Modelos Computacionais de Projeções de Indicadores Sociais para Municípios do Brasil

Estudos de previsão estatística e aprendizado de máquina

**Vellenich**, Danton F.**, Meneguzzi,** Leandro**, Araujo**, Iago Breno

Artigo submetido em 18 de Dezembro de 2020

# 

# Abstract

Os indicadores sociais são dados estatísticos que acompanham as evoluções econômicas e sociais de uma determinada população. A coleta de dados e estudos é conduzida na maioria dos países por órgãos governamentais. Esta pesquisa se dedica a investigar algumas técnicas de aprendizado de máquina e modelos computacionais que sejam capazes de projetar estatisticamente a evolução da população e PIB (Produto Interno Bruto) de municípios brasileiros baseados em dados de anos anteriores coletados pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

*© Universidade de São Paulo, All rights reserved.*

*Keywords:* Socioeconomic indicators, machine learning, statistical projection



# Introdução

Uma definição de “Desenvolvimento Sustentável” é “o desenvolvimento capaz de suprir as necessidades da geração atual, sem comprometer a capacidade de atender as necessidades das futuras gerações”. Este conceito foi proposto na Comissão Mundial sobre Meio Ambiente e Desenvolvimento. Esta comissão tem como objetivo gerar discussões e propostas para equilibrar o desenvolvimento econômico e conservação ambiental, e foi criada pelas Nações Unidas.

O monitoramento do desenvolvimento econômico e das condições sociais, que guiam o entendimento sobre as necessidades atuais, são apresentadas globalmente pelos indicadores sociais. Os indicadores sociais são dados estatísticos de uma população, geridos por órgãos governamentais na maioria dos países pertencentes às Nações Unidas.

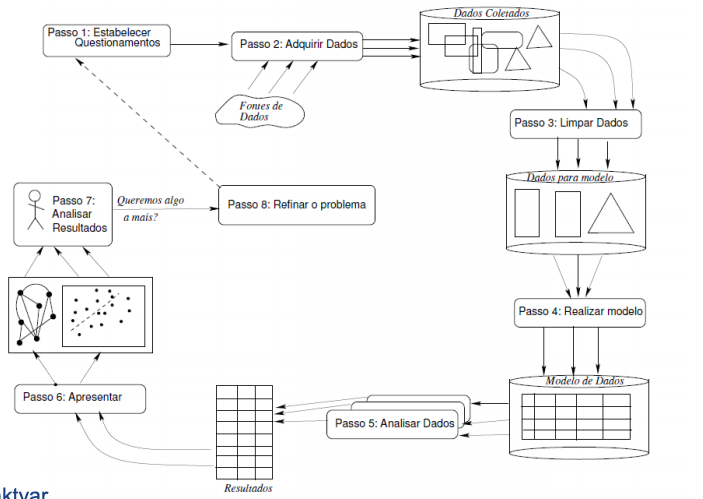
No Brasil, o órgão responsável por estes dados é o Instituto Brasileiro de Geografia e Estatística - IBGE. O IBGE foi fundado na década de 1940 e desde então realiza o Censo Demográfico a cada 10 anos, salvo excepcionalidades. O Censo Demográfico tem destaque, dentre as dezenas de outras pesquisas realizadas por este órgão, pela sua regularidade e abrangência do estudo, capaz de entrar em contato com todas as residências brasileiras.

No período entre as pesquisas do Censo Demográfico no Brasil, os indicadores populacionais são estimados estatisticamente.   
A pergunta que este trabalho de pesquisa busca responder é se é possível treinarmos e utilizarmos um modelo computacional de aprendizado de máquina que possa, a partir dos dados de estimativa populacional e de demais pesquisas do IBGE, projetar os indicadores de anos futuros e demonstrar tendências para unidades geográficas municipais.

Seu objetivo é investigar algumas técnicas de modelos computacionais de aprendizado de máquina que sejam capazes de projetar estatisticamente a partir de dados históricos dos municípios brasileiros, indicadores socioeconômicos População e Produto Interno Bruto e alcançar no mínimo 80% de similaridade com o indicador efetivamente mensurado pelo IBGE, que pode ser calculado pelo erro quadrático médio.

A razão pela escolha do domínio municipal brasileiro para análise deste trabalho é o fato que é a menor granularidade possível de dados disponíveis pelo IBGE para a variável de produto interno bruto, além da composição de outros conjuntos de informações, mantendo o propósito interpretativo e analítico do dado.

Este trabalho de pesquisa fundamenta-se no conceito do ciclo de Ciência dos Dados (Alex Dekhtyar, 2016), ao estabelecer a questão declarada acima e sequencialmente seguirá por aquisição de dados, limpeza, modelamento, análise, apresentação e análise dos resultados, detalhados nos capítulos seguintes.

  
**Figura 1**- Ciclo de Ciência dos Dados por Alex Dekhtyar, 2016

Este artigo integra o grupo de pesquisa internacional nomeado PARSEC Project ([www.parsecproject.org](http://www.parsecproject.org)), que tem como objetivo o desenvolvimento de novas ferramentas para o compartilhamento e reuso de dados através de pesquisa transnacional sobre o impacto socioeconômico das áreas protegidas, criado pelo Belmont Fórum e no Brasil, financiado pela FAPESP (2018/24017-3).

# Métodos

## Dados

A pesquisa utiliza dados originalmente do IBGE, disponibilizados no sistema Web SIDRA (Sistema IBGE de Recuperação Automática). Os dados foram obtidos a partir da pesquisa com o foco na esfera municipal. Inicialmente, buscou-se obter variáveis definidas no escopo do projeto e associadas à responder às questões. Essas variáveis são PIB, IDH e população das cidades. O conjunto de dados é descrito como uma série temporal contendo os dados disponíveis para cada variável. Como está sendo considerado as unidades geográficas municipais, cada exemplo ou instância do conjunto representa um município com um código identificador e os valores quantitativos para cada ano do conjunto.

Foram consideradas também variáveis que apresentaram preenchimento satisfatório e não possuem restrições de privacidade à nível municipal.

Além das variáveis de interesse do estudo, outras variáveis foram pesquisadas e agregadas ao conjunto inicial com o objetivo de compreender a relação entre elas. Estas variáveis são:

* IDH global;
* IDH na dimensão renda;
* IDH na dimensão longevidade;
* IDH na dimensão educação;
* Número de casamentos;
* Salários e outras remunerações;
* Unidade empresariais;
* Pessoas empregadas;
* Pessoas assalariadas.

Esse conjunto de dados utilizado na pesquisa foi registrado e disponibilizado em um D.O.I. (Document Object Identifier) correspondente (10.5281/zenodo.4323383), para garantir a reprodutibilidade deste experimento.

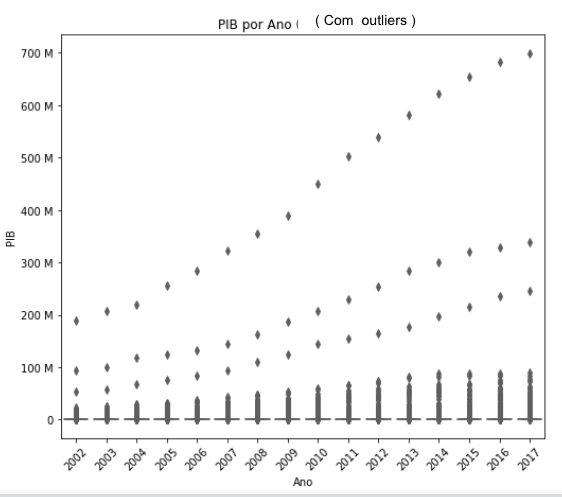
Ainda relacionado à gestão dos dados, descrito pelo DMPTool, template USP, vale ressaltar que os dados utilizados nesta pesquisa são públicos e as questões éticas relacionadas à privacidade e identificação já são tratados pelo próprio IBGE.

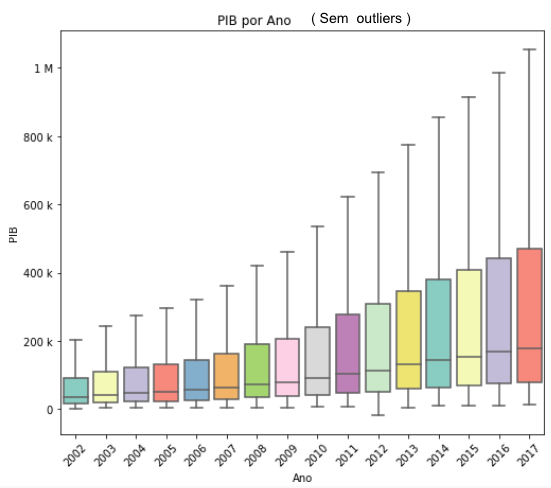
## Análise Exploratória dos Dados

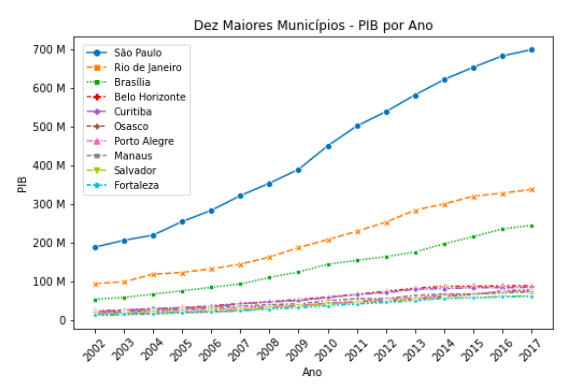
Analisando os dados disponíveis, podemos encontrar 5570 municípios brasileiros, com dados do período de 2002 a 2017.

Das principais variáveis de interesse, podemos analisar seus valores absolutos e distribuição dentro do período.

Iniciando a análise sobre o Produto Interno Bruto do Município (PIB), que representa a soma das receitas auferidas no município e contabilizadas no Sistema de Contas da União.



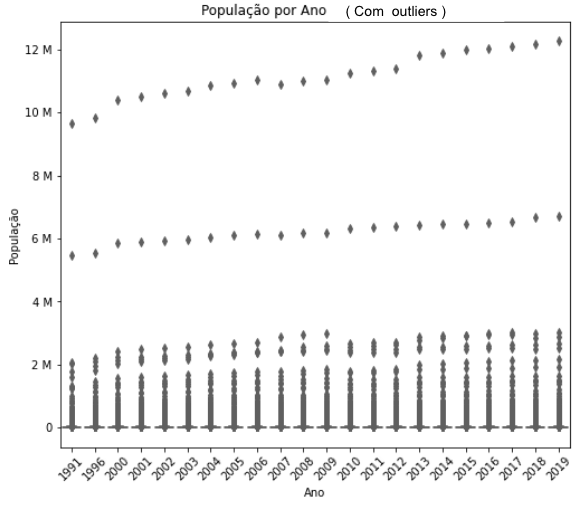


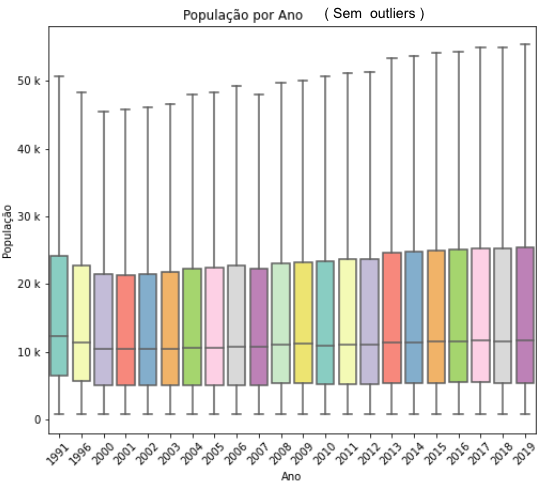


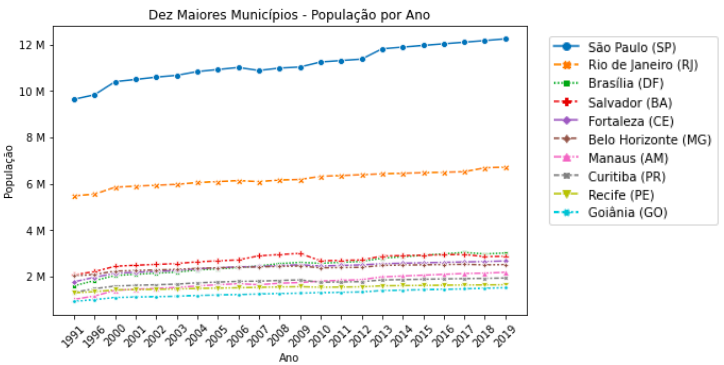
**Figura 2**- Conjunto de Gráficos de Produto Interno Bruto do Município

Os gráficos acima demonstram que o PIB de 3 municípios brasileiros (São Paulo, Rio de Janeiro e Brasília) se distanciam em centenas de milhões de reais do restante dos 5570 municípios. Analisando, através dos box plots, percebemos que os quartis dos municípios estão presentes dentro da faixa até 400 mil reais em 2017.

Seguindo o mesmo modelo de análise para população, podemos analisar os seguintes gráficos.







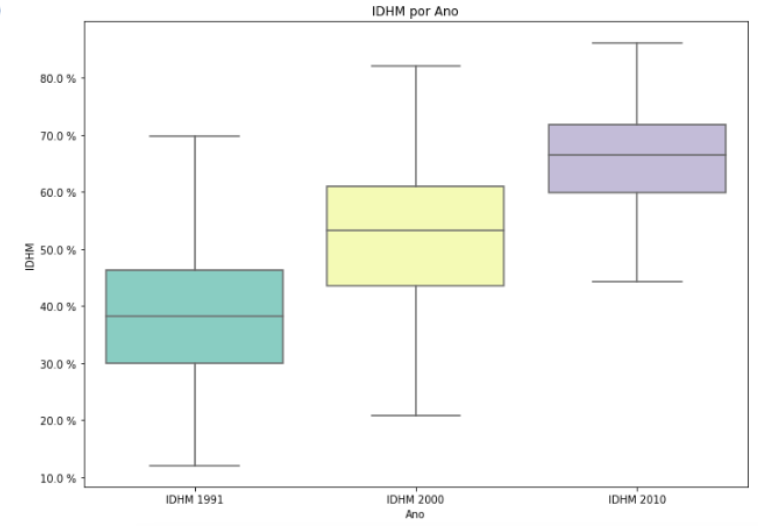
**Figura 3**- Conjunto de Gráficos de População Municipal

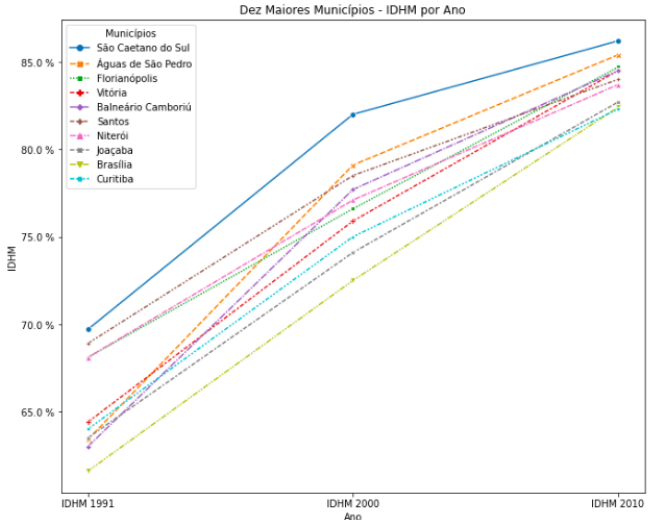
A população dos municípios brasileiros, assim como o produto interno bruto municipal, tem crescimento anual entre os anos de 2002 e 2017. Entretanto, a taxa de crescimento anual da população é comparativamente inferior à taxa do PIB.

Além disso, é possível analisar através do boxplot que em média o município brasileiro tem a ordem de 10 mil habitantes e que 75% dos municípios estão na faixa de 5 e 25 mil habitantes.

Esta ordem de grandeza é bastante distante do município de São Paulo, com 12 Milhões de habitantes e do município do Rio de Janeiro com 6 Milhões de habitantes em 2019.

O Índice de Desenvolvimento Humano (IDH) é um indicador utilizado globalmente para a medição das condições de vida e composto por 3 variáveis: Renda, Longevidade e Educação. No Brasil, a análise do IDH por município é realizada a cada 10 anos, em sincronia com o Censo Demográfico, utilizado para formatação da informação.

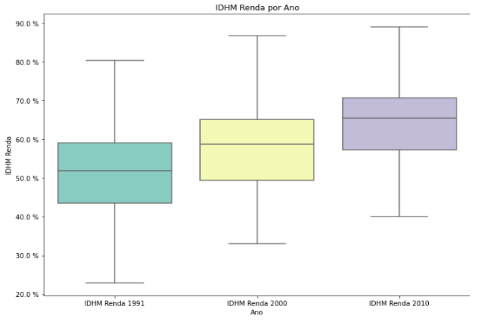


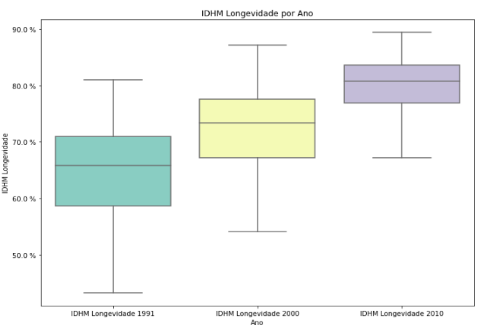


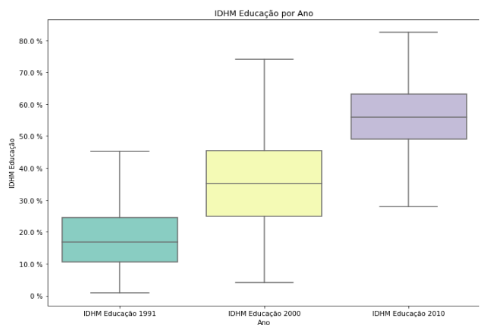
**Figura 4**- Distribuição do IDH por município brasileiro e TOP 10 municípios por IDH.

Analisando a distribuição e evolução dos gráficos da Figura 4, podemos perceber que tanto a média quanto os quartis cresceram significativamente, representando melhoria nas condições de vida nos municípios brasileiros nas últimas 3 décadas.

Podemos analisar independentemente cada critério que compõe o IDH, seguindo a figura abaixo:



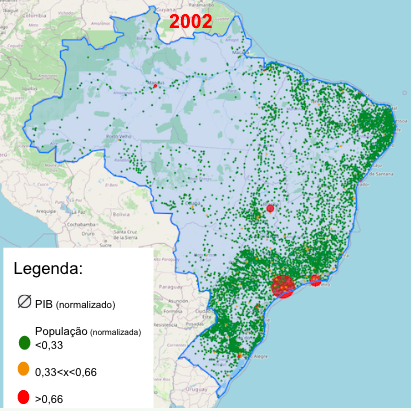


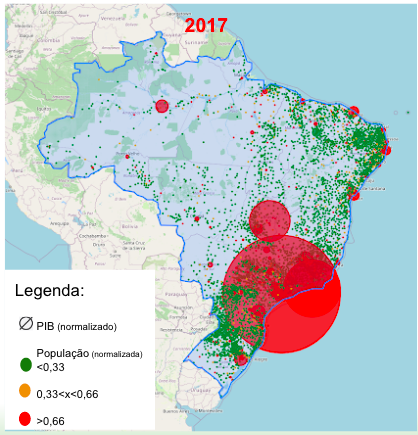


**Figura 5** - Distribuição dos 3 critérios do IDH por município brasileiro

Os gráficos demonstram que todos os critérios evoluíram nos últimos 30 anos significativamente. O critério de Educação teve em termos absolutos o maior salto ainda que a média do ano de 2010 na faixa de 60% está 20 ppc da média da renda, demonstrando que existe espaço para melhoria.

Dos dados anteriores, localizar geograficamente cada município brasileiro, é possível gerar as seguintes visualizações:





**Figura 6** - Geolocalização das informações de PIB e população dos municípios brasileiros

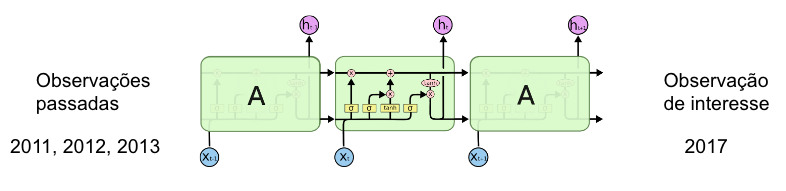
Analisando geograficamente os dados de população e produto interno bruto, é possível constatar visualmente o conceito do grande número de municípios pequenos em termos de população (verde) e PIB e a grande distorção produzida pelos municípios de São Paulo, Rio de Janeiro, Brasília e demais capitais estaduais brasileiras.

Analisando as diferenças dos períodos entre 2002 e 2017 é possível notar a evolução de municípios do nordeste, do sul do Pará e norte de Mato Grosso, com municípios trocando de categoria de população (de verde para laranja) e crescimento dos raios, representando evolução do PIB.

Todas as análises até aqui representam entendimento a respeito do dado histórico. Para análises de futuro, utilizaremos a partir de agora modelos computacionais para prever o comportamento destas variáveis nos próximos anos.

## Modelos

Os experimentos realizados neste trabalho investigam resultados de projeção estatística utilizando aprendizado de máquina em redes neurais profundas. Essa modelagem considera uma transformação inicial dos dados temporais em um escopo de aprendizado supervisionado. Essa decisão se baseia nos resultados de sucesso do treinamento supervisionado encontrado na literatura em diversas áreas de aplicação para as redes neurais profundas. Além dessa motivação, um subproduto dessa transformação dos dados é a amplitude de uso de diversas arquiteturas que são comumente definidas e propostas para aprender de forma supervisionada. Portanto, a modelagem nessa perspectiva poderá ser amplamente avaliada incluindo redes neurais recorrentes. Especificamente, considerando as limitações das redes neurais recorrentes tradicionais em relação à sua capacidade de manter informações relevantes anteriores cada vez maiores, esse trabalho focou em investigar a perspectiva de uso de redes de memória longa de curto prazo (LSTMs). As LSTMs são redes neurais recorrentes que possuem uma memória longa que é implementada a partir de um mecanismo mais complexo na construção de seus módulos. Enquanto a rede neural recorrente padrão possui uma única camada com uma função de ativação não-linear, as LSTMs possuem quatro camadas em seu módulo que interagem para que o contexto de aprendizado considere mais informações relevantes. Além da definição da arquitetura, outros aspectos envolvidos na definição dos modelos também poderão ser considerados na avaliação, como os otimizadores e as funções objetivo.



**Figura 7** - Estratégia de modelagem utilizando redes neurais profundas recorrentes (LSTMs)

A estratégia de modelagem utilizada pelos experimentos, representada na Figura 7, é eleger uma variável de análise (ora PIB, ora População) e treinar a rede neural (LSTM) oferecendo como entrada da rede os dados históricos de períodos anteriores, utilizando o método de aprendizado supervisionado com dados de um período conhecido futuro aos de entrada.

Além da modelagem, existem opções de arquitetura da rede neural e treinamento que foram testadas nos experimentos:

* Arquiteturas: Bidirecional, Stacked e Single-Layer;
* Otimizadores: Adam e RMSPro;
* Funções objetivo: MSE e MAE.
* Treinamentos: Diferentes opções de épocas e early stopping.

As arquiteturas consideradas são tipos distintos de LSTMs. O objetivo de usar tais variações foi entender como as características específicas de cada uma delas se ajustava ou não ao problema e qual das arquiteturas oferecia o melhor resultado.

## Equipamento

Para realizar os experimentos, foi utilizada a linguagem de programação Python e suas bibliotecas científicas. Para a análise exploratória dos dados, foram consideradas as bibliotecas mais utilizadas para manipulação e visualização de dados dentro da linguagem Python: “Pandas”, “NumPy”, “Matplotlib” e “Folio”. Para a modelagem serão consideradas as bibliotecas “statsmodels”, “scikit-learn”, “TensorFlow” e “Keras”.

## Estatísticas

A avaliação foi realizada considerando prioritariamente a métrica mais difundida na literatura em problemas de regressão, o erro quadrático médio (RMSE). Outras métricas foram consideradas para certificar a performance do modelo, como o erro absoluto média (MAE).

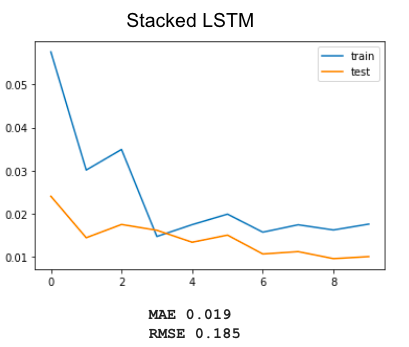
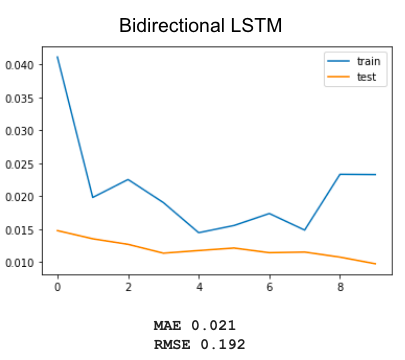
# Resultados

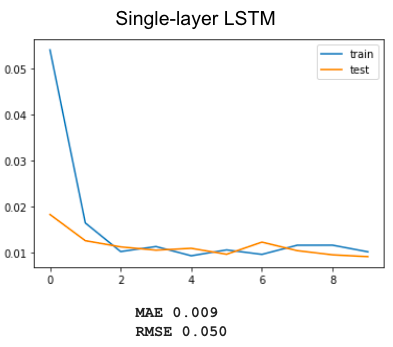
*Resultados dos Modelos*

Os resultados encontrados nos modelos foram os seguintes.

*Predição do PIB*

Para o PIB, foi observado que a melhor arquitetura foi a arquitetura mais simples, a LSTM com uma única camada. Pode-se observar na Figura 8, os resultados para as três arquiteturas. Para esse conjunto de dados, o otimizador Adam foi o melhor para todos os modelos treinados e a função objetivo MSE contribuiu melhor para a minimização dos erros no treinamento. Apesar de serem arquiteturas com maior capacidade, a Stacked LSTM e a Bidirecional LSTM tiveram dificuldades para se ajustar aos dados. Para a Bidirecional LSTM, seu caráter de considerar as duras direções temporais para predição pode não ser tão efetiva por causa da natureza dos dados. A hipótese seria que o contexto do futuro para o passado, que algumas tarefas de processamento natural de linguagem se beneficiam, pode não relacionar bem os dados do PIB. Intuitivamente, uma informação posterior em um texto pode ajudar a entender uma informação mencionada anteriormente. No caso da evolução do produto interno bruto municipal, essa relação pode não existir dado que o contexto que melhor explicaria seriam os avanços de um ano para o outro. Para a Stacked LSTM, mais camadas trouxeram mais complexidade em um contexto relativo de poucos dados.

**

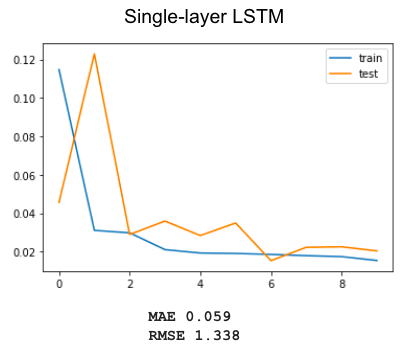
**

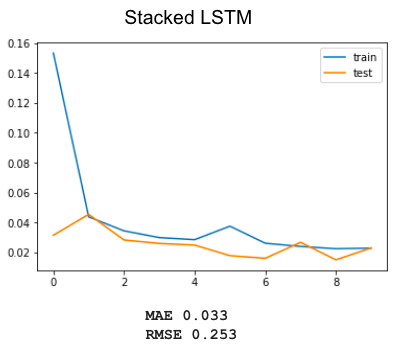
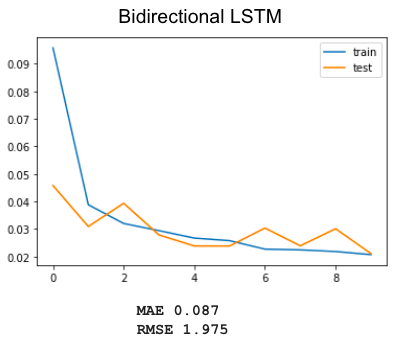
**Figura 8** - Resultados dos modelos de previsão do PIB.

*Predição da População*

Para os dados da população, onde mais anos foram considerados, temos resultados distintos. Para esse caso, na Figura 9 pode-se observar as curvas das funções objetivo de treino e de validação com os respectivos resultados das métricas na avaliação no conjunto de teste, para as arquiteturas investigadas. O melhor modelo foi o baseado na arquitetura Stacked LSTM. Nesse contexto, mais camadas ajudaram o modelo a treinar melhor e obter resultados melhores. Isso pode ter se dado por causa de mais features utilizadas representada no maior conjunto de anos. Assim, a maior capacidade do modelo se ajustou melhor. Para o Single-layer LSTM, pode-se observar maiores valores no erro de validação e nos resultados finais das métricas. Esses resultados concordam com os resultados anterior. A menor quantidade de camadas se ajustou menos. Para o modelo Bidirectional LSTM, a sua capacidade maior demonstrou um bom ajuste com os dados, mas a sua especificidade se ajustou menos em relação à perspectiva temporal padrão.

Considerando os resultados obtidos com o Stacked LSTM, pode-se observar que melhores resultados podem ser obtidos a partir de testes empíricos onde o número de camadas e o número de unidades por camada podem ser testados. Além disso, outros tipos de camada poderiam ser testados como camadas convolucionais.

**

**

**Figura 9** - Resultados dos modelos de previsão de População.

*Referência das Previsões*

Os resultados dos modelos permitem comparar as previsões estatísticas com a realidade factual do dado.

Para avaliar uma projeção de um período que ainda não aconteceu, utilizamos uma referência do estudo realizado pela Oxford Economics para o município de São Paulo no ano 2030, publicado na Revista Exame em 2016, que estima o Produto Interno Bruto de R$ 2,447 B.

A previsão do modelo com arquitetura LSTM Single Layer, otimizador Adam, que ofereceu melhor resultado na variável do PIB, encontramos o valor de R$ 2,064 B.

Se considerarmos o valor do estudo da Oxford Economics como verdade, a diferença entre os dois está em 15%, que atenderia em tese o objetivo do estudo.

# Conclusão

Este trabalho portanto demonstrou o estudo de modelos de aprendizado de máquina com redes neurais profundas para projeção da População e Produto Interno Bruto de um município brasileiro a partir de dados históricos.

Os objetivos propostos foram alcançados parcialmente uma vez que os modelos projetando a variável do Produto Interno Bruto atingiram erros quadráticos médio inferiores a 20% do valor real medido e os modelos para variável de população não atingiram níveis inferiores a este objetivo.

Para futuros trabalhos, sugere-se:

1. o estudo de generalização do modelo com dados de outras regiões e outros períodos para análise de viés e sobreajuste.

2. análise de modelos que considerem multivariáveis versus modelos utilizados neste trabalho que consideram única variável para projeção do seu próprio valor futuro.

3. considerar utilizar novas variáveis e fontes de dados, além das utilizadas neste trabalho.

A expectativa é que este trabalho dê suporte ao avanço das técnicas de identificação de características socioeconômicas, que possam suportar tomadores de decisão para aplicações de políticas de desenvolvimento sustentável no Brasil e em outros países.

# 

# Agradecimentos

Nós gostaríamos de agradecer ao Prof. Pedro Corrêa Piziggatti e Dra. Jeaneth Machicao pela orientação, apoio e incentivo.

Aos Profissionais do IBGE, Dra. Nadya Deps e Msc. Miguel Suarez Xavier Penteado pelo apoio na realização da pesquisa.

E aos parceiros de pesquisa da FAPESP e PARSEC project.

# Referências

* Vivian, Gláucio & Cervi, Cristiano. (2016). “Utilizando Técnicas de Data Science para Definir o Perfil do Pesquisador Brasileiro da Área de Ciência da Computação”. ERBD 2016. https://www.researchgate.net/profile/Glaucio\_Vivian/publication/299418337\_Utilizando\_Tecnicas\_de\_Data\_Science\_para\_Definir\_o\_Perfil\_do\_Pesquisador\_Brasileiro\_da\_Area\_de\_Ciencia\_da\_Computacao/links/56f54e4408ae81582bf20f51.pdf
* Silvano, Tiago Prudencio & Correa, Bryan Maia & Barbosa, Ivanildo, 2020 “Análise da distribuição espacial de indicadores sociais e demográficos: uma abordagem baseada em mineração de dados “, Revista Brasileira de Cartografia, <http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/download/50970/28548/>
* JOAQUIM, Maura da Silva Costa Furtado et al. MODELAMENTO DO FUNDO DE PARTICIPAÇÃO DOS MUNICÍPIOS POR SÉRIES TEMPORAIS. http://www2.ime.unicamp.br/sinape/sites/default/files/Artigo\_Maura\_SINAPE\_0.pdf
* Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.
* Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-17.
* https://exame.com/economia/sp-fica-em-13a-entre-cidades-que-mais-vao-crescer-ate-2030/#:~:text=O%20estudo%20foi%20elaborado%20pela,alta%20de%20US%24%20335%20bilh%C3%B5es.