia de 04 a 12 (no Google Colab)

Obs: É importante destacar que, pela natureza do estudo (adoção do FPM como “*benchmark*”) não foram considerados os valores dos demais repasses/transferências constitucionais (FEX, ITR, Fundeb, Fundef, LC 87/96 (Lei Kandir), CIDE-Combustíveis, IOF-Ouro, Royalties e AFM) constantes do arquivo.

3. Processamento/Tratamento de Dados

O estudo foi desenvolvido utilizando-se a linguagem de programação interpretada Python, no Google Colaboratory (que é um ambiente de notebooks Jupyter que não requer configuração e é executado na nuvem), utilizando as bibliotecas citadas abaixo:

```
import pandas as pd
import geopandas as gpd
from google.colab import drive
import logging
import matplotlib.pyplot as plt
import folium
import numpy as np
import scipy
-

import io
import html
-

from scipy.stats import variation
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.svm import LinearSVR
from sklearn.neighbors import KNeighborsRegressor
from statsmodels.api import OLS
-

from sklearn.model_selection import KFold
from sklearn.model_selection import ShuffleSplit
-

from sklearn.metrics import mean_squared_error
-

from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_percentage_error
-

import statsmodels.api as sm
-

from autorank import autorank, plot_stats, create_report, latex_table
```

Figura 3.0 – Principais bibliotecas, funções e classes utilizadas no projeto.

Estas bibliotecas, funções e classes foram carregadas ao longo do desenvolvimento do trabalho.

Preparação dos Dados e Pré-processamento:

A primeira fase do trabalho foi a extração/baixa, no site do Portal da Transparência, dos dados relativos aos benefícios de cunho assistencial (AE, BPC e BF) por município (do Estado do Rio de Janeiro), considerando o período do pagamento do AE (de abril a dezembro/2020). O próprio site disponibiliza uma opção de extração/baixa dos dados (por campos/atributos de interesse) no formato CSV (*Comma-Separated Values*). Os dados foram consolidados em um objeto *dataframe* da biblioteca *pandas* segmentado por município, mês/ano, valor (por tipo de benefício).

Ainda nesta primeira fase, o desafio maior foi o “pré-tratamento” dos dados do FPM obtidos no Portal Brasileiro de Dados Abertos, pois foi necessária uma programação adicional para preparação dos dados de trabalho visto que os arquivos disponibilizados no Portal são segmentados por decêndio, dentro do mês, além de incluir vários outros repasses que extrapolavam o escopo da investigação. Os dados também foram baixados no formato CSV.

Durante o processo de leitura e concatenação dos arquivos (8) do FPM, foi necessário ajustar o parâmetro *encoding* para “ISO-8859-1”, pois o padrão dessa função, UTF-8, apresentou erro. Neste processo, além de ajustar o mês para 2 dígitos, também foi necessário acertar a grafia de 2 municípios (de “PARATI” para “PARATY” e de “ARMAÇÃO DE BÚZIOS” para “ARMAÇÃO DOS BÚZIOS”) que apresentaram divergência em relação à grafia dos arquivos base (dos benefícios assistenciais).

A partir daí temos o “Pré-processamento”, com a consolidação dos arquivos num único *dataframe*, que também demandou uma transformação ou ajuste dos valores para *float*.

	Município	Mês/Ano	BPC	BOLSA FAMILIA	AUXÍLIO EMERGENCIAL	FPM
0	COMENDADOR LEVY GASPARIAN	12/2020	78375.00	127396.00	1040044.00	914002.63
1	COMENDADOR LEVY GASPARIAN	10/2020	79420.00	127436.00	862543.00	430504.81
2	COMENDADOR LEVY GASPARIAN	11/2020	79434.18	127437.00	1120382.00	570601.49

Figura 3.2 – “Dump” do arquivo geral do IBGE tratado com dados gerais dos municípios do Rio de Janeiro

Ainda no ambiente do IBGE, foram obtidos, pela opção de download própria (e parametrizada), no Sistema IBGE de Recuperação Automática – SIDRA (<https://sidra.ibge.gov.br/tabela/5938>), os dados do PIB dos municípios do Rio de Janeiro (atualizado até 2018). A baixa foi feita como *PIB dos Municípios 2018 Tabela 5938_IBGE.csv*, no Google Colab. Houve a necessidade de ajustar os nomes dos municípios para maiúsculo e a retirada da referência da UF “(RJ)”.

Município	PIB (2018)
ANGRA DOS REIS	8936325
APERIBÉ	181448
ARARUAMA	2671410
...	...

Figura 3.3 – “Dump” do arquivo de PIB dos municípios (base IBGE)

Finalizando esta fase de complementação das informações, foram obtidos no Portal da Secretaria de Estado de Fazenda do Rio de Janeiro (http://www.fazenda.rj.gov.br/sefaz/faces/menu_structure/servicos?_afLoop=43027467854185016&datasource=UCMServer%23dDocName%3AWCC275242&_adf.ctrl-state=12wwwrwmnf_186), de forma direta, os valores de arrecadação do ICMS dos municípios fluminenses, que foram carregados (no Google Colab), como *icms_XXXX_modificado.csv* (onde XXXX indica os anos de 2019 e 2020). Neste arquivo, houve a necessidade de excluir manualmente o cabeçalho com informações/títulos como 'Governo do Estado do Rio de Janeiro' e as linhas que separavam os dados por (9) regiões.

Município	Janeiro	Fevereiro	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro	Total Anual
ARARUAMA	R\$ 2.063.886,94	R\$ 1.556.280,63	R\$ 1.735.636,50	R\$ 1.333.368,64	R\$ 1.849.151,85	R\$ 1.814.926,58	R\$ 1.996.168,42	R\$ 1.943.139,80	R\$ 2.515.786,24	R\$ 2.318.472,37	R\$ 2.445.751,00	R\$ 2.174.517,11	R\$ 23.747.086,08
ARMAÇÃO DOS BÚZIOS	R\$ 650.436,92	R\$ 677.400,24	R\$ 569.249,91	R\$ 281.072,82	R\$ 327.138,82	R\$ 298.011,61	R\$ 316.675,94	R\$ 421.198,59	R\$ 349.485,56	R\$ 418.500,73	R\$ 544.579,87	R\$ 464.186,41	R\$ 5.317.937,42

ARRAIAL DO CABO	R\$ 278.131,00	R\$ 270.669,61	R\$ 231.316,25	R\$ 172.449,53	R\$ 147.884,34	R\$ 192.509,13	R\$ 141.749,93	R\$ 157.039,97	R\$ 158.708,25	R\$ 157.464,82	R\$ 159.389,78	R\$ 187.640,51	R\$ 2.254.953,12
CABO FRIO	R\$ 5.265.748,44	R\$ 4.011.183,02	R\$ 3.348.951,56	R\$ 1.928.653,65	R\$ 3.205.321,00	R\$ 2.879.085,17	R\$ 5.295.967,59	R\$ 4.856.657,93	R\$ 5.236.438,51	R\$ 4.523.758,56	R\$ 5.589.750,64	R\$ 5.498.503,22	R\$ 51.640.019,29
...

Figura 3.4 – “Dump” do arquivo da arrecadação geral de ICMS distribuída por município (Exercício 2020)

Análise Gráfica e Georeferenciamento:

Ainda nesta fase preliminar de estudo (ou exploração preliminar dos dados), optou-se por uma análise gráfica (utilizando o mapa do Estado do Rio de Janeiro) da representatividade ou impacto dos benefícios assistenciais e repasses constitucionais para cada município. Para tal, fizemos uso do Shapefile (que é um formato aberto e popular de arquivo contendo dados geoespaciais em forma de vetor usado por Sistemas de Informações Geográficas - SIG) do IBGE, com informações de coordenadas dos municípios fluminenses (no http://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2020/UFs/RJ/), em conjunto com o GeoPandas (que é uma extensão do Pandas, para possibilitar e facilitar o trabalho com dados espaciais e georreferenciados).

Nesta fase de geração dos mapas, por conta de inconsistências (entre a base de estudo e do IBGE) na grafia do município de “TRAJANO DE MORAIS”. O nome foi retificado por “TRAJANO DE MORAES”.

Efetuada a correção da inconsistência, o passo posterior foi obter a média (dos 8 meses) dos benefícios assistenciais e repasses constitucionais feitos por município.

Nesta pré-análise gráfica, já temos uma sinalização (visual) dos maiores recebedores de valores e dos padrões de distribuição no Estado.

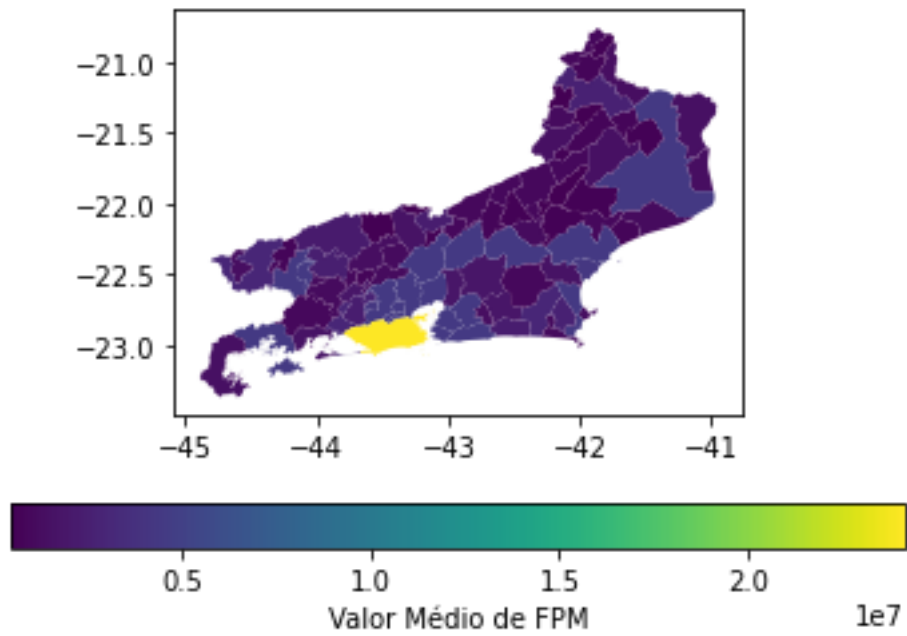


Figura 3.5 – Visão das transferências constitucionais do FPM pelos municípios do RJ (de abril a dezembro/20) em 10 milhões (R\$)

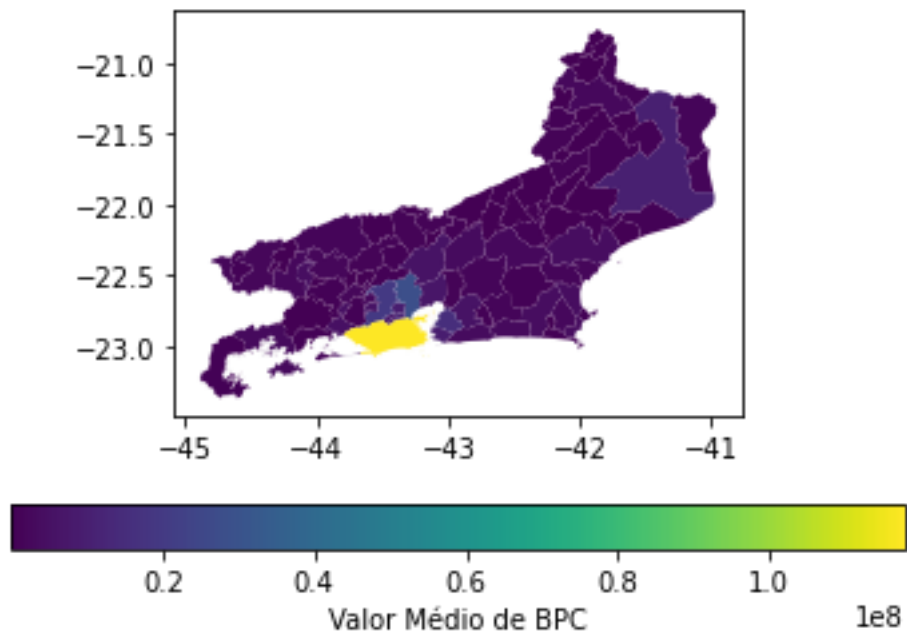


Figura 3.6 – Visão dos pagamentos do BPC pelos municípios do RJ (de abril a dezembro/20) em 100 milhões (R\$)

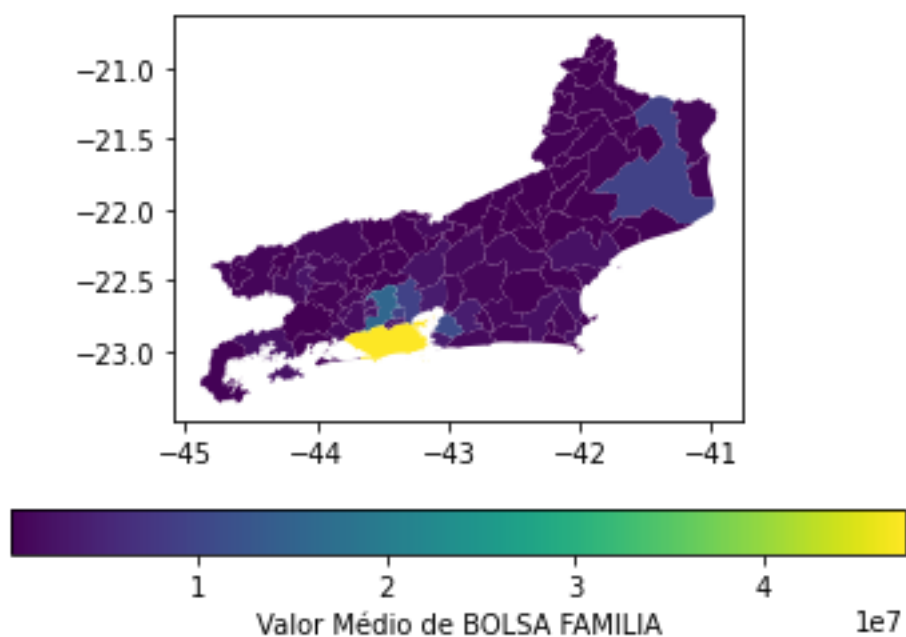


Figura 3.7 – Visão dos pagamentos do Bolsa Família pelos municípios do RJ (de abril a dezembro/20) em 10 milhões (R\$)

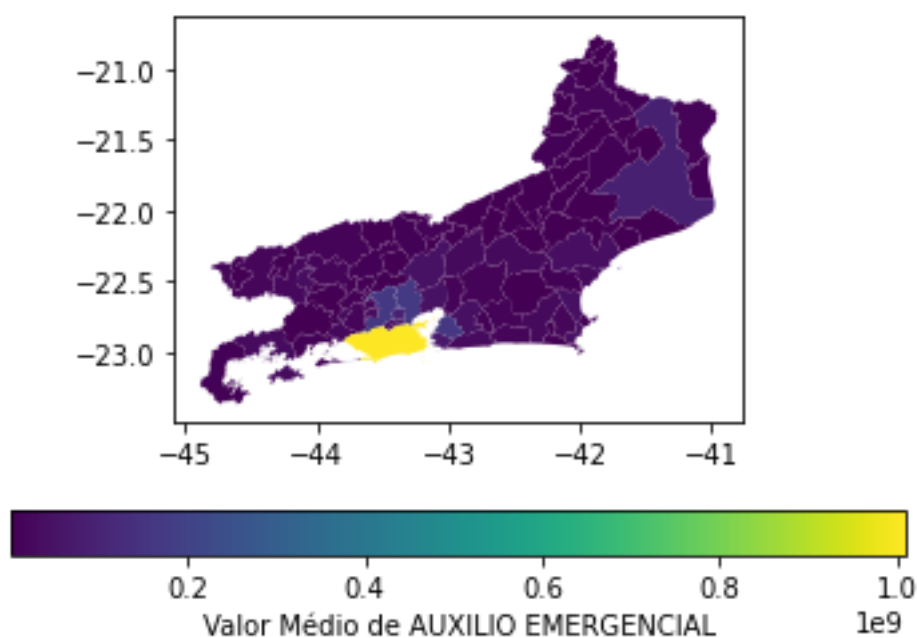


Figura 3.8 – Visão dos pagamentos do AE pelos municípios do RJ (de abril a dezembro/20) em bilhões (R\$)

Na busca por elementos significativos no estudo do AE, ainda a partir dos dados do IBGE, foi possível algumas visualizações como a observação do Índice de

Desenvolvimento Humano Municipal (IDHM) fluminense. Embora prejudicado pela não realização do último Censo (em 2020), o IDHM, do Programa das Nações Unidas para o Desenvolvimento (PNUD), é um índice composto que agrega 3 das mais importantes dimensões do desenvolvimento humano: a oportunidade de viver uma vida longa e saudável, de ter acesso ao conhecimento e ter um padrão de vida que garanta as necessidades básicas, representadas pela saúde, educação e renda.

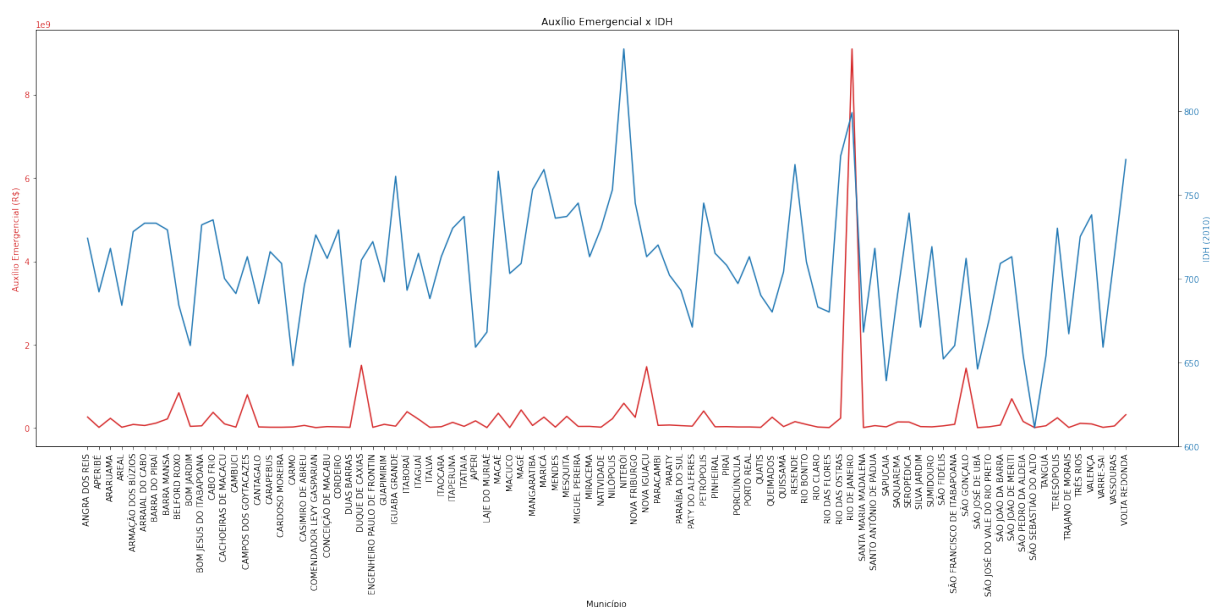


Figura 3.9 – Visão, em vermelho, do AE pagos nos municípios do RJ (de abril a dezembro/20) em bilhões (R\$) e, em azul, do IDHM, do PNUD

Nesta busca por correlações adicionais, uma nova visualização foi a do AE x PIB (Produto Interno Bruto), que mensura a atividade econômica de uma região, no caso dos municípios, por meio de cálculos de oferta e demanda de bens e serviços. No caso, podemos comparar o desempenho econômico dos municípios fluminenses com os valores percebidos pelo auxílio. A defasagem de data do PIB (2018) com o AE (2020) não permitiu uma análise mais consistente.

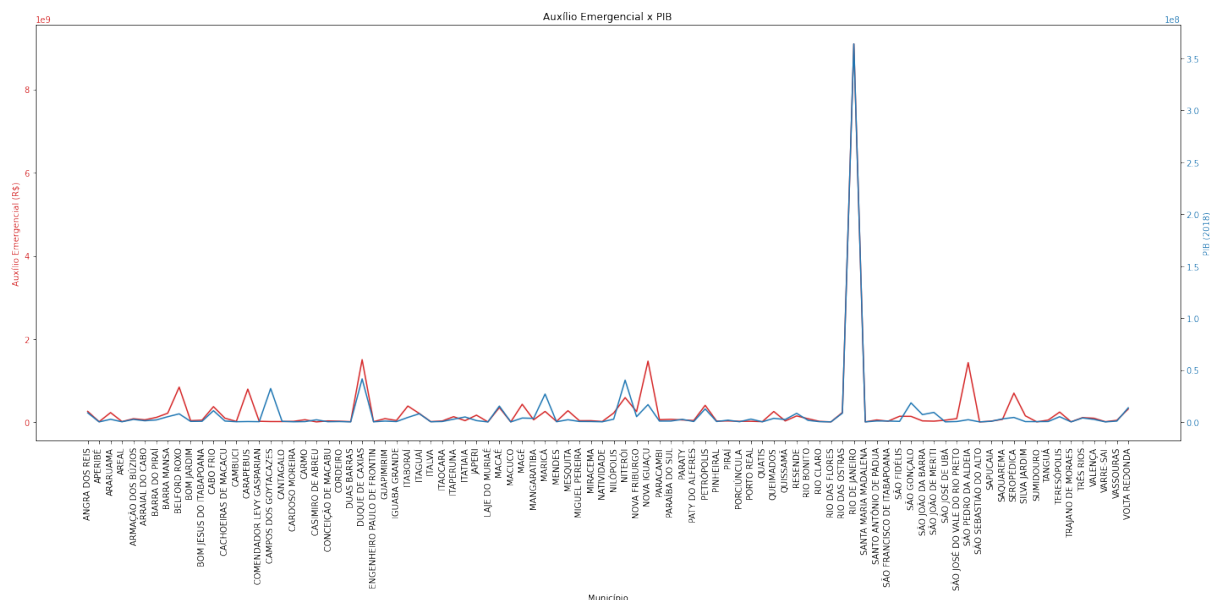


Figura 3.10 – Visão do AE pagos nos municípios do RJ (de abril a dezembro/20), em bilhões (R\$), e do PIB municipal (referência de 2018), em escala de 100 milhões (R\$).

4. Análise e Exploração dos Dados

Análises Estatísticas

Após a análise dos mapas e visualizações, ainda numa fase inicial de análise dos dados nos utilizamos de um Boxplot para aprofundar a avaliação da distribuição dos valores médios de repasses (pagamentos e repasses constitucionais) entre os municípios.

Neste momento, cabe uma breve explanação sobre o Boxplot ou Diagrama de Caixa. Trata-se de uma representação gráfica (comparativa) que, além de apresentar uma importante medida posição relativa – Quartis (que dividem a amostra/população em 4 partes iguais) -, se destaca por evidenciar os valores discrepantes (*outliers*). As medidas de estatísticas descritivas como o Mínimo (limite inferior sem os *outliers*), Máximo (limite superior sem os *outliers*), Primeiro Quartil, Segundo Quartil ou Mediana (que divide a amostra/população em duas partes iguais) e o Terceiro Quartil são apresentadas no Boxplot:

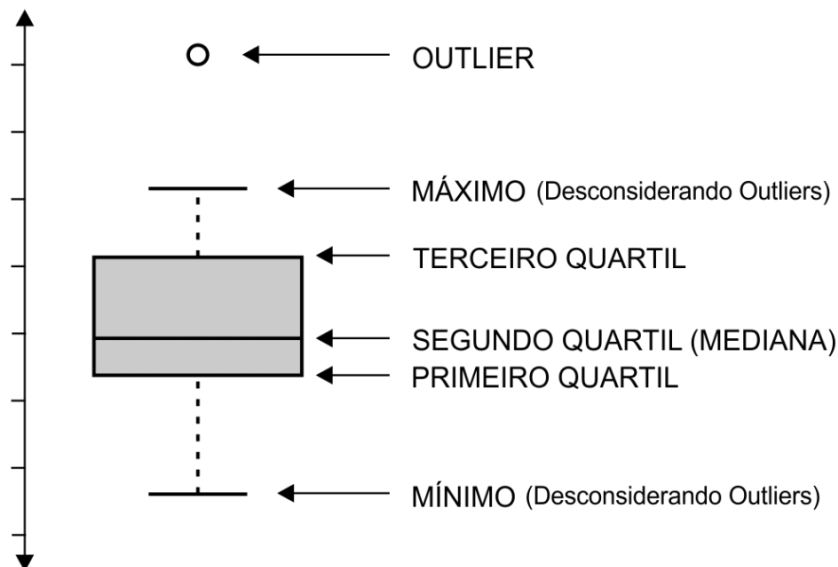


Figura 4.0 –Representação do Boxplot com a haste vertical (de baixo para cima) indicando as medidas de posição relativa.

O retângulo no meio (da haste vertical) possui três linhas horizontais:

- a inferior indica o Primeiro Quartil – PQ (25% da amostra/população);
- a superior indica o Terceiro Quartil - TQ (75% da amostra/população). A diferença entre o PQ e o TQ (tamanho da caixa), chamada de Distância Interquartil (DIQ), indica a dispersão dos dados ou intervalo interquartilico (que não é influenciado pelos *outliers*);

- a linha interna indica o Segundo Quartil ou Mediana, que também é uma medida de tendência central ou posicional (50% da amostra/população). A Mediana também sinaliza a distribuição do conjunto de dados: se está no centro do retângulo: Simétrica, - se a Mediana está próxima do TQ: Assimétrica Positiva, e – se a Mediana está próxima do PQ: Assimétrica Negativa;

Além disso, evidencia os valores extremos ou *outliers* (possíveis valores discrepantes). O limite de detecção de *outliers* é construído utilizando o intervalo interquartilico, sendo:

$$\text{Limite Inferior} = \text{PQ} - 1,5 * (\text{TQ} - \text{PQ}); \text{ e}$$

$$\text{Limite Superior} = \text{TQ} + 1,5 * (\text{TQ} - \text{PQ}).$$

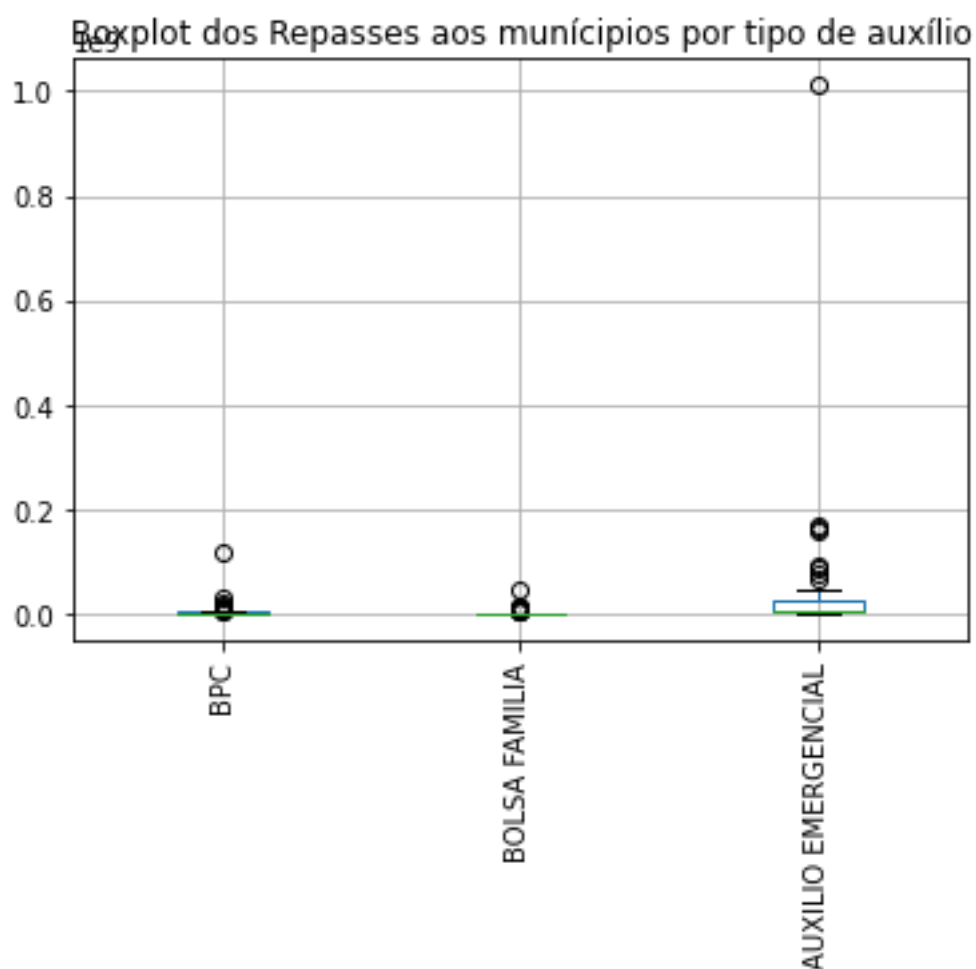


Figura 4.1 – Diagrama de Caixa dos pagamentos dos benefícios assistenciais (BPC, BF e AE) pelos municípios do RJ (de abril a dezembro/20) em bilhões (R\$), destacando o outlier do AE

Percebemos, pelo Bloxpot acima, que um município recebeu repasses de Auxílio Emergencial muito acima dos demais, além disso, este repasse é significativamente maior do que todos os outros (BF e BPC). Aprofundando esta análise, teremos, na sequência, uma visão geral dos pagamentos e repasses e os Boxplots dos valores pagos individualmente.

Fizemos uso de um gráfico para uma visualização “macro” (espécie de “*big picture*”) dos repasses constitucionais e pagamentos de benefícios assistenciais (BPC, BF e AE) por município, nestes 8 meses (abril a dezembro de 2020). O gráfico antecipa visualmente os maiores beneficiários neste período. Considerando impacto do AE, na sequência, também faremos a análise visual sem este benefício.

Figura 4.2 – “Big Picture” da distribuição dos benefícios assistenciais (BPC, BF e AE) e repasses constitucionais do FPM (de abril a dezembro/20) em bilhões (R\$)

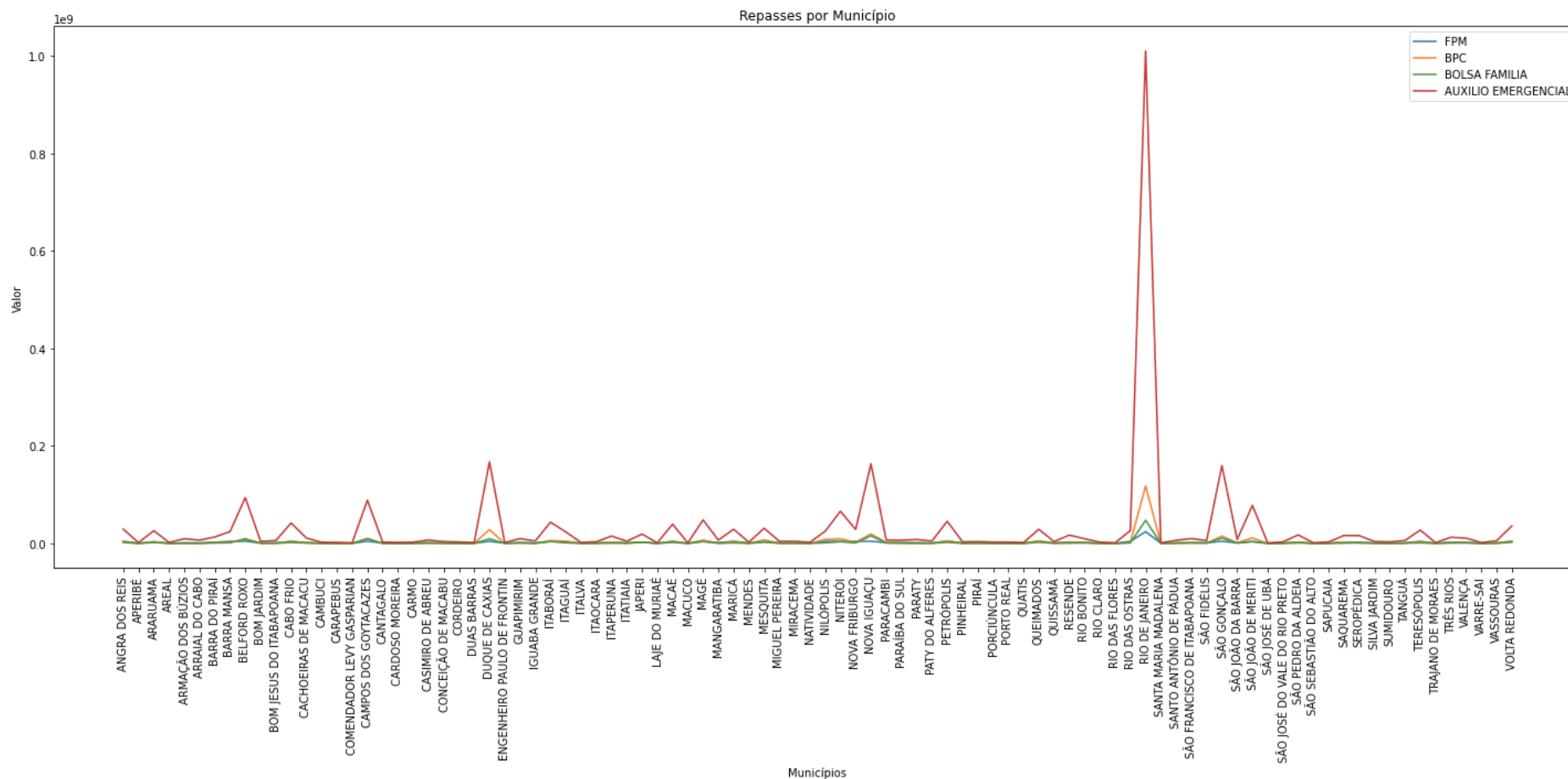
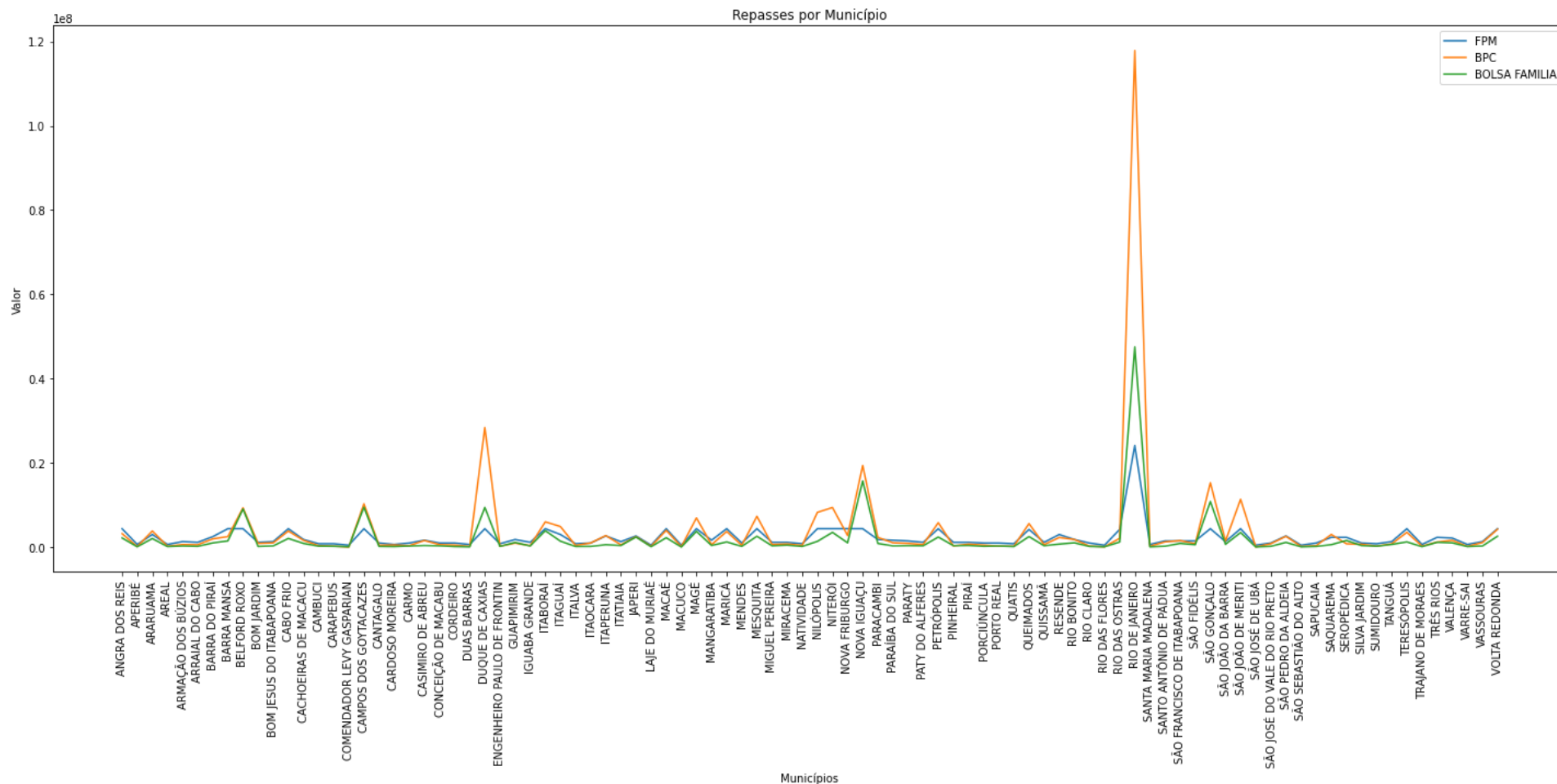


Figura 4.3 – “Big Picture” da distribuição dos benefícios assistenciais (BPC e BF) e repasses constitucionais do FPM (de abril a dezembro/20) em 100 milhões (R\$)



Cabe destacar que uma importante análise, relacionada ao fomento da economia dos municípios fluminenses, diz respeito aos entes cujo repasse dos benefícios assistenciais superaram as transferências constitucionais do FPM às prefeituras. No caso do Estado do Rio de Janeiro, em que a totalidade (100%) dos municípios já recebem, por exemplo, mais de benefícios previdenciários (conforme já citado no estudo da ANFIP) do que repasses constitucionais, 25% dos municípios também possuem o BF e o BPC como fonte superior ao ingresso do FPM, demonstrando a relevância destes benefícios assistenciais como elemento indutor das economias locais. Pelo gráfico acima, podemos ter uma ideia da distribuição dos repasses do FPM por prefeitura (linha azul), bem como, na lista abaixo, observar os municípios que recebem valores de benefícios assistenciais superiores a este repasse.

2	ARARUAMA
8	BELFORD ROXO
16	CAMPOS DOS GOYTACAZES
24	DUQUE DE CAXIAS
28	ITABORAÍ
29	ITAGUAÍ
31	ITAOCARA
32	ITAPERUNA
38	MAGÉ
42	MESQUITA
46	NILÓPOLIS
47	NITERÓI
49	NOVA IGUAÇU
50	PARACAMBI
54	PETRÓPOLIS
60	QUEIMADOS
63	RIO BONITO
67	RIO DE JANEIRO
70	SÃO FRANCISCO DE ITABAPOANA
72	SÃO GONÇALO
73	SÃO JOÃO DA BARRA
74	SÃO JOÃO DE MERITI
80	SAQUAREMA

Figura 4.4 – Lista dos municípios (23) que receberam (de abril a dezembro/20) valores de benefícios assistenciais (BPC e BF) superiores aos repasses constitucionais do FPM.

Na sequência, vamos analisar os Boxplots dos valores pagos individualmente.

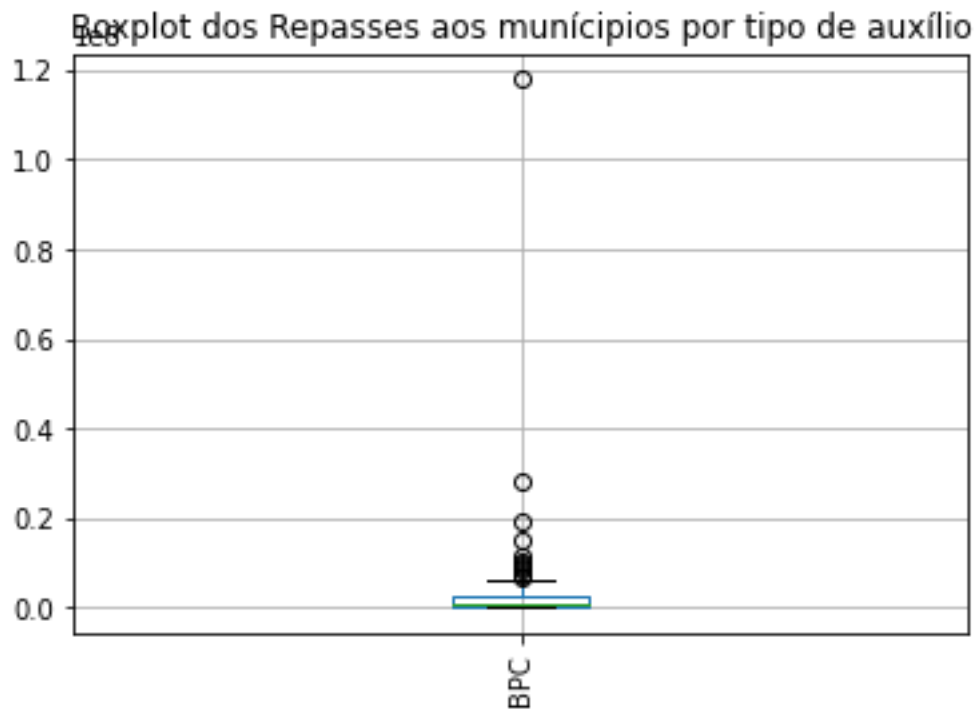


Figura 4.6 – Diagrama de Caixa dos pagamentos do BPC nos municípios do RJ (de abril a dezembro/20) em milhões (R\$)

```

count      BPC
count      92.00
mean       3891339.76
std        12765946.49
min         82092.13
25%        444367.39
50%        929720.53
75%        2860229.11
max        117852859.67
Coeficiente de Variação: 3.2627263487796383

```

Pelo Boxplot do BPC, observamos que a distribuição apresenta uma assimetria positiva(superior) e alguns municípios discrepantes ($> Q3 + 1,5 \times DIQ$). É importante destacar que a média (*mean*) é influenciada pelos discrepantes (*outliers*).

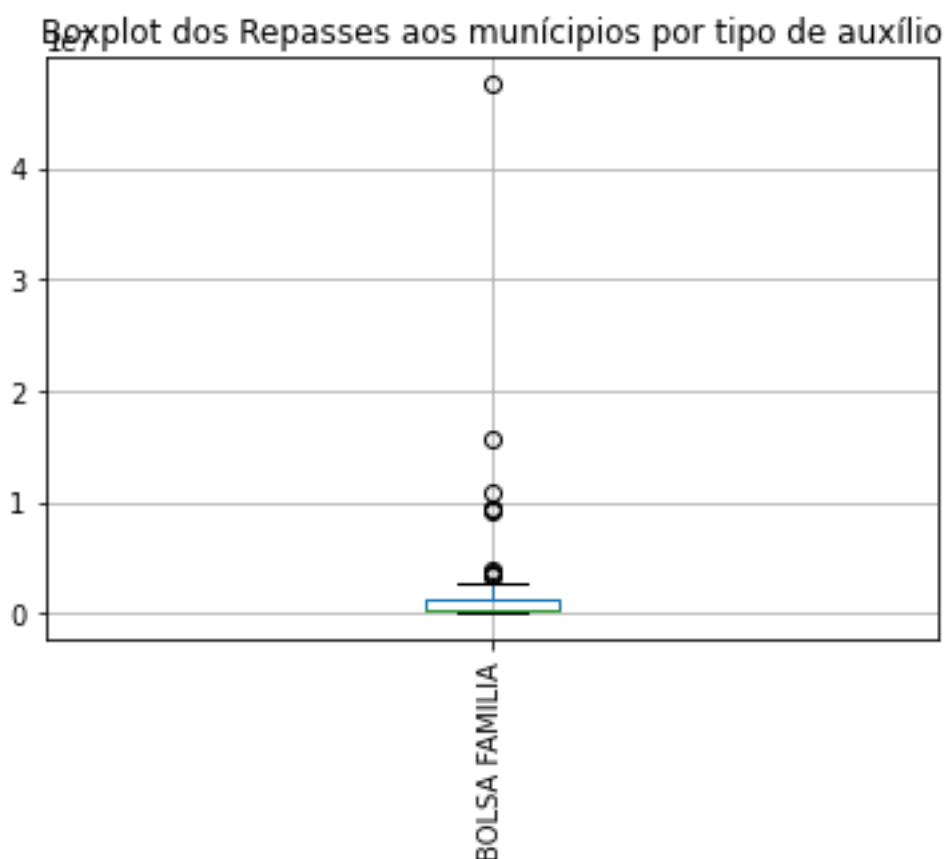


Figura 4.7 – Diagrama de Caixa dos pagamentos do BF nos municípios do RJ (de abril a dezembro/20) em 10 milhões (R\$)

```

BOLSA FAMILIA
count          92.00
mean         1905535.30
std          5437716.81
min           69426.78
25%          216876.00
50%          425620.94
75%          1310976.00
max          47533532.78
Coeficiente de Variação: 2.838091346781201

```

O Boxplot do BF apresenta, a exemplo do BPC, uma distribuição assimétrica positiva e alguns *outliers*. Comparando o Coeficiente de Variação (CV), temos que o BF apresenta maior homogeneidade dos dados.

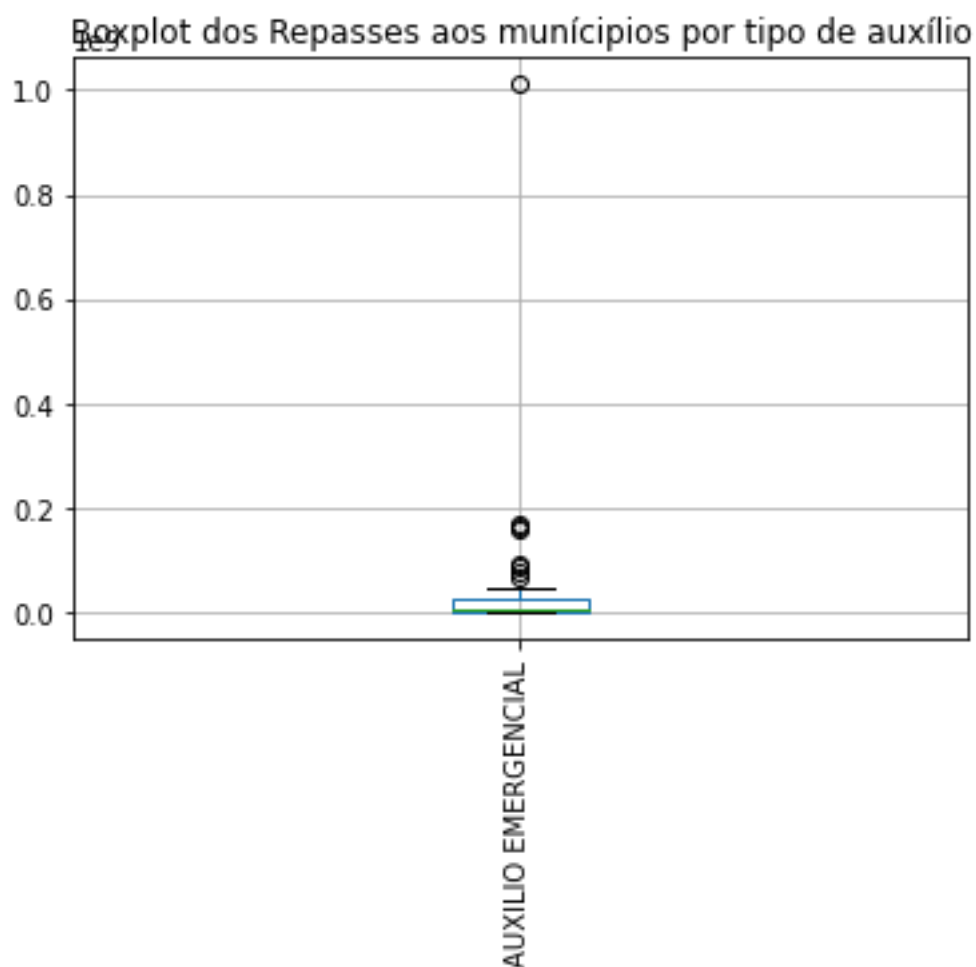


Figura 4.8 – Diagrama de Caixa dos pagamentos do AE nos municípios do RJ (de abril a dezembro/20) em bilhões (R\$)

```

AUXILIO EMERGENCIAL
count          92.00
mean          30120723.67
std           108314324.22
min            1197989.00
25%            2952253.08
50%            6537236.78
75%            24606667.42
max           1010316090.67
Coeficiente de Variação: 3.5764097908494126

```

O Boxplot do AE segue a mesma distribuição assimétrica positiva dos demais benefícios assistenciais com alguns *outliers* ($> Q3 + 1,5 \times DIQ$). Comparando o CV dos vários benefícios, temos que o AE apresenta maior variabilidade dos dados. Os altos valores de desvio padrão (*std*) indicam uma grande dispersão nos dados

Maiores e menores repasses:

Aprofundando o estudo, iniciado com a análise visual dos mapas, podemos investigar os municípios que, além da capital do estado, receberam valores acima da média de benefícios assistenciais e repasses, indicando os *outliers*.

	NM_MUNICIP	BPC
67	RIO DE JANEIRO	117.852.859,67
24	DUQUE DE CAXIAS	28.416.651,60
49	NOVA IGUAÇU	19.415.190,33
72	SÃO GONÇALO	15.310.780,65
74	SÃO JOÃO DE MERITI	11.389.637,04

Figura 4.9 – Tabela com os municípios (5) que receberam mais BPC (de abril a dezembro/20)

	NM_MUNICIP	BOLSA FAMILIA
67	RIO DE JANEIRO	47.533.532.78
49	NOVA IGUAÇU	15.739.079.11
72	SÃO GONÇALO	10.856.986.67
16	CAMPOS DOS GOYTACAZES	9.514.689.78
24	DUQUE DE CAXIAS	9.458.914.56

Figura 4.10 – Tabela com os municípios (5) que receberam mais BF (de abril a dezembro/20)

	NM_MUNICIP	AUXÍLIO EMERGENCIAL
67	RIO DE JANEIRO	1.010.316.090.67
24	DUQUE DE CAXIAS	167.405.029.11
49	NOVA IGUAÇU	163.562.353.22
72	SÃO GONÇALO	159.548.020.56
8	BELFORD ROXO	94.120.832.22

Figura 4.11 – Tabela com os municípios (5) que receberam mais AE (de abril a dezembro/20)

	NM_MUNICIP	FPM
67	RIO DE JANEIRO	24.149.461.22
42	MESQUITA	4.413.445.33
46	NILÓPOLIS	4.413.445.12

24	DUQUE DE CAXIAS	4.413.445.12
85	TERESÓPOLIS	4.413.445.12

Figura 4.12 – Tabela com as prefeituras (5) que receberam mais FPM (de abril a dezembro/20)

Também elencamos os municípios que menos receberam as menores médias de benefícios assistenciais e repasses.

	NM_MUNICIP	BPC
1	APERIBÉ	172080.00
37	MACUCO	155823.78
65	RIO DAS FLORES	126099.10
75	SÃO JOSÉ DE UBÁ	104384.76
15	COMENDADOR LEVY GASPARIAN	82092.13

Figura 4.13 – Tabela com os municípios (5) que receberam menos BPC (de abril a dezembro/20)

	NM_MUNICIP	BOLSA FAMILIA
78	SÃO SEBASTIÃO DO ALTO	116806.33
23	DUAS BARRAS	112691.44
68	SANTA MARIA MADALENA	106868.22
75	SÃO JOSÉ DE UBÁ	96283.00
37	MACUCO	69426.78

Figura 4.14 – Tabela com os municípios (5) que receberam menos BF (de abril a dezembro/20)

	NM_MUNICIP	AUXÍLIO EMERGENCIAL
68	SANTA MARIA MADALENA	1384082.33
35	LAJE DO MURIAÉ	1336401.00
37	MACUCO	1309969.33
15	COMENDADOR LEVY GASPARIAN	1287305.00
75	SÃO JOSÉ DE UBÁ	1197989.00

Figura 4.15 – Tabela com os municípios (5) que receberam mais AE (de abril a dezembro/20)

	NM_MUNICIP	FPM
37	MACUCO	506171.76
65	RIO DAS FLORES	506171.76
35	LAJE DO MURIAÉ	506171.76
75	SÃO JOSÉ DE UBÁ	506171.76
15	COMENDADOR LEVY GASPARIAN	506171.76

Figura 4.16 – Tabela com as prefeituras (5) que receberam menos FPM (de abril a dezembro/20)

Análise de Correlação

Uma análise que precede a análise preditiva (ou de regressão) é a análise de correlação. Esta análise permite avaliar a relação entre diferentes variáveis e os coeficientes de correlação mensuram essa relação e, em nosso caso de estudo, indicam se os valores dos repasses ou pagamentos estão correlacionados. É importante destacar que correlação não implica causalidade, ou seja, não há uma relação de causa e efeito entre as variáveis.

Neste particular, considerando o caráter linear da relação, utilizaremos o coeficiente de correlação de Pearson. Trata-se de um teste que mede a relação estatística entre duas variáveis, que pode assumir um intervalo de valores de +1 a -1. Um valor de 0 indica que não há associação entre as duas variáveis. Um valor maior que 0 indica uma associação positiva. Isto é, à medida que o valor de uma variável aumenta, o mesmo acontece com o valor da outra variável. Sendo que, se assumir o valor 1, teremos uma correlação perfeita positiva. Um valor menor que 0 indica uma associação negativa. Isto é, à medida que o valor de uma variável aumenta, o valor da outra diminui. Por analogia, se o coeficiente assumir o valor -1, teremos uma correlação perfeita negativa.

A partir daí, analisamos os nossos dados:

	BPC	FPM	BOLSA FAMÍLIA	AUXÍLIO EMERGENCIAL
BPC		0.92324116	0.97137647	0.99468916

FPM	0.92324116		0.90843526	0.92150122
BOLSA FAMÍLIA	0.97137647	0.97137647		0.97730426
AUXÍLIO EMERGENCIAL	0.99468916	0.92150122	0.97730426	

Figura 4.17 – Tabela de correlações entre os repasses e pagamentos (de abril a dezembro/20)

De fato, a correlação de Pearson entre os valores médios de repasses por município mostra uma correlação positiva (ou diretamente proporcional) forte, ou seja, quanto maior o valor de um tipo de repasse, maior o valor de outro tipo de repasse. Podemos observar que a relação entre o BPC e o AE é quase perfeita, pois se aproxima do valor limite (=1). A relação do AE com o BF e os repasses do FPM também é bastante significativa.

5. Criação de Modelos de Machine Learning

A partir da análise de correlação, que nos permitiu entender como as variáveis do estudo se comportam num cenário de variação das outras, passamos agora para o estudo dos Modelos de *Machine Learning*.

Utilizaremos neste estudo de Predição as técnicas de Regressão Linear, que se enquadram na categoria dos algoritmos de aprendizado de Máquina Supervisionado, ou seja, o conjunto de dados de treinamento tem entradas, bem como a saídas desejadas. Durante a sessão de treinamento, o modelo ajustará suas variáveis para mapear as entradas para as saídas correspondentes.

Ainda sobre a Análise de Regressão, vale destacar que é uma abordagem para modelar a relação entre um conjunto de variáveis **y** (variáveis dependentes) e variáveis explicativas **x** (chamadas preditoras ou independentes)

Separação dos Dados

Em primeiro lugar, antes da geração dos nossos modelos, precisamos fazer a separação ou *split* dos dados de treinamento dos dados de teste). Nesta fase, utilizaremos a técnica de Validação Cruzada (*Cross Validation*). Esta técnica reduz a

probabilidade de aleatoriedade, trazendo um resultado mais robusto. Em essência, o *Cross Validation* visa entender como seu modelo generaliza, ou seja, como ele se comporta quando vai prever um dado que nunca viu.

Partimos então para dividir os dados que vamos treinar e os dados que serão utilizados na predição (ou teste), ou seja TRAINING SET e TEST SET. Além disso, precisamos de ter um processo que nos auxilie a escolher qual modelo se ajusta melhor aos nossos dados como um todo. Apesar de podermos treinar os modelos escolhidos com todos os dados de treinamento e avaliá-lo apenas com o conjunto de teste, esse processo pode esconder algumas “armadilhas”, pois há modelos um pouco mais sofisticados que podem encontrar um “atalho” nos dados que o faça ter um desempenho melhor que os outros modelos. O desafio é evitar estas “armadilhas”, num ambiente de produção, pois os dados usados para predição podem ser exatamente os dados que esse modelo “ignorou” ao tomar os “atalhos”, que davam um melhor desempenho no conjunto de teste.

Para mitigar esse risco, podemos fazer uso do processo *K-Cross Validation*, onde avaliamos um mesmo modelo, em diferentes recortes dos dados de treinamento antes de avaliá-lo nos dados de teste.

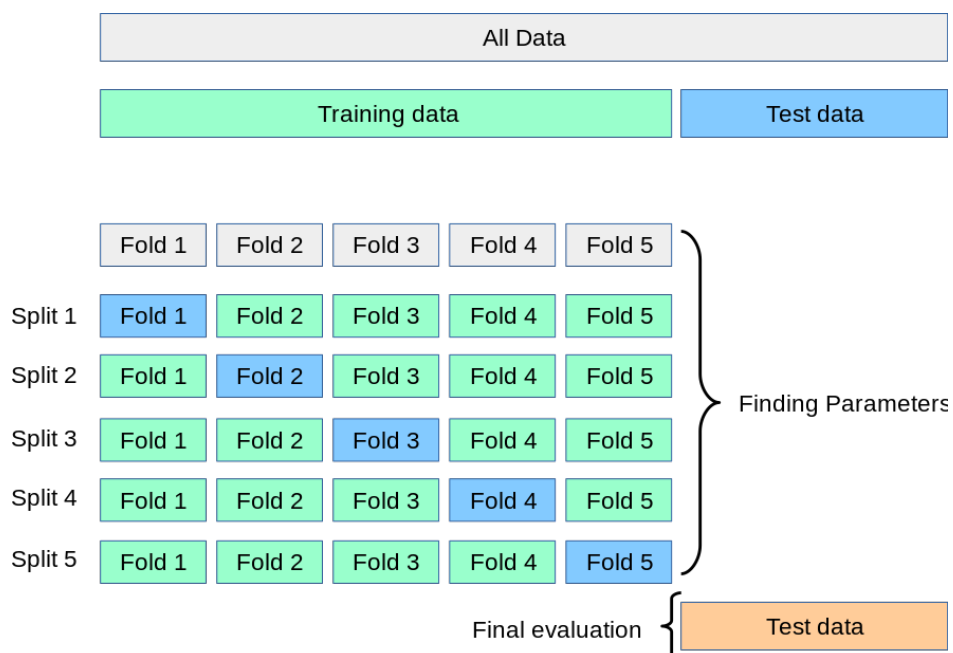


Figura 5.0 – Quadro que exemplifica o funcionamento da Validação Cruzada, com os splits (ou iterações) e as divisões (Folds)

Em seguida, vamos trabalhar com os modelos de Regressão. É importante destacar que as bases de treinamento serão os dados de abril a novembro de 2020 e a base de teste será dezembro de 2020. A princípio utilizaremos 4 modelos básicos que vão nos permitir 6 análises distintas. São eles:

- **OLS** - *Ordinary Least Squares*
- **Linear Regression**
- **KNR** - *KNeighborsRegressor*
- **SVR** - *Support Vector Regression* (*kernels*= [RBF, polinomial e linear])

OLS: Iniciaremos pelo método de estimação mais comum - o OLS (*Ordinary Least Squares*) ou Método dos Mínimos Quadrados (MMQ), também conhecido como Mínimos Quadrados Ordinários (MQO). Trata-se de uma técnica de estimação/otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados (tais diferenças são chamadas resíduos).

Linear Regression: Os modelos OLS (do statsmodel) e *Linear Regression* (do sklearn) seguem basicamente o mesmo procedimento, no entanto, a fórmula da regressão utilizada tem uma pequena diferença de tratamento (<https://stats.stackexchange.com/a/482178>). Por exemplo, enquanto no OLS uma fórmula seria utilizada do seguinte modo:

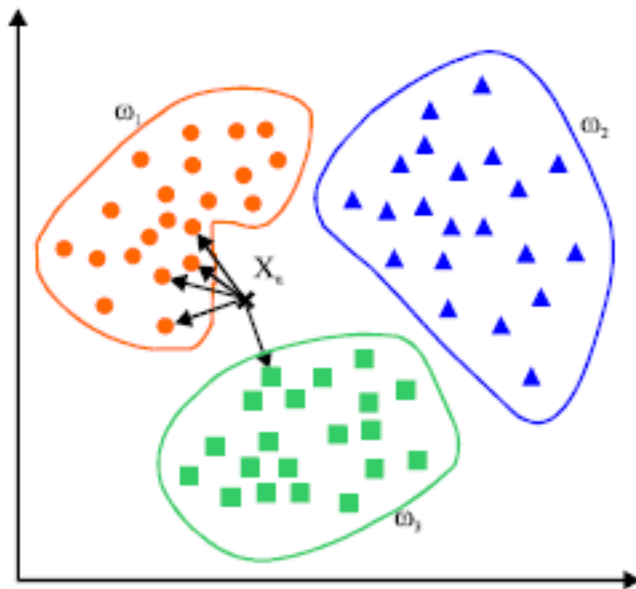
$$y = B_0 + B_1x_1$$

o *Linear Regression* acrescentaria um termo:

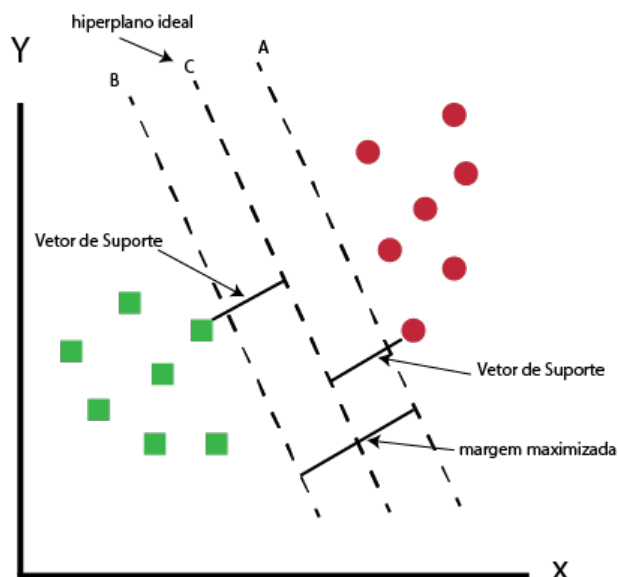
$$y = B_0x_0 + B_1x_1$$

o que em casos de borda podem alterar os resultados obtidos, como será visto nos resultados do K-Fold Cross Validation

KNR: Na sequência, utilizaremos o KNR (*KNeighborsRegressor*). Trata-se de um algoritmo relativamente simples que prevê pontos de dados desconhecidos a partir dos seus vizinhos mais próximos. O valor de k é um fator crítico aqui quanto à precisão da predição. Ele determina o mais próximo ao calcular a distância usando funções básicas de distância como Euclidean.



SVR: Por fim, utilizaremos o modelo de regressão *Support Vector Regression* (SVR) baseado na máquina de vetores de suporte ou SVM - sigla do inglês, *Support Vector Machine*. O funcionamento básico das SVM consiste em ajustar a equação de uma reta, denominada hiperplano de tal forma que a distância entre ela e os pontos com características diferentes seja maximizada. Uma vantagem importante na utilização dos modelos baseados em SVM é que eles não são sensíveis aos *outliers*, ou seja, valores extremos não causam ruído no treinamento.



A ideia básica do SVR é mapear um conjunto de dados “ \mathbf{x} ” em um espaço multidimensional por meio de um mapeamento não-linear (usualmente

utilizando *kernels*) e então realizar uma regressão linear neste espaço multidimensional transformado.

Os *kernels* são responsáveis por criar uma transformação dos dados a partir de uma função, que são responsáveis por maximizar as margens dos vetores de suporte. A maioria das bibliotecas de ML, já possuem *kernels* implementados e permitem a integração de outras funções customizadas. Os principais *kernels* utilizados são:

- **Linear:** é eficiente para problemas linearmente separáveis, uma vez que seu ajuste se por meio da equação de uma reta;
- **Polinomial:** Os *kernels* polinomiais permitem adicionar curvas aos hiperplanos;
- **Função gaussiana de base radial:** Os *kernels* RBF (*Radial Basis Function*), como também são chamados, são recomendados quando não se tem um conhecimento prévio acerca dos dados. Esse *kernel* realiza uma transformação dos pontos utilizando uma função gaussiana.

Validação de Regressão

A fim de validar o desempenho ou precisão dos modelos, selecionamos 4 métricas de avaliação dos modelos, que devem ser observadas:

- **R² score**
- **MSE** - *Mean Square Error*;
- **RMSE** - *Root Mean Square Error*;
- **MAPE** - *Mean Absolute Percentage Error*.

O **R²** ou **Coefficiente de Determinação** é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. O **R²** varia entre 0 e 1 e expressa a quantidade da variância dos dados que é explicada pelo modelo, ou seja, quanto a variância da variável dependente se explica a partir da variável independente. Sendo que quanto mais próximo de 1 (ou 100%) melhor é o modelo.

O **Erro Quadrático Médio**, em inglês **Mean Square Error (MSE)**, é a métrica mais utilizada, que consiste na média do erro das previsões ao quadrado. Em outras

palavras, pega-se a diferença entre o valor predito pelo modelo e o valor real, eleva-se o resultado ao quadrado, faz-se a mesma coisa com todos os outros pontos, soma-os, e divide-se pelo número de elementos preditos. Quanto menor esse número, melhor o modelo (mais próximo do valor mínimo 0). Neste sentido, diferenças menores têm menos importância, enquanto diferenças maiores recebem mais peso.

A **Raiz do Erro Quadrático Médio**, em inglês **Root Mean Square Error (RMSE)**, é comumente usada para expressar a “acurácia” dos resultados numéricos do modelo. Uma característica do RMSE é que os erros (reais - previsões) são elevados ao quadrado antes de ter a média calculada. Portanto, pesos diferentes serão atribuídos à soma e, conforme os valores de erros das instâncias aumentam o índice do RMSE aumenta consideravelmente. Assim, quanto maior seu valor, pior o desempenho do modelo, sendo o ideal próximo de 0.

O **Erro Percentual Absoluto Médio**, em inglês **Mean Absolute Percentage Error (MAPE)**, é outra importante métrica de adequação dos modelos de Regressão que representa a média percentual da divisão entre erro de previsão e o valor real. Valores pequenos para o MAPE determinam precisão nos dados previstos

Inicialmente, fizemos uma avaliação dos modelos usando o *KFold Cross Validation* com $K = 10$ (= número de *splits* e *folds*), escolha arbitrária:

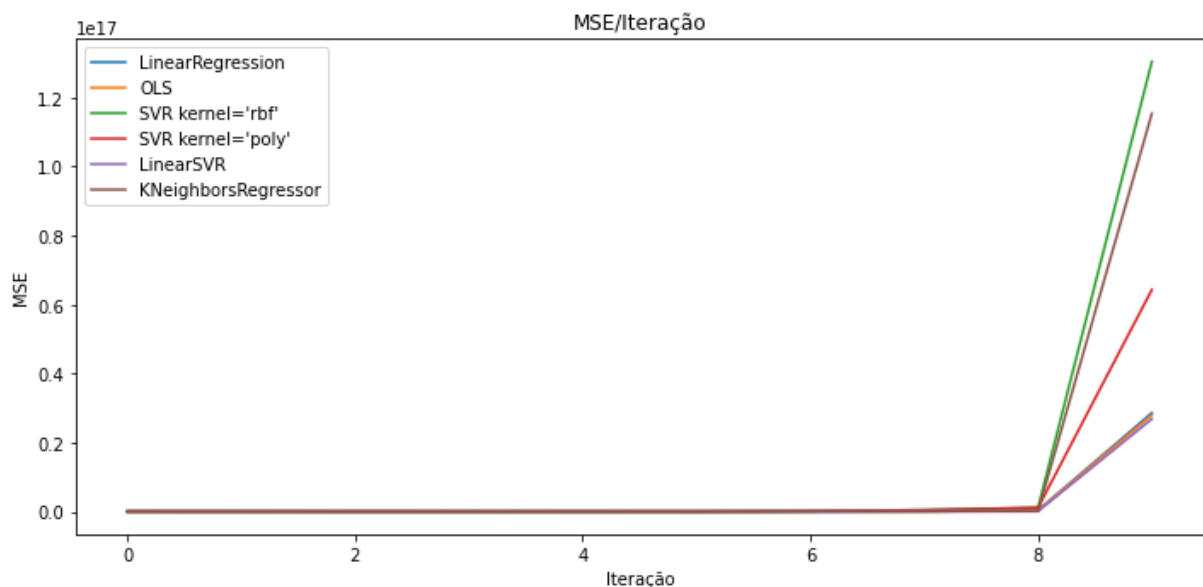


Figura 5.1 – Gráfico comparativo do MSE para os modelos adotados com 10 iterações (em 100 quatrilhões)

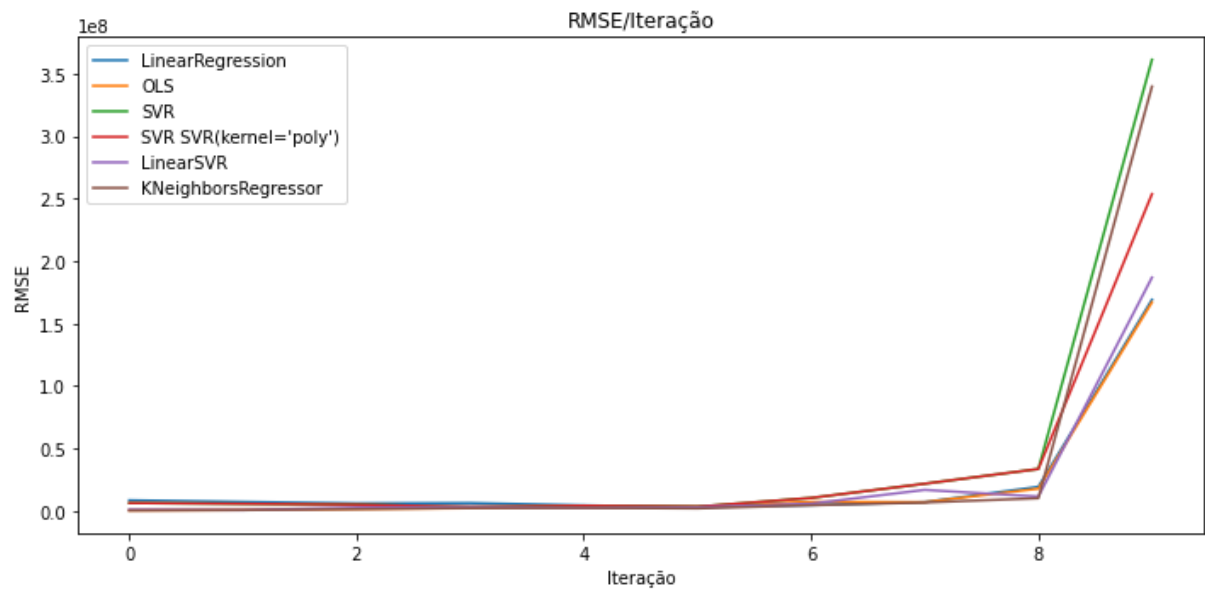


Figura 5.2 – Gráfico comparativo do RMSE para os modelos adotados com 10 iterações (em 100 milhões)

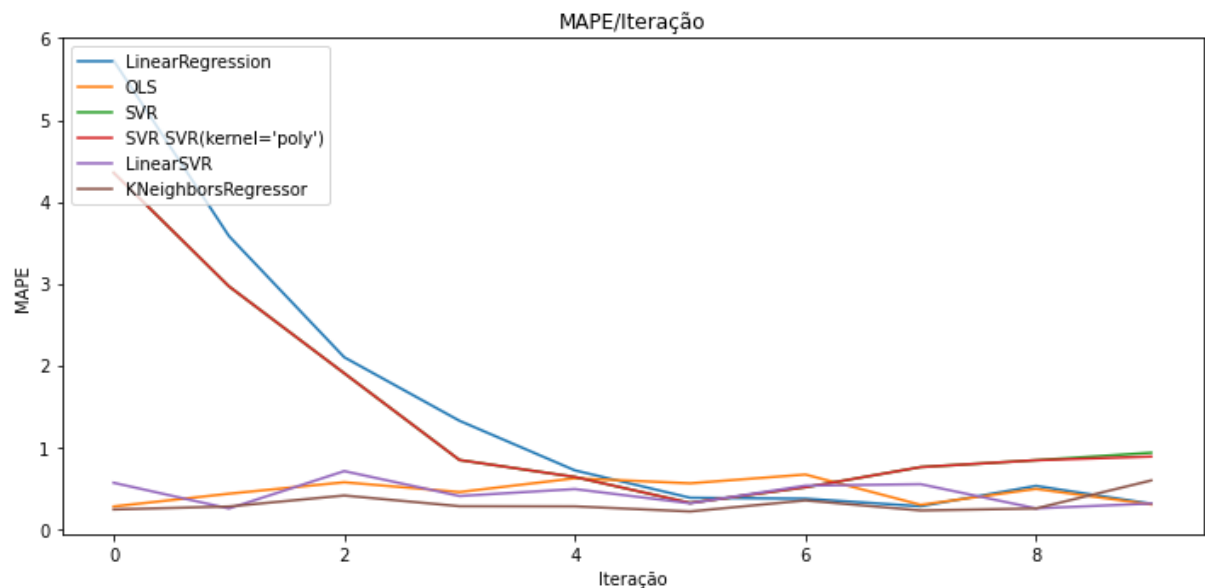


Figura 5.3 – Gráfico comparativo do MAPE score para os modelos adotados com 10 iterações

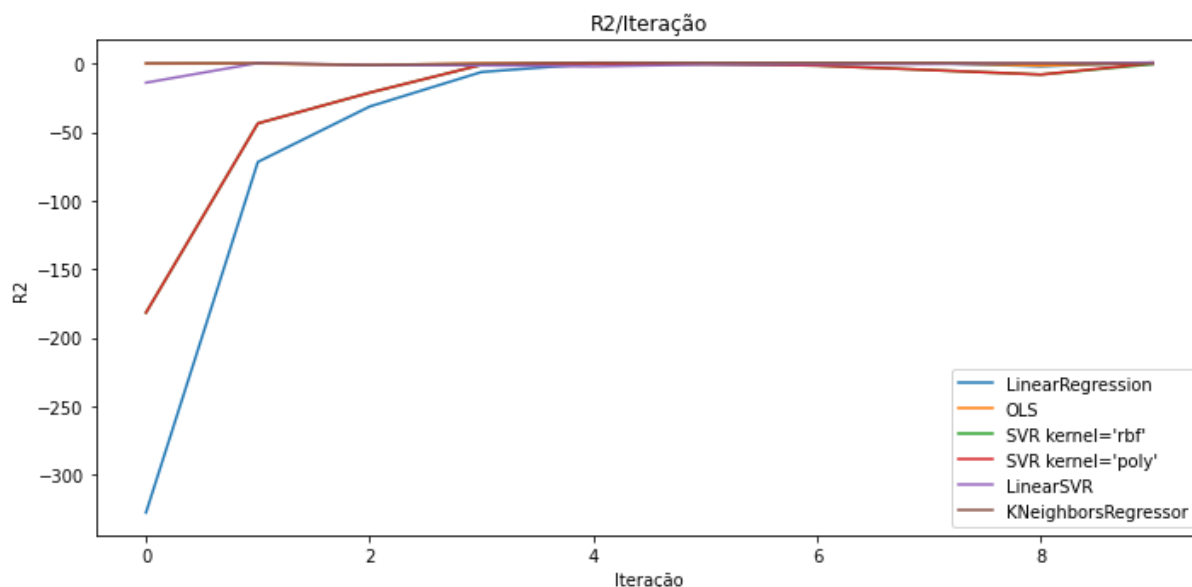


Figura 5.4 – Gráfico comparativo do R2 score para os modelos adotados com 10 iterações

Podemos observar, avaliando o MSE e o RMSE, que todos os modelos têm um comportamento similar nos primeiros *Folds* (mais próximos de 0), mas nos *Folds* (ou divisões) finais começam a aumentar os valores. O MSE apresenta piora nos 2 últimos *Folds* e o RMSE nos últimos 4 *Folds*. Contrariamente, o MAPE inicia apresentando valores ruins que melhoram nas interações finais. O OLS também apresenta a mesma tendência, aumentando a qualidade a partir do 3º *Fold*. No geral, os melhores modelos foram o OLS, Linear Regression e o SVR com *kernel linear*.

Contudo, devemos observar se não temos um desbalanceamento em função do *split* dos dados em 10 *Folds* sequenciais, pois o número de *Folds* é maior que o número de meses dos dados. Nesta situação, podemos ter um *Fold* com uma variância muito diferente da distribuição original dos dados. Provavelmente, essa diferença entre a variância dos dados de um *Fold* e os dados originais podem justificar os resultados ruins de alguns modelos nos *Folds* finais.

Na sequência, optamos por realizar o processo de validação cruzada com um $k=8$. Assim, teremos *Folds* que coincidem com os meses dos dados de treinamento, e, portanto, os dados não estarão desbalanceados.

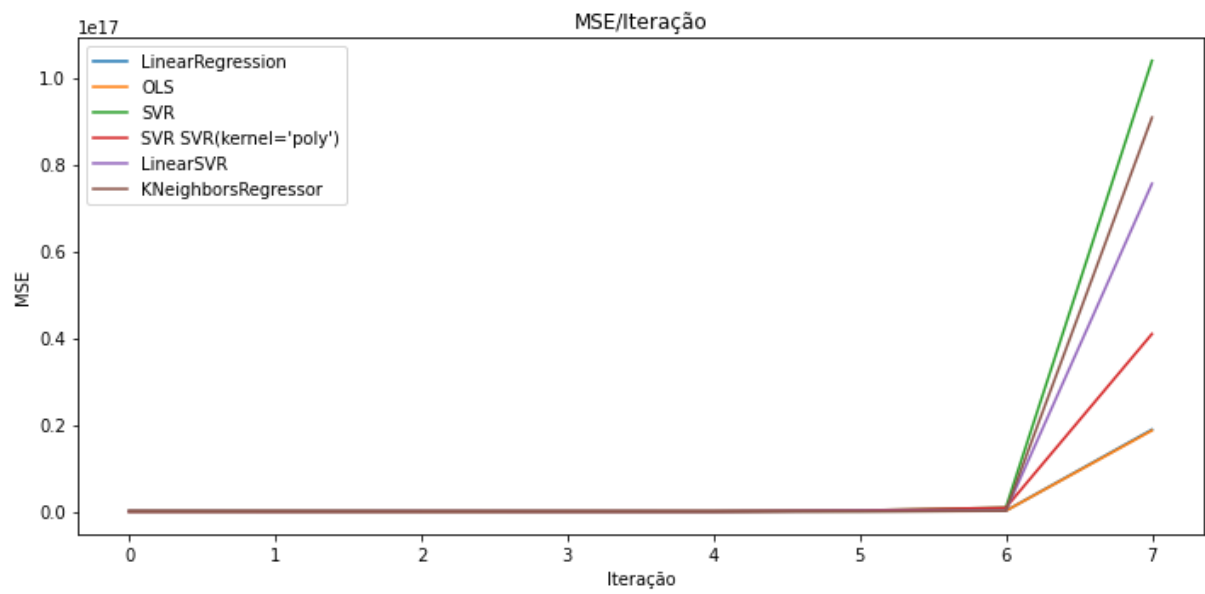


Figura 5.5 – Gráfico comparativo do MSE para os modelos adotados com 8 iterações (em 100 quatrilhões)

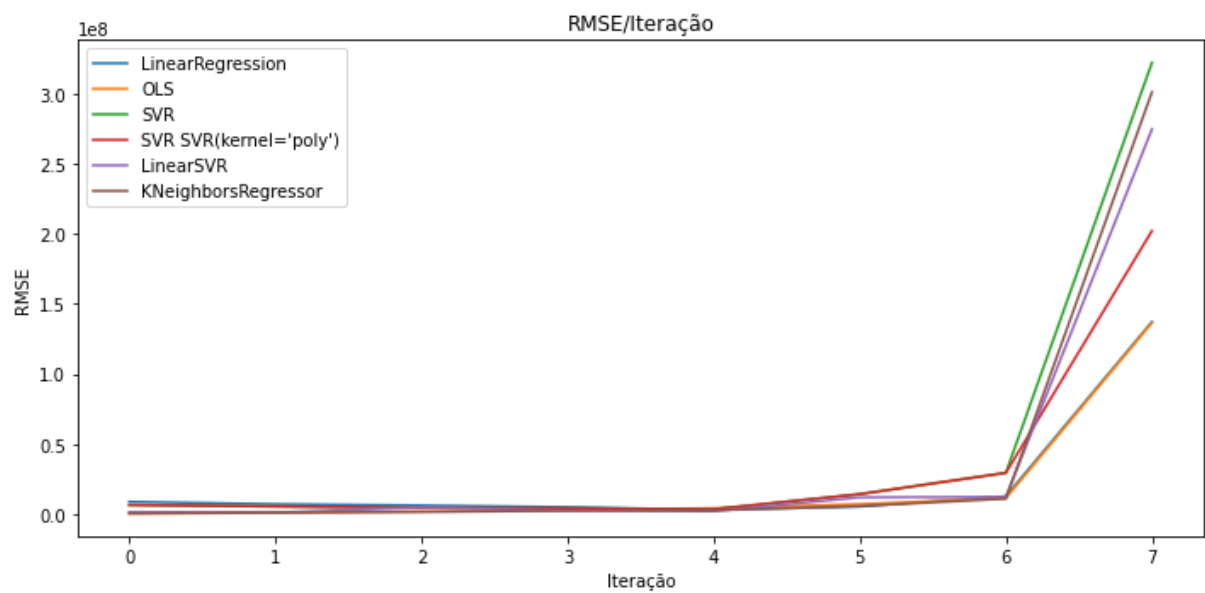


Figura 5.6 – Gráfico comparativo do RMSE para os modelos adotados com 8 iterações (em 100 milhões)

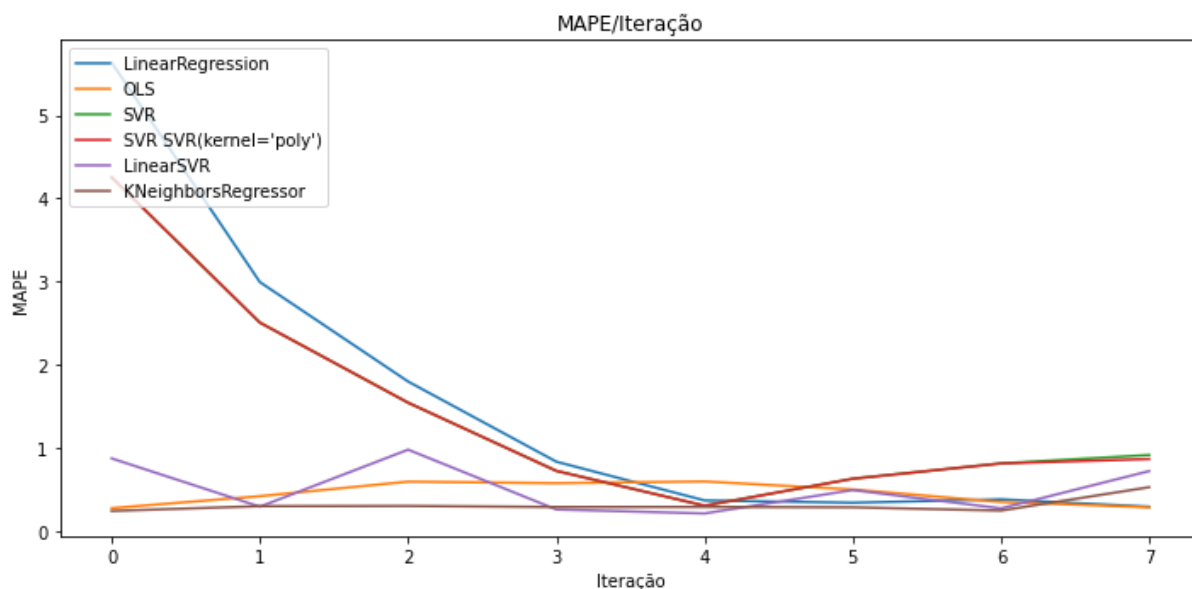


Figura 5.7– Gráfico comparativo do MAPE para os modelos adotados com 8 iterações

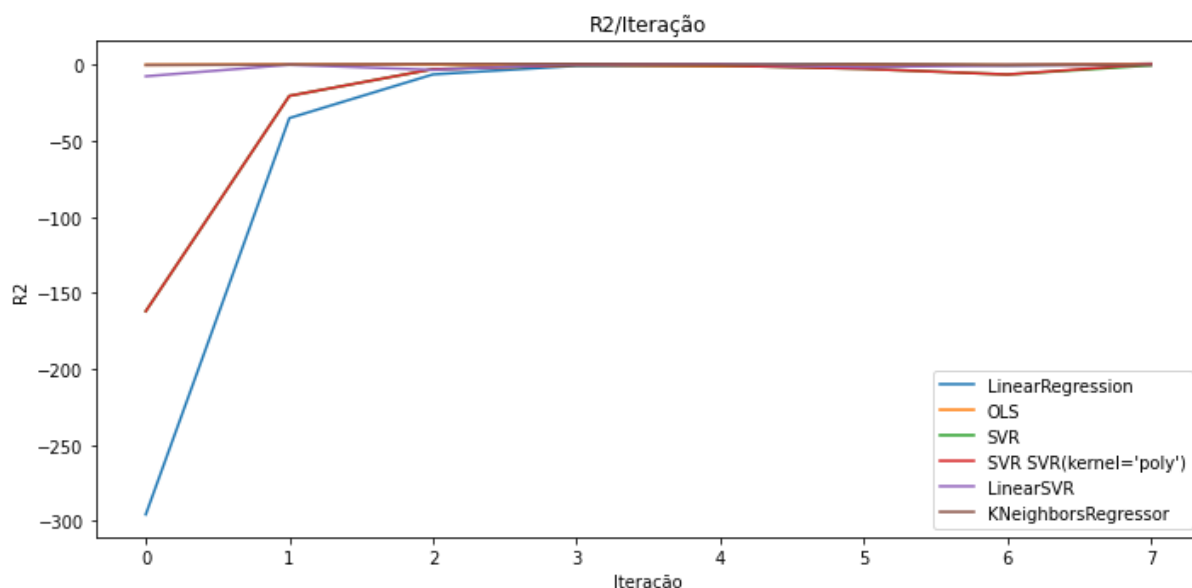


Figura 5.8– Gráfico comparativo do R2 score para os modelos adotados com 8 iterações

Pelo gráfico, embora o desenho geral tenha sido similar (ao K=10), percebemos (pelo MSE) que diminuir a quantidade de *Folds* (e consequentemente diminui a quantidade de dados de treinamento em cada iteração) afetou positivamente o desempenho do modelo SVR com *kernel linear*, tornando-o o modelo com melhores resultados. No entanto, podemos também alterar os *Folds* da validação cruzada para usarem os dados de forma aleatória e não sequencial. Isso é desejável, já que a Regressão não tem como premissa a dependência temporal.

Iniciaremos a validação (utilizando o parâmetro: *shuffle=True*) com 10 *Folds*.

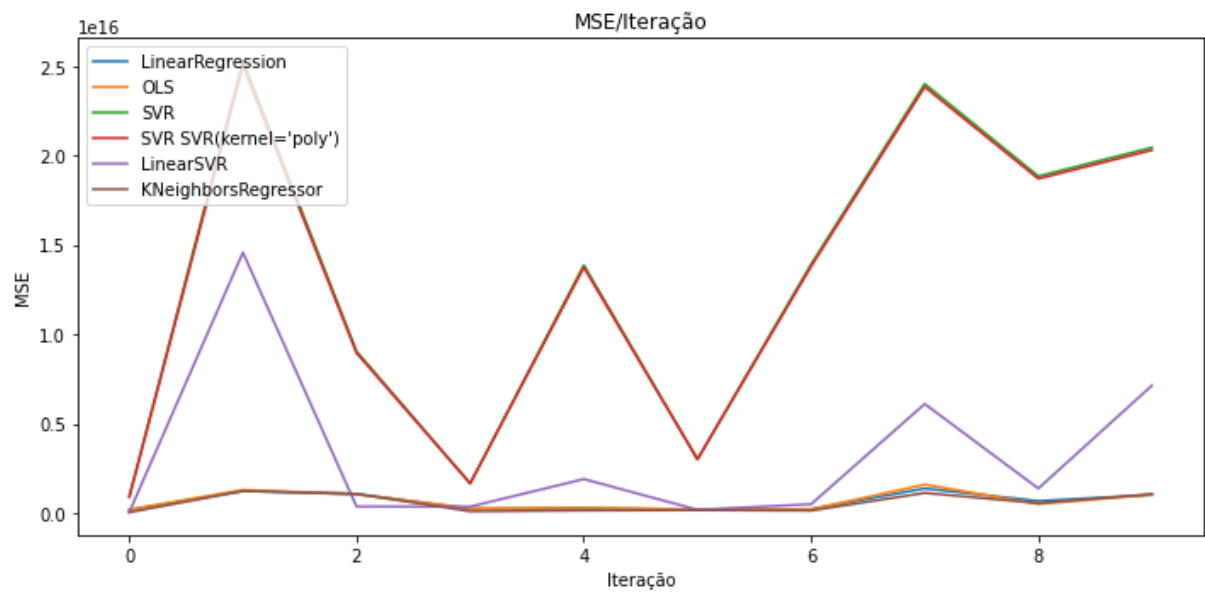


Figura 5.9 – Gráfico comparativo do MSE para os modelos adotados com 10 iterações e divisão aleatória (*shuffle=True*) (em 10 quatrilhões)

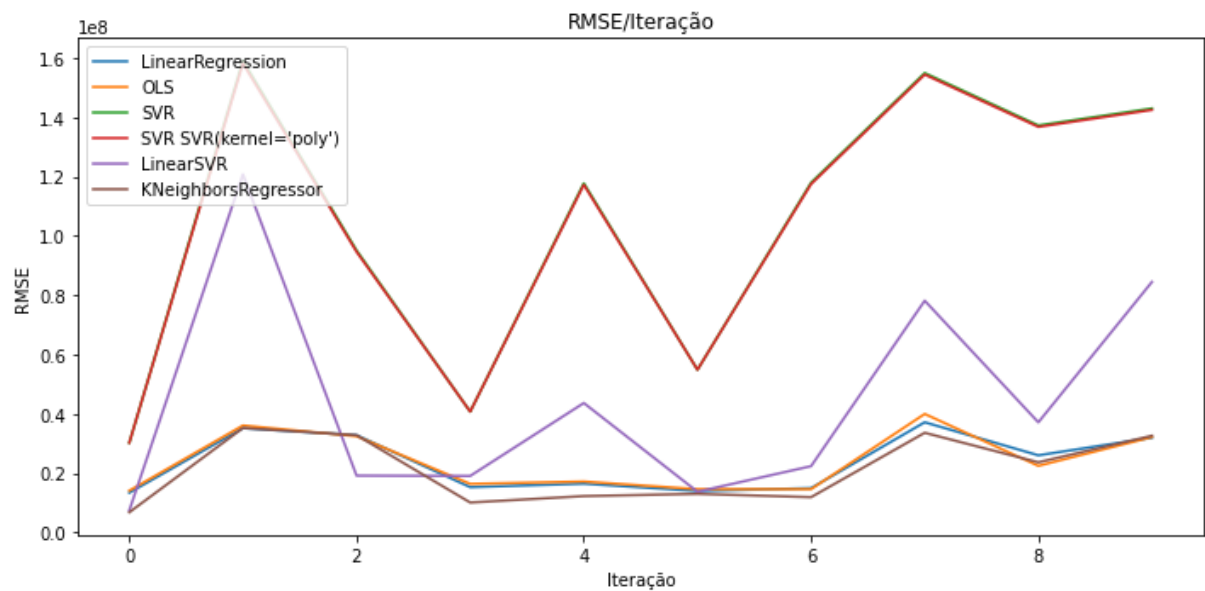


Figura 5.10 – Gráfico comparativo do RMSE para os modelos adotados com 10 iterações e divisão aleatória (*shuffle=True*) (em 100 milhões)

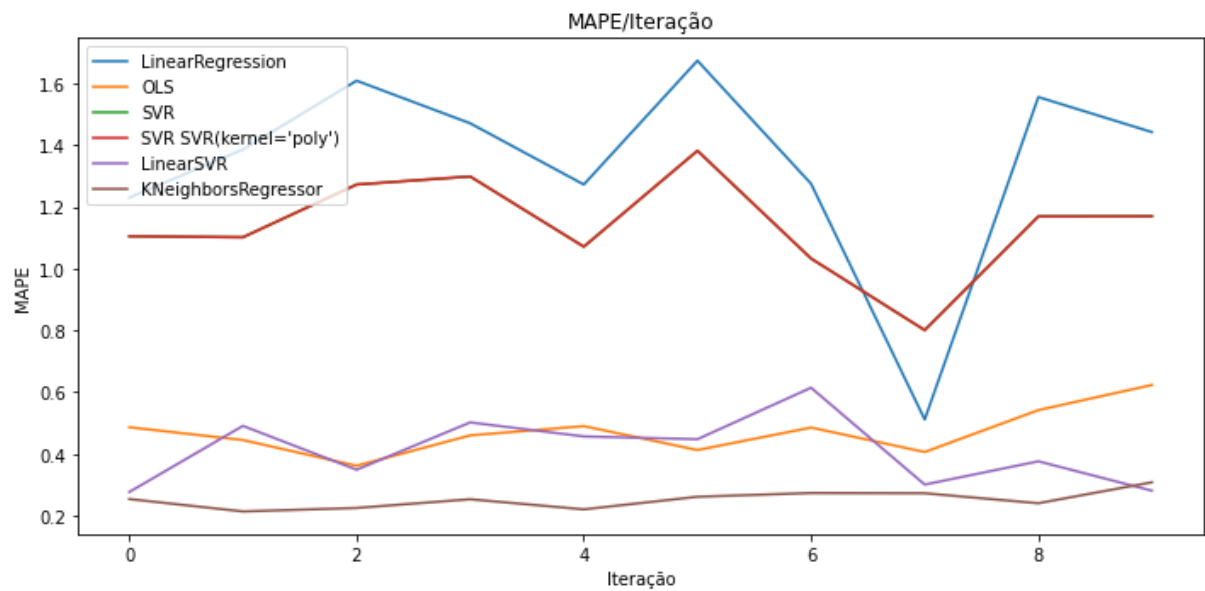


Figura 5.11 – Gráfico comparativo do MAPE para os modelos adotados com 10 iterações e divisão aleatória (shuffle=True)

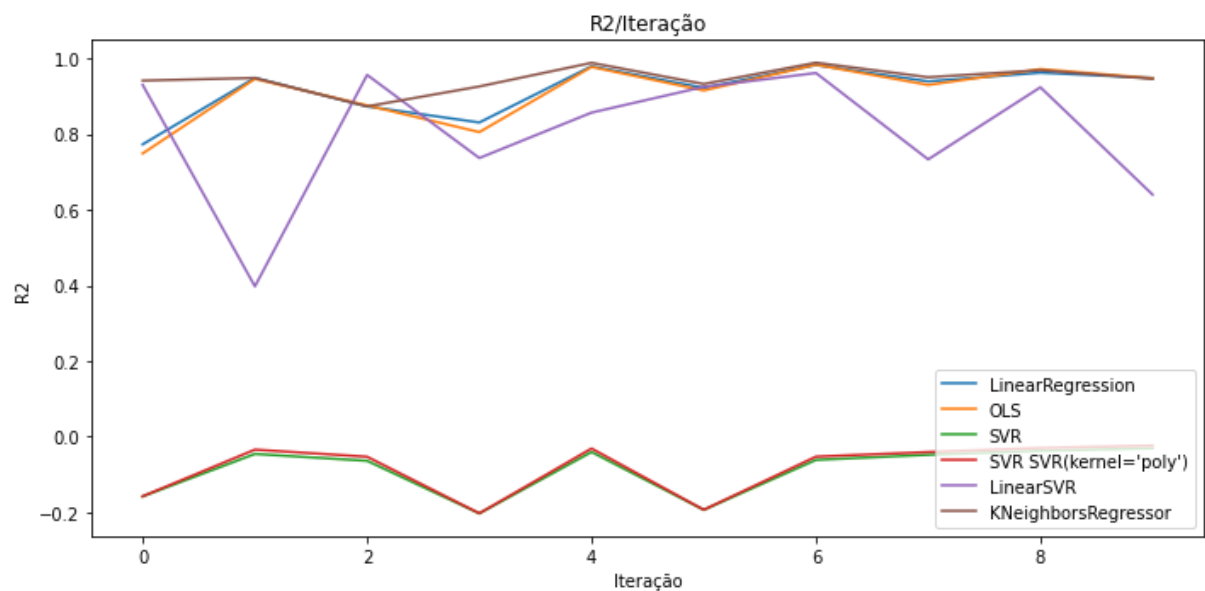


Figura 5.12 – Gráfico comparativo do R2 score para os modelos adotados com 10 iterações e divisão aleatória (shuffle=True)

Nesta avaliação (com divisão aleatória), observamos uma mudança significativa dos gráficos/resultados. O SVR linear que havia apresentado uma boa performance nas duas primeiras avaliações, apresentou uma piora em todas as avaliações (MSE, RMSE, MAPE e R2). Observamos que o KNeighborsRegressor apresentou uma

melhora em todas as avaliações. LinearRegression e OLS apresentaram resultados positivos no MSE e RMSE, mas não tiveram resultados satisfatórios no MAPE

Na sequência, ajustaremos para 8 *Folds* (= número de meses), com divisão aleatória:

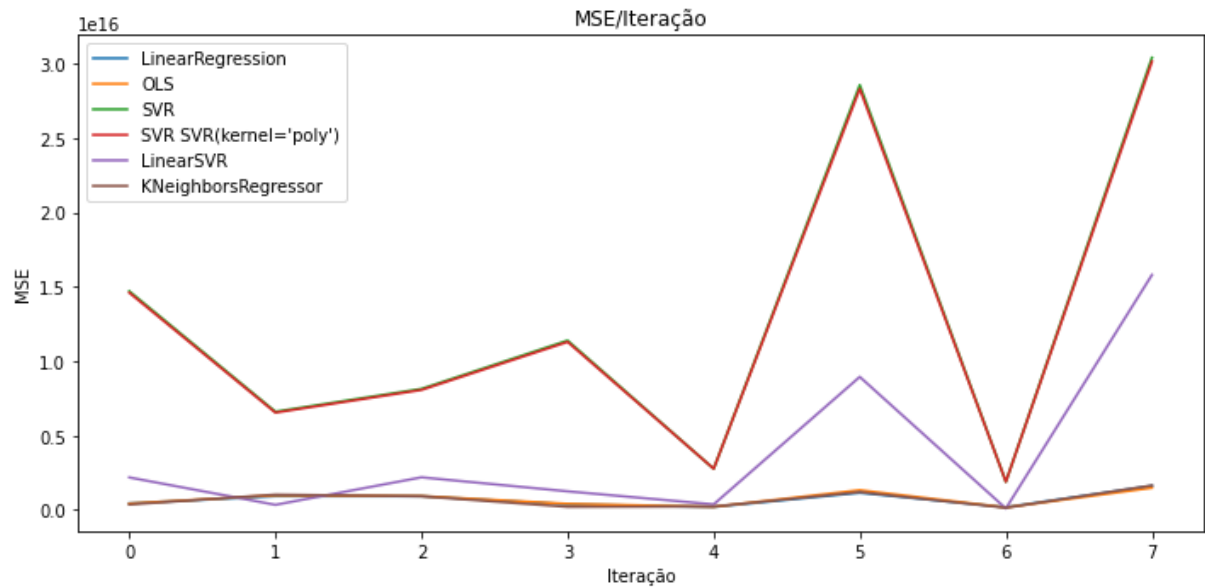


Figura 5.13 – Gráfico comparativo do MSE para os modelos adotados com 8 iterações e divisão aleatória (*shuffle=True*)(em 10 quadrilhões)

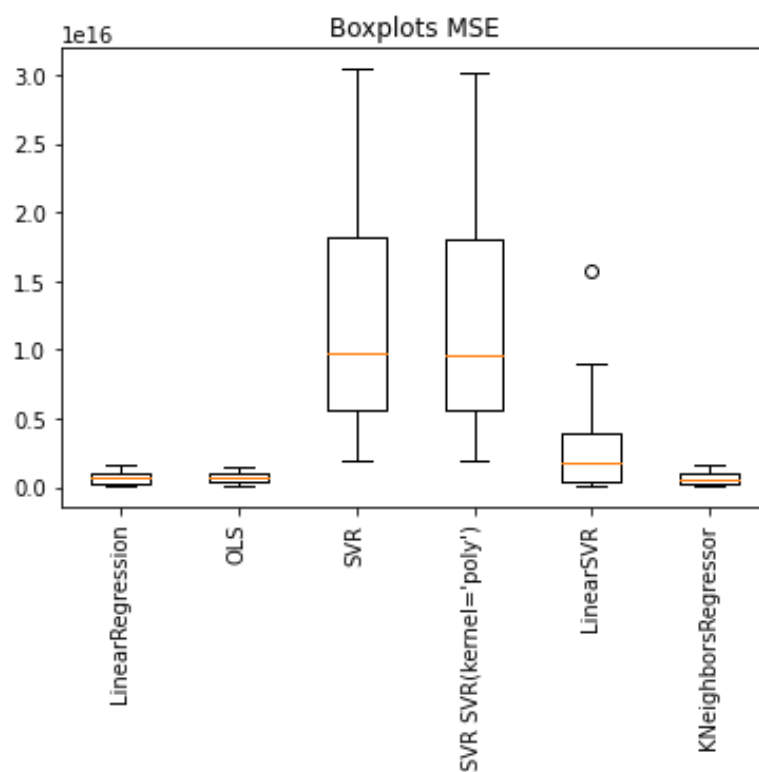


Figura 5.14 – Boxplot comparativo do MSE para os modelos adotados com 8 iterações e divisão aleatória (shuffle=True)

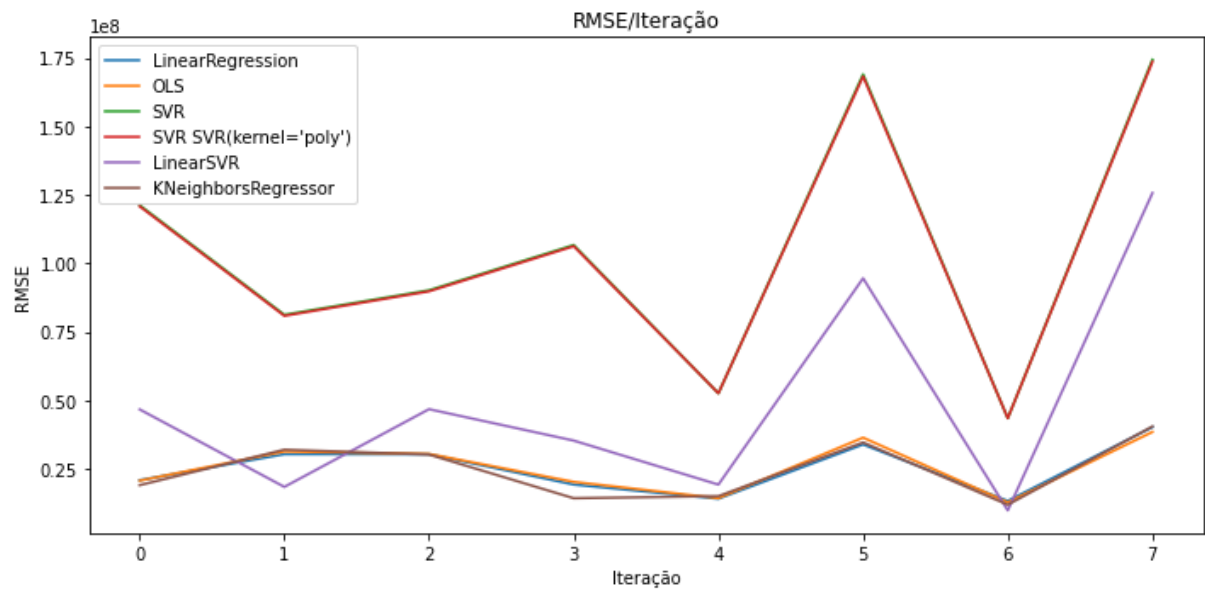


Figura 5.15 – Gráfico comparativo do RMSE para os modelos adotados com 8 iterações e divisão aleatória (shuffle=True) (em 100 milhões)

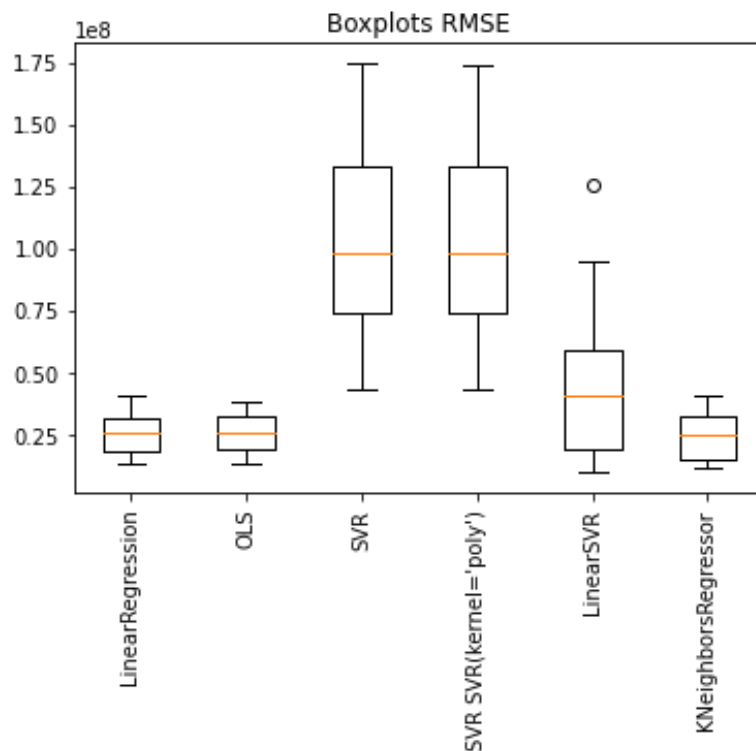


Figura 5.16 – Boxplot comparativo do RMSE para os modelos adotados com 8 iterações e divisão aleatória (shuffle=True)

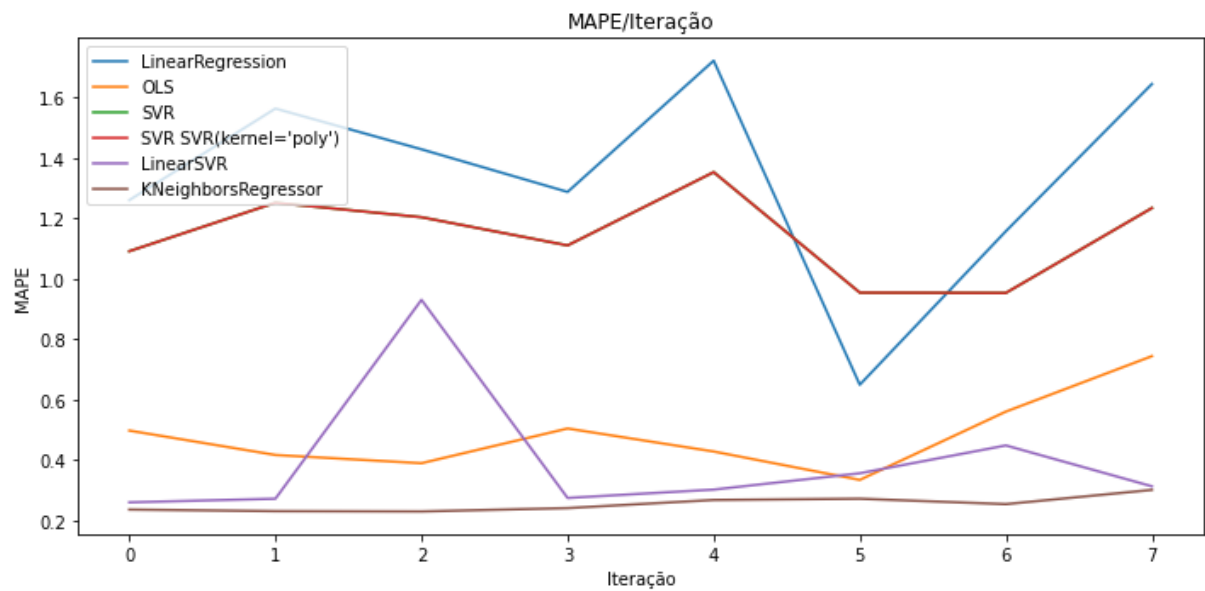


Figura 5.17 – Gráfico comparativo do MAPE para os modelos adotados com 8 iterações e divisão aleatória (*shuffle=True*)

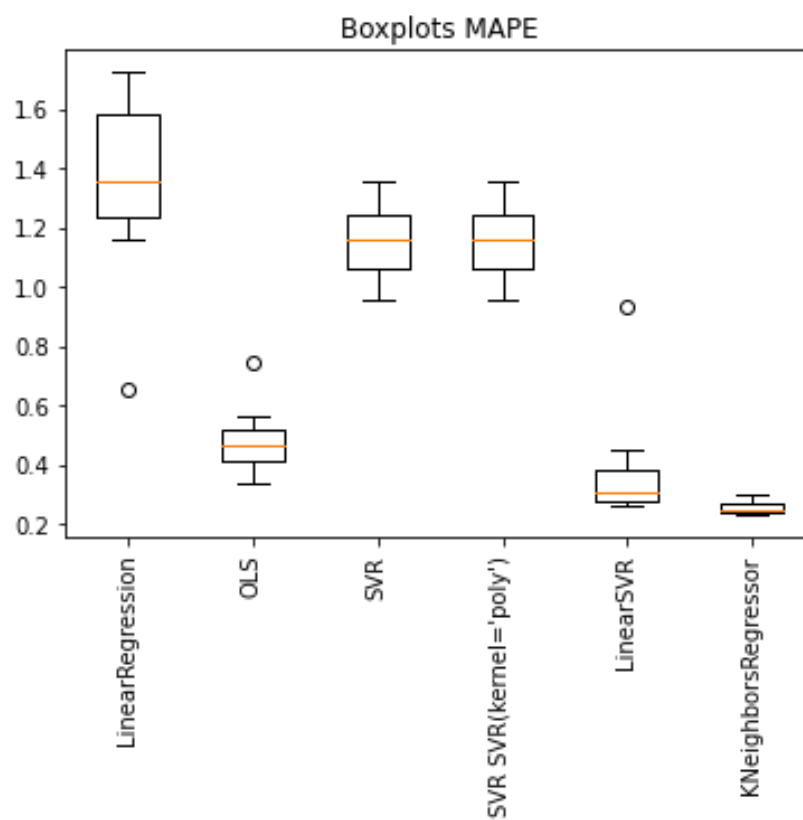


Figura 5.18 – Boxplot comparativo do MAPE para os modelos adotados com 8 iterações e divisão aleatória (*shuffle=True*)

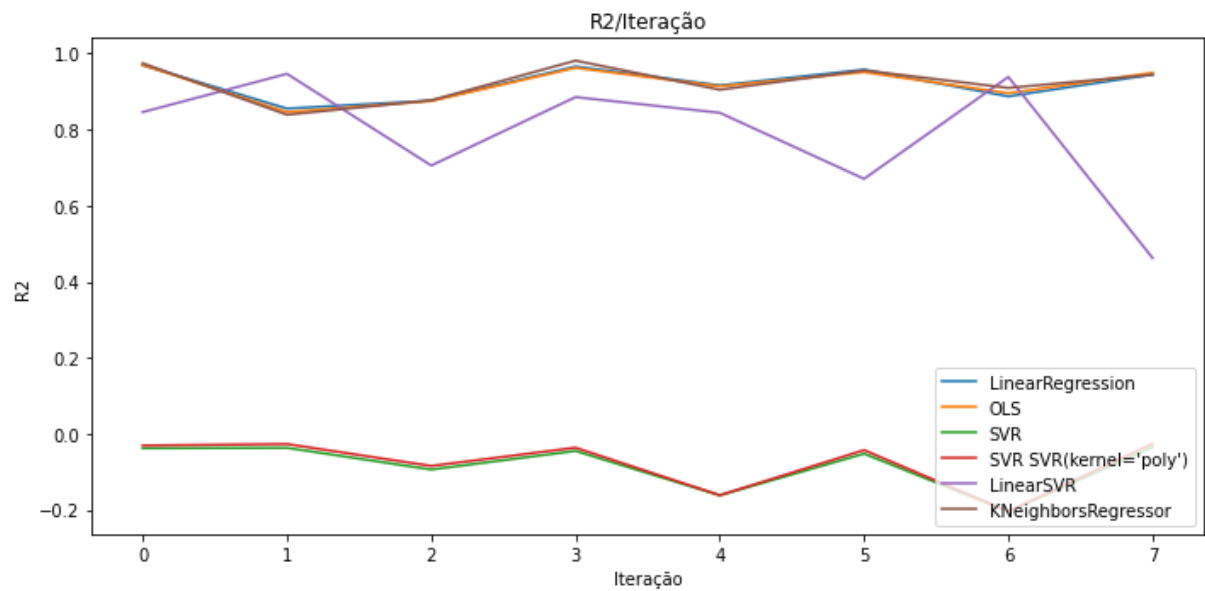


Figura 5.19– Gráfico comparativo do R2 score para os modelos adotados com 8 iterações

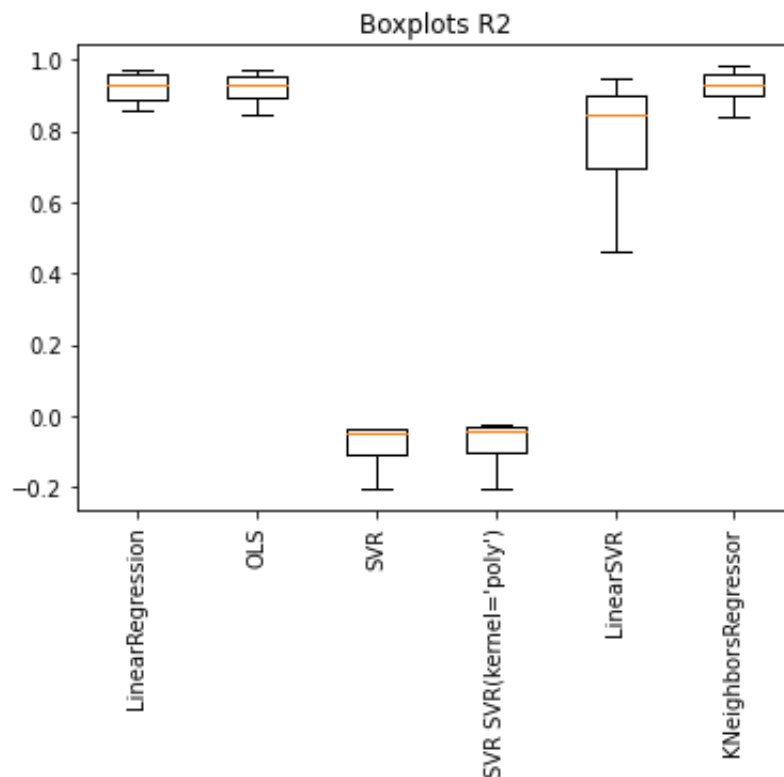


Figura 5.20 – Boxplot comparativo do R2 score para os modelos adotados com 8 iterações

Avaliando os gráficos dos resultados usando 10 e 8 *folds* (da divisão aleatória), podemos observar o seguinte:

- O modelo KNR apresentou, no geral, o melhor resultado seguido do OLS e *LinearRegression* que apresentaram variações significativas no MAPE. Estes

modelos foram bastante superiores aos SVR com *kernel*s rbf e polinomial para esses dados;

- O modelo SVR linear oscilou bastante com a divisão aleatória, indicando sua maior sensibilidade a distribuição dos dados.

Embora a análise gráfica já nos sinalize os melhores modelos, devemos realizar a avaliação com teste de hipótese das diferenças entre os modelos (usando o *8-Fold Cross Validation* aleatório), para melhorar a nossa escolha.

Teste de Hipóteses

Apesar de criarmos diferentes gráficos e computar diferentes métricas, para afirmarmos que um modelo é melhor (ou mais adequado do ponto de vista estatístico) que outro precisamos realizar um teste de hipóteses. Vamos realizar o teste de hipótese com o auxílio da biblioteca *autorank*, a partir da métricas MSE calculadas com o *8-Fold Cross Validation* aleatório.

Numa breve explanação, dentro a Inferência Estatística, o Teste de Hipóteses é uma metodologia que nos fornece um instrumento que permite rejeitar ou não rejeitar uma hipótese estatística por meio da evidência fornecida pela amostra. Neste sentido, temos uma Hipótese Nula, denotada por H_0 (que contém uma afirmativa de igualdade “=”), e uma hipótese complementar chamada de Hipótese Alternativa, denotada por H_a (que se assume como verdadeira, se H_0 for rejeitada).

Diante disso, iniciamos o teste assumindo, independentemente de qual das hipóteses representa a alegação (a ser provada), que a condição de igualdade na H_0 é verdadeira. Assim, quando realizarmos um teste de hipóteses, teremos duas decisões possíveis:

- 1) Rejeitar H_0 , indicando que H_a é verdadeira; ou
- 2) Aceitar H_0 .

Neste contexto, não podemos desprezar a possibilidade de se tomar uma decisão errada. O quadro a seguir mostra os quatro resultados possíveis de um teste de hipóteses:

Decisão:	Situação Real (da Hipótese Nula):	
	Ho é Verdadeira	Ho é Falsa
Aceitar Ho	Decisão Correta (1- α)	Erro tipo II (β)
Rejeitar Ho	Erro tipo I (α)	Decisão Correta (1- β)

Figura 5.21 – Quadro dos Tipos de Erros

Temos então:

· <u>Erro tipo I (denotada por α)</u> : Ocorre se a hipótese nula (Ho) for rejeitada quando ela for realmente verdadeira. É o chamado falso positivo	· <u>Erro tipo II (denotada por β)</u> : Ocorre se a hipótese nula (Ho) não for rejeitada quando ela for realmente falsa. É o falso negativo.
--	--

A partir daí, devemos observar o nível de significância (α) de um teste que é a probabilidade de uma hipótese nula (Ho) ser rejeitada quando ela é verdadeira. Em geral, os programas de estatística definem como 5% ou 0,05 o erro máximo tolerável no seu experimento.

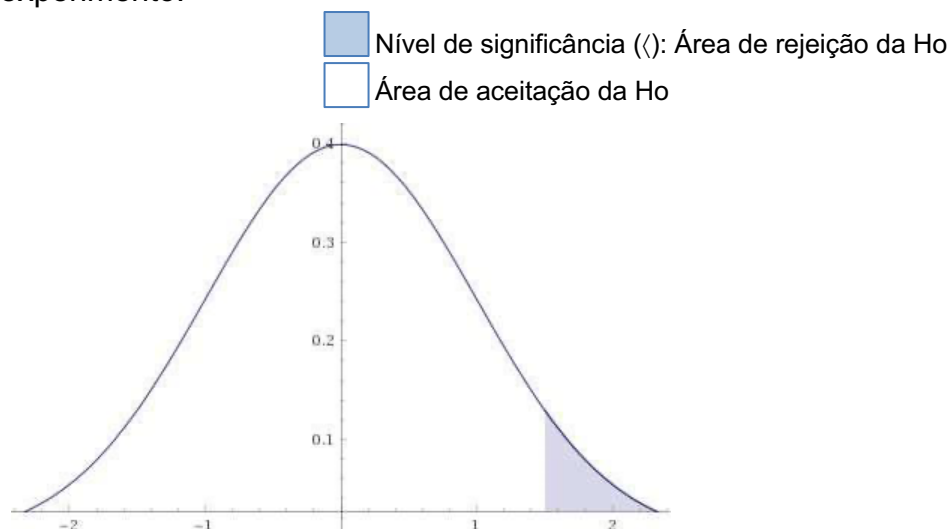


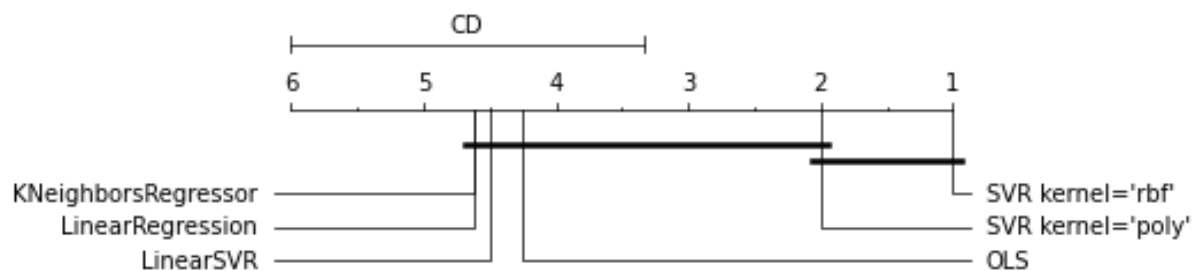
Figura 5.22 – Distribuição amostral representando a estatística de teste, indicando as áreas de aceitação separadas pelo valor crítico (ponto divisório)

Dentre os “caminhos” possíveis para conclusão do Teste de Hipóteses, utilizaremos o Valor p, que é adotado pela maioria dos softwares estatísticos. Esta abordagem reflete a probabilidade da Ho ser (de fato) verdadeira, ou seja, quantifica o erro cometido ao rejeitar a hipótese nula. Um Valor p muito pequeno sugere que os resultados amostrais

são muito improváveis sob a hipótese nula, ou seja, constitui evidência contra a hipótese nula. Ou seja:

· <u>Rejeitar a hipótese nula (Ho)</u> se o Valor p é no máximo igual ao nível de significância (α)	· <u>Aceitar a hipótese nula (Ho)</u> se o Valor p é maior do que o nível de significância (α).
--	--

Após esta fundamentação teórica, seguimos na avaliação dos modelos:



Resultado Estatístico:

The statistical analysis was conducted for 6 populations with 8 paired samples.

The family-wise significance level of the tests is $\alpha=0.050$.

We failed to reject the null hypothesis that the population is normal for all populations (minimal observed p-value=0.016). Therefore, we assume that all populations are normal.

We applied Bartlett's test for homogeneity and reject the null hypothesis ($p=0.000$) that the data is homoscedastic. Thus, we assume that our data is heteroscedastic.

Because we have more than two populations and the populations are normal but heteroscedastic, we use the non-parametric Friedman test as omnibus test to determine if there are any significant differences between the mean values of the populations. We use the post-hoc Nemenyi test to infer which differences are significant. We report the mean value (M), the standard deviation (SD) and the mean rank (MR) among all populations over the samples. Differences between populations are significant, if the difference of the mean rank is greater than the critical distance $CD=2.666$ of the Nemenyi test.

We reject the null hypothesis ($p=0.000$) of the Friedman test that there is no difference in the central tendency of the populations SVR ($M=13062432500707424.000$, $SD=10980485243431714.000$, $MR=1.000$), SVR SVR(kernel='poly') ($M=12956726946740730.000$, $SD=10879987434402318.000$, $MR=2.000$), LinearSVR ($M=1808003386005340.000$, $SD=2156322527834716.500$, $MR=4.125$), OLS ($M=741108657780631.000$, $SD=508888207887439.438$, $MR=4.375$), LinearRegression ($M=725055525222255.500$, $SD=513652906371172.562$, $MR=4.750$), and KNeighborsRegressor ($M=714820849981967.375$, $SD=558086045992135.938$, $MR=4.750$). Therefore, we assume that there is a statistically significant difference between the median values of the populations.

Based on the post-hoc Nemenyi test, we assume that there are no significant differences within the following groups: SVR and SVR SVR(kernel='poly'); SVR SVR(kernel='poly'), LinearSVR, and OLS; LinearSVR, OLS, LinearRegression, and KNeighborsRegressor. All other differences are significant.

Considerando os resultados do Teste de Hipótese, podemos entender que os modelos significativamente relevantes são os contidos no intervalo da distância crítica (CD): KNR, *LinearRegression*, *LinearSVR* e OLS. Por estar no limite, optamos por não seguir com o SVR com *kernel* polinomial.

6. Apresentação dos Resultados

A partir do teste de hipóteses, pré-selecionamos 4 modelos: *LinearRegression*, OLS, *LinearSVR* e *KNeighborsRegressor*

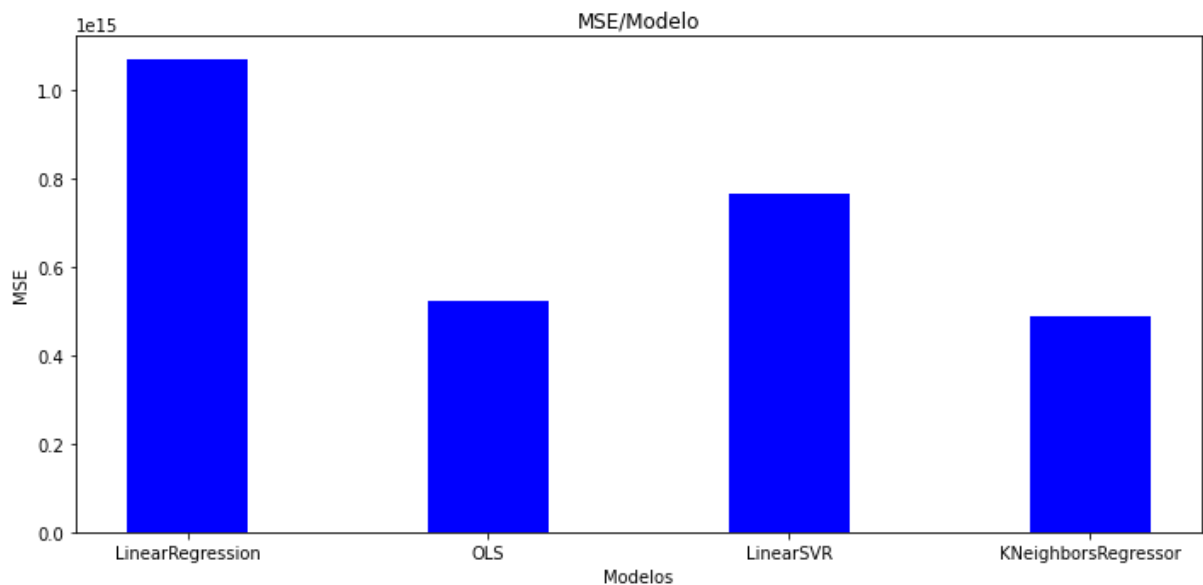


Figura 6.0 – Gráfico de colunas com os valores dos modelos selecionados no MSE (em quatrilhão)

OLS e KNR se destacaram na validação MSE.

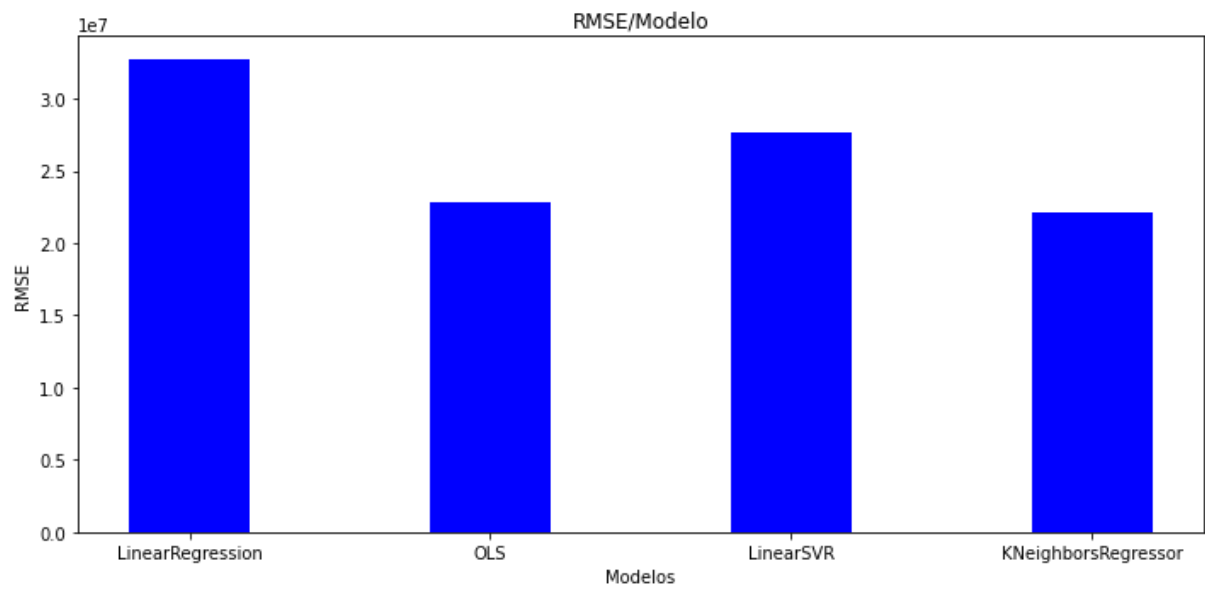


Figura 6.1 – Gráfico de colunas com os modelos selecionados no RMSE (em 10 milhões)

OLS e KNR mantiveram os melhores resultados na validação MSE.

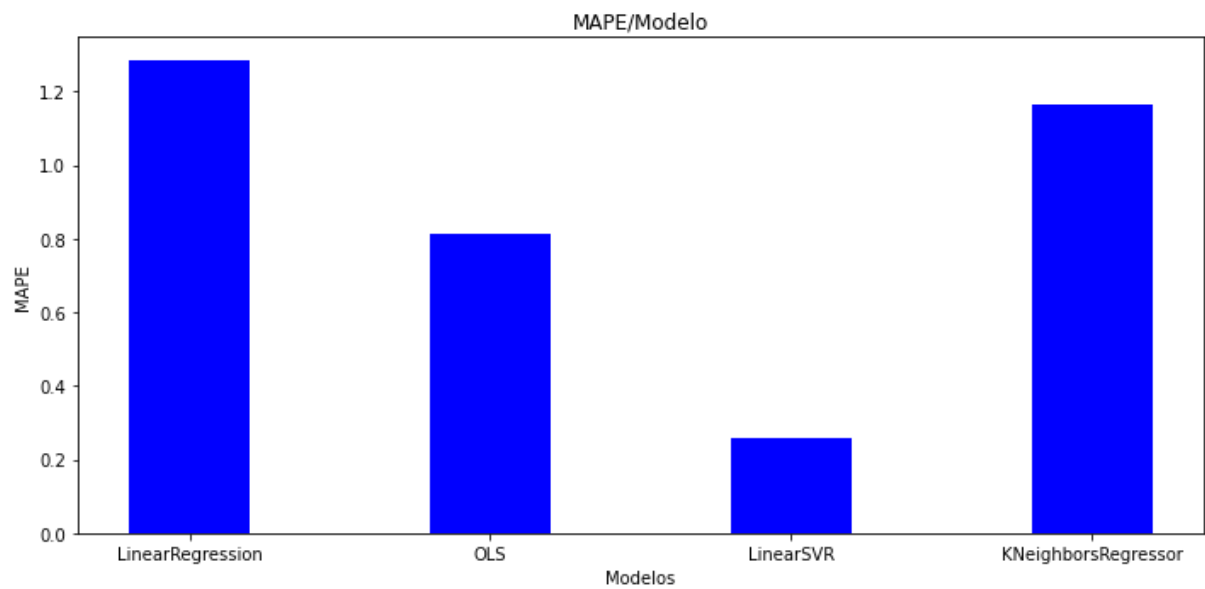


Figura 6.2 – Gráfico de colunas com os modelos selecionados no MAPE

OLS manteve o bom resultado das análises anteriores na validação MAPE e o SVR linear se destacou positivamente.

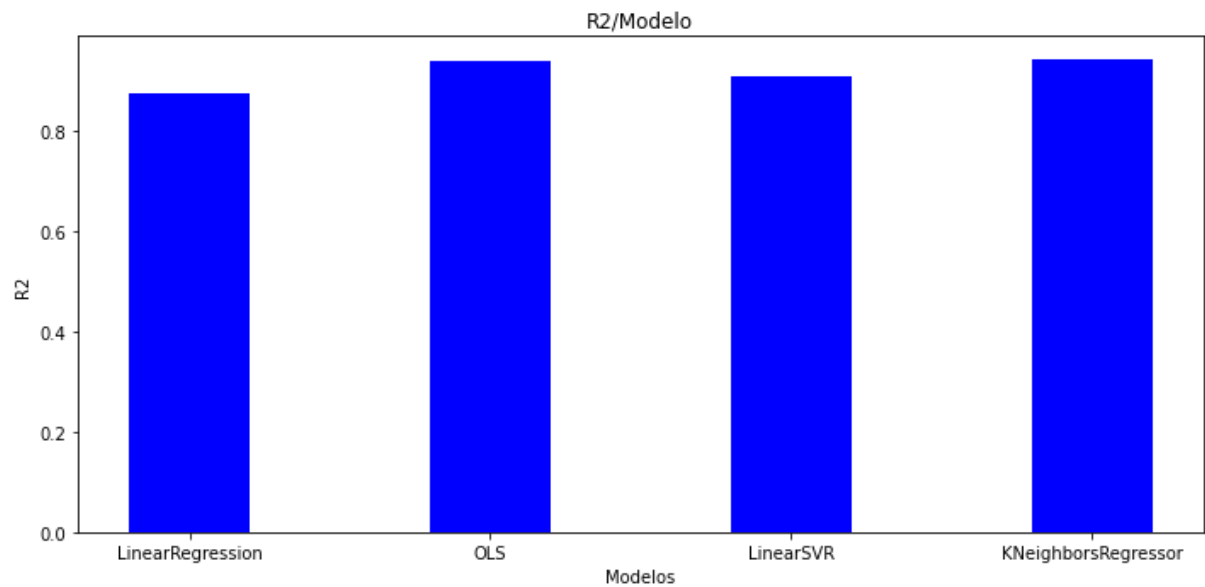


Figura 6.3 – Gráfico de colunas com os modelos selecionados no R2

OLS e KNR foram ligeiramente melhores (mais próximos de 1), seguidos do SVR linear e do LinearRegression.

Avaliação do Predito x Real

A partir da avaliação dos modelos, obtemos os gráficos comparativos dos resultados dos modelos (treinados) induzidos com o resultado real de dezembro/20 (mês de testes).

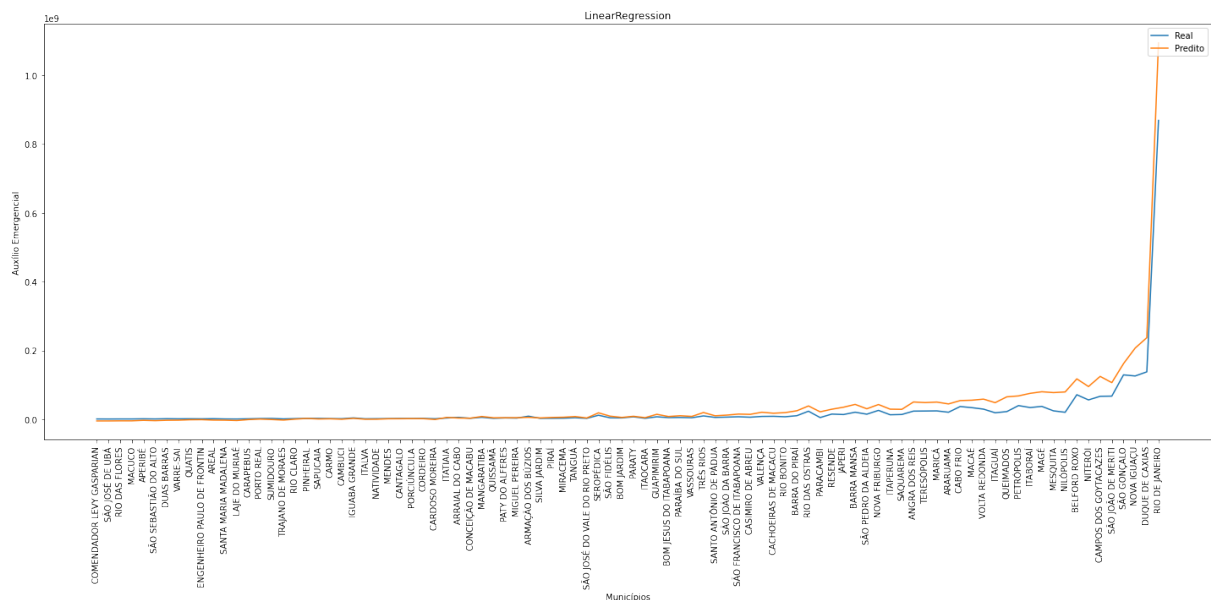


Figura 6.4 – Gráfico do modelo LinearRegression comparativo do valor predito x real em dezembro/20 (em bilhões)

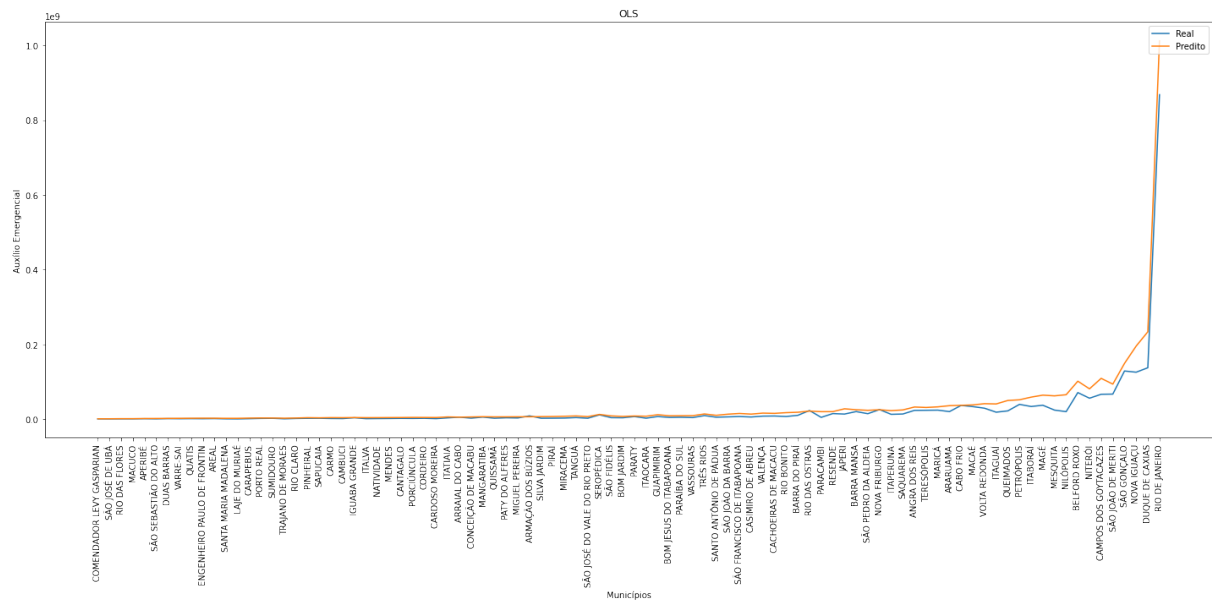


Figura 6.5 – Gráfico do modelo OLS comparativo do valor predito x real em dezembro/20 (em bilhões)

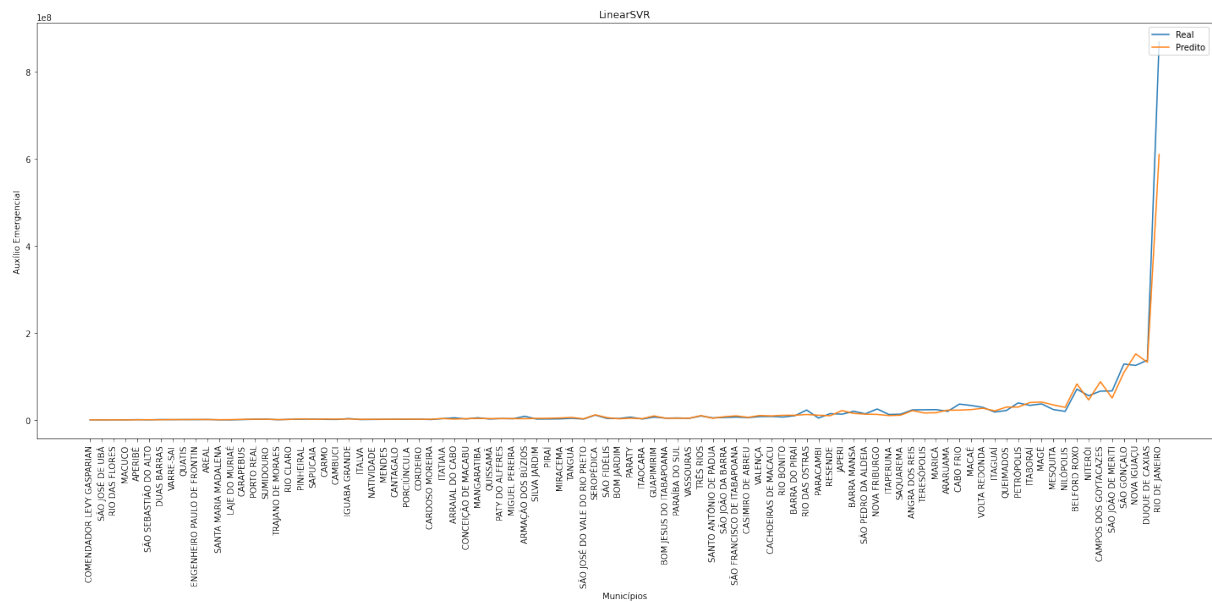


Figura 6.6 – Gráfico do modelo LinearSVR comparativo do valor predito x real em dezembro/20 (em bilhões)

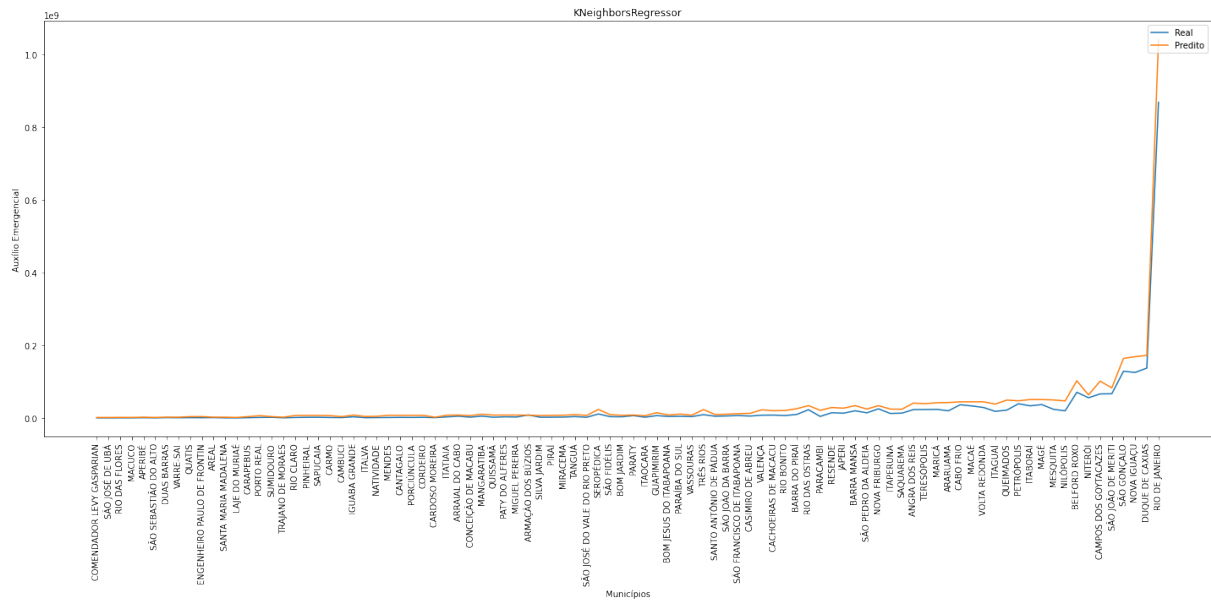


Figura 6.7 – Gráfico do modelo KNR comparativo do valor predito x real em dezembro/20 (em bilhões)

Indução dos Modelos de Regressão

Após a indução dos modelos, que pode ser observada nos gráficos da seção acima, o próximo passo é obter as funções dos modelos. Neste contexto, é importante destacar que o modelo KNR, que é um dos mais utilizados e simples em *Machine Learning*, é conhecido como um algoritmo de aprendizado lento ou, melhor dizendo, “lazy” (preguiçoso), não tendo função associada.

OLS - Ordinary Least Squares

Iniciamos a análise pelo OLS (em português: Mínimos Quadrados Ordinários (MQO)), com todo o conjunto de treinamento. O modelo gera uma variável resposta (no caso, AUXÍLIO_EMERGENCIAL) a partir de um conjunto de variáveis explicativas (ou preditoras). Na análise comparativa, foi o modelo que apresentou um dos melhores ajustes para a Regressão.

OLS Regression Results

Dep. Variable: y **R-squared (uncentered):** 0.958
Model: OLS **Adj. R-squared (uncentered):** 0.958
Method: Least Squares **F-statistic:** 5591.
Date: Thu, 27 May 2021 **Prob (F-statistic):** 0.00
Time: 14:26:45 **Log-Likelihood:** -13543.
No. Observations: 736 **AIC:** 2.709e+04
Df Residuals: 733 **BIC:** 2.710e+04
Df Model: 3
Covariance Type: nonrobust

	coef	std err	t	P> t	[0.025	0.975]
x1	6.9257	0.290	23.901	0.000	6.357	7.495
x2	0.2822	0.498	0.567	0.571	-0.695	1.259
x3	3.6995	0.683	5.420	0.000	2.359	5.040

Omnibus: 380.990 **Durbin-Watson:** 1.700
Prob(Omnibus): 0.000 **Jarque-Bera (JB):** 185962.733
Skew: 0.938 **Prob(JB):** 0.00
Kurtosis: 80.849 **Cond. No.** 12.4

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Recordando a questão do Teste de Hipóteses, temos que o “t” (divisão do “coef” pelo erro padrão ou “std erro”) viabiliza, a partir da consulta a tabela de distribuição normal, o teste da Hipótese Nula (H_0) que, no caso, seria de o valor teórico do coeficiente ser igual a zero. Tal hipótese, pode ser rejeitada se o Valor p (associado ao nível de significância) for inferior 0,05, indicando que a H_0 pode ser rejeitada com 95% de certeza.

A partir da tabela anterior, temos que os valores de coeficiente obtidos pela Regressão Linear com o modelo OLS (do *statsmodel*) assim como seus respectivos Valor p (“p-valor”) são:

Coeficiente	Mapeamento	Valor	p-valor
x1	BPC	6.9257	0.000
x2	FPM	0.2822	0.571
x3	BOLSA FAMILIA	3.6995	0.000

Analisando os dados destacados pela tabela acima, percebemos que o valor de FPM não é estatisticamente significativo para o modelo, pois possui um Valor p > 5%. Neste caso, o ideal a ser feito é induzir novamente o modelo sem essa variável.

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.958
Model:	OLS	Adj. R-squared (uncentered):	0.958
Method:	Least Squares	F-statistic:	8394.
Date:	Thu, 27 May 2021	Prob (F-statistic):	0.00
Time:	14:26:45	Log-Likelihood:	-13543.
No. Observations:	736	AIC:	2.709e+04
Df Residuals:	734	BIC:	2.710e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
x1	6.9475	0.287	24.201	0.000	6.384	7.511
x2	3.7869	0.665	5.698	0.000	2.482	5.092

Omnibus:	372.047	Durbin-Watson:	1.704
Prob(Omnibus):	0.000	Jarque-Bera (JB):	189130.674
Skew:	0.866	Prob(JB):	0.00
Kurtosis:	81.513	Cond. No.	11.9

Warnings:

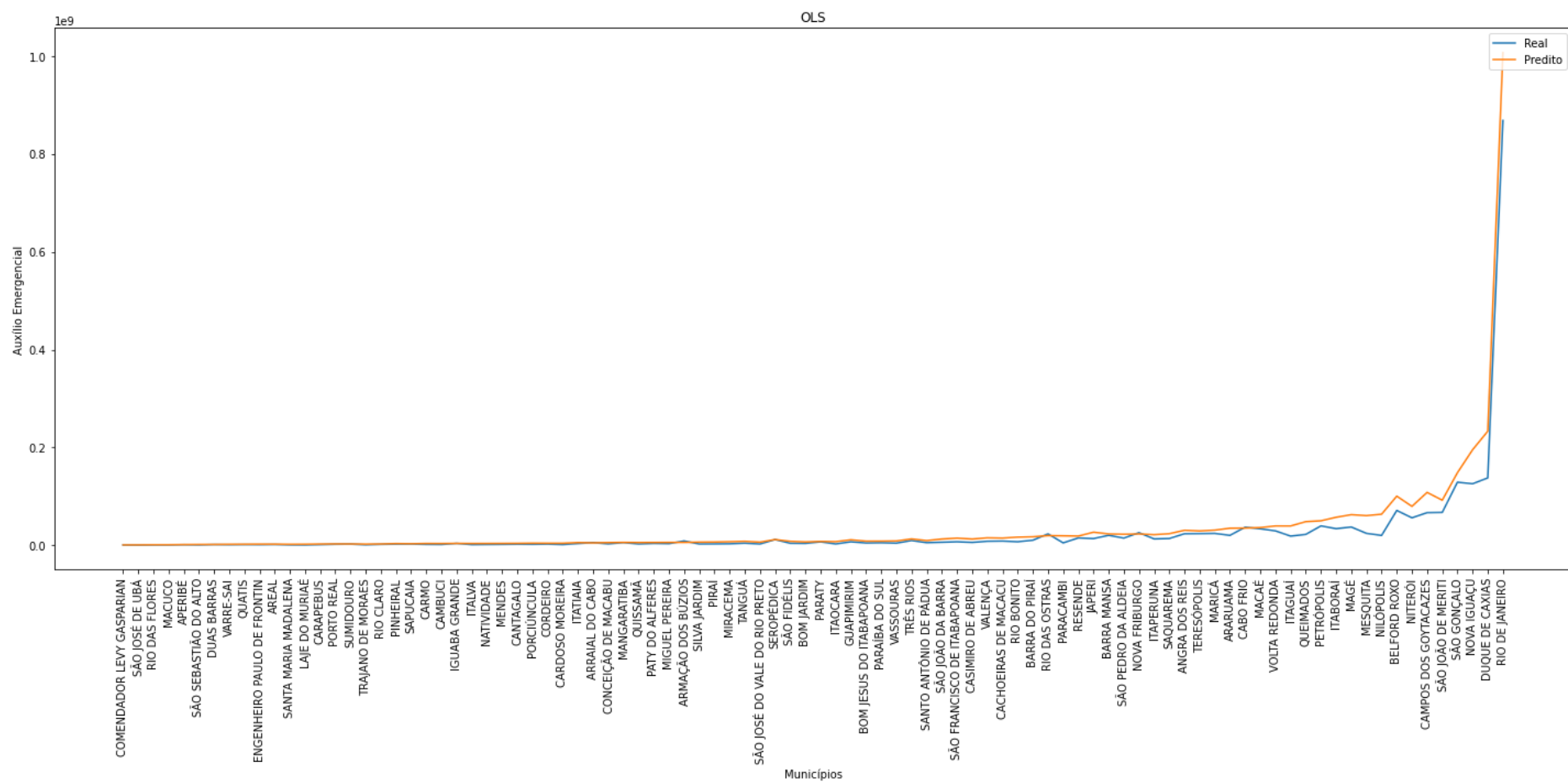
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Note que os valores dos coeficientes foram levemente alterados. Daí podemos escrever a função de regressão induzida pelo modelo como:

$$AUXILIO_EMERGENCIAL = 6.9475 \cdot BPC + 3.7869 \cdot BOLSA_FAMILIA$$

A partir do ajuste temos uma nova visão gráfica do modelo:

Figura 6.8 – Gráfico OLS ajustado (em bilhões)



LinearSVR

Realizamos a análise do *LinearSVR* induzido, que se destacou no MAPE, com todo o conjunto de treinamento:

$$AUXILIO_EMERGENCIAL = 1.02850565 \cdot BPC + 0.44230761 \cdot FPM + 4.26802729 \cdot BOLSA_FAMILIA - 0.09803584$$

Nesta função de regressão induzida pelo Linear SVR, é dado mais peso para o BOLSA_FAMILIA, diminuindo significativamente o peso do BPC, o que reflete numa constante negativa bem menor também. O FPM, que foi excluído do modelo OLS, tem uma pequena relevância em relação ao peso das demais variáveis BPC e BF.

Linear Regression

Realizamos a análise do Linear Regression (do *sklearn*) induzido com todo o conjunto de treinamento

$$AUXILIO_EMERGENCIAL = 6.25515798 \cdot BPC + 3.81634711 \cdot FPM + 4.03283863 \cdot BOLSA_FAMILIA - 9182542.084970292$$

É interessante notar que o coeficiente relativo ao BPC manteve a faixa de valor quando comparado ao valor obtido com o OLS (do *statsmodel*). Enquanto isso, o coeficiente relativo ao BOLSA_FAMILIA teve sua diferença menor que a diferença entre o BPC deste modelo para o modelo do OLS. Por fim, cabe ressaltar que o uso da variável FPM com um valor significativo resultou na adição de uma constante negativa de grande ordem.

Conclusão e Resumo no Canvas

Por fim, podemos concluir, após a “jornada” da predição descrita no modelo conceitual *Business Model Canvas*, que foi assertiva a escolha das variáveis preditoras (BF, BPC e FPM) e dos algoritmos de Regressão Linear selecionados (*OLS*, *LinearSVR*, *Linear Regression* e *KNR*), após a avaliação dos modelos pelas 4 métricas aplicadas (MSE, RMSE, MAPE e R2). O desempenho dos modelos foi relativamente próximo. Podemos dizer que, embora não sejam diferentes estatisticamente, numericamente o

OLS (cuja função e gráfico da indução foram apresentados nesta seção) apresentou um ligeiro destaque, seguido do KNR (que é um modelo “*lazy*”). O LinearSVR e o Linear Regression completaram o rol (cujas funções e gráficos da indução foram apresentados nesta seção).

Título (da Predição): A importância de estimar o Auxílio Emergencial no contexto econômico dos Municípios Fluminenses		
<p>Problema Proposto: <i>Que problema se está tentando resolver? Quais os principais pontos a serem tratados?</i></p> <ul style="list-style-type: none"> - Estimar, neste cenário pandêmico, o valor de um novo Auxílio Emergencial (AE) a ser pago nos municípios (92) do Estado do Rio de Janeiro, considerando a necessidade de manutenção do equilíbrio econômico destes entes e a sustentabilidade social. 	<p>Resultados/Previsões: <i>Que previsões você está tentando fazer? Identificar as variáveis preditoras a variável alvo.</i></p> <ul style="list-style-type: none"> - Variável alvo: valor de um novo Auxílio Emergencial (AE); - Variáveis preditoras/explicativas: valores dos benefícios assistenciais – Bolsa Família (BF) e Benefício de Prestação Continuada da LOAS (BPC) – e dos repasses da União aos municípios – Fundo de Participação dos Municípios (FPM). 	<p>Coleta de Dados: <i>Quais são as fontes para os seus dados? Há dados suficientes? Você pode trabalhar os dados?</i></p> <ul style="list-style-type: none"> - Fonte dos dados: <ul style="list-style-type: none"> - Portal da Transparência - valor do Auxílio Emergencial (AE), Bolsa Família (BF) e Benefício de Prestação Continuada da LOAS (BPC); - Portal de Dados Abertos – valor dos repasses da União aos municípios – Fundo de Participação dos Municípios (FPM); - Site do Instituto Brasileiro de Geografia e Estatística - IBGE – dados gerais (geográficos, demográficos, econômicos etc.); - Portal da Secretaria de Estado de Fazenda do Rio de Janeiro – dados sobre a arrecadação do ICMS nos municípios.
<p>Preparação dos Dados: <i>O que você precisa fazer em seus dados de modo a aplicar seus modelos e obter os resultados?</i></p> <ul style="list-style-type: none"> - Após a baixa/carga alguns dados precisaram ser ajustados (corrigidos, concatenados ou transformados); - Foi realizada uma exploração preliminar dos dados, que incluiu o uso de georeferenciamento e uma análise gráfica, a partir de dados complementares do IBGE e SEFAZ-RJ; - Foram realizadas algumas análises estatísticas dos dados; - Foi realizada uma análise de correlação (de Pearson) das variáveis (AE, BF, BPC e FPM) para predição; - Antes da predição ppd, foi adotada a técnica de Validação Cruzada, que demandou a divisão dos dados de treinamento e teste. 	<p>Criando o Modelo: <i>Que modelos são mais apropriados diante dos resultados esperados?</i></p> <ul style="list-style-type: none"> - Foram adotados 4 modelos de Regressão Linear que nos permitiram 6 análises distintas: <ul style="list-style-type: none"> - OLS - Ordinary Least Squares; - Linear Regression; - KNR - KNeighborsRegressor; - SVR - Support Vector Regres (kernels=[RBF, polinomial e linear]). 	<p>Avaliação do Modelo: <i>Como você pode avaliar a performance do seu modelo?</i></p> <ul style="list-style-type: none"> - Foram adotados 4 métricas de avaliação dos modelos de Regressão Linear: <ul style="list-style-type: none"> - R² score; - MSE - Mean Square Error; - RMSE - Root Mean Square Error; - MAPE - Mean Absolute Percentage Error. - Foi realizado um Teste de Hipóteses para entender os modelos significativamente relevantes.

Figura 6.9 – modelo conceitual Business Model Canvas (proposto por Jasmine Vasandani) que resume as principais fases do trabalho, os insumos necessários e os resultados esperados em cada uma delas.

7. Links

Nesta seção, indicamos o link para a apresentação resumo da predição, além dos scripts e bases do Projeto.

Link para o vídeo: <https://youtu.be/UaUO8v7MaW0>

Link para o repositório: <https://github.com/RogérioSM66/Projeto-TCC-RSM-PUCMinas.git>

REFERÊNCIAS

BRASIL. Lei nº 13.982, de 2 de abril de 2020. Diário Oficial da República Federativa do Brasil, Brasília, DF, 02 abr. 2020. Disponível em: <https://www.in.gov.br/en/web/dou/-/lei-n-13.982-de-2-de-abril-de-2020-250915958> Acesso em: 29 abr. 2021.

BRASIL. Lei nº 8.666, de 21 de junho de 1993. Diário Oficial da República Federativa do Brasil, Brasília, DF, 22 jun. 1993. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm Acesso em: 29 abr. 2021.

FIRJAN. Federação das Indústrias do Estado do Rio de Janeiro – Nota Técnica - Resultados e perspectivas para o PIB do RJ. FIRJAN Publicações. Rio de Janeiro, RJ, 2021, Disponível em: <https://www.firjan.com.br/publicacoes/publicacoes-de-economia/pib-brasil-e-rio-de-janeiro-resultados-e-projecoes.htm> Acesso em: 29 jun. 2021

FRANÇA, Álvaro Sólón. A Previdência Social e a Economia dos Municípios. Brasília, DF: ANFIP, 2019. Disponível em: <https://www.anfip.org.br/wp-content/uploads/2019/04/2019-Economia-dos-munici%C3%A7%C3%B5es.pdf> Acesso em: 28 dez. 2020.

GONZALES, Lauro; OLIVEIRA, Leonardo. NOVO AUXÍLIO EMERGENCIAL- CENÁRIOS E EFEITOS SOBRE RENDA. Rio de Janeiro, RJ: FGV/EAESP 2020. Disponível em: <https://eaesp.fgv.br/sites/eaesp.fgv.br/files/u676/aefevereiro2021vfinal.pdf> Acesso em: 20 jun. 2021.

IBPT. Instituto Brasileiro de Planejamento e Tributação – QUEDA DA ARRECAÇÃO TRIBUTÁRIA EM DECORRÊNCIA DOS EFEITOS DA PANDEMIA DO CORONAVÍRUS. IBPT Blog Estudos. Curitiba, PR, 2020, Disponível em: <https://ibpt.com.br/queda-da-arrecadacao-tributaria-em-decorrencia-dos-efeitos-da-pandemia-do-coronavirus> Acesso em: 27 jun. 2021

MENEZES, Naércio; KOMATSU, Bruno K.; ROSA, João Pedro Rosa. Reducing Poverty and Inequality during the Coronavirus Outbreak: The Emergency Aid Transfers in Brazil. São Paulo, SP, 2021: INSPER Policy Paper. Disponível em: https://www.insper.edu.br/wp-content/uploads/2021/02/Policy_Paper_54.pdf Acesso em: 19 mar. 2021.