Médias Móveis em contagem de infectados por Covid-19 em Curitiba

Bruno Luvizotto Carli

Universidade Federal do Paraná

1 Introdução

O presente artigo pretende investigar os resultados da aplicação de técnicas de pré-processamento de dados sob uma base de dados de fonte pública. O tema escolhido foi Infecções confirmadas por Covid-19 no Município de Curitiba. A prefeitura disponibiliza uma base de dados (OLIVEIRA, 2021), de casos confirmados da doença, de onde se pretende extrair os dados, identificar uma variável preditora, aplicar um modelo de machine learning e identificar os resultados de melhoria que as técnicas de escalonamento e padronização dos dados na fase de pré-processamento podem propiciar ao modelo.

2 Trabalhos relacionados

Rizatti et al. (2020) realizaram um mapeamento da pandemia em Santa Maria/RS ao qual apresentou ideias enaltecedoras na resolução da visualização do nível de infecção nas zonas urbanas da cidade, motivação suficiente para buscar a geolocalização dos distritos utilizados na base de dados deste trabalho a fim de promover uma ilustração semelhante par aa cidade de Curitiba.

Porém a essência da proposta deste trabalho somente pode se desenrolar sob a luz iluminadora de Guizelini et al. (2020) cujo foi a maçã de newton para este estudo, possibilitando revelar a variável preditora escondida como agulha no palheiro dos dados escolhidos para o estudo. Guizelini et al. (2020) do Laboratório de Inteligência Artificial Aplicada a Bioinformática (AIBIA) da Universidade Federal do Paraná (UFPR) utilizaram algoritmos para prever as médias móveis de casos de infecção por Covid no territorio nacional. Estes, justificam a usabilidade dos dados escolhidos, cujos apresentaram grande resiliência à análise, cabendo ao pesquisador desbobrar os dados para a extração da informação que será utilizada como combustível dos algoritmos de machine learning e consequentemente, ilustrar a porposta deste trabalho: avaliar a influência da padronização e escalonamento dos dados sob o resultado da previsão do modelo.

3 Metodologia

Conforme Harrison (2020) "O CRISP-DM (Cross-Industry Standard Process for Data Mining) ou processo para fazer mineração de dados (data mining), o qual contém vários passos que podem ser seguidos para uma melhoria contínua" elencando-os:

- Entendimento do negócio;
- Entendimento dos dados;
- Preparação dos dados;
- Modelagem;
- Avaliação;
- Implantação.

Para fins deste estudo vamos apenas focar nas fases de entendimento e preparação dos dados, modelagem e avaliação do modelo. Após a limpeza inicial será calculado com base nas datas de ocorrência as médias móveis para número de infectados e para o número de óbitos. A partir desta informação utilizarse-a de modelos de regrassão para estimar os valores futuros para infectados e óbitos. Salienta-se aqui que o objetivo do estudo em si, não é prever com eficiência e precisão os dados futuros, mas sim observar o impacto do pré-processamento, comparado os resultados das regressões sob modelos treinados nos dados brutos e modelos trinados sob dados escalonados e padronizados.

4 Base de dados

A base de dados utilizada foi a Casos de COVID-19 em Curitiba Oliveira (2021), cuja possui 234125 instâncias para 7 colunas de características, das quais descreve-se a data de notificação, idade, gênero, região sanitária (bairro), data de óbito (se em caso de óbito) e estado do indivíduo (se recuperado ou óbito). Todos as ocorrências são casos em que a doença foi confirmada.

Tabela 1: Dados brutos.

	notification_date	class	age	gender	district	death_date	status
0	11/03/2020	CONFIRMADO	54	M	DSMZ	NaN	RECUPERADO
1	12/03/2020	CONFIRMADO	43	M	DSBQ	NaN	RECUPERADO
2	12/03/2020	CONFIRMADO	15	M	DSBQ	NaN	RECUPERADO
3	12/03/2020	CONFIRMADO	25	F	DSMZ	NaN	RECUPERADO
4	12/03/2020	CONFIRMADO	58	M	DSMZ	NaN	RECUPERADO

Fonte: Oliveira (2021)

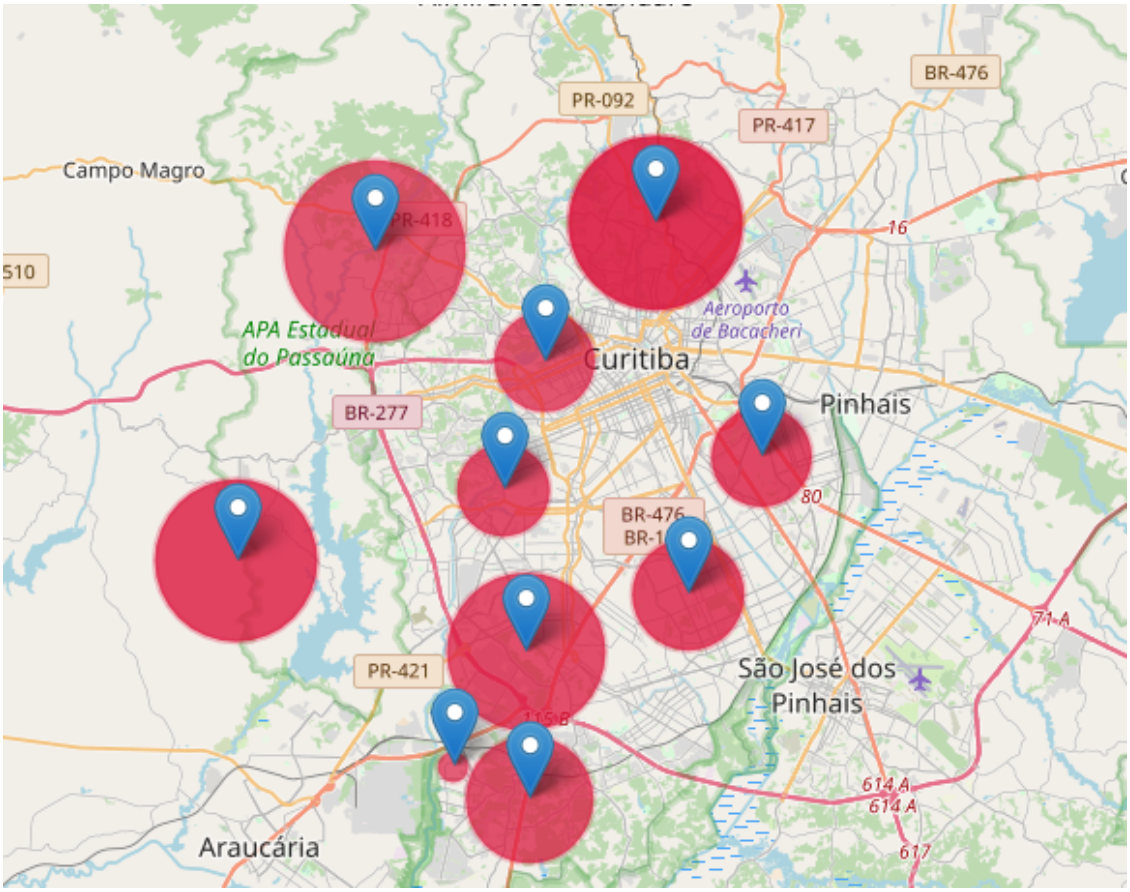
Ao verificar por linhas que contém valores ausentes, determinou-se que a coluna idade apresenta duas linhas à qual não contém a informação, portanto pode-se remover estas linhas. As colunas `district` e `status` apresenta muitos dados para que seja removido, podendo impactar nos resultados. A coluna `death_date` e `status` apresentam ausencia devido a binaridade da informação, pois as ocorrências que não apresentam data de óbito, consequentemente, não foram a óbito, portanto optou-se por manter estes registros, removendo apenas as instâncias com dados ausentes para idade.

Os gêneros cujos estavam definidos como atributos categóricos (`M` e `F`) foram substituidos por valores binários, sendo `1` para masculino e `0` para feminino.

O mesmo foi realizado para o *status* final do infectado, onde substituiu-se o `status` por variáveis *dummy* representando se o indivíduo se recuperou ou se foi a óbito, estes foram dados fundametais para contabilizar as médias de infectados e óbitos.

Na Imagem 1 a seguir pode-se observar as regiões sanitárias (distritos ou bairros) da cidade de Curitiba cujas apresentaram registros do conjunto de dados o qual correspondem á região geográfica da cidade, onde o Boa Vista apresentou maior número de ocorrências contando com 34774 regisdtros de casos confirmados, seguida pelo CIC com 26996 registros e o Boqueirão com 24562 registros de casos confirmados. Após as análises iniciais e limpeza dos dados, pôde-se extrair os valores para calcular as médias móveis, informação principal à qual compele o âmago deste estudo.

Imagem 1: Raio de ocorrências por distrito



Fonte: O autor.

5 Médias móveis

A partir das datas de registro de ocorrência, foram extraídas as médias móveis para períodos de 14 dias, período em que pode-se dizer que os sintomas da doença começam a aparecer e o indivíduo busca orientação médica:

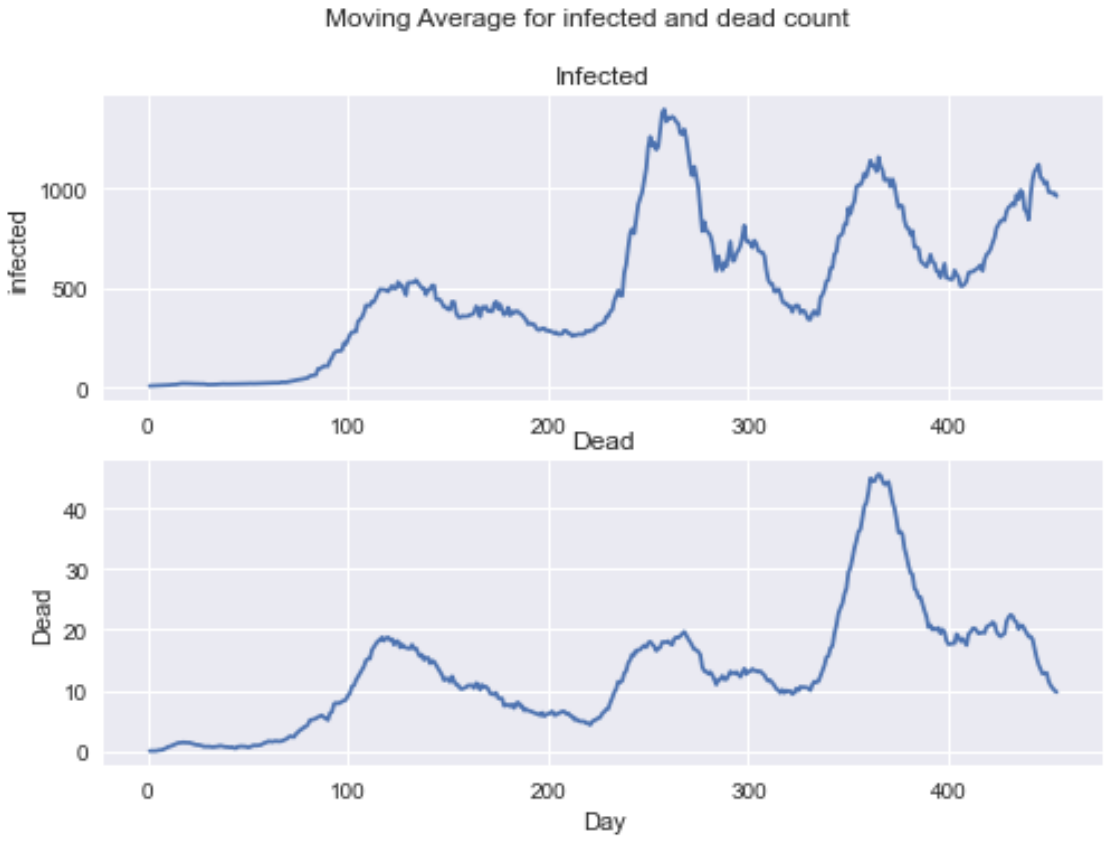


Gráfico 1: Médias móveis para infectados e óbitos

6 Particionando os dados

De acordo com Jain (2021) é tradicionalmente comum entre cientistas de dados, dividir aleatoriamente o conjunto de dados, testar e aplicar uma validação cruzada. Mas ao invés disso pode-se utilizar uma outra bordagem chamada time-base splitting (JAIN, 2021) onde aplica-se um recorte temporal no conjunto. Esta aboragem cabe perfeitamente ao problema proposto. Para este estudo, separou-se o conjunto em 90% dos dados para treino dos modelos, separando 10% dos dados para teste:

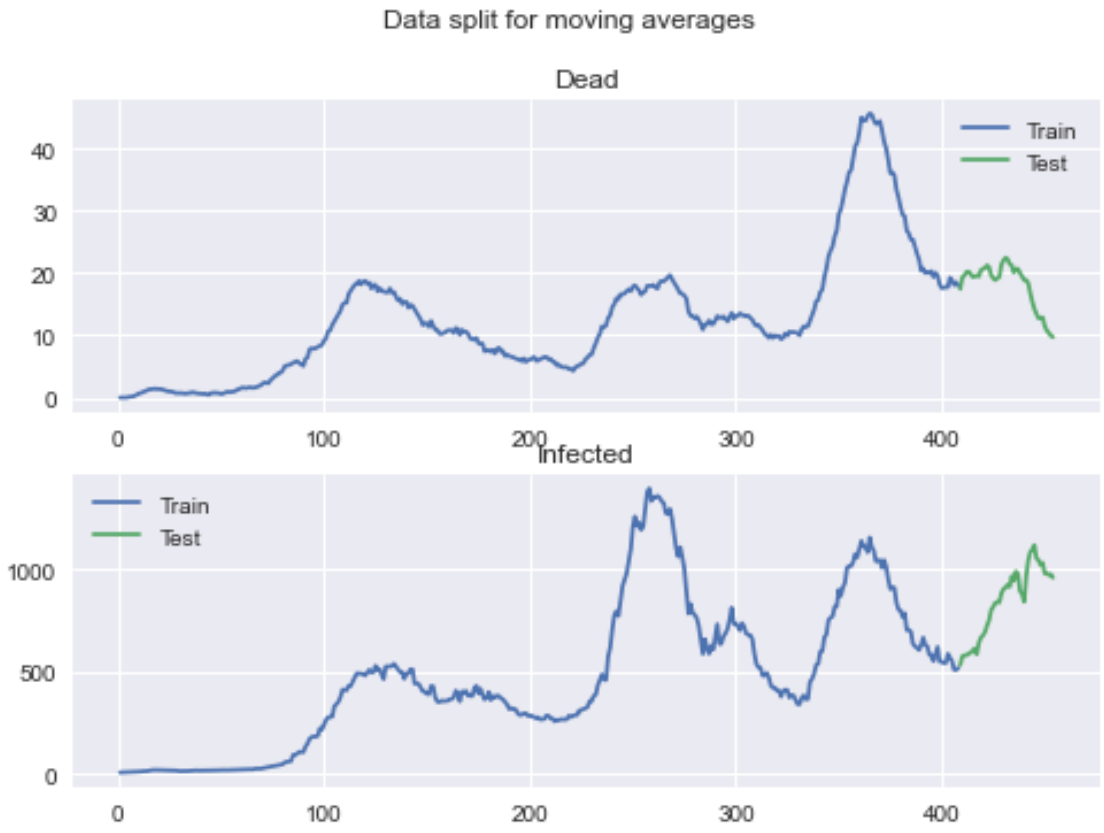


Grafico 2: Divisão de treino e teste

7 Treinando os modelos iniciais

Foram treinados quatro modelos de regressão básicos sob os dados particionados para infectados e registros de óbito, como pode-se ver no Gráfico 3 e 4 a Floresta Aleatória apresenta sobreajuste sobre os dados de treino de contagemd e infectados, tendendo a linearidade no teste, de mesmo modo, a Perceptron Multi Camada também tende a linearidade após convergir. O modelo mais flexível nos testes iniciais é a SVR. Nas contagems de óbito os modelos apresentam as mesmas características

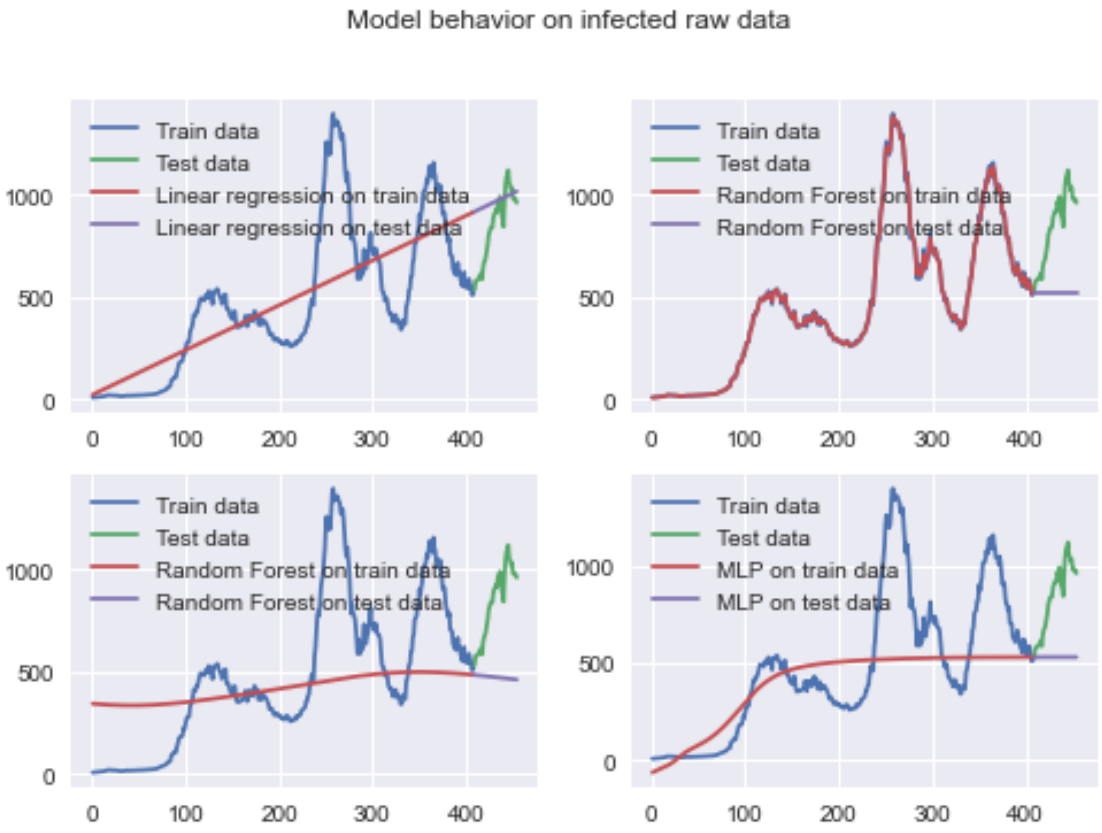


Gráfico 3: Modelos de regressão iniciais n contagem de infectados

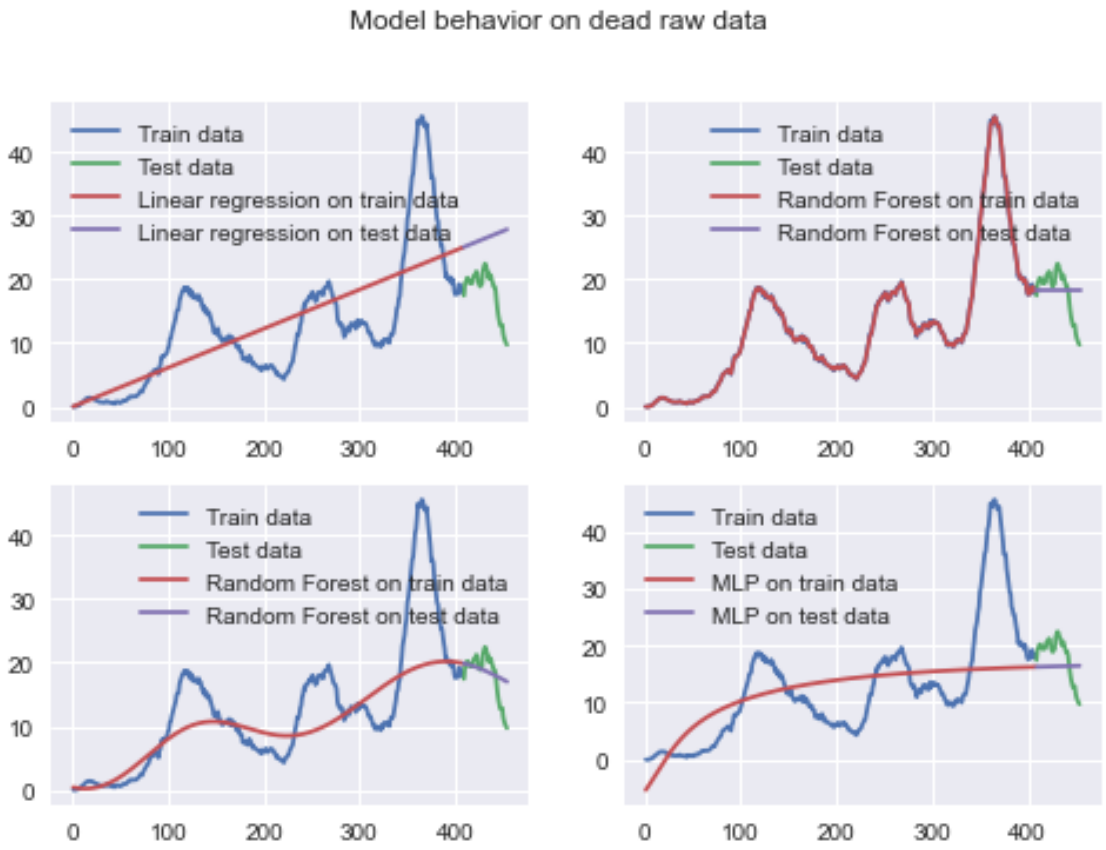


Gráfico 4: Modelos de regressão iniciais n contagem de infectados

8 Escalonando os dados

De acordo com Géron (2019, p.68) "Uma das transformações mais importantes que você precisa aplicar aos seus dados é o escalonamento das características. Com poucas exceções, os algoritmos de Aprendixado de Máquina não funcionam bem quando atributos numéricos de entrada tem escalas muito diferentes" portanto será aplicado um pipeline cujo transformará os dados aplicando um imputer e em seguida um Escalonador Padrão (StandardScaler).

imputer: O imputer calcula a média de cada atributo

Escalonador Padrão O escalonamento padrão subtrai o valor médio e divide pela variância:

$$z = \frac{(x - avg)}{std} \tag{1}$$

9 Ajustando os parâmetros dos modelos

Para ajustar os modelos foi aplicado um Grid Search que encontrou parâmetros cujos melhoram o desempenho dos modelos nos dados. Após a identificação dos parÂmetros e normalização dos dados, os modelos foram retreinados e os seguintes resultados foram obtidos conforme gráficos 6 e 7 e tabelas 2 e 3:

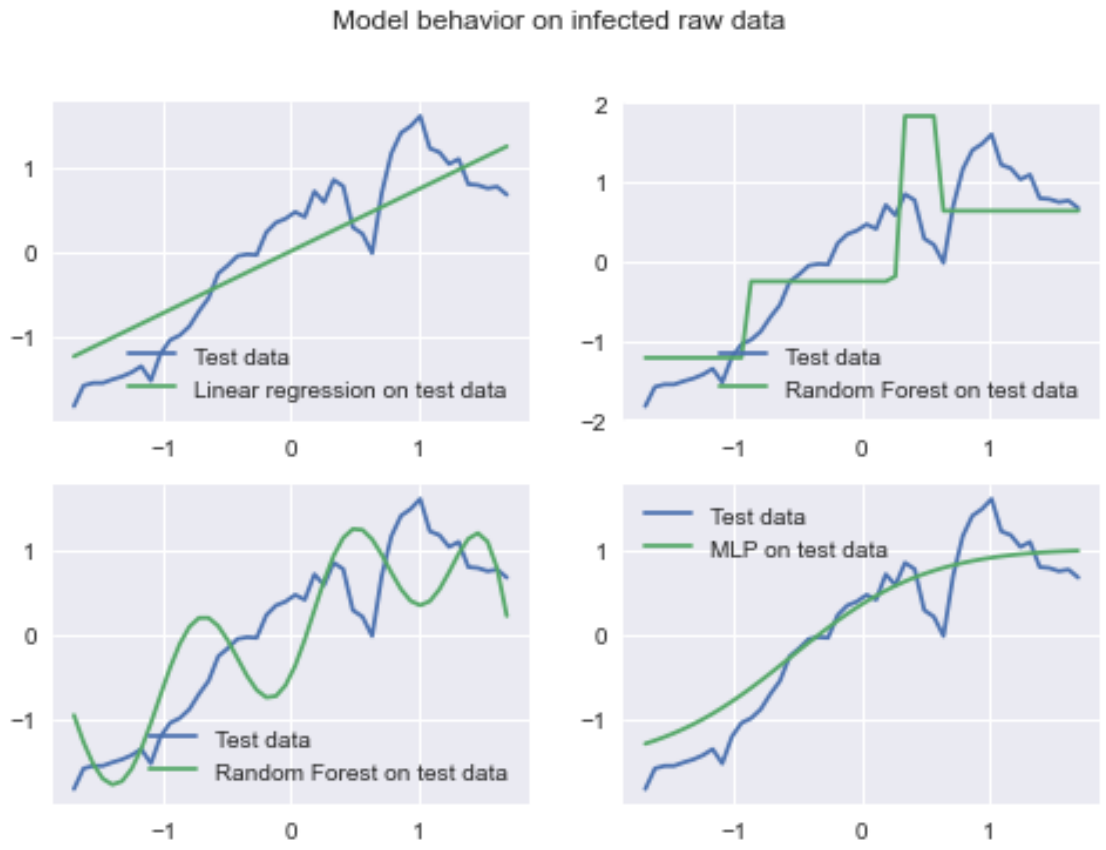


Gráfico 6: Ajuste dos modelos na contagem de infectados após normalização e ajuste dos parâmetros

Tabela 2: Comparação de resultados dos modelos no conjunto de infectados

	Sc -	Sc +	R2 -	R2 +	AE -	AE +	RMSE -	RMSE +
LR	0.43	0.80	0.43	0.80	0.58	0.39	1.61	0.43
RF	0.08	0.62	0.08	0.62	0.77	0.48	1.78	0.60
SVR	0.19	-1.63	0.19	-1.63	0.72	1.38	1.68	0.64
MLP	0.44	-1.47	0.44	-1.47	0.58	1.29	1.59	0.34

Legendas: - (Antes da normalização), + (Após normalização), LR(Linear Regression), RF(Random Forest), SVR(Support Vector Regressor), MLP(Multi Layer Perceptron)

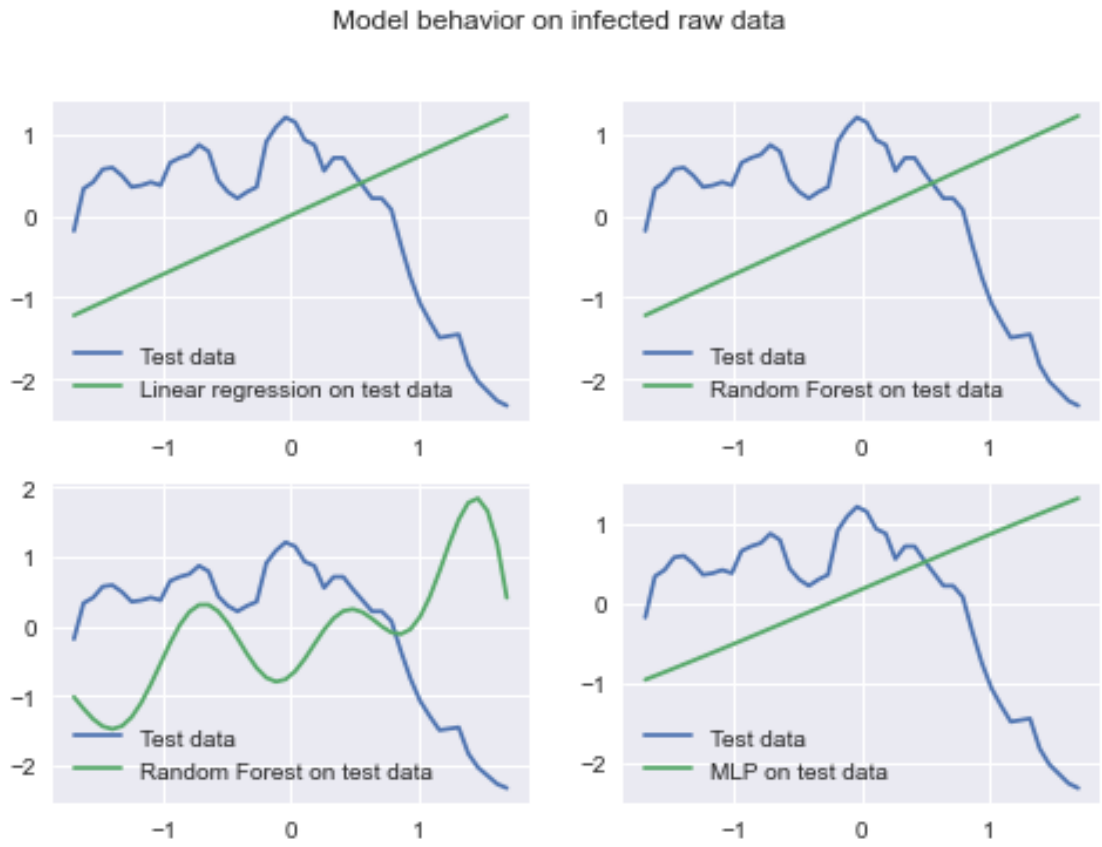


Gráfico 7: Ajuste dos modelos na contagem de óbitos após normalização e ajuste dos parâmetros

Tabela 3: Comparação de resultados dos modelos no conjunto de óbitos

	Sc -	Sc +	R2 -	R2 +	AE -	AE +	RMSE -	RMSE +
LR	-4.14	-1.53	-4.14	-1.53	7.76	1.33	9.03	1.59
RF	-0.11	-2.14	-0.11	-2.14	2.93	1.35	9.03	1.77
SVR	0.30	-1.92	0.30	-1.92	2.60	1.36	3.32	1.71
MLP	-0.22	-1.44	-0.22	-1.44	3.96	1.25	4.41	4.56

Legendas: - (Antes da normalização), + (Após normalização), LR(Linear Regression), RF(Random Forest), SVR(Support Vector Regressor), MLP(Multi Layer Perceptron)

10 Conclusão

Pode-se notar que após aplicar a normalização dos dados através de um imputer e um StandardScaling e escolher uma melhor combinação de parâmetros para os modelos de teste, pode-se obter resultados melhores, minimzando a taxa de erros e maximizando a pontuação do modelo. Mesmo que os resultados das previsões ainda sejam insatisfatorios, é inquestionável que a aplicação da normalização minimiza moderadamente a taxa de erros. A regressão inicial aplicada aos de infec-

tados dados não escalonados apresentou pontuação de 40% e após da normalização chegou a 80%. A Floresta aleatória atingiu 62% de pontuação após a normalização de uma pontuiação inicial de 8%. Todos os modelos apresentaram redução na taxa de erros. Convém-se, como conclussão deste estudo, que a normalização dos dados e escolha dos parâmetros são uma etapa fundamental na fase de pré-processamebnto dos dados, podendo impactar fortemente no resultado final.

11 Referências

CURITIBA, Prefeitura Municipal de. **Endereços da Vigilância Sanitária Municipal**. 2021. Online, disponível em: https://www.curitiba.pr.gov.br/servicos/enderecos-da-vigilancia-sanitaria-municipal/729 Acesso em: 21/07/2021;

GÉRON, Aurélien. Mãos à Obra Aprendizado de máquina com Scikit-Learn & tensorFlow: Conceitos, Ferramentas e técnicas para a construção de Sistemas Inteligentes. Rio de Janeiro: Alta Books, 2019;

GUIZELINI, Dieval et al. Algoritmos Genéticos e média móvel para ajuste da curva de predição da infecção por SARS-Cov-2. In: Algoritmos Genéticos e média móvel para ajuste da curva de predição da infecção por SARS-Cov-2. [S. l.], 2020. Disponível em: https://www.bioinfo.ufpr.br/covid19/index.html. Acesso em: 22 jul. 2021.

HARRISON, Matt. Machine Learning: Guia de Referência Rápida. Novatec, 1 edição. 2020;

JAIN, Akshay P. Time Series Forecasting – Data, Analysis, and Practice. Neptune ai, 2021. Online, Disponível em: https://neptune.ai/blog/time-series-forecasting. Acesso em: 22 jul. 2021.

OLIVEIRA, Alcides Augusto Souto de. **Casos de COVID-19 em Curitiba**. Prefeitura de Curitiba - Dados abertos, 2021; Online: Disponível em: https://www.curitiba.pr.gov.br/dadosabertos/busca/?pagina=2 Último acesso em: 21/07/2021;

RIZZATTI, Maurício et al. MAPEAMENTO DA COVID-19 POR MEIO DA DENSIDADE DE KERNEL. Metodologias e Aprendizado, [S. l.], v. 3, p. 44-53, 24 maio 2020. DOI https://doi/10.21166/metapre.v3i0.1312. Disponível em: https://publicacoes.ifc.edu.br/index.php/metapre/article/view/1312/1020. Acesso em: 22 jul. 2021.