GKS - Análise e previsão de quantidade de unidades locais de restaurantes

Marcelo A. S. Fantini

25 de setembro de 2023

Sumário

Introdução

O objetivo deste estudo é analisar e prever a quantidade de unidades locais de restaurantes em cada região do Brasil. Para isso, utilizamos o modelo de gradiente descendente e construímos um modelo de regressão para cada região. Neste artigo, apresentamos as previsões para 2022 e discutimos as atividades mais correlacionadas com a quantidade de unidades locais de restaurantes em cada região.

O primeiro modelo criado é o modelo de regressão linear. É o modelo mais simples, porém com alto poder de previsão se os dados possuírem uma tendência linear. Verificamos para cada região se existe tal tendência, apresentando a acurácia do modelo e previsões da quantidade de unidades locais de restaurantes para os anos de 2022 e 2023.

O segundo modelo é o modelo de gradiente descendente, com o qual fazemos três previsões para o ano de 2022 para cada região: uma considerando que os valores para cada atividade serão iguais à média dos valores dos anos anteriores, outra considerando que os valores serão iguais à média dos valores dos anos anteriores mais o desvio padrão, e outra considerando que os valores serão iguais à média dos valores dos anos anteriores menos o desvio padrão. Essa hipótese de assumir valores para as atividades de 2022 como a média dos anteriores é necessária para que o modelo de gradiente descendente faça as previsões. Além disso, é uma estimativa razoável para o ano seguinte, e as previsões adicionando ou subtraindo o desvio padrão nos permitem introduzir um intervalo de confiança na previsão.

Comparação entre quantidades nas grandes regiões do Brasil

Observe o gráfico abaixo.

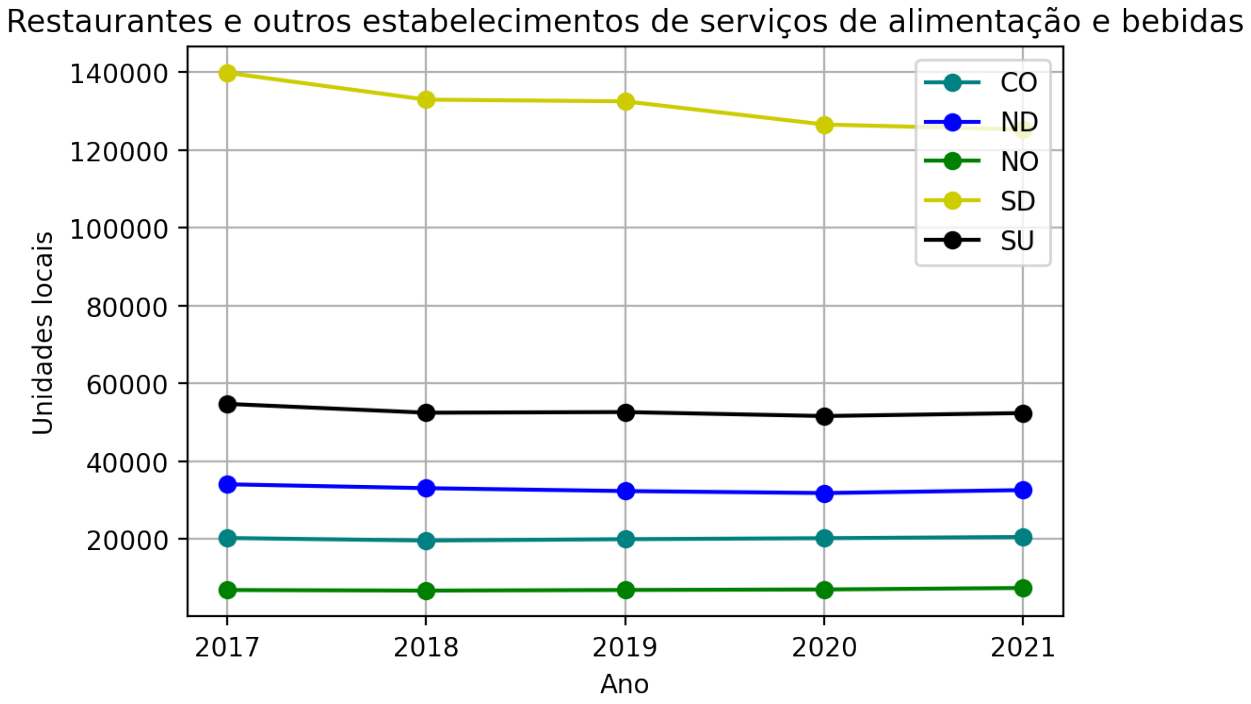

Figura 1: Comparação em mesma escala das unidades locais de restaurantes em cada grande região do Brasil.

A região Sudeste lidera a quantidade de unidades locais de restaurantes. Também é possível observar que a quantidade de unidades locais da região Sudeste diminui com o passar dos anos. Em síntese, a ordem decrescente das regiões é:

1. Sudeste
2. Sul
3. Nordeste
4. Centro-Oeste
5. Norte

Entretanto, não é possível inferir o crescimento ou decrescimento das regiões por esse gráfico. Como próximo passo, construíremos simultaneamente gráficos para mostrar a tendência dos dados para cada região, resolvendo este problema, e um modelo de regressão linear, ajustado para cada região. Usaremos cada modelo para prever as quantidades de unidades locais para os anos de 2022 e 2023.

Modelo preditivo: Regressão linear

Como primeiro modelo, utilizamos uma regressão linear, que é um modelo simples e facilmente ajustável. Esse modelo é comumente utilizado quando a tendência dos dados apresenta uma característica linear, ou seja, pode ser descrita por uma reta. Após ajustarmos o modelo aos dados, plotamos os dados originais juntamente com a reta que representa a regressão linear dos dados no mesmo gráfico. Em seguida, avaliamos a eficiência do modelo naquele caso específico com base na métrica[1] do valor de R^2.

A métrica R^2, também conhecida como coeficiente de determinação, é uma medida de quão bem um modelo de regressão linear se ajusta aos dados. O valor de R^2 é calculado como a proporção da variância total dos dados que é explicada pelo modelo. Em outras palavras, R^2 mede a fração da variação dos dados que é explicada pela regressão linear.

Após a avaliação da métrica R^2 fazemos previsões da quantidade de unidades locais de restaurantes para os anos de 2022 e 2023.

[1] Se o valor da métrica R^2 é 1, isso significa que o modelo explica 100% da variabilidade dos dados, enquanto que se o valor da métrica for 0 o modelo explica 0% da variabilidade dos dados.

Região Sudeste

O gráfico com os dados originais da grande região Sudeste e a reta que representa a regressão linear está abaixo.

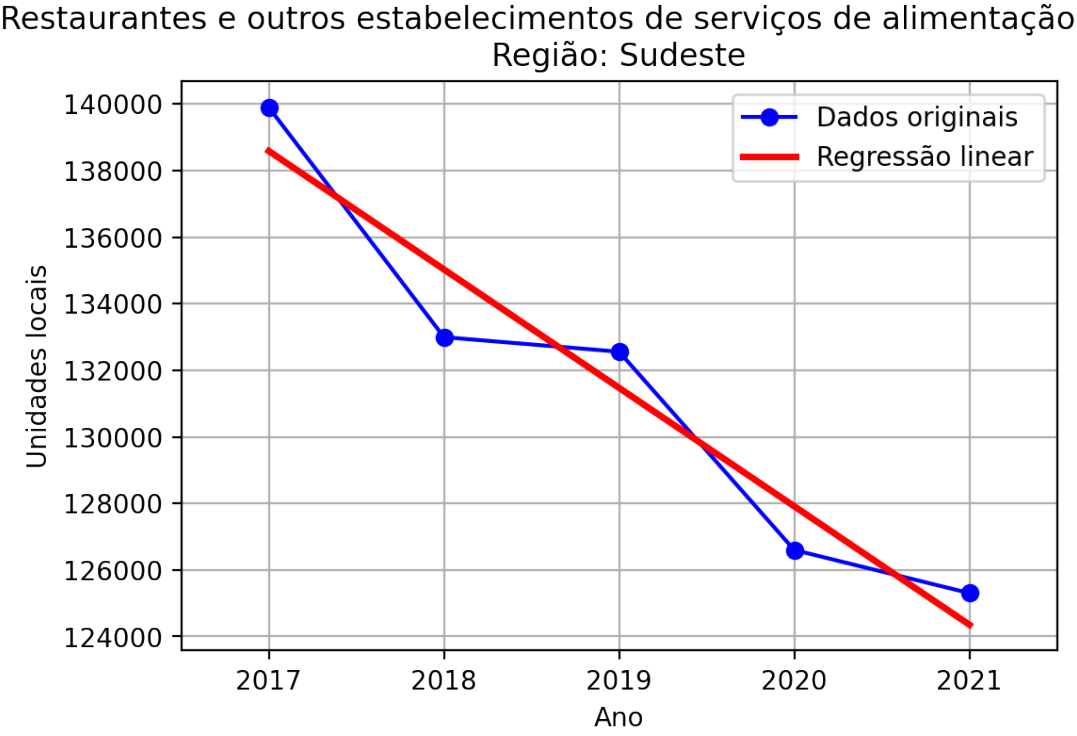

Figura 2: Gráfico com os dados originais de quantidade de unidades locais na grande região Sudeste (em azul) e a reta que representa a regressão linear obtida com esses dados (em vermelho). As linhas em azul foram adicionadas para facilitar a visualização da tendência dos dados.

No caso da grande região Sudeste o modelo obteve métrica R^2 de valor 0.929, que significa que o modelo explica 92,9% da variabilidade dos dados.

De todas as regiões esta foi a região com melhor variação explicada pelo modelo da regressão linear. Pelo gráfico observamos que os pontos originais, em azul, possuem uma tendência linear de decrescimento que foi bem descrita pela reta da regressão linear, em vermelho.

Com a pandemia do ano de 2020 observamos que a região Sudeste manteve a tendência de decrescimento em 2021.

Usando este modelo as previsões para quantidades de unidades locais de restaurantes em 2022 é de 120.765 unidades, e a previsão para 2023 é de 117.236 unidades[2].

[2] Os valores foram arrendondados para baixo nesta região e nas regiões subsequentes.

Região Norte

O gráfico com os dados originais da grande região Norte e a reta que representa a regressão linear está abaixo.

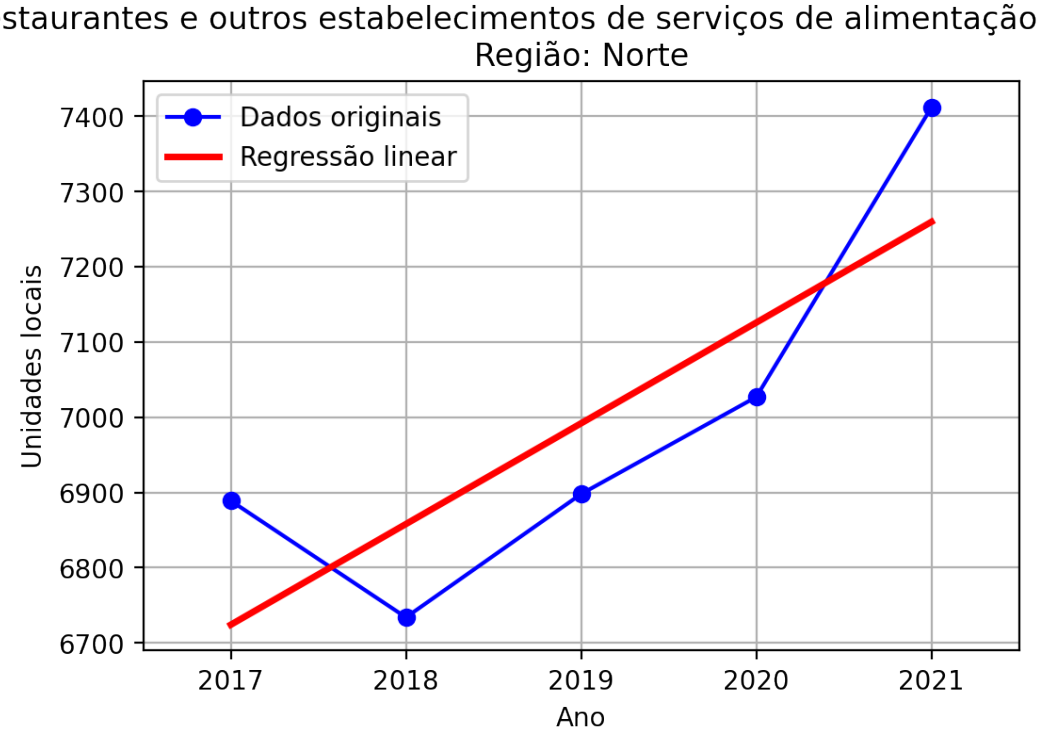

Figura 3: Gráfico com os dados originais de quantidade de unidades locais na grande região Norte (em azul) e a reta que representa a regressão linear obtida com esses dados (em vermelho). As linhas em azul foram adicionadas para facilitar a visualização da tendência dos dados.

No caso da grande região Norte o modelo obteve métrica R^2 de valor 0.680, que significa que o modelo explica 68,0% da variabilidade dos dados.

Esta região foi a região com segunda melhor variação explicada pelo modelo da regressão linear. É possível observar que existe uma diferença sensível entre a tendência nos dados da região Sudeste e da região Norte, o que acarreta a diferença na métrica R^2 para as duas regiões. Neste caso é aconselhável utilizar outro modelo para previsão dos anos seguintes.

Observamos que a região Norte manteve a tendência de crescimento em 2021, apesar das políticas de isolamento social e fechamento do comércio em 2020. Este aumento na quantidade de unidades locais de restaurantes provavelmente também deve ser explicado pela falta de auxílio financeiro durante a época da pandemia e o não respeito à política de permanecer em casa.

Usando este modelo as previsões para quantidades de unidades locais de restaurantes em 2022 é de 7.393 unidades, e a previsão para

2023 é de 7.527 unidades.

Região Nordeste

O gráfico com os dados originais da grande região Nordeste e a reta que representa a regressão linear está abaixo.

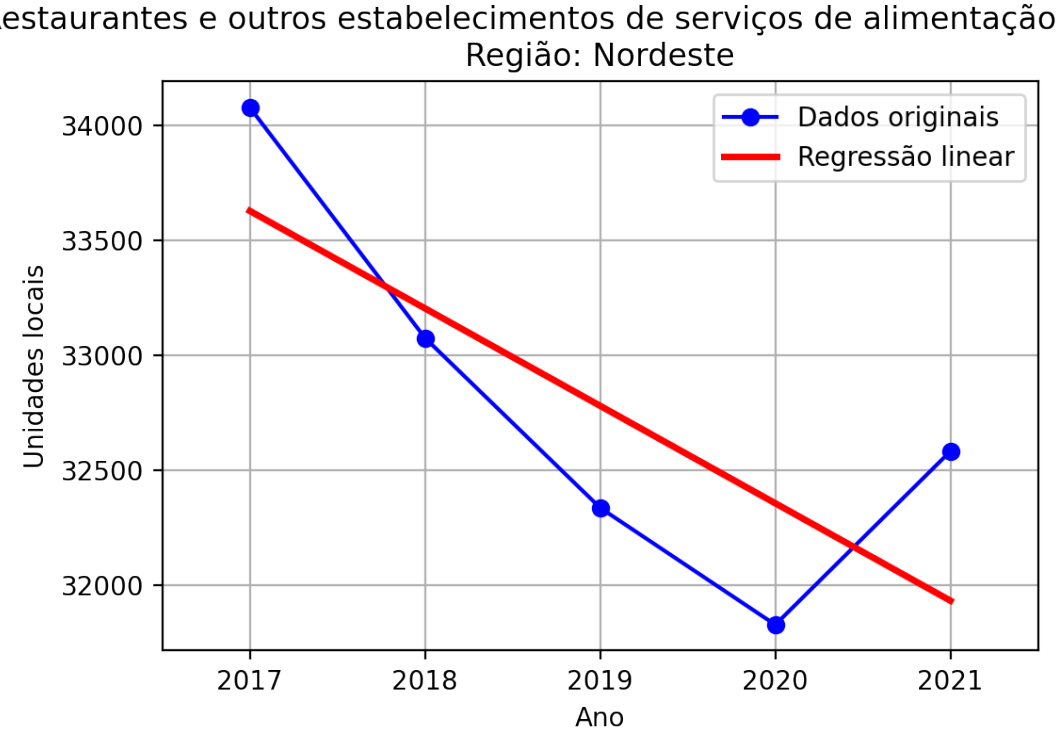

Figura 4: Gráfico com os dados originais de quantidade de unidades locais na grande região Nordeste (em azul) e a reta que representa a regressão linear obtida com esses dados (em vermelho). As linhas em azul foram adicionadas para facilitar a visualização da tendência dos dados.

No caso da grande região Sudeste o modelo obteve métrica R^2 de valor 0.614, que significa que o modelo explica 61,4% da variabilidade dos dados.

Esta região foi a região com terceira melhor variação explicada pelo modelo da regressão linear. Assim como o caso da região Norte, existe uma tendência que não é linear nos dados. Isso explica a forte redução na métrica R^2 na mudança da região Sudeste para a região Nordeste. Neste caso é aconselhável utilizar outro modelo para previsão dos anos seguintes.

Esta região apresenta um aumento surpreendente do ano de 2020 para o ano de 2021, apesar da pandemia de 2020. O modelo não é capaz de perceber o efeito drástico de aumento causado pós-pandemia e indica um contínuo decrescimento, apesar do dado de 2021 contradizer este fato.

Usando este modelo as previsões para quantidades de unidades locais de restaurantes em 2022 é de 31.508 unidades, e a previsão para 2023 é de 31.084 unidades.

Região Sul

O gráfico com os dados originais da grande região Sul e a reta que representa a regressão linear está abaixo.

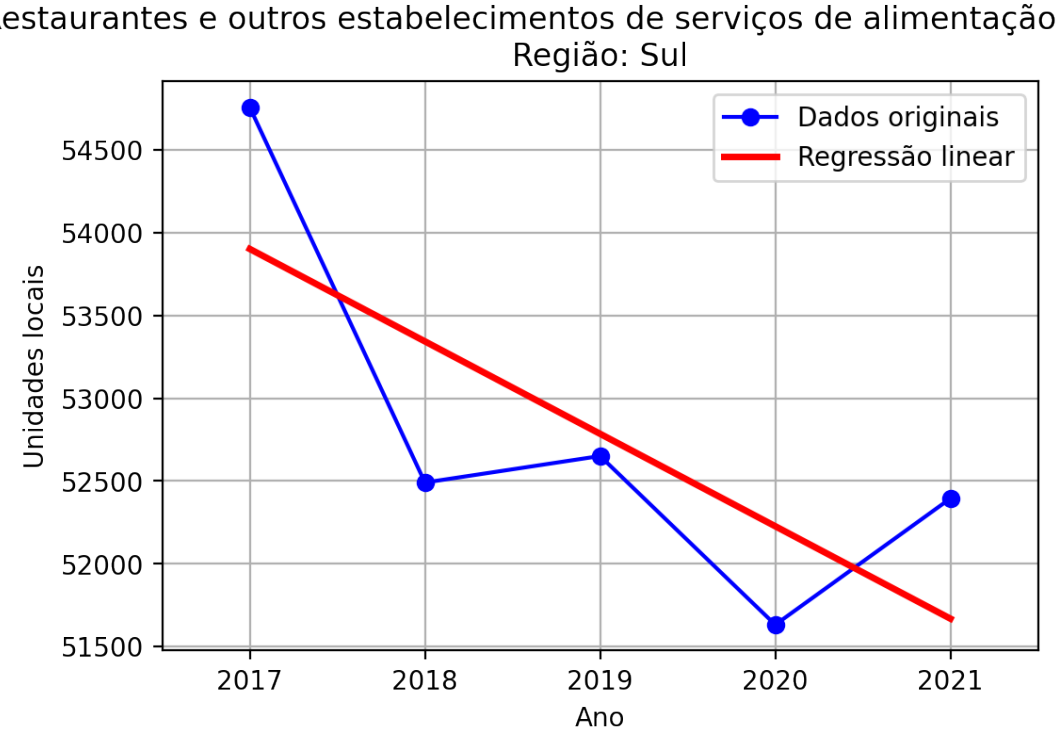

Figura 5: Gráfico com os dados originais de quantidade de unidades locais na grande região Sul (em azul) e a reta que representa a regressão linear obtida com esses dados (em vermelho). As linhas em azul foram adicionadas para facilitar a visualização da tendência dos dados.

No caso da grande região Sudeste o modelo obteve métrica R^2 de valor 0.569, que significa que o modelo explica 56,9% da variabilidade dos dados.

Esta região foi a região com quarta melhor variação explicada pelo modelo da regressão linear. Neste caso existe uma tendência nos dados como uma curva serrilhada, que não é bem descrita pela reta da regressão linear. Neste caso é aconselhável utilizar outro modelo para previsão dos anos seguintes.

O modelo não é capaz de descrever o movimento aparentemente comum de redução na quantidade de restaurantes seguido, no ano seguinte, por um aumento na quantidade. O modelo parece ter seguido a forte tendência de decrescimento até o ano da pandemia de 2020, não sendo capaz de incorporar a mudança drástica pós-pandemia no ano de 2021.

Usando este modelo as previsões para quantidades de unidades locais de restaurantes em 2022 é de 51.110 unidades, e a previsão para 2023 é de 50.552 unidades.

Região Centro-Oeste

O gráfico com os dados originais da grande região Centro-Oeste e a reta que representa a regressão linear está abaixo.

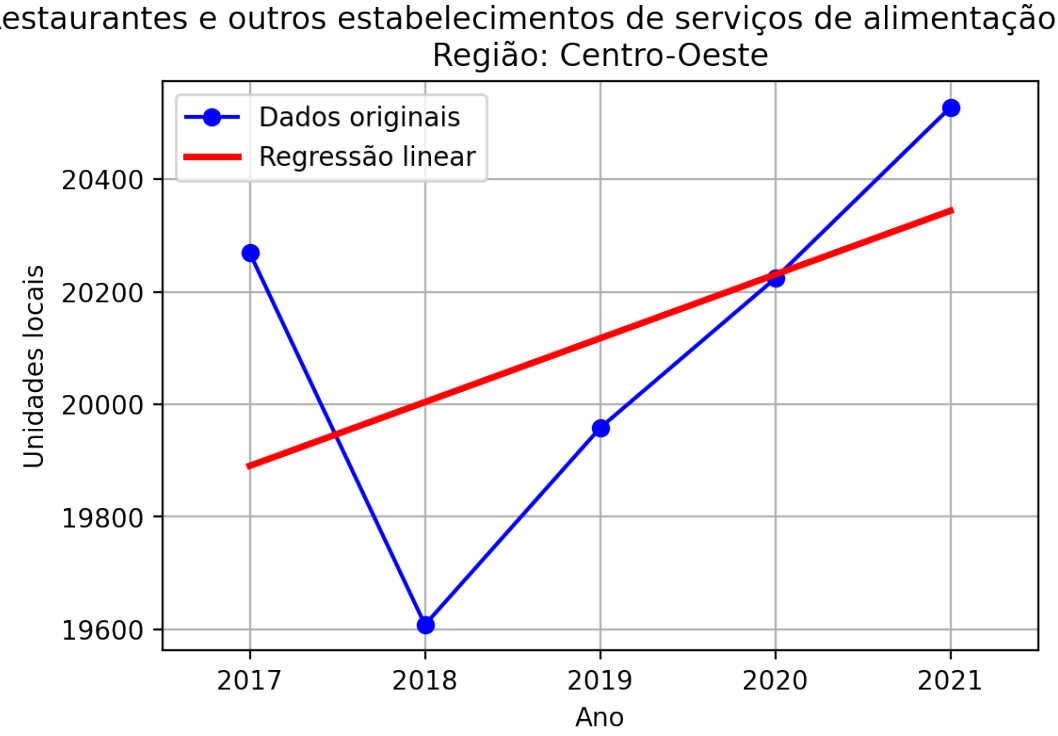

Figura 6: Gráfico com os dados originais de quantidade de unidades locais na grande região Centro-Oeste (em azul) e a reta que representa a regressão linear obtida com esses dados (em vermelho). As linhas em azul foram adicionadas para facilitar a visualização da tendência dos dados.

No caso da grande região Sudeste o modelo obteve métrica R^2 de valor 0.263, que significa que o modelo explica 26,3% da variabilidade dos dados.

Esta região foi a região com pior variação explicada pelo modelo da regressão linear. Não captura a tendência dos dados, que talvez pudessem ser descritos por uma reta mas apenas a partir de 2018. Neste caso é aconselhável utilizar outro modelo para previsão dos anos seguintes, a não ser que escolha-se analisar apenas a partir de 2018.

Apesar de o modelo apontar o crescimento nos anos de 2018 a 2021, a redução drástica do ano de 2017 para o ano de 2018 fez com que ele subestimasse a taxa de crescimento, indicando valores menores que a tendência dos dados.

Usando este modelo as previsões para quantidades de unidades locais de restaurantes em 2022 é de 20.457 unidades, e a previsão para 2023 é de 20.571 unidades.

Preparação para modelos com gradiente descendente

A construção do próximo modelo preditivo, o de regressão com gradiente descendente, exige uma maneira diferente de observar os dados. Assim, foi necessário transformar a tabela do formato original em um novo formato, em que as colunas são as atividades e o índice são os anos. Isto foi feito para cada região. Assim, podemos analisar as atividades como características de cada ano e criar um modelo para prever a característica Restaurantes e outros estabelecimentos de serviços de alimentação e bebidas em função das demais atividades da tabela.

Correlação com outras atividades

Como são 55 atividades, não é viável usar todas as 54 demais atividades como preditoras da quantidade de unidades locais de restaurantes. Então, fizemos uma tabela de correlação entre as atividades e construímos dois gráficos para cada região: um gráfico com as atividades com maiores correlações positivas e um gráfico com com as atividades com menores correlações negativas.

Correlação é uma medida estatística que indica como e se duas variáveis estão alinhadas. Essa medida varia de -1 a 1. Uma correlação positiva significa que quando uma medida aumenta a outra também tende a aumentar. Em outras palavras, a tendência é como se as duas medidas fossem diretamente proporcionais. Uma correlação negativa significa que quando uma medida aumenta a outra tende a diminuir. Em outras palavras, a tendência é como se as duas medidas fossem inversamente proporcionais.

Uma correlação neutra indica que não há relação entre as duas variáveis. Por exemplo, a cor dos olhos e a altura de uma pessoa têm uma correlação neutra, pois não há relação entre essas duas variáveis. A correlação é uma ferramenta importante na análise de dados, pois permite identificar relações entre variáveis e entender como elas se relacionam. No entanto, é importante lembrar que a correlação não implica causalidade, ou seja, apenas porque duas variáveis estão correlacionadas, não significa que uma causa a outra.

Correlações para região Sudeste

O gráfico abaixo mostra as atividades com correlação maior que 0.7 com a atividade de restaurantes.

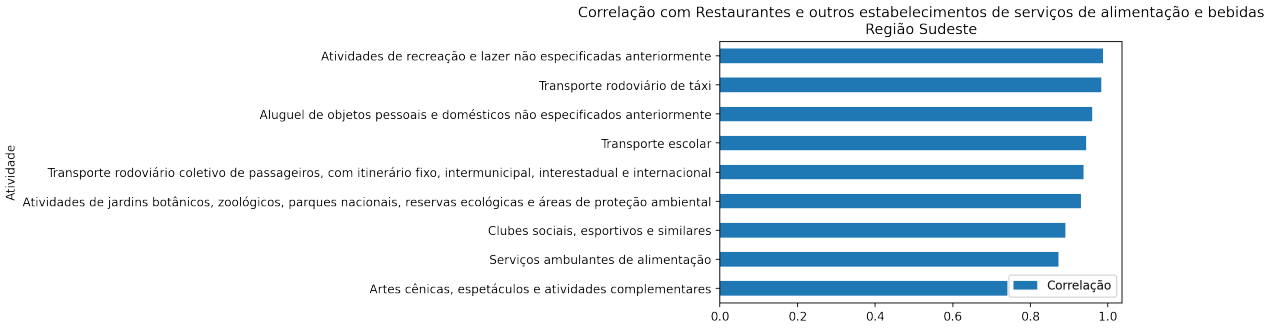

Figura 7: Gráfico das atividades com correlação superior a 0.7 em relação aos restaurantes. As barras são mostradas em ordem decrescente.

Observamos algumas relações interessantes. Transporte rodoviário de táxi é a segunda maior correlação positiva, quase máxima, assim como transporte rodoviário coletivo de passageiros é a quinta maior correlação, aproxidamente 0.9.

O gráfico abaixo mostra as atividades com correlação menor que -0.85 com a atividade de restaurantes.

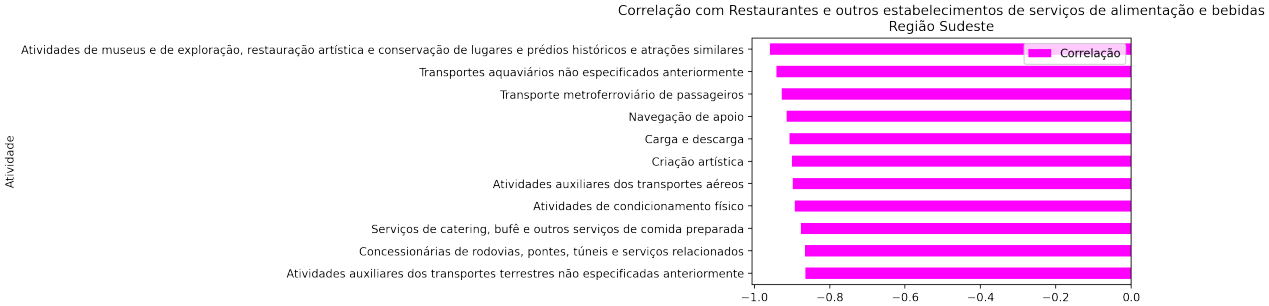

Figura 8: Gráfico das atividades com correlação inferior a -0.85 em relação aos restaurantes. As barras são mostradas em ordem crescente.

Observamos que atividades de museus e de exploração representam a correlação mais negativa, o que parece surpreendente. Outras correlações do gráfico não parecem tão surpreeendentes assim, como transporte metroferroviário de passageiros e atividades de condicionamento físico.

As atividades de correlação positiva no gráfico serão as atividades que utilizaremos para construir o modelo de gradiente descendente para a região Sudeste.

Correlações para região Norte

O gráfico abaixo mostra as atividades com correlação maior que 0.9 com a atividade de restaurantes.

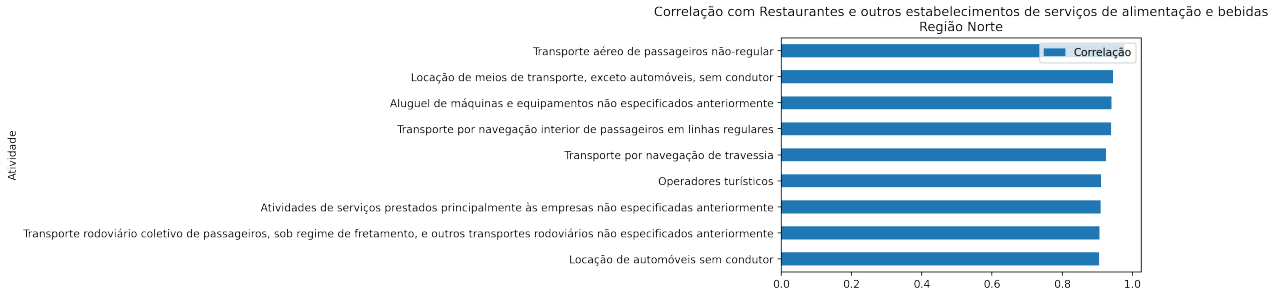

Figura 9: Gráfico das atividades com correlação superior a 0.9 em relação aos restaurantes. As barras são mostradas em ordem decrescente.

Observamos que transporte aéreo de passageiros não-regular é a atividade mais positivamente correlacionada com restaurantes na região Norte. No geral, as atividades que aparecem com maior correlação positiva indicam que a quantidade de restaurantes está positivamente correlacionadas com acesso por diversos meios de transporte por conta do turismo.

O gráfico abaixo mostra as atividades com correlação menor que -0.5 com a atividade de restaurantes.

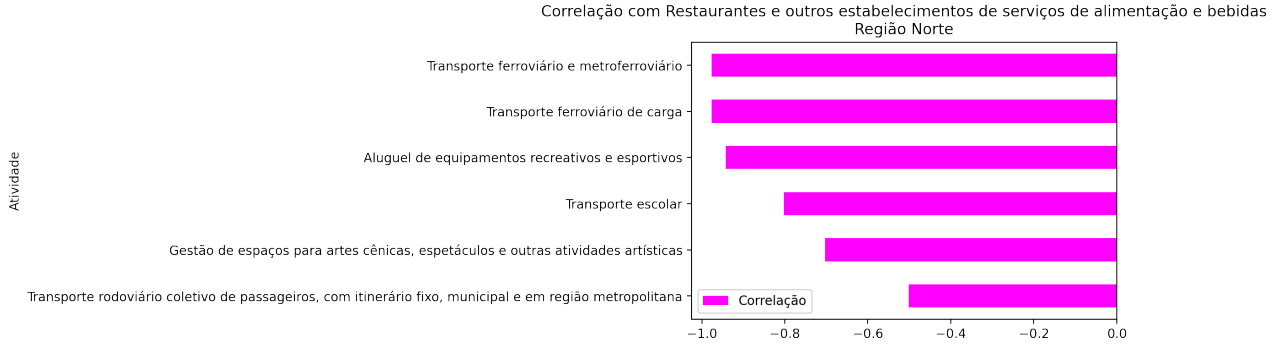

Figura 10: Gráfico das atividades com correlação inferior a -0.5 em relação aos restaurantes. As barras são mostradas em ordem crescente.

Observamos que as categorias que principalmente aparecem com correlação negativa envolvem transporte sem passageiros, formas de transporte praticamente inexistentes para a região, ou transporte tradicionalmente utilizado por pessoas de menor renda, como transporte rodoviário coletivo de passageiros.

As atividades de correlação positiva no gráfico serão as atividades que utilizaremos para construir o modelo de gradiente descendente para a região Norte.

Correlações para região Nordeste

O gráfico abaixo mostra as atividades com correlação maior que 0.7 com a atividade de restaurantes.

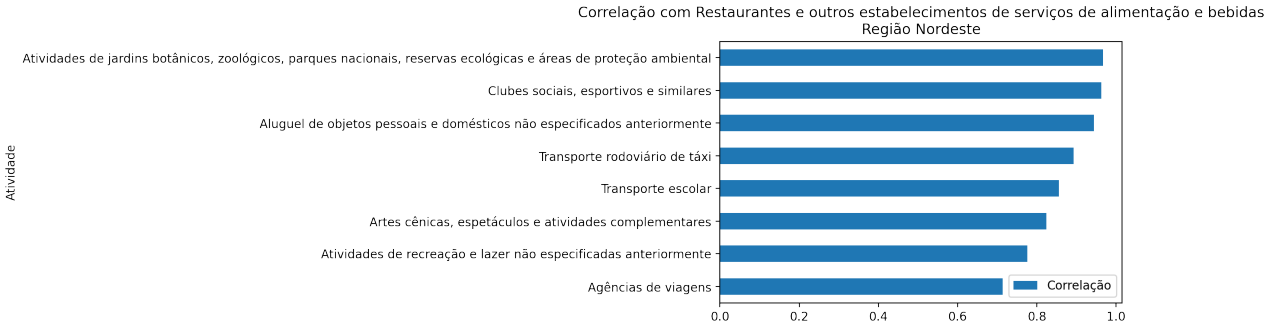

Figura 11: Gráfico das atividades com correlação superior a 0.7 em relação aos restaurantes. As barras são mostradas em ordem decrescente.

Observamos que a atividade mais positivamente correlacionada é a atividade de áreas de proteção ambiental, provavelmente devido a beleza artística das regiões protegidas do Nordeste. Outros notáveis são artes cênicas e clubes sociais. A relação surpreendente é de transporte escolar, com pouco mais de 0.8 de correlação.

O gráfico abaixo mostra as atividades com correlação menor que -0.7 com a atividade de restaurantes.

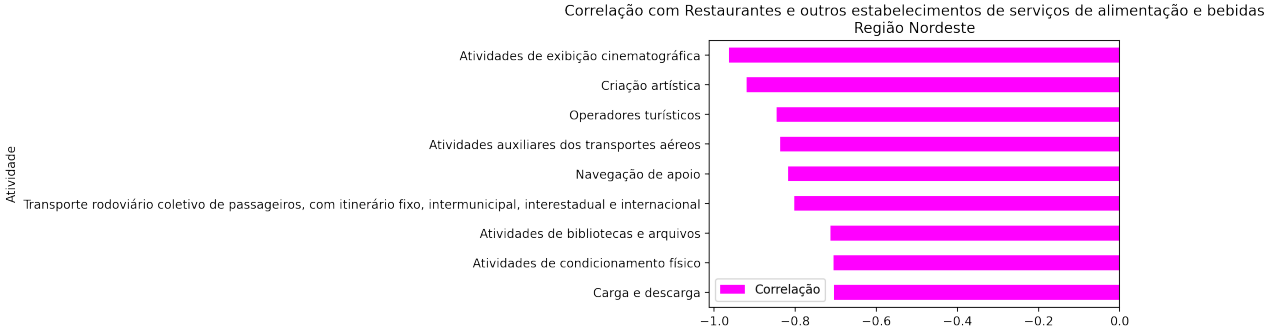

Figura 12: Gráfico das atividades com correlação inferior a -0.7 em relação aos restaurantes. As barras são mostradas em ordem crescente.

Observamos que atividades de exibição cinematográfica é a atividade mais negativamente correlacionada. Algumas atividades surpreendentemente correlacionadas negativamente são criação artística e operadores turísticos.

As atividades de correlação positiva no gráfico serão as atividades que utilizaremos para construir o modelo de gradiente descendente para a região Nordeste.

Correlações para região Sul

O gráfico abaixo mostra as atividades com correlação maior que 0.7 com a atividade de restaurantes.

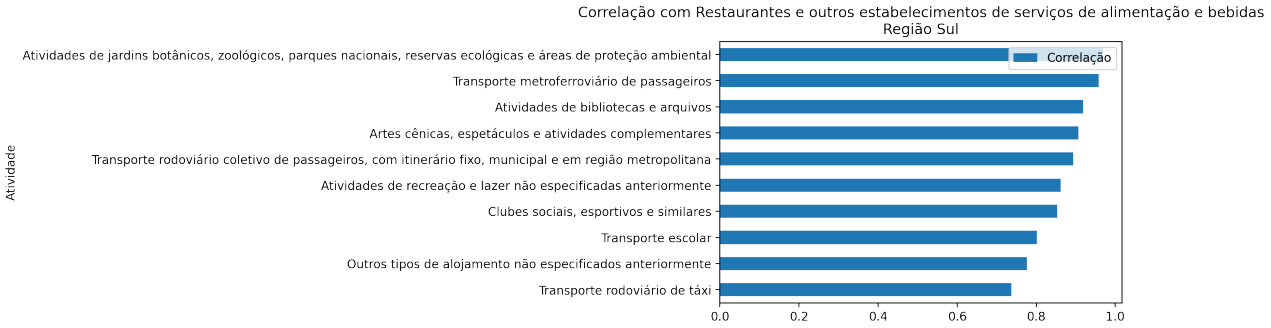

Figura 13: Gráfico das atividades com correlação superior a 0.7 em relação aos restaurantes. As barras são mostradas em ordem decrescente.

Observamos que assim como na região Nordeste existe forte correlação em áreas de proteção ambiental. Interessante notar as altas correlações com atividades de bibliotecas e artes cênicas.

O gráfico abaixo mostra as atividades com correlação menor que -0.6 com a atividade de restaurantes.

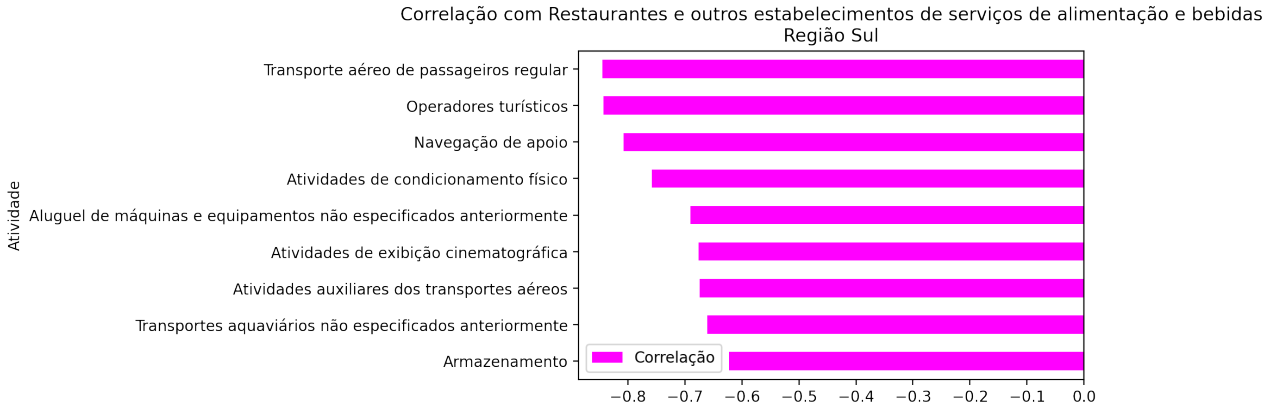

Figura 14: Gráfico das atividades com correlação inferior a -0.6 em relação aos restaurantes. As barras são mostradas em ordem crescente.

Observamos que a atividade mais negativamente correlacionada é transporte aéreo de passageiros regular, o que é um pouco surpreendente. As demais atividades não apresentam tanta surpresa.

As atividades de correlação positiva no gráfico serão as atividades que utilizaremos para construir o modelo de gradiente descendente para a região Sul.

Correlações para região Centro-Oeste

O gráfico abaixo mostra as atividades com correlação maior que 0.8 com a atividade de restaurantes.

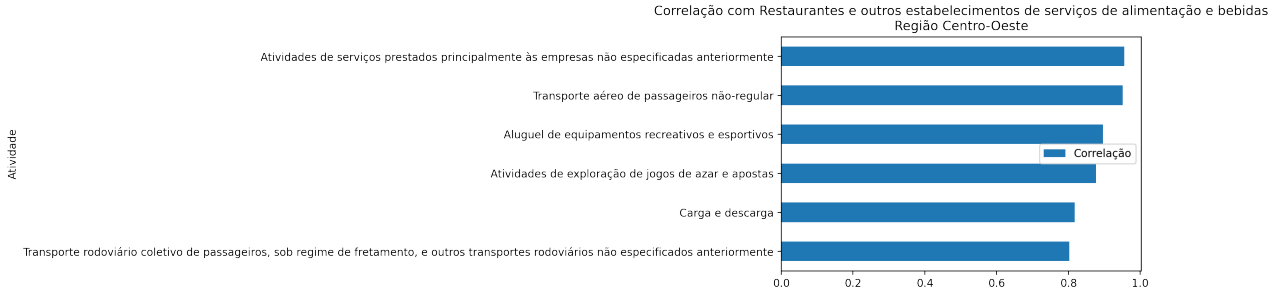

Figura 15: Gráfico das atividades com correlação superior a 0.8 em relação aos restaurantes. As barras são mostradas em ordem decrescente.

Observamos que a atividade mais positivamente correlacionada é a categoria Atividades de serviços prestados principalmente às empresas não especificadas anteriormente, que é surpreendente pois é altamente não específica. Também temos aluguel de equipamentos esportivos e atividades de jogos de azar com correlações positivas maiores que 0.8.

O gráfico abaixo mostra as atividades com correlação menor que -0.25 com a atividade de restaurantes.

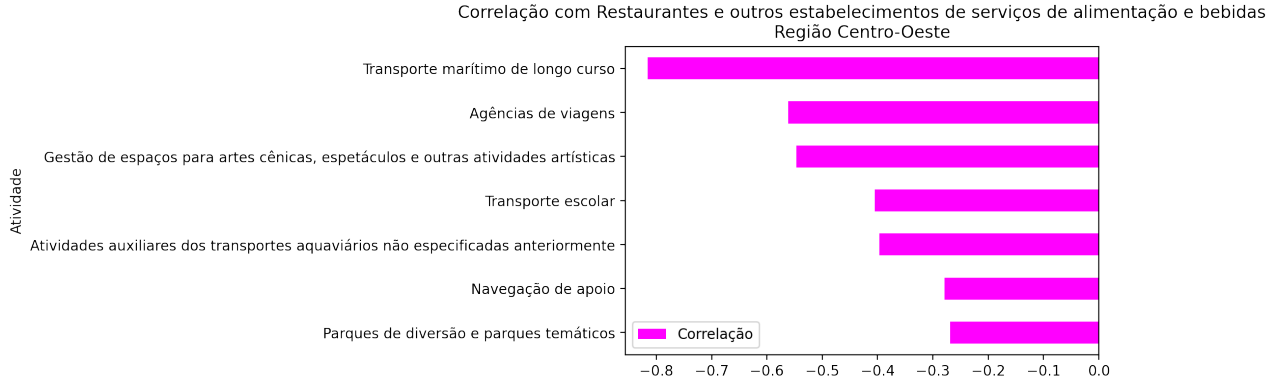

Figura 16: Gráfico das atividades com correlação inferior a -0.25 em relação aos restaurantes. As barras são mostradas em ordem crescente.

Observamos que as atividades com correlação mais negativa é transporte marítimo, seguida de agência de viagens e gestão de espaços para artes cênicas. É uma coleção peculiar de correlações negativas, da qual é difícil extrair um elemento unificador.

As atividades de correlação positiva no gráfico serão as atividades que utilizaremos para construir o modelo de gradiente descendente para a região Centro-Oeste.

Modelo preditivo: Gradiente descendente

Após a determinação de um subconjunto de atividades mais positivamente correlacionadas com a quantidade de unidades locais de restaurantes, feita na seção anterior, construímos para cada região um modelo de regressão utilizando gradiente descendente para prever valores para 2022.

Diferentemente da análise da regressão linear, o modelo do gradiente descendente não recebe apenas o ano como entrada. A entrada são novos valores **para cada atividade na tabela, exceto restaurantes.** Isso significa que, para produzir uma previsão para cada região, será necessário produzir valores para cada conjunto de atividades mais correlacionadas e então utilizar esses valores como entrada para o modelo.

Isso torna o modelo mais complexo e difícil de ser usado, porém é mais flexível e mais poderoso. Nas subseções seguintes mostraremos como o novo modelo se adapta aos dados anteriores e na última subseção discutiremos como foi feita a previsão em cada caso, juntamente com seus valores.

Modelo da região Sudeste

O gráfico abaixo mostra os dados originais e os pontos obtidos pelo gradiente descendente para a região Sudeste.

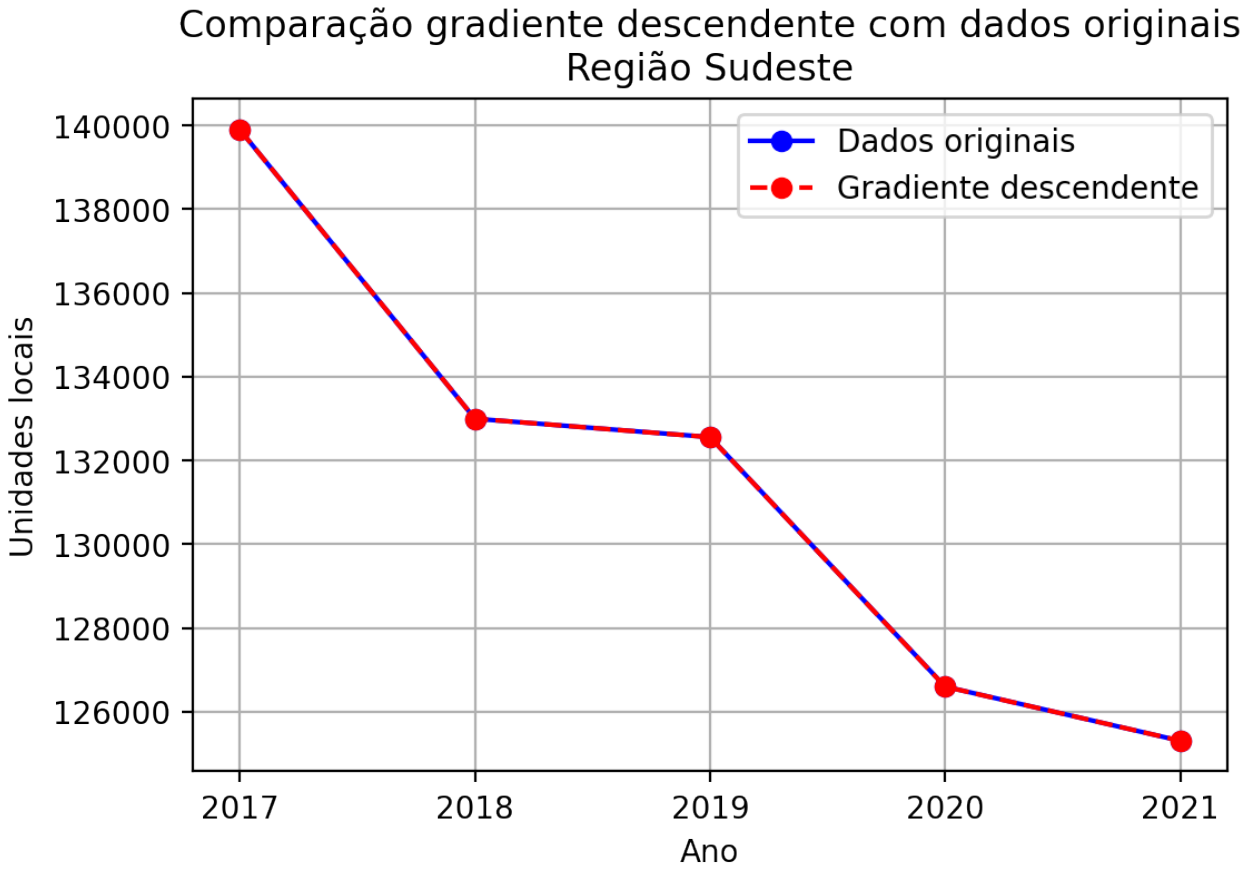

Figura 17: Gráfico dos pontos originais, representados por pontos azuis, e gráfico dos pontos obtidos pelo gradiente descendente, representados por pontos vermelhos. Os pontos azuis são conectados por linhas cheias para representar a tendência de um ano para o outro, enquanto que a tendência representada pelo gradiente descendente está tracejada em vermelho.

A avaliação deste modelo foi de 0.99998, o que significa que nos dados utilizados para treinamento possui acurácia de 99.99%. É evidente pelo gráfico que o modelo segue a mesma tendência dos dados, praticamente copiando os valores. A diferença, nos casos em que apareceu, foi de apenas uma ou duas unidades.

Modelo da região Norte

O gráfico abaixo mostra os dados originais e os pontos obtidos pelo gradiente descendente para a região Norte.

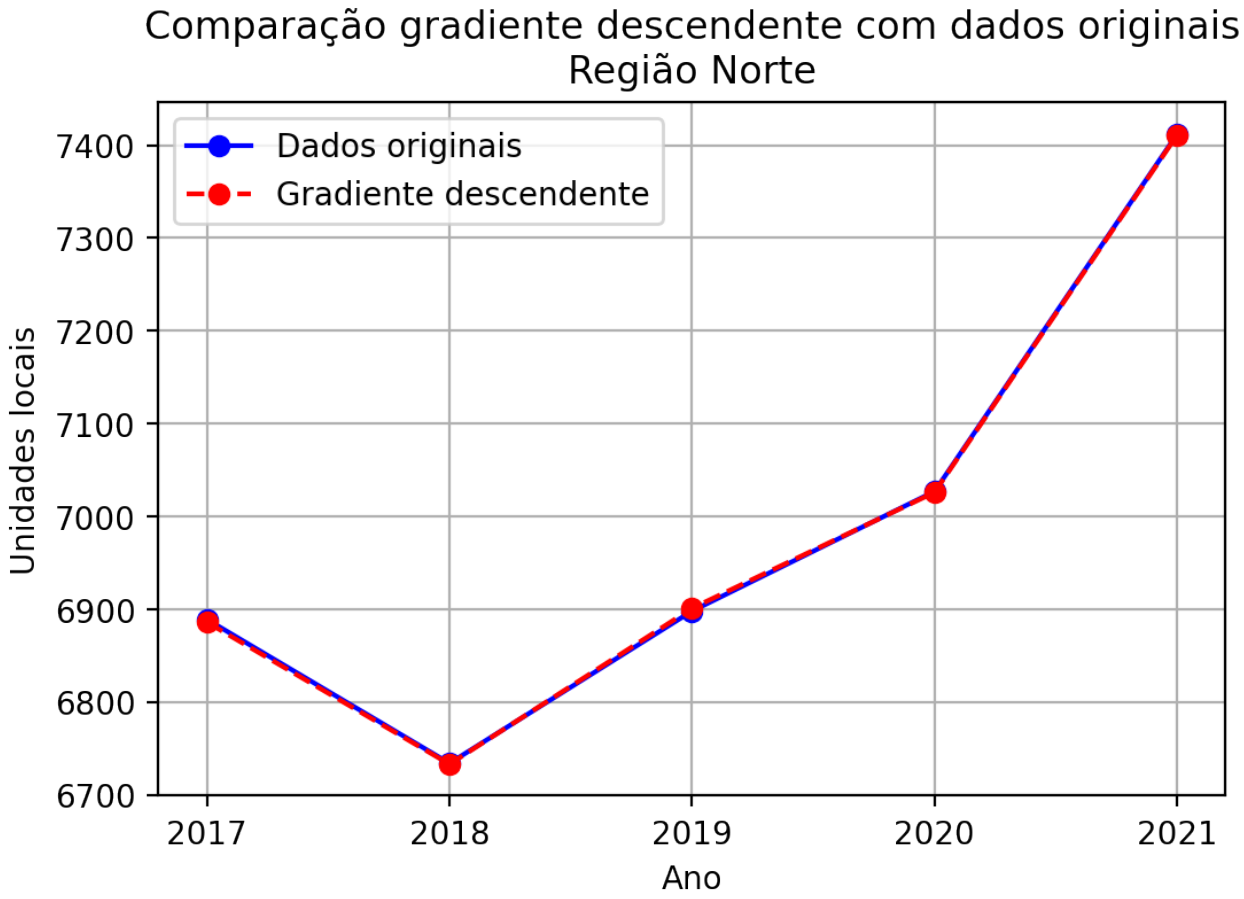

Figura 18: Gráfico dos pontos originais, representados por pontos azuis, e gráfico dos pontos obtidos pelo gradiente descendente, representados por pontos vermelhos. Os pontos azuis são conectados por linhas cheias para representar a tendência de um ano para o outro, enquanto que a tendência representada pelo gradiente descendente está tracejada em vermelho.

A avaliação deste modelo foi de 0.99919, o que significa que nos dados utilizados para treinamento possui acurácia de 99.91%. É evidente pelo gráfico que o modelo segue a mesma tendência dos dados, praticamente copiando os valores. A diferença, nos casos em que apareceu, foi de apenas uma ou duas unidades.

Modelo da região Nordeste

O gráfico abaixo mostra os dados originais e os pontos obtidos pelo gradiente descendente para a região Nordeste.

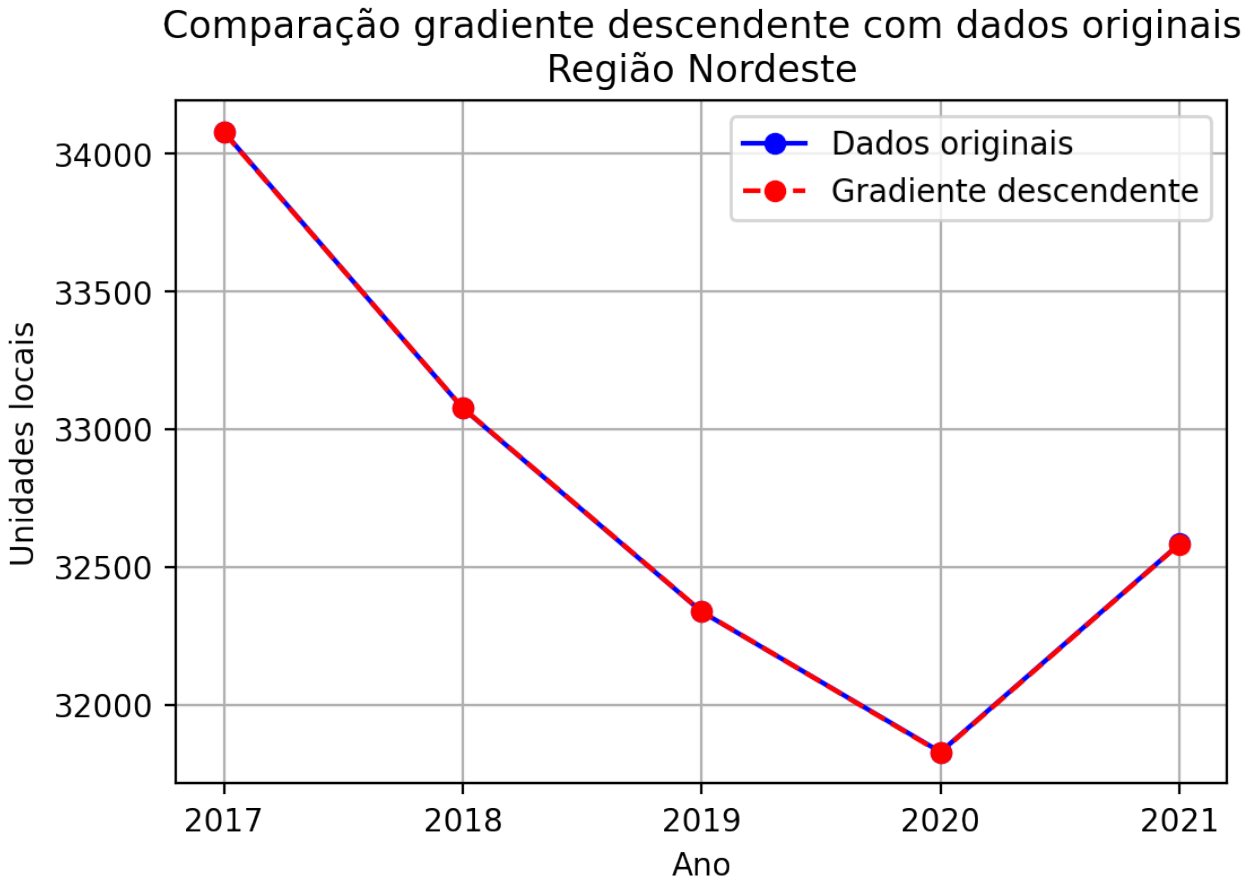

Figura 19: Gráfico dos pontos originais, representados por pontos azuis, e gráfico dos pontos obtidos pelo gradiente descendente, representados por pontos vermelhos. Os pontos azuis são conectados por linhas cheias para representar a tendência de um ano para o outro, enquanto que a tendência representada pelo gradiente descendente está tracejada em vermelho.

A avaliação deste modelo foi de 0.99988, o que significa que nos dados utilizados para treinamento possui acurácia de 99.98%. É evidente pelo gráfico que o modelo segue a mesma tendência dos dados, praticamente copiando os valores. A diferença, nos casos em que apareceu, foi de apenas uma ou duas unidades.

Modelo da região Sul

O gráfico abaixo mostra os dados originais e os pontos obtidos pelo gradiente descendente para a região Sul.

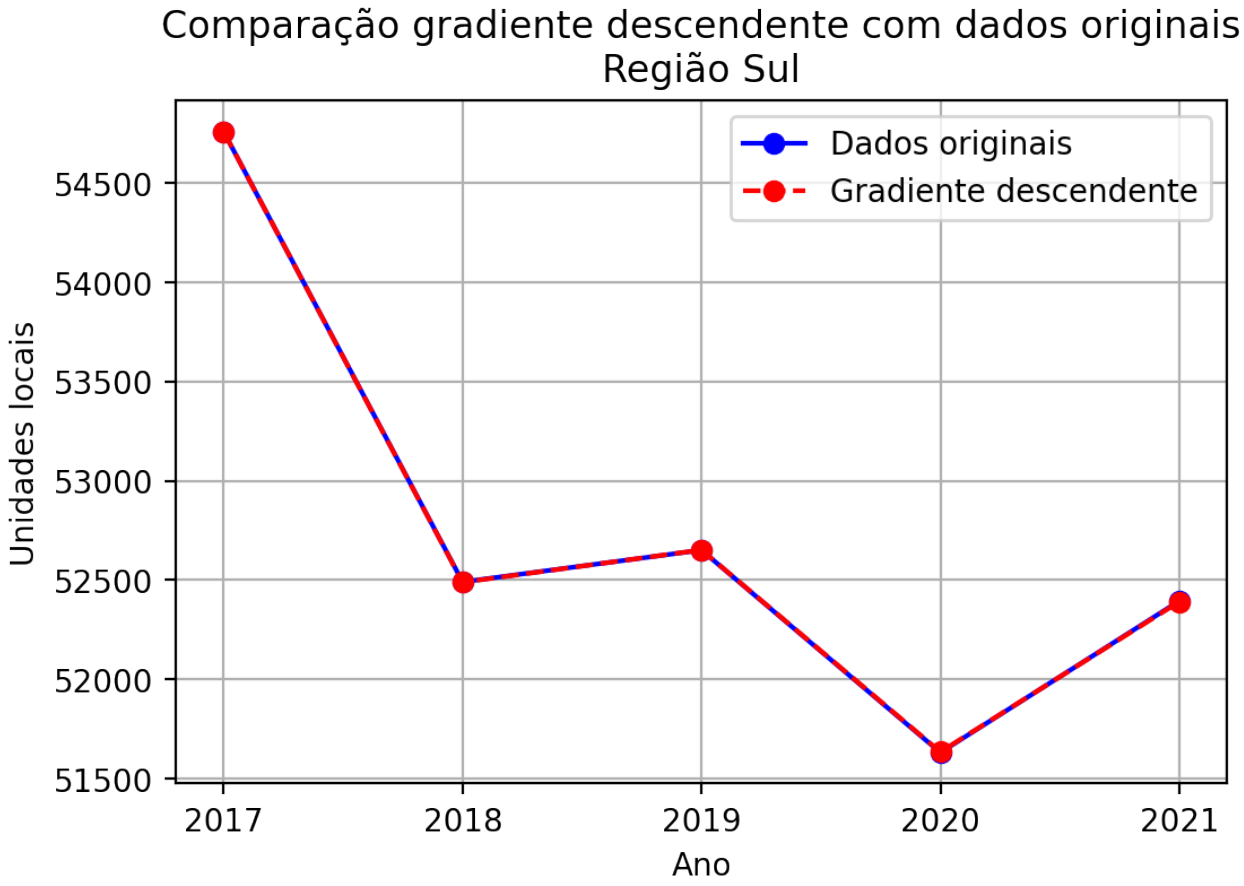

Figura 20: Gráfico dos pontos originais, representados por pontos azuis, e gráfico dos pontos obtidos pelo gradiente descendente, representados por pontos vermelhos. Os pontos azuis são conectados por linhas cheias para representar a tendência de um ano para o outro, enquanto que a tendência representada pelo gradiente descendente está tracejada em vermelho.

A avaliação deste modelo foi de 0.99925, o que significa que nos dados utilizados para treinamento possui acurácia de 99.92%. É evidente pelo gráfico que o modelo segue a mesma tendência dos dados, praticamente copiando os valores. A diferença, nos casos em que apareceu, foi de apenas uma ou duas unidades.

Modelo da região Centro-Oeste

O gráfico abaixo mostra os dados originais e os pontos obtidos pelo gradiente descendente para a região Centro-Oeste.

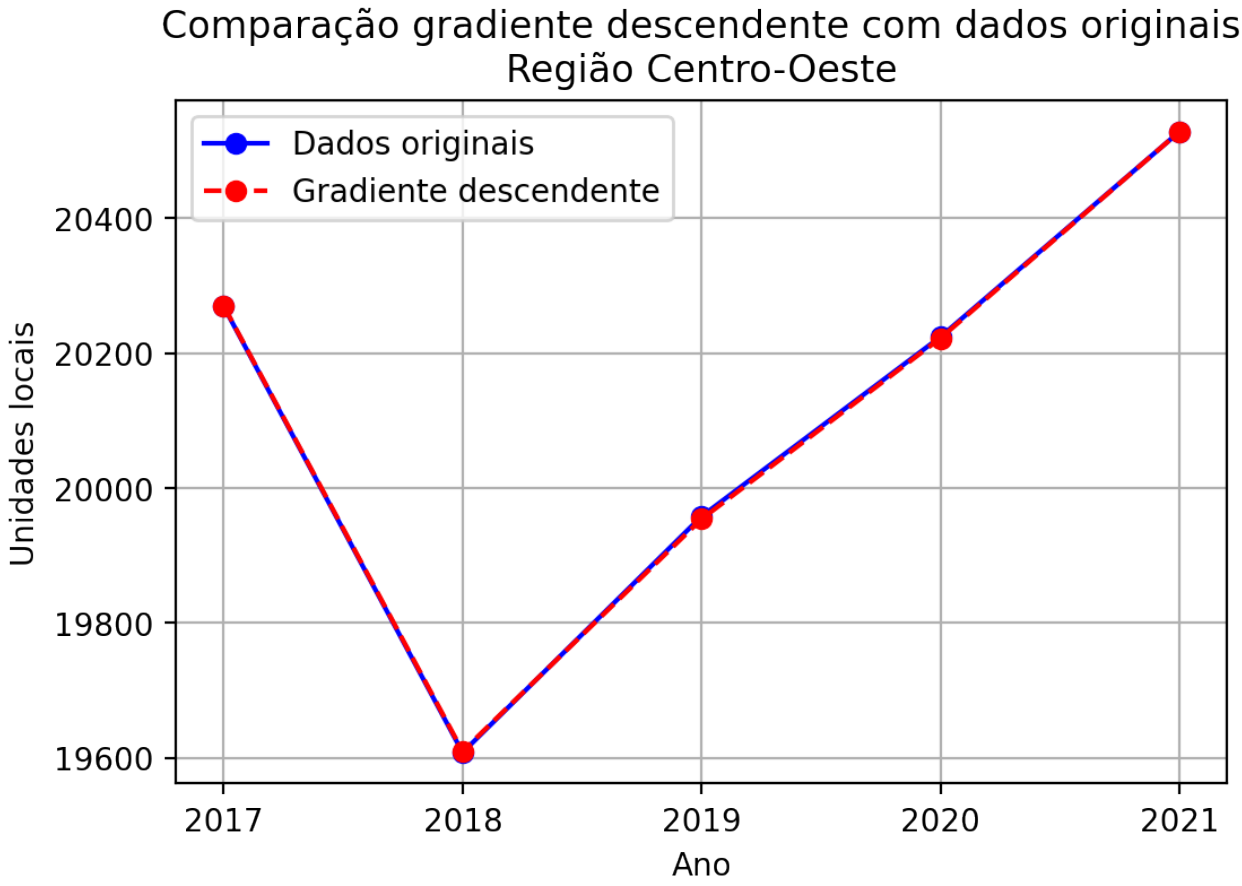

Figura 21: Gráfico dos pontos originais, representados por pontos azuis, e gráfico dos pontos obtidos pelo gradiente descendente, representados por pontos vermelhos. Os pontos azuis são conectados por linhas cheias para representar a tendência de um ano para o outro, enquanto que a tendência representada pelo gradiente descendente está tracejada em vermelho.

A avaliação deste modelo foi de 0.99999, o que significa que nos dados utilizados para treinamento possui acurácia de 99.99%. É evidente pelo gráfico que o modelo segue a mesma tendência dos dados, praticamente copiando os valores. A diferença, nos casos em que apareceu, foi de apenas uma ou duas unidades.

Discussão de validade e previsões

Os modelos de gradiente descendente obtidos para cada região são extremamente próximos dos valores originais. Pela quantidade pequena de dados, é possível afirmar que existe *overfitting*[3] do modelo nos dados. Entretanto, mesmo assim é possível fazer previsões.

Faremos três previsões para 2022 em cada região:

- uma previsão assumindo que os valores para cada atividade é igual à média dos valores dos anos anteriores;
- uma previsão assumindo que os valores para cada atividade é igual à média dos valores dos anos anteriores mais o desvio padrão;
- uma previsão assumindo que os valores para cada atividade é igual à média dos valores dos anos anteriores menos o desvio padrão;

Essa hipótese de assumir valores para as atividades de 2022 como a média dos anteriores é necessária pois precisamos imputar dados nas atividades para que o modelo de gradiente descendente faça as previsões. Além disso, é uma estimativa razoável para o ano seguinte, e as previsões adicionando ou subtraindo o desvio padrão nos permitem introduzir um intervalo de confiança na previsão.

[3] Overfitting é quando o modelo é ajustado demais aos dados de treinamento, capturando o ruído e as variações aleatórias dos dados. Isso também costuma significar que o modelo não generaliza bem para novos dados.

Região	$\mu - \sigma$	μ	$\mu + \sigma$
Sudeste	125.667	131.466	137.264
Norte	6.714	6.964	7.214
Nordeste	31.955	32.777	33.600
Sul	51.663	52.739	53.814
Centro-Oeste	19.706	20.100	20.494

Tabela 1: Previsões para o ano de 2022, com média μ e desvio padrão σ.

Conclusão

Este estudo analisou e previu a quantidade de unidades locais de restaurantes em cada grande região do Brasil utilizando dois modelos preditivos: regressão linear e gradiente descendente.

A seguinte tabela resume a análise para o modelo de regressão linear:

Região	Acurácia	Previsão 2022	Previsão 2023
Sudeste	92,9%	120.765	117.236
Norte	68,0%	7.393	7.527
Nordeste	61.4%	31.508	31.084
Sul	56,9%	51.110	50.552
Centro-Oeste	26,3%	20.457	20.571

Tabela 2: Previsões para os anos de 2022 e 2023 obtidos com o modelo de regressão linear, juntamente com a acurácia do modelo para cada região.

Essas previsões aparecem na ordem de mais confiável a menos confiável. Elas também evidenciam uma fraqueza do modelo de regressão linear: uma vez que o modelo determina os coeficientes de previsão, os valores podem apenas aumentar (se o modelo for crescente) ou diminuir (se o modelo for decrescente). Torna-se difícil de prever qualquer mudança drástica baseado na variabilidade de dados anteriores, ou mesmo adaptar-se a dados com grande variabilidade, como evidente nas regiões Sul e Centro-Oeste, regiões com piores acurácias. A pandemia do ano de 2020 é um exemplo de mudança drástica que um modelo como a regressão linear apresenta pouca ou nenhuma possibilidade de incorporar sua variabilidade confiavelmente.

A seguinte tabela resume a análise para o modelo de gradiente descendente:

Região	Acurácia	Previsão (μ)	$[\mu - \sigma, \mu + \sigma]$
Sudeste	99,99%	131.466	[125.667, 137.264]
Norte	99,91%	6.964	[6.714, 7.214]
Nordeste	99,98%	32.777	[31.995, 33.600]
Sul	99,92%	52.739	[51.663, 53.814]
Centro-Oeste	99,99%	20.100	[19.706, 20.494]

Tabela 3: Previsões e intervalo de confiança da previsão para o ano de 2022, com média μ e desvio padrão σ, juntamente com a acurácia do modelo para cada região.

As acurácias tão altas não são confiáveis, pois com poucos dados o modelo de gradiente descendente é altamente suscetível a sofrer de *overfitting*, o que aconteceu para cada região. Entretanto, ainda assim é possível criar um intervalo de confiança.

Para que o modelo pudesse funcionar confiavelmente como preditor dos anos seguintes o ideal seria possuir mais dados estatísticos, como valores de mais anos, ou mais características, como impacto no PIB ou outras variáveis econômicas.

Limitações e direções futuras

Regressão linear

O modelo de regressão linear é simples e fácil de aplicar, porém com resultados pouco aplicáveis exceto em situações como a da região Sudeste, em que os dados mostraram uma tendência linear. Essa limitação impede que o modelo de regressão linear seja mais amplamente utilizado. Seria possível fazer uma regressão multi-linear caso existissem mais dados ou mais variáveis para inclusão no modelo.

Gradiente descendente

Uma das limitações deste estudo é a falta de dados mais recentes e de outras variáveis além da quantidade de unidades locais, o que afeta a precisão das previsões. Isto dificultou a implementação do modelo de gradiente descendente, uma vez que o método comum é separar o conjunto de dados em pelo menos 2 partes, denominadas treinamento e teste[4].

Porém, com apenas 5 dados é impossível fazer essa separação, forçando o modelo a não ter um conjunto de testes e apenas fazer o treinamento com os dados disponíveis.

Para trabalhos futuros, sugere-se a inclusão de outras variáveis no conjunto de dados, como o PIB per capita e a densidade populacional, para melhorar a precisão das previsões e das análises. Também seria interessante utilizar modelos mais avançados, como redes neurais, para comparar com os modelos utilizados neste estudo.

Outra possibilidade seria analisar a relação entre a quantidade de unidades locais de restaurantes e a qualidade dos serviços oferecidos, como a avaliação dos clientes em plataformas online.

[4] O ideal é separar o conjunto de dados em 3 partes, denominadas treinamento, validação, e teste. O conjunto de validação serve para otimizar os parâmetros do modelo após o ajuste do modelo ao conjunto de treinamento, com o objetivo de garantir melhor generalização do modelo para dados novos. Essa generalização é verificada através da previsão do modelo no conjunto de dados teste.

Considerações finais

Poderíamos ter utilizado a técnica de *bootstrapping* para criar mais dados artificialmente, se houvesse tempo, e refazer as análises com o modelo de gradiente descendente (que sofreu *overfitting*) para melhorar a precisão das previsões obtidas. Também não houve tempo para procurar melhores métodos estatísticos para tratar dos dados.

Referências

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

C. R. Harris et al. Numpy. `https://numpy.org/devdocs/reference/index.html`, 2020.

J. D. Hunter. Matplotlib. `https://matplotlib.org/stable/`, 2007.

W. McKinney et al. Pandas. `https://pandas.pydata.org/docs/reference/index.html`, 2010.

Gottfried Noether. *Introduction to Statistics*. Wiley, 1990.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn. `https://scikit-learn.org/stable/`, 2011.

M. Waskom. Seaborn. `https://seaborn.pydata.org/`, 2014.