

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

**Previsão de consumo de cimento nos
estados do Brasil usando métodos de
aprendizado automático**

Julia Leite da Silva

MONOGRAFIA FINAL

MAC 499 — TRABALHO DE
FORMATURA SUPERVISIONADO

Supervisor: Prof. Dr. Marcelo Finger

São Paulo
2022

*O conteúdo deste trabalho é publicado sob a licença CC BY 4.0
(Creative Commons Attribution 4.0 International License)*

*Dedico este trabalho ao meu pai e à minha mãe que
têm iluminado meu dia e me dado força, alegria e paz.*

Agradecimentos

It is good to go into the fight with determination, embrace life and live with passion. Lose with class and win with daring, because the world belongs to those who dare and life is much more to be insignificant.

— Charles Chaplin

Queria agradecer, antes de tudo, a Deus por tudo que Ele tem me dado. Uma vez minha mãe me disse que Deus abençoa o trabalho das nossas mãos, que várias vezes onde ela pensou que não daria certo ou não daria conta, ela pedia ajuda e, no final, tudo se endireitava. Ela me disse também que Deus colocava pessoas na vida dela que a ajudavam e orientavam a encontrar um caminho. Fico muito grata de enxergar isso acontecendo na minha vida também.

Agradeço aos meus pais, que têm sido minha inspiração, força, alegria e lar. Agradeço aos meus amigos, amigas e namorado, por colorirem meus dias e encherem meu rosto de sorrisos.

Quero agradecer ao Marcelo Finger, meu orientador, pelo acompanhamento e apoio neste trabalho. Agradeço também ao Sérgio Goldbaum pelo apoio e àqueles que contribuíram para a pesquisa: Marcelo Chamma, Eliana Taniguti e Valter Frigieri, obrigada pela confiança e parceria. Por fim, agradeço ao Felipe Peressim pelo apoio em relação ao projeto e ao Gabriel Fallaci pelo apoio em relação à escrita deste trabalho.

Resumo

Julia Leite da Silva. **Previsão de consumo de cimento nos estados do Brasil usando métodos de aprendizado automático**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

O cimento é um dos principais insumos da indústria da construção civil que é uma das maiores impulsionadoras da economia brasileira. A indústria cimenteira apresenta a demanda por um método bem fundamentado para prever o consumo do cimento no Brasil, para melhor direcionar estratégias e investimentos. Este trabalho, então, propõe o uso de aprendizado de máquina para prever o consumo de cimento no país, além de investigar o desempenho dos modelos para identificar o mais eficiente para a tarefa. Foram testadas regressão linear, redes *feed forward*, redes recorrentes e redes bidirecionais, além disso, durante o trabalho foi realizado um pré-processamento dos dados para garantir a granularidade necessária, tratar de valores faltantes e assegurar que foram utilizados apenas dados do passado ao realizar a previsão. Por fim, nos experimentos realizados de forma a identificar o modelo com melhor desempenho, destaca-se o aumento na assertividade das previsões – mensurada a partir de métricas de erro absoluto e percentual – à medida que se eleva a complexidade do modelo. Assim, o trabalho em questão investiga se é possível prever a demanda por cimento com assertividade ao utilizar modelos de aprendizado de máquina.

Palavras-chave: Cimento. Consumo de Cimento. Economia Brasileira. Construção Civil. Aprendizado de máquina. Redes Neurais. Regressão Linear. Inteligência Artificial. Previsão.

Abstract

Julia Leite da Silva. **Cement consumption forecasting at brazilian states using machine learning**. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2022.

Cement is one of the main inputs of the civil construction industry, which is one of the biggest input of the Brazilian economy. The cement industry presents the demand of a way of forecasting cement consumption in Brazil, in order to target strategies and investments. This work, then, proposes the use of machine learning to forecasting cement consumption in the country, in addition to investigating the performance of the models to identify the most efficient one for the task. They were tested linear regression, feed forward networks, recurrent networks and bidirectional networks, in addition, during the work was carried out a pre-processing of the data to guarantee the necessary granularity, treat of missing values and ensure that only past data was used when making the prediction. Finally, in the experiments carried out in order to identify the model with the best performance, the increase in the performance of forecasts – measured from absolute and percentage error metrics – to as the complexity of the model increases. Thus, the work in question investigates whether it is possible to predict the demand of cement with good performance when using machine learning models.

Keywords: Cement. Cement Consumption. Brazilian Economy. Machine Learning. Civil Construction. Machine Learning. Neural Networks. Linear Regression. Artificial Intelligence. Forecasting.

Lista de Abreviaturas

IA	Inteligência Artificial
ML	<i>Machine Learning</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
FGV	Fundação Getúlio Vargas
PIB	Produto Interno Bruto
BACEN	Banco Central do Brasil
IPEA	Instituto de Pesquisa Econômica Aplicada
RMSE	<i>root-mean-square error</i>
MAPE	<i>mean absolute percentual error</i>
MAE	<i>mean absolute error</i>
PNAD	Pesquisa Nacional por Amostra de Domicílios
INCC	Índice Nacional de Custo da Construção Civil
IPCA	Índice Nacional de Preços ao Consumidor Amplo
IGP	Índice Geral de Preços
NFSP	Necessidade de Financiamento do Setor Público
IDH	Índice de Desenvolvimento Humano
SNIC	Sindicato Nacional da Indústria do Cimento
LSTM	<i>Long Short Term Memory</i>
GRU	<i>Gated Recurrent Units</i>

Lista de figuras

1.1	Relação entre inteligência artificial, <i>machine learning</i> e aprendizado profundo (PATTERSON e GIBSON 2017, p.4)	5
1.2	Exemplo de um modelo simples de regressão linear. Imagem retirada de HYNDMAN e ATHANASOPOULOS (2021) e disponível em https://otexts.com/fpp3/regression-intro.html	7
1.3	Ilustração de um neurônio biológico (PATTERSON e GIBSON 2017, p.44) . .	8
1.4	Neurônios artificiais	9
1.5	Ilustração de uma rede neural obtida de NIELSEN (2015)	10
1.6	Rectified linear unit (ReLU) (PATTERSON e GIBSON, 2017)	11
1.7	Função <i>swish</i> (RAMACHANDRAN <i>et al.</i> , 2017)	12
1.8	Arquitetura de uma rede <i>feed forward</i>	13
1.9	Arquitetura das redes recorrentes	14
1.10	Célula de memória de rede LSTM (DSA, 2022)	14
1.11	Unidade da rede GRU (DSA, 2022)	15
1.12	Estrutura de rede bidirecional (ALLA, 2021)	16
2.1	Gráfico <i>boxplot</i> do PIB da construção civil construído a partir dos dados disponíveis no portal do IPEA, conforme descrito na tabela 2.1	19
2.2	Comparação dos gráficos <i>boxplot</i> entre as regiões	20
2.3	Matriz de correlação	20
4.1	Evolução do consumo mensal de cimento em São Paulo	32
4.2	Distribuição da variação percentual das previsões da regressão linear . .	33
4.3	Previsão realizada pela regressão linear para o consumo de cimento em São Paulo	33
4.4	Distribuição da variação percentual das previsões das redes <i>feed forward</i> .	34
4.5	Previsão realizada pelas redes <i>feed forward</i> em São Paulo	34
4.6	Distribuição da variação percentual das previsões das redes recorrentes .	35
4.7	Previsão realizada pelas redes recorrentes em São Paulo	35

4.8	Distribuição da variação percentual das previsões das redes bidirecionais	36
4.9	Previsão realizada pelas redes bidirecionais em São Paulo	36

Lista de tabelas

2.1	Indicadores utilizados no trabalho	18
3.1	Experimentos realizados na regressão linear	25
3.2	Desempenho dos modelos de regressão linear	26
3.3	Parâmetros testados nas redes <i>feed forward</i>	26
3.4	Experimentos com melhor desempenho segundo o número de camadas das redes <i>feed forward</i>	27
3.5	Parâmetros testados nas redes recorrentes	28
3.6	Experimentos com melhor desempenho segundo o número de camadas das redes recorrentes	28
3.7	Parâmetros testados nas redes bidirecionais	29
3.8	Experimentos com melhor desempenho das redes bidirecionais	29
4.1	Comparação do desempenho dos modelos	31

Sumário

Introdução	1
Motivação	1
Objetivos	2
1 Fundamentação teórica	3
1.1 Inteligência artificial	3
1.2 Modelos de Aprendizado Automático	5
1.3 Modelos utilizados	6
1.3.1 Regressão linear	6
1.3.2 Redes neurais	7
1.3.3 Redes neurais recorrentes	12
1.3.4 Redes bidirecionais	15
2 Metodologia	17
2.1 Dados	17
2.1.1 Fontes	17
2.1.2 Preparação dos dados	18
2.1.3 Pré-processamento de dados	21
2.2 Avaliação de performance	22
2.2.1 <i>Mean absolute error</i> (MAE)	22
2.2.2 <i>Root mean squared error</i> (RMSE)	22
2.2.3 <i>Mean absolute percentage error</i> (MAPE)	23
2.2.4 Variação percentual	23
2.3 Avaliação do desempenho do modelo	23
2.4 Tecnologias utilizadas	24
3 Experimentos	25
3.1 Regressao linear	25
3.2 Redes <i>feed forward</i>	26

3.3	Redes recorrentes	27
3.4	Redes bidirecionais	28
4	Resultados	31
4.1	Regressão linear	32
4.2	Redes <i>feed forward</i>	34
4.3	Redes recorrentes	35
4.4	Redes bidirecionais	36
5	Discussão	39
6	Conclusão	41
	Referências	43

Introdução

O que prédios, pontes, hidrelétricas e aeroportos têm em comum? Todos são frutos da indústria da construção civil, um importante componente do investimento brasileiro e, consequentemente, uma das grandes engrenagens responsáveis por movimentar a atividade econômica no Brasil. Em 2021, por exemplo, o Produto Interno Bruto (PIB) desse setor teve destaque com alta de 9,7%, enquanto o PIB do Brasil cresceu 4,6% e o do Agronegócio registrou queda de 0,2%. Dessa forma, e em consonância com [VASCONCELOS \(2022\)](#), o setor de construção figura como importante impulsionador da economia do país e como fonte de renda e de empregos.

Tendo isso em vista, o cimento¹, que é uma das principais matérias primas da indústria da construção civil, é um pó com propriedades aglomerantes que endurece quando é submetido à água. Após endurecer, não é mais decomposto, mesmo em contato com a água, além disso, ao ser misturado com "[...]areia, a pedra britada, o pó-de-pedra, a cal"[ABCP \(2002\)](#), este cimento se torna argamassa ou concreto, utilizados em construções.²

Motivação

É necessário considerar que a ampliação da produção de cimento representa para a empresa um alto investimento de recursos financeiros e de tempo a longo prazo, um exemplo que ilustra essa necessidade é o da fábrica construída pela Votorantim Cimentos no Complexo Industrial e Portuário do Pecém, no Ceará, cuja construção demorou três anos e custou cerca de 200 milhões de reais, conforme noticiado no portal do Instituto Brasileiro de Mineração³. Além disso, o cimento é um produto que não suporta longos períodos de tempo em estoque, visto que a norma brasileira recomenda o uso em até 90 dias após a fabricação, o que reforça o fato de que a produção do material deva estar fortemente alinhada com o seu consumo⁴.

Tendo em vista o que foi falado, um modelo que permitisse prever a demanda por

¹ O material conhecido na construção civil como cimento é denominado mundialmente como cimento de *portland*. Em 1824, construções com pedra de *portland* eram comuns na Inglaterra, por isso, o inventor do cimento, Joseph Aspdin, ao notar que sua invenção tinha aspecto similar ao material tão difundido na época, optou por registrar a patente como "cimento de *portland*"([ABCP, 2002](#)).

² As informações técnicas sobre o processo de fabricação do cimento foram retiradas de [ABCP, 2002](#).

³ As informações sobre a construção dessa fábrica pela Votorantim foram retiradas de [IBRAM, 2021](#).

⁴ As informações sobre o período adequado de armazenamento de cimento foram retiradas de [NORMAS TÉCNICAS \(2018\)](#).

cimento em cada estado brasileiro poderia auxiliar gestores a tomar melhores decisões e a estruturar as suas estratégias administrativas de forma mais embasada, reduzindo os riscos do setor. Com isso, seria possível apoiar órgãos governamentais a direcionar ações para mitigar o impacto ambiental advindo da fabricação desse produto⁵.

Objetivos

Este trabalho propõe o uso de modelos de *machine learning* para comprovar a possibilidade de prever o consumo mensal de cimento nos estados brasileiros e, também, discorrer mais especificamente sobre as previsões realizadas para o estado de São Paulo. Isso de forma a propor uma solução para a necessidade da indústria de prever a demanda do material e investigar o desempenho de diferentes modelos para determinar qual é o mais eficiente. Os algoritmos avaliados em ordem crescente de complexidade e robustez são: regressão linear e redes neurais *feed forward*, recorrentes e bidirecionais.

⁵ É importante ter em mente o que foi pontuado por CARVALHO (2008), de que a fabricação do cimento é uma grande emissora de gases de efeito-estufa em escala mundial

Capítulo 1

Fundamentação teórica

1.1 Inteligência artificial

Atualmente, a inteligência artificial (IA) permeia diversos momentos do cotidiano, um exemplo é a empresa norte-americana de *streaming* Netflix, que utiliza um conjunto de técnicas de inteligência artificial para recomendar conteúdo personalizado aos usuários da plataforma, de acordo com os interesses particulares de cada um.

Considerando o que diz [STECK et al. \(2021\)](#), especificamente sobre a Netflix, não há um modelo ou algoritmo único utilizado para todas as recomendações de conteúdo, essa tarefa é dividida em subtarefas realizadas por diferentes modelos, isso de acordo com a atividade a ser realizada e os dados disponíveis. Por exemplo, a escolha de qual vídeo será exibido para o usuário ao logar no perfil da plataforma é executada por um modelo diferente do que o que elenca os vídeos já assistidos que o membro pode continuar a ver.

Ainda no que se refere à Netflix, a empresa proporciona uma experiência única a cada indivíduo que acessa a plataforma. Essa estratégia tem o objetivo de aumentar a satisfação a longo prazo dos clientes e garantir a permanência dos assinantes, uma vez que a plataforma é monetizada com assinaturas mensais. Por fim, ressalta-se que a estratégia de utilizar inteligência artificial para as recomendações tem se refletido ao longo dos anos em uma melhora na taxa de retenção dos usuários, diante disso:

[...]the value of a recommender system can be measured by the increase in member retention. Over years of the development of personalization and recommendation technologies, we have been able to repeatedly create meaningful improvements in retention ([STECK et al., 2021](#)).

Com base no que até agora foi exposto, o que seria inteligência artificial? O termo, do inglês *artificial intelligence*, foi elaborado por John McCarthy e utilizado oficialmente pela primeira vez em 1956 no seminário de Dartmouth, um *workshop* sobre a área que reuniu os maiores estudiosos do ramo durante dois meses, de acordo com [RUSSELL e NORVIG, 1995](#). Em 1950, entretanto, Alan Turing já se perguntava se máquinas poderiam pensar e desenvolvia estudos e conceitos sobre o tema que permanecem relevantes, como o Teste

de Turing¹. O termo *artificial intelligence* pode ser utilizado com várias conotações, uma vez que não apresenta uma definição única e aceita². Uma possível definição do termo é a conferida por McCarthy: "The science and engineering of making intelligent machines, especially intelligent computer programs. [...] Intelligence is the computational part of the ability to achieve goals in the world"(McCarthy 2007, p.2)³.

Conforme relembra PATTERSON e GIBSON (2017), o pioneiro em IA, Arthur Samuel⁴, descreveu em 1959 *machine learning* como o setor de pesquisa que proporciona aos computadores a capacidade realizar tarefas sem serem diretamente programados para isso. *Machine learning*, então, configura-se como uma sub-área da inteligência artificial e compreende sistemas capazes de adquirir seu próprio conhecimento a partir de dados, como explica GOODFELLOW *et al.* (2016).

A área de *machine learning* ainda apresenta uma categoria particular de modelos que realizam um processamento de dados inspirado no funcionamento do cérebro humano, as redes neurais. De acordo com DSA (2022), esses algoritmos são conhecidos como *deep learning* e repassam a informação por meio de neurônios matemáticos de camada em camada até obter a resposta final.

A relação entre inteligência artificial, aprendizado de máquina e aprendizado profundo pode ser vista na imagem 1.1:

¹ Em 1950, Alan Turing propôs o Teste de Turing com a intenção de determinar se uma máquina era inteligente. Esse teste é uma variação do Jogo da Imitação, em que um entrevistador deve fazer perguntas a dois jogadores, um humano e uma máquina, sem qualquer distinção. Ao final, o entrevistador deve descobrir qual dos jogadores é uma máquina e qual é a pessoa. Se a máquina fosse capaz de enganar o entrevistador, seria considerada inteligente.(TURING, 1950)

² Para um maior aprofundamento sobre a questão, WANG, Pei. On defining artificial intelligence. **Journal of Artificial General Intelligence**. Vol. 10, n. 2, p. 1-37, 2019. DOI: doi:10.2478/jagi-2019-0002. Disponível em: <https://sciendo.com/article/10.2478/jagi-2019-0002>.

³ "A ciência e engenharia de construir máquinas inteligentes, em especial programas de computador. [...] Inteligência é o aspecto computacional da habilidade de atingir os objetivos no mundo (McCarthy 2007, p.2, tradução nossa).

⁴ Arthur Samuel foi um engenheiro e um dos pioneiros em inteligência artificial, tendo desenvolvido um programa que jogava damas com humanos e aprendia com cada jogada dos oponentes. PATTERSON e GIBSON (2017) ainda explicam que o programa tornava suas jogadas mais assertivas ao calcular as probabilidades de cada jogada e é considerado uma das primeiras aplicações de *machine learning*.

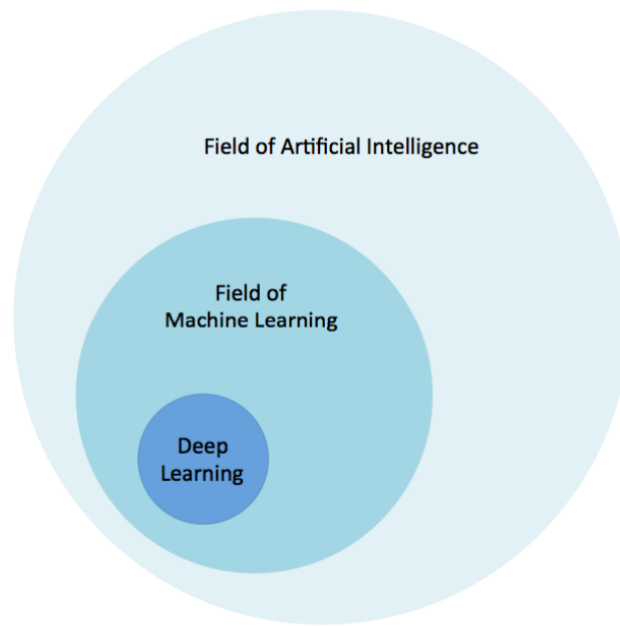


Figura 1.1: Relação entre inteligência artificial, machine learning e aprendizado profundo (PATTERSON e GIBSON 2017, p.4)

1.2 Modelos de Aprendizado Automático

Os modelos de *machine learning* podem realizar várias categorias de tarefas, dentre elas a regressão e a classificação, a depender da atividade realizada pelo algoritmo. Os modelos de regressão têm como objetivo é prever um valor real a partir dos dados de entrada, segundo GOODFELLOW *et al.* (2016). É possível, então, descrever cada entrada (*input*) dos modelos como um vetor x com n atributos (*features*)⁵ tal que $x \in \mathbb{R}^n$, $x = \{x_1, x_2, \dots, x_n\}$ e o processamento realizado pelos modelos de regressão com a função⁶ $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

Os modelos de classificação, doutro modo, têm como objetivo determinar a qual das k categorias disponíveis um *input* pertence, conforme explica PATTERSON e GIBSON (2017). Então, é utilizada uma função $f : \mathbb{R}^n \rightarrow \{1, \dots, k\}$, quando $f(x) = y$, o vetor de entrada x foi classificado na categoria y . Um exemplo de tarefa de classificação seria determinar se uma operação com cartão de crédito é fraudulenta ou não; neste trabalho, contudo, não foi empregado tal tipo de modelo, apenas modelos de regressão.

Ainda conforme GOODFELLOW *et al.* (2016), os problemas de *machine learning* também podem ser divididos entre aprendizado não-supervisionado e supervisionado. No primeiro, o modelo recebe um conjunto de dados (*dataset*) não rotulado e aprende propriedades de sua estrutura. Um exemplo de aprendizado não supervisionado é o *clustering*, que consiste em dividir o conjunto de dados em agrupamentos (*clusters*) com amostras similares. No segundo, por outro lado, os dados de entrada estão associados a resultados conhecidos,

⁵ Neste estudo, os atributos utilizadas são indicadores econômicos, monetários, sociais e da construção civil descritos na seção 2.1.

⁶ As equações apresentadas neste subcapítulo foram retiradas de GOODFELLOW *et al.*, 2016.

chamados de *labels* ou rótulos. Neste estudo, utiliza-se aprendizado supervisionado para prever o consumo de cimento mensal nos estados da União a partir dos dados de entrada e compará-los com o valor real do consumo e, assim, calcular a precisão do modelo.

1.3 Modelos utilizados

Para a produção desta pesquisa, foram utilizadas três categorias de modelos de *machine learning* para prever a demanda por cimento: regressão linear, redes neurais *feed forward* e redes recorrentes. Foram testados, também, métodos de pré-processamento de dados⁷ e diferentes arquiteturas de redes neurais ao alterar o número de camadas, a quantidade de neurônios em cada camada, o tipo de camada, a função de ativação, entre outras configurações, tudo isso com o objetivo de comparar o desempenho dos modelos e encontrar o que apresenta menor erro na previsão.

1.3.1 Regressão linear

A regressão linear é um modelo de *machine learning* que assume um relacionamento linear entre a variável que será prevista (*target*) e os dados de entrada, conforme explica [HYNDMAN e ATHANASOPOULOS \(2021\)](#). Desse modo, o intuito é construir uma função que, para cada par⁸ (x, y) , recebe como entrada o vetor x correspondente às variáveis de *input*, $x \in \mathbb{R}^k$, $x = \{x_1, x_2, \dots, x_k\}$ e calcula coeficientes $\beta = \{\beta_0, \beta_1, \dots, \beta_k\}$ para cada atributo de x , além da constante β_0 . O algoritmo, então, utiliza esses coeficientes para prever o valor da variável *target*, y , sendo que $y \in \mathbb{R}$. Seja, então, \hat{y} o valor previsto pelo modelo para um par (x, y) , a função que descreve a regressão linear, então, é dada pela equação 1.1:

$$\hat{y} = f(x) = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \dots + \beta_k x_{k_i} \quad (1.1)$$

Logo, é possível contruir uma matriz X em que a linha i corresponde ao vetor x_i dos dados de entrada e cada coluna j representa uma *feature*. Pode-se, além disso, construir a matriz β dos coeficientes associados a cada elemento da matriz X . Assim, o modelo é dado por:

$$y = X\beta + \epsilon \quad (1.2)$$

Na equação 1.2, ϵ é o vetor com o erro associado a cada umas das previsões; lembrando que ϵ_i é a i -ésima entrada do vetor ϵ , assim como $\epsilon_i = y_i - \hat{y}_i = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$. Os coeficientes β são calculados durante o treinamento do modelo utilizando o método do *least squares estimation*, que visa minimizar a soma do erro quadrado associado às previsões, como descrito em:

⁷ Os métodos de pré-processamento de dados são descritos na seção 2.1.3.

⁸ No caso deste estudo, o par (x, y) é tal que x representa o valor de cada indicador descrito na seção 2.1 em um estado, mês e ano, enquanto y corresponde ao número de toneladas de cimento consumidas por esse estado nessa data.

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}))^2 \quad (1.3)$$

Os coeficientes $\beta_1, \beta_2, \dots, \beta_k$ determinam como cada atributo de entrada afeta a previsão, no presente caso, o consumo de cimento. Por exemplo, se o coeficiente da variável x_i for $\beta_i = 0$, essa variável não terá influência no valor previsto pelo modelo. Caso o coeficiente seja positivo um aumento no valor de x_i resultará no aumento do valor previsto \hat{y}_i , já se β_i for negativo, um aumento no valor de x_i refletirá na diminuição no valor de \hat{y}_i .

Na imagem 1.2, há um exemplo de um modelo de regressão linear com apenas uma variável:

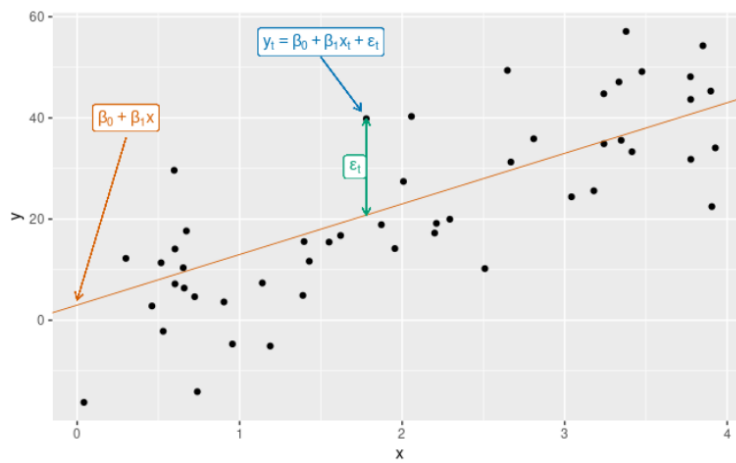


Figura 1.2: Exemplo de um modelo simples de regressão linear. Imagem retirada de *HYNDMAN e ATHANASOPOULOS (2021)* e disponível em <https://otexts.com/fpp3/regression-intro.html>

Na figura 1.2, as observações y_i estão representadas pelos pontos pretos, enquanto a linha em laranja corresponde à previsão realizada pelo modelo. Observa-se que o modelo não prevê com total exatidão os dados observados, isso porque há um erro associado a cada previsão — o que é um comportamento esperado, já que fenômenos previstos são sujeitos a fatores externos e não lineares —, como o destacado em verde na ilustração.

Por se tratar de um modelo mais simples, os resultados obtidos com a regressão linear são utilizados neste trabalho como base para comparar o desempenho de modelos mais robustos.

1.3.2 Redes neurais

Redes neurais são modelos de *machine learning* inspirados no funcionamento do cérebro animal, formado por neurônios que se conectam para transmitir informações sem a necessidade de controle central, segundo *DSA (2022)* e *PATTERSON e GIBSON (2017)*. Um neurônio biológico, então, é uma célula nervosa que se comunica com outros neurônios por meio de impulsos eletroquímicos. Essa comunicação é denominada sinapse e ocorre

apenas se o impulso for forte o bastante para ativar a liberação de químicos na fenda sináptica.

Um neurônio biológico é composto por vários dendritos, um axônio e um corpo celular, como ilustrado na figura 1.3:

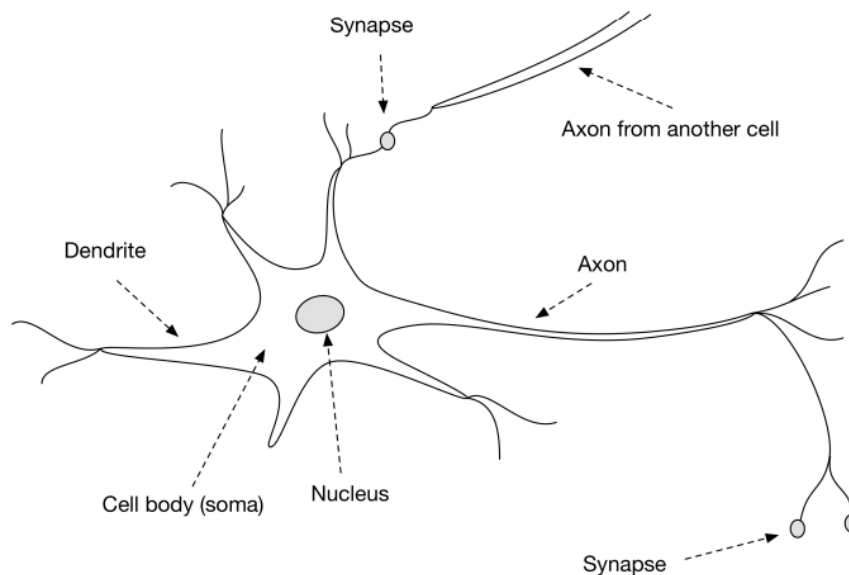


Figura 1.3: Ilustração de um neurônio biológico (PATTERSON e GIBSON 2017, p.44)

Destaca-se na figura 1.3 a informação chegando a um dendrito do neurônio em destaque por meio de uma sinapse, também está representada a comunicação com outra célula por meio de outra sinapse iniciada no axônio do neurônio em questão.

O processo de propagação da informação nos neurônios envolve as três partes da célula: dendritos, corpo celular e axônio. Os dendritos recebem informações de neurônios vizinhos na forma de impulsos elétricos e são responsáveis por conduzi-las até o corpo celular. Ao chegar ao local, a informação é processada, e novos impulsos são gerados e repassados a outro neurônio por meio do axônio no processo de sinapse, segundo NICHOLAS LACASCIO (2017) e DSA (2022). A estrutura e funcionamento dos neurônios biológicos inspiraram os cientistas ao projetarem neurônios artificiais, como os *perceptrons*.

Neurônios artificiais

O *perceptron* foi desenvolvido em 1957 por Frank Rosenblatt, inspirado nos trabalhos de Warren McCulloch e Walter Pitts⁹. Trata-se de um modelo linear de classificação que recebe n entradas e produz uma saída binária, como mostrado na ilustração simplificada 1.4a. Ainda de acordo com DSA (2022), esse modelo inicial apresentava limitações e passou por evoluções com o passar do tempo. Apesar disso, as redes neurais atualmente utilizam, em geral, outro modelo de neurônio, como o ilustrado na imagem 1.4b:

⁹ Em 1943, Warren McCulloch e Walter Pitts apresentaram a primeira ideia de neurônio artificial. (WALTER PITTS, 1943)

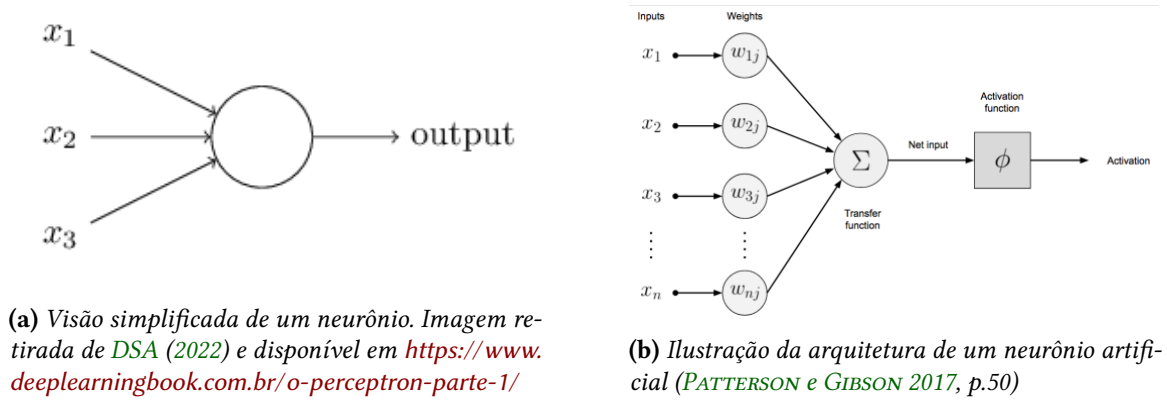


Figura 1.4: Neurônios artificiais

A figura 1.4b mostra o funcionamento de um neurônio artificial. O neurônio apresenta n entradas x_i , cada uma associada a um peso w_i , que expressa a importância da respectiva entrada para o valor de saída¹⁰, produto escalar entre os pesos e as respectivas entradas passa por uma função de ativação¹¹ ϕ que determina a saída do neurônio. Além disso, ainda segundo DSA (2022) e PATTERSON e GIBSON (2017), um valor de *bias* ou polarização é adicionado ao produto escalar e possibilita que um neurônio com todas as entradas nulas apresente saída não nula, de modo a aumentar a capacidade de aproximação da rede.

Seja x o vetor das n entradas do neurônio, $x \in \mathbb{R}^n$, $x = \{x_1, x_2, \dots, x_n\}$, e seja w o vetor com os pesos associados a cada entrada, $w \in \mathbb{R}^n$, $w = \{w_1, w_2, \dots, w_n\}$. Além disso, seja b o valor de *bias* e ϕ a função de ativação. Dessa forma, o *output* h de um neurônio é dado por:

$$h_{w,b} = \phi(w \cdot x + b) \quad (1.4)$$

Essa saída é utilizada como uma das entradas dos neurônios na camada seguinte, de modo a formar a estrutura das redes neurais.

Os neurônios, então, formam a unidade que compõe as redes neurais artificiais, como ilustrado na figura 1.5:

¹⁰ O peso atribuído a uma das entradas expressa a influência desta no *output* do nó, de modo similar aos coeficientes calculados para cada atributo da regressão linear, explicada na seção 1.3.1

¹¹ Um detalhamento sobre as funções de ativação pode ser encontrado na subseção 1.3.2

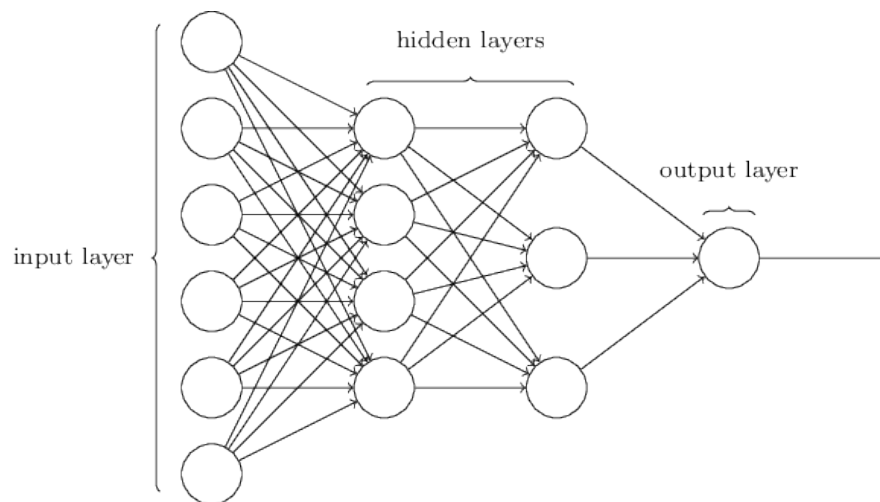


Figura 1.5: Ilustração de uma rede neural obtida de [NIELSEN \(2015\)](#)

A imagem 1.5 representa uma simplificação de uma rede neural. A informação chega a uma camada da rede, é processada pelos neurônios e repassada para a camada seguinte, até obter a saída final. Em particular, é representada uma rede *feed forward*, uma vez que não há conexões entre neurônios de uma mesma camada nem conexões entre uma camada e a anterior. A informação então se propaga apenas no sentido da camada de entrada em direção à final, de saída.

Além disso, em 1.5, é ilustrada uma rede neural *multilayer*. Conforme a figura, a camada mais à esquerda da rede é denominada *input layer*, a camada mais à direita, de *output layer*, e as camadas intermediárias são chamadas *hidden layers*. Por razões históricas, de acordo com [NIELSEN \(2015\)](#), é possível encontrar referências a essas redes como *multilayer perceptron*, contudo, as redes em geral utilizam neurônios sigmóides ao invés de *perceptrons*.

Função de ativação

A função de ativação determina se um neurônio será ativado, ou seja, se a saída será propagada para a camada seguinte. Enquanto os pesos e o *bias* realizam uma transformação linear nos dados de entrada, a função de ativação aplica uma transformação não linear, assim, possibilita que a rede neural resolva problemas não lineares e complexos, como reconhecer padrões de escrita, de acordo com [DSA \(2022\)](#) e [ZHANG et al. \(2021b\)](#).

A função de ativação é um atributo de cada uma das camadas da rede e é escolhida de acordo com a tarefa que será executada. Por exemplo, a função sigmóide é recomendada para problemas de classificação. Neste estudo, foram utilizadas as funções *rectified linear unit* (ReLU) e *swish*.

Rectified linear unit (ReLU)

A função ReLU, do inglês *rectified linear unit*, é a função de ativação mais popular atualmente, uma vez que apresenta bom desempenho em diferentes tarefas, conforme [PATTERSON e GIBSON \(2017\)](#). A função ReLU é dada por:

$$f(x) = \max(0, x) \quad (1.5)$$

Ainda segundo PATTERSON e GIBSON (2017), por possuir derivada igual a zero ou a uma constante, a ReLU não sofre do problema da dissipação do gradiente¹² como outras funções de ativação, uma vez que a derivada é utilizada nos modelos de *machine learning* para atualizar os pesos e *bias* no treinamento da rede. O gráfico da função ReLU está ilustrado na figura 1.6:

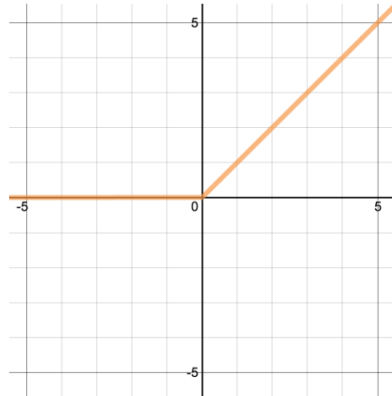


Figura 1.6: Rectified linear unit (ReLU) (PATTERSON e GIBSON, 2017)

Swish

A função *swish* foi proposta por pesquisadores da Google em 2017, com a promessa de apresentar desempenho igual ou superior à ReLU em redes neurais profundas, conforme RAMACHANDRAN *et al.*, 2017. A *swish* é uma função não monótona¹³ e suave¹⁴, cuja fórmula é dada por:

$$f(x) = x \cdot \text{sigmoid}(x) = \frac{x}{1 + e^{-x}} \quad (1.6)$$

O gráfico da função é similar ao da ReLU e pode ser observado na imagem 1.7:

¹² Um detalhamento sobre o problema da dissipação do gradiente será dado na subseção 1.3.2.

¹³ Uma função definida como $f : I \rightarrow \mathbb{R}$ é monótona se é não crescente no intervalo I ou não decrescente em I .

¹⁴ Uma função f é suave se possui derivada de todas as ordens.

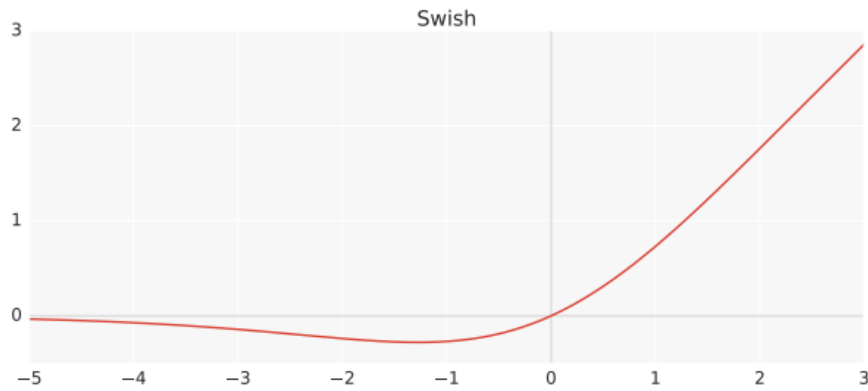


Figura 1.7: Função swish (RAMACHANDRAN *et al.*, 2017)

Problema da Dissipação do Gradiente

O aprendizado de uma rede neural ocorre ao alterar os valores dos pesos e *bias* dos neurônios de acordo com a função de custo ou *loss function* (NIELSEN, 2015). Sejam x , w e b os conjuntos de entradas, pesos e *bias* da rede, respectivamente. Então, a função de custo C é dada por:

$$C(w, b) = \frac{1}{2n} \sum_x \|y(x) - a\|^2 \quad (1.7)$$

Na equação acima, $y(x)$ representa o vetor com as respostas que a rede tenta prever, a é o vetor de *output* da rede e n é o número total de entradas de treino. Assim, é possível calcular o gradiente δ_j^i relativo ao j -ésimo neurônio da i -ésima camada:

$$\delta_j^i = \frac{\partial C}{\partial b_j^i} \quad (1.8)$$

Seja δ^i o vetor cujos elementos estão associados aos neurônios da camada i . Como os gradientes definem a alteração nos pesos e *bias* de cada neurônio, δ^i determina a velocidade de aprendizado dessa camada.

O problema da dissipação do gradiente¹⁵ ocorre quando os gradientes das camadas iniciais ficam com valores muito próximos a zero e atinge, em especial, modelos de *deep learning* com muitas camadas e redes recorrente. Por conta disso, os pesos e *bias* não são atualizados de forma eficiente nessas camadas, de tal forma que o treinamento e aprendizado tornam-se demasiadamente lentos.

1.3.3 Redes neurais recorrentes

Redes recorrentes são uma subclasse das redes neurais e possuem a capacidade de capturar o contexto como diferencial frente às redes *feed forward* tradicionais. Por levar

¹⁵ O problema da dissipação do gradiente é conhecido em inglês como *the vanishing gradient problem*.

em consideração o presente e o passado recente ao realizar uma previsão, apresentam bom desempenho na previsão de séries temporais ou sequências. Há várias implementações possíveis de redes recorrentes. Neste trabalho testaram-se as redes LSTM e GRU.

Redes LSTM

As redes *Long Short Term Memory* (LSTM) foram introduzidas em 1997 por Hochreiter e Schmidhuber¹⁶. Atualmente, são a variação mais utilizada de redes neurais recorrentes, em particular, na classificação de séries temporais e reconhecimento de fala. Segundo DSA (2022), redes LSTM apresentam células de memória conectadas cujo conteúdo é modulado pelo *input gate* e o *forget gate*¹⁷. Por exemplo, se ambos estiverem fechados em um instante, o conteúdo da célula permanecerá o mesmo no instante em questão e no próximo. Essa estrutura permite que a informação seja retida ao longo do tempo e evita o problema do *vanishing gradient* que ocorre com a maioria das redes recorrentes.

As redes neurais LSTM apresentam uma arquitetura diferente das redes tradicionais, uma vez que há ciclos de *feedback* nas conexões entre as células. Para melhor ilustrar essa arquitetura, PATTERSON e GIBSON, 2017 utilizam a visualização *flat* (achatada) das redes neurais, na qual as células de uma mesma camada da rede estão representadas como um único nó. Pode-se observar essa representação em uma rede tradicional *feed forward* na imagem 1.8b e, em 1.8a, a representação tradicional:

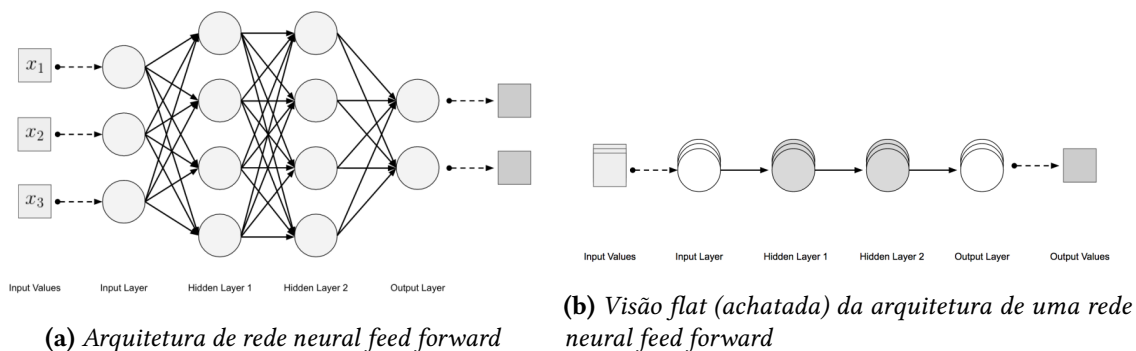


Figura 1.8: Arquitetura de uma rede feed forward

As redes LSTM introduzem o conceito de conexão entre a saída de uma camada oculta da rede neural e a entrada da mesma camada. A partir desse ciclo, obtém-se *outputs* de tempos anteriores como parte da informação que chega ao tempo atual. Na figura 1.9a, essas conexões recorrentes são representadas como as setas que saem de uma célula e atingem a mesma célula, uma vez que se utiliza a representação achatada. A imagem 1.9b, por sua vez, representa a rede LSTM desenrolada através do eixo do tempo:

¹⁶ As redes neurais LSTM foram apresentadas no artigo HOCHREITER e SCHMIDHUBER, 1997.

¹⁷ Um melhor detalhamento sobre o *input gate* e o *forget gate* será dado mais adiante nesta seção.

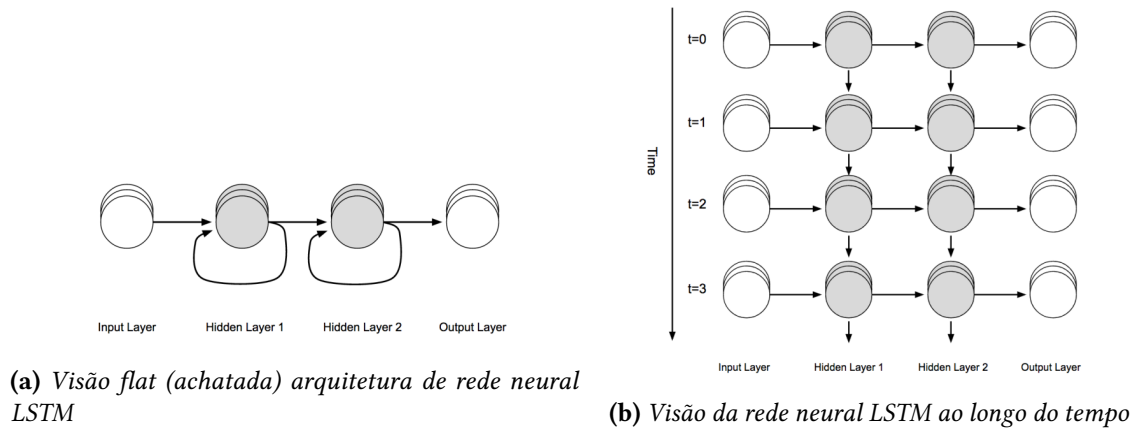


Figura 1.9: Arquitetura das redes recorrentes

As unidades que formam cada uma das camadas de uma LSTM são uma variação dos neurônios artificiais clássicos. Essas unidades, representadas em 1.10, permitem que a rede mantenha o estado ao longo do tempo e apresentem conexões vindas da camada anterior e de *outputs* dessa mesma unidade em tempos anteriores:

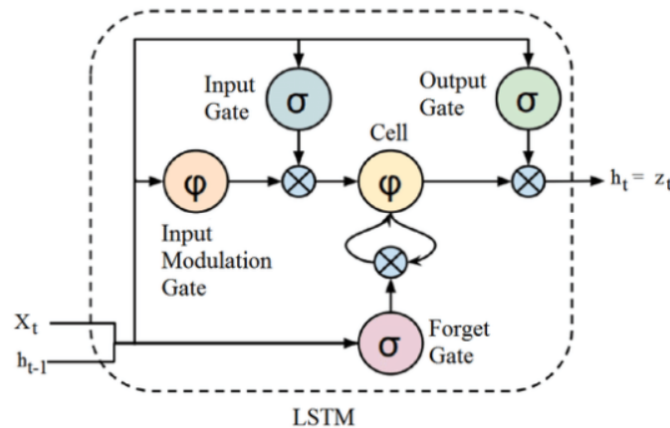


Figura 1.10: Célula de memória de rede LSTM (DSA, 2022)

A unidade LSTM recebe duas entradas: a observação no tempo atual e a saída do último estado oculto. Nela, a informação é retida nas células e as manipulações de memória são realizadas nos *gates* (portões). O *forget gate* é responsável por remover informações que não são mais úteis para a célula, ou seja, pelo esquecimento. Essa operação é realizada por meio de multiplicação de matrizes de pesos, adição de parâmetro de viés e aplicação de uma função de ativação que fornece uma saída binária. O *input gate*, por sua vez, adiciona informações úteis ao estado da célula a partir da aplicação de funções sigmóides e tangente hiperbólica, além da multiplicação de vetores. Por fim, o *output gate* extrai informações úteis ao estado da célula atual para apresentar como saída da célula e entrada para a seguinte, com o auxílio de funções tangente hiperbólica e sigmoide, além da multiplicação de vetores. (DSA, 2022)

Redes GRU

As redes *gated recurrent units* são redes recorrentes similares às LSTM e foram introduzidas por (CHO *et al.*, 2014). As unidades das redes LSTM apresentam dois estados passados entre as células: um que carrega memória de longo prazo e outro, de curto prazo. As unidades GRU, por sua vez, apresentam apenas um estado oculto carregado ao longo do tempo, capaz de manter as dependências de curto e longo prazo (DSA, 2022). A estrutura das unidades GRU pode ser vista em 1.11:

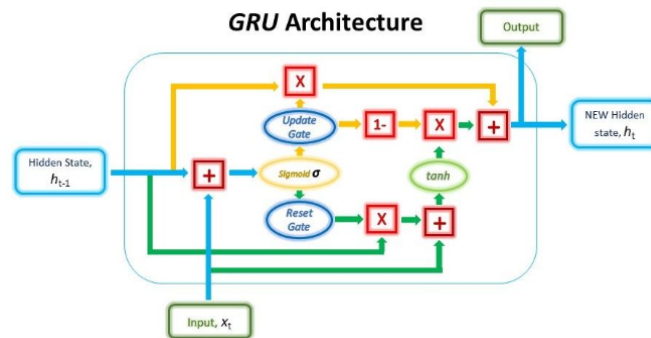


Figura 1.11: Unidade da rede GRU (DSA, 2022)

As unidades GRU apresentam portões, assim como as unidades LSTM, para controlar o fluxo de informações. O *reset gate*, ou portão de redefinição, é responsável por controlar quais informações das etapas anteriores serão mantidas na etapa atual, por meio da função sigmoide e multiplicação de vetores. O *update gate*, por sua vez, tem como objetivo determinar quanto das informações armazenadas no estado oculto atual são apenas cópias do estado anterior. (ZHANG *et al.*, 2021a)

Por possuírem arquitetura mais simples, as redes GRU apresentam um treinamento mais rápido que as LSTM.

1.3.4 Redes bidirecionais

As redes bidirecionais foram introduzidas em 1997 por Schuster e Paliwal com a proposta de utilizar uma entrada para influenciar ao mesmo tempo as previsões do futuro e do passado (SCHUSTER e PALIWAL, 1997). Redes recorrentes, como as LSTM e GRU, analisam uma entrada com base nos eventos passados e presentes. As redes bidirecionais, por outro lado, inovam ao considerar o passado, presente e o futuro ao realizar uma previsão. Por isso, apresentam bom desempenho em atividades como reconhecimento de escrita a mão, reconhecimento de fala, processamento de linguagem natural. (ALLA, 2021)

As redes bidirecionais combinam duas redes recorrentes para realizar a previsão. Uma rede analisa a sequência $x^{(i)}$ fornecida no *input* do início x_0 em direção ao final x_i , a outra é treinada no sentido oposto. As redes utilizadas podem ser simples redes recorrentes, redes LSTM ou GRU. Na figura 1.12, pode-se observar o treinamento das redes nos dois sentidos ocorrendo simultaneamente:



Figura 1.12: Estrutura de rede bidirecional (ALLA, 2021)

Capítulo 2

Metodologia

Nesta seção serão detalhadas as fontes dos dados utilizados no projeto e as estratégias de preparação e pré-processamento de dados. Serão descritas, ainda, as tecnologias utilizadas na implementação, as métricas estatísticas usadas para mensurar o desempenho dos modelos e o método de avaliação dos modelos.

2.1 Dados

Os algoritmos de *machine learning* utilizam dados para fazer previsões. Neste trabalho, optou-se por utilizar dados de 2003 até 2019 para treinamento e avaliação dos modelos, em virtude da disponibilidade dos dados de consumo mensal de cimento. Além disso, adotou-se granularidade mensal e por estados de dados com o intuito de aumentar a quantidade de entradas disponíveis para treinamento e avaliação dos modelos. Essa estratégia de aumento de dados, de acordo com [GOODFELLOW *et al.* \(2016\)](#), visa diminuir o risco de *underfitting*, que acontece quando o modelo não é capaz de aprender com os dados e resulta em altos erros nas etapas de treinamento e teste.

2.1.1 Fontes

O modelo utiliza dados econômicos, sociais e da construção civil para estimar a demanda por cimento. Na tabela [2.1](#), são apresentados os dados utilizados, juntamente com a fonte, a granularidade e o período em que estavam disponíveis:

Dado	Fonte	Período disponível	Granularidade
PIB a preços constantes	IBGE ^a	1983 até 2019	anual por estado
PIB a preços de mercado	IBGE ^a	1985 até 2019	anual por estado
PIB <i>per capita</i>	IBGE ^a	1985 até 2019	anual por estado
PIB da construção civil	IBGE ^a	1985 até 2019	anual por estado
Desemprego	IBGE ^a	1991 até 2022	irregular ^b
IPCA	IBGE ^c	1981 até 2021	mensal para o Brasil
INCC	FGV ^d	1980 até 2021	mensal para o Brasil
IGP	FGV ^a	1944 até 2021	mensal para o Brasil
Taxa Selic	IBGE ^e	1986 até 2022	mensal para o Brasil
NFSP	BACEN ^a	1991 até 2022	mensal para o Brasil
Estoque líquido de capital fixo	IPEA ^a	1947 até 2019	anual para o Brasil
População	IBGE ^f	1991 até 2021	anual por estado
IDH	IBGE ^a	1991 até 2017	irregular ^g
Produção mensal de cimento	SNIC ^h	2003 até 2022	mensal por estado
Valor médio do cimento ⁱ	SNIC ^h	1947 até 2019	anual para o Brasil
Consumo de cimento em ton.	SNIC ^h	2003 até 2019	mensal por estado

Tabela 2.1: Indicadores utilizados no trabalho

^a Dado retirado do portal do Ipeadata em <http://www.ipeadata.gov.br/Default.aspx>

^b Havia dados de 1992 até 2014 com granularidade anual e por estado. A partir de 2012, foram disponibilizados dados mensais a nível de Brasil por conta da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) realizada pelo IBGE. Neste trabalho, utilizou-se os dados anuais até 2012 e, após 2012, os dados provenientes da PNAD Contínua.

^c Dado retirado do IBGE em <https://sidra.ibge.gov.br/tabela/1737>

^d Dado obtido a partir do portal da FGV em <https://www.debit.com.br/tabelas/tabela-completa-pdf.php?indice=incc>

^e Dado obtido em <https://www.debit.com.br/tabelas/tabela-completa.php?indice=selic>

^f Dado obtido do portal Base dos Dados em <https://basedosdados.org/dataset/br-ibge-populacao>

^g Os indicadores de IDH (Renda, Longevidade e Educação) estão disponíveis a nível de estado da União em anos de censo do IBGE (1990, 2000, 2010). Há dados, também, de 2014 a 2017 por conta da PNAD Contínua.

^h Dados retirados do portal <http://www.cbicdados.com.br/menu/materiais-de-construcao/cimento>

ⁱ Evolução do valor médio/mediano do cimento Portland 32 em US\$/Tonelada

Na tabela, são utilizadas siglas para melhorar a legibilidade; uma descrição das abreviações utilizadas no trabalho pode ser encontrada no início do documento, mais especificamente na Lista de Abreviaturas.

2.1.2 Preparação dos dados

Com o intuito de direcionar a estratégia de preparação de dados, foi realizada uma análise exploratória dos dados de entrada e da variável resposta. Na análise em questão, foi identificada uma alta taxa de variação nos atributos, sendo possível a observância de um grande número de *outliers* presente nos dados. De acordo com HOAGLIN (2013) e TUKEY (1977), *outliers* são observações discrepantes do restante dos dados que podem interferir

no processo de previsão.

Um exemplo de atributo com alta variação é o PIB da construção civil, que apresenta desvio padrão maior que o valor médio dessa variável. Para ilustrar a distribuição dos pontos de dados desse indicador, foi utilizado um gráfico *boxplot* na figura 2.1:

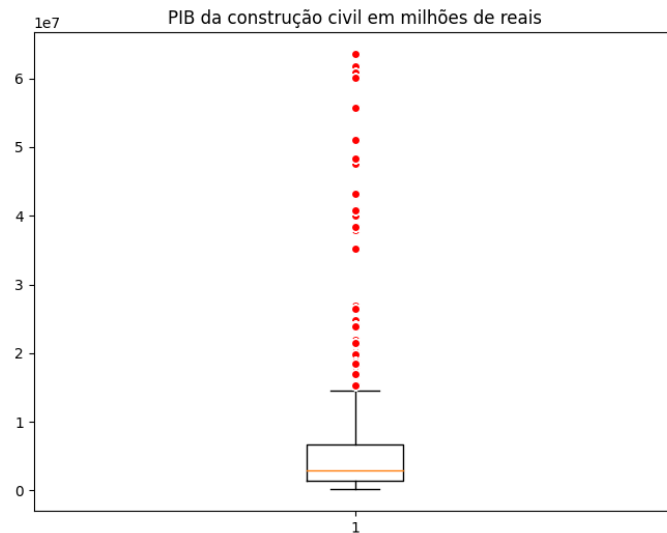


Figura 2.1: Gráfico *boxplot* do PIB da construção civil construído a partir dos dados disponíveis no portal do IPEA, conforme descrito na tabela 2.1

Segundo WILLIAMSON *et al.* (1989), o *boxplot* é uma técnica estatística utilizada para identificar visualmente padrões nos dados. Na figura acima, a linha em laranja corresponde à mediana¹ dos dados, os limite inferior e o superior do retângulo representam o primeiro e terceiro quartis², respectivamente. As observações fora do intervalo dos limites superior e inferior, representadas com círculos vermelhos na figura, são *outliers*. Pode-se observar, então, a alta incidência de dados discrepantes nesse indicador.

Parte do alto volume de *outliers* nos atributos se deve às diferenças econômicas, geográficas e sociais entre os estados do Brasil. A saber, o valor médio do PIB da construção civil no estado de São Paulo é de 48.96 milhões, enquanto em Santa Catarina é de 7.1 milhões e em Roraima, de 0.4 milhões. Aliás, ao analisar o PIB da construção civil nas regiões do país, há uma significativa redução no número de dados discrepantes, como é possível pode observar na figura abaixo:

¹ Mediana é valor que fica no meio quando os dados estão ordenados ou a média dos dois valores centrais se o número de pontos de dados for par. (ESSELMAN, 2022)

² O primeiro quartil marca a mediana relativa aos valores superiores à mediana destacada na imagem. O terceiro quartil, de maneira análoga, assinala a mediana dos valores inferiores à mediana destacada. Assim, entre o primeiro e terceiro quartis está contida metade dos dados.

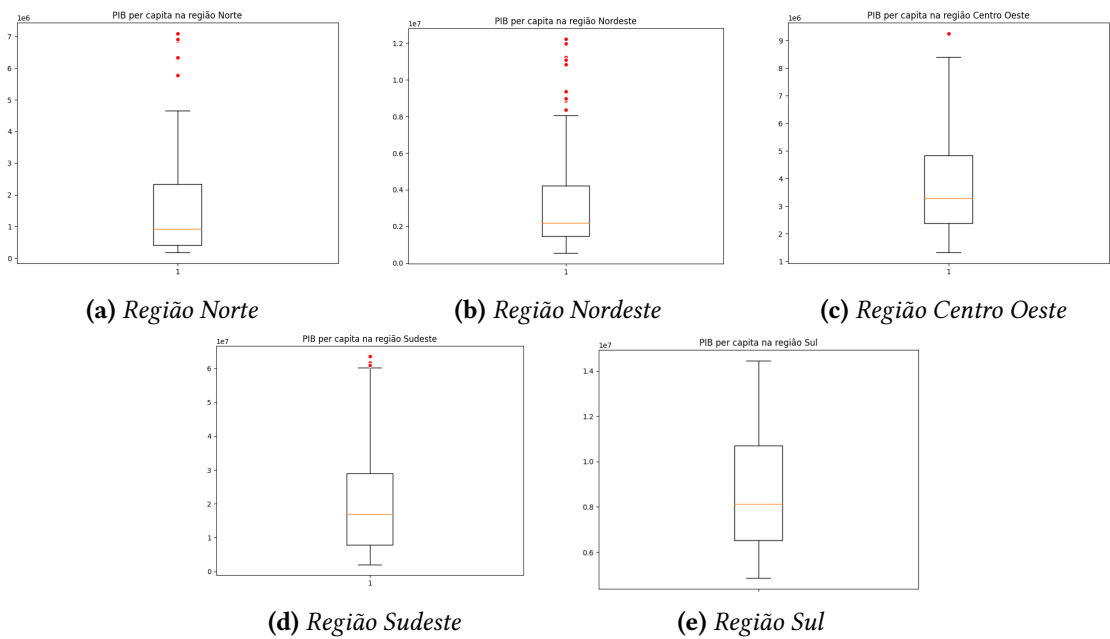


Figura 2.2: Comparação dos gráficos boxplot entre as regiões

Além disso, a correlação entre as variáveis de entrada foi analisada com o auxílio de uma matriz de correlação, o que gerou a identificação de uma alta correlação entre os indicadores do PIB do estado, o PIB da construção civil e o tamanho da população. Por fim, também foi observada a correlação entre os três indicadores de IDH (longevidade, saúde e renda) e entre o preço do saco de cimento e o preço do quilograma, como pode-se validar na figura 2.3:

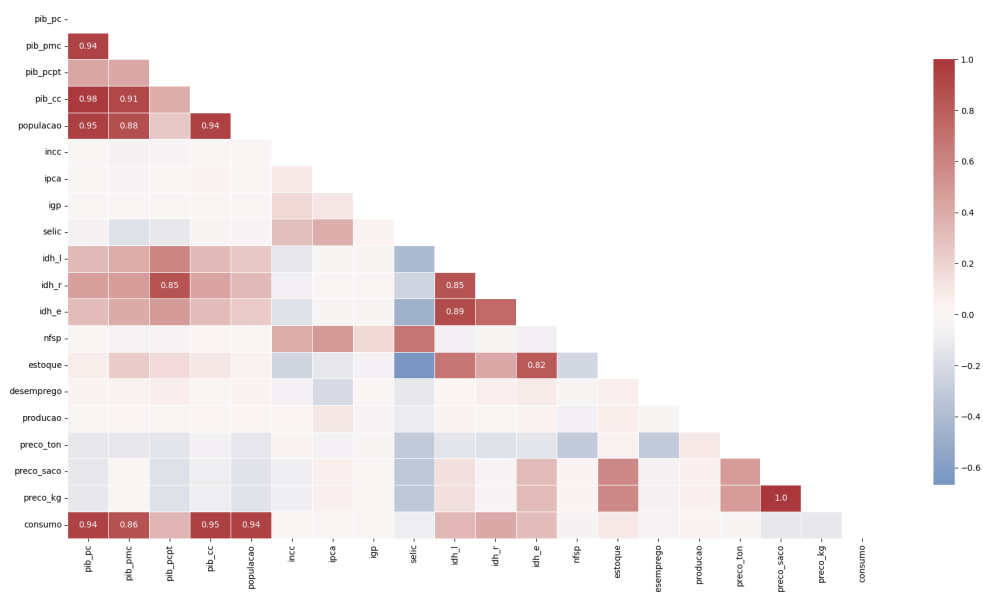


Figura 2.3: Matriz de correlação

Foram adotadas estratégias para garantir que os dados estivessem na granularidade³ mensal e por estado. Caso os indicadores apresentassem granularidade anual, o valor de cada medição seria dividido por 12 de modo a obter a média mensal utilizada em todos os meses do ano correspondente. Caso a granularidade fosse a nível de Brasil, o valor apresentado seria repetido para todos os estados no mês e ano correspondentes. As exceções foram os indicadores de IDH, por apresentarem valores apenas em anos específicos, conforme detalhado na tabela 2.1, quando não havia dados para um determinado ano, o valor da última medição foi repetido.

Também foi necessário lidar com dados faltantes e, para isso, o método utilizado foi repetir o último valor disponível nos dados de entrada para preencher a ocorrência faltante. Contudo, para a produção mensal de cimento e para os indicadores de evolução do preço do cimento, foi utilizado um valor não presente no intervalo dos dados de entrada (-1) para marcar as ocorrências faltantes como nulo, uma vez que essas variáveis não apresentavam valores mais antigos.

Além disso, garantiu-se que a previsão utilizasse dados dos meses anteriores e não do mês alvo da previsão, uma vez que o objetivo do projeto é prever a demanda por cimento em um mês a partir dos dados disponíveis no mês anterior. Por isso, nos dados com granularidade anual, foi realizado um deslocamento de modo a associar as entradas de um ano ao consumo no ano seguinte. Nos dados mensais, analogamente, deslocaram-se as entradas para associá-las ao consumo no mês seguinte.

Por fim, o estado correspondente à medição foi usado como dado de entrada. Como os modelos de inteligência artificial aceitam apenas caracteres numéricos, utilizou-se o método de codificação *one hot* para criar 27 colunas, uma para cada estado, nas quais o valor é um quando a linha possui dados daquele estado e é zero em caso contrário.

2.1.3 Pré-processamento de dados

Foram utilizadas técnicas de pré-processamento de dados com o intuito de garantir que as variáveis de entrada estivessem na mesma escala. Esse processo tem a finalidade de permitir a comparação entre variáveis com diferentes unidades e melhorar o funcionamento dos algoritmos de *machine learning*, uma vez que se as *features* estiverem em escalas diferentes, alguns pesos podem ser atualizados mais rápidos que outros, segundo RASCHKA (2014). Tendo o que se disse em mente, foram testadas no presente trabalho: normalização, *min-max scaler* e *power transformer*.

Normalização

A normalização garante que as variáveis de entrada estejam em uma escala com as propriedades de uma distribuição normal: média (μ) igual a zero e variância (σ) igual a um. Dessa forma, a operação realizada para aplicar o processo em uma entrada x é:

$$x_{norm} = \frac{x - \mu}{\sigma} \quad (2.1)$$

³ A granularidade original dos dados está na tabela 2.1.

Min-Max scaler

De acordo com [RASCHKA \(2014\)](#), o *min-max scaler* é uma abordagem alternativa à normalização e transforma os dados de modo que a escala tenha um intervalo definido, em geral, entre 0 e 1. A operação aplicada em uma entrada x para é:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.2)$$

Sendo x_{norm} a versão normalizada da entrada x , e x_{min} e x_{max} os valores mínimo e máximo assumidos por x .

Power transformer

O *power transformer* é outra alternativa para aproximar os dados de uma distribuição normal, esse método visa estabilizar a variância e diminuir a assimetria dos dados por meio da aplicação da transformação de [YEO e JOHNSON \(2000\)](#) nas entradas.

2.2 Avaliação de performance

Para comparar a eficiência dos modelos, mede-se os erros de cada previsão, ou seja, a distância entre o valor previsto pelo algoritmo e o valor do dado real. Neste trabalho, utilizou-se as seguintes métricas estatísticas para mensurar o desempenho: *mean absolute error* (MAE), *root mean square error* (RMSE) e *mean absolute percentage error* (MAPE). Além disso, foi utilizada a variação percentual (Δ) para avaliar se o modelo tende a subestimar ou superestimar o valor previsto.

2.2.1 Mean absolute error (MAE)

De acordo com [HEWAMALAGE et al. \(2022\)](#), a *mean absolute error* (MAE) mede o erro absoluto de cada previsão. Seja \hat{y}_i o valor previsto pelo modelo e y_i o valor real, a fórmula da MAE é dada por:

$$MAE = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n} \quad (2.3)$$

Essa métrica não leva em consideração a proporção do erro em relação ao valor real, apenas o valor absoluto da diferença, dessa forma não expressa a ordem de grandeza do erro em relação ao valor real.

2.2.2 Root mean squared error (RMSE)

A RMSE, sigla para *root mean squared error*, é uma métrica de erro semelhante à MAE. Sendo \hat{y}_i o valor previsto pelo modelo para um *output* y_i , a fórmula da RMSE é dada por:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (2.4)$$

A RMSE é influenciada por erros mais significativos, dessa forma, é mais sensível a *outliers* que a MAE, segundo HEWAMALAGE *et al.* (2022).

2.2.3 Mean absolute percentage error (MAPE)

Foi utilizada também a *mean absolute percentage error*, para mensurar a magnitude do erro em relação ao tamanho das medições, a MAPE calcula a proporção do erro em relação ao valor que o modelo tentou prever. Sendo \hat{y}_i o valor previsto pelo modelo para um *output* y_i , a fórmula da MAPE é dada por:

$$MAPE = \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (2.5)$$

A MAPE proporciona uma visão do tamanho do erro em relação ao valor que se deseja prever.

2.2.4 Variação percentual

A variação percentual, Δ , é utilizada para mensurar se o modelo apresenta tendência de subestimar ou superestimar a variável. Sendo \hat{y}_i o valor previsto pelo modelo para um *output* y_i , a fórmula da variação percentual Δ é dada por:

$$\Delta = \frac{\hat{y}_i - y_i}{y_i} \quad (2.6)$$

2.3 Avaliação do desempenho do modelo

Neste trabalho, separou-se a massa de dados em conjunto de treino com 85% das entradas e de teste com 15% do conjunto, a metodologia empregada teve como objetivo avaliar a capacidade de generalização do modelo e, por medir o desempenho em dados desconhecidos pelo modelo, também permite avaliar a ocorrência de *overfitting*⁴. Levando isso em conta, não foi utilizado embaralhamento de dados, logo, os modelos foram treinados com dados de janeiro de 2003 até junho de 2017, e testados com dados de julho de 2017 até dezembro de 2019.

⁴ O *overfitting* ocorre quando um modelo tem baixa capacidade de generalização, então, apesar de apresentar baixa taxa de erro nos dados de treino, apresenta desempenho ruim em dados não conhecidos.

2.4 Tecnologias utilizadas

Neste projeto⁵ foi utilizada a linguagem Python e o ambiente de desenvolvimento interativo fornecido pelo Jupyter Notebook⁶. Foram utilizadas as bibliotecas Pandas⁷ e NumPy⁸ para realizar a preparação dos dados, além da TensorFlow⁹ e Scikit-learn com o intuito de treinar e avaliar os modelos. A maior parte dos gráficos foi gerada por meio das bibliotecas Seaborn e Matplotlib, contudo, alguns foram contruídos por meio da plataforma Microsoft Excel.

⁵ Os códigos utilizados no projeto estão disponíveis em <https://github.com/LeiteJu/TCC>.

⁶ O projeto Jupyter, cujo nome é originado da junção do nome das linguagens Julia, Python e R, provê um ambiente de desenvolvimento interativo baseado em *web*. Nos *notebooks*, é possível executar células de código em várias linguagens, além de inserir células de Markdown para auxiliar na documentação, explicação e organização do código. Essas informações foram retiradas de <https://jupyter-notebook.readthedocs.io/en/stable/notebook.html>.

⁷ A Pandas é uma ferramenta *open source* para análise e manipulação de dados em Python, informação retirada da documentação oficial em <https://pandas.pydata.org/about/index.html>.

⁸ Numpy é um pacote que provê operações matemáticas em vetores, segundo a documentação <https://numpy.org/doc/stable/>

⁹ TensorFlow é uma biblioteca que permite a criação de modelos de aprendizado de máquina, informação retirada de <https://www.tensorflow.org/learn?hl=pt-br>.

Capítulo 3

Experimentos

Para cada um dos modelos selecionados, foram realizados diferentes experimentos alterando parâmetros com o objetivo de obter um melhor desempenho. Nesta seção estão descritos os experimentos realizados de acordo com o modelo.

3.1 Regressao linear

A regressão linear é utilizada neste estudo como base de comparação de desempenho dos outros modelos por se tratar de um modelo mais simples de *machine learning*. Dito isso, foram testados métodos de normalização e transformação de dados, além da remoção de variáveis de alta correlação¹ conforme descrito na tabela 3.1:

Experimento	Transformação nos dados	Período de dados	Remoção de variáveis
A	nenhum	de 2003 até 2019	não
B	<i>Standard scaler</i>	de 2003 até 2019	não
C	<i>MinMax scaler</i>	de 2003 até 2019	não
D	<i>Power Transformer</i>	de 2003 até 2019	não
E	nenhum	de 2003 até 2019	sim

Tabela 3.1: Experimentos realizados na regressão linear

O desempenho do modelo em cada um dos experimentos encontra-se na tabela 3.2:

¹ As variáveis removidas foram: PIB *per capita*, INCC, IGP, taxa Selic, IDH Educação e IDH Longevidade, NFSP, preço do saco de cimento e preço da tonelada de cimento.

Experimento	MAE	RMSE	MAPE
A	35148.6	55391.1	0.6
B	35148.6	55391.1	0.6
C	35148.6	55391.1	0.6
D	49073.6	67446.7	1.05
E	26435.6	38970.6	0.32

Tabela 3.2: Desempenho dos modelos de regressão linear

Observa-se que não houve alteração na performance ao normalizar os dados com *standard scaler* ou *minmax scaler*, ao utilizar *power transformer*, contudo, houve piora no desempenho. Ressalta-se ainda a grande melhoria nas métricas ao remover variáveis de alta correlação tanto nas métricas de erro quantitativo, como o MAE e RMSE, quanto proporcionalmente ao atingir 32% de erro percentual médio.

Cumpramos ressaltar também que foi realizado o teste de remover variáveis de alta correlação, pois segundo [JIANG et al., 2022](#), esse tipo de *feature* pode prejudicar o modelo por aumentar a variabilidade dos coeficientes em relação à amostra, dessa forma, os coeficientes calculados pelo modelo podem ter um alto grau de variação de uma amostra para outra.

No capítulo 4, estão apresentados os resultados e previsões do modelo de regressão linear com melhor desempenho dentre os testados no modelo, o experimento E, sem normalização de dados e removendo atributos com alta correlação.

3.2 Redes *feed forward*

As redes neurais *feed forward* podem ser construídas com várias arquiteturas e configurações. Neste estudo, testou-se alterar a quantidade de camadas da rede, o número de neurônios em cada camada, a função de ativação utilizada, além da quantidade de *epoch*² utilizada no treinamento.

Os valores testados para cada um dos parâmetros são dados na tabela 3.3:

Parâmetro	Valores
Número de camadas	1,2,3 e 4
Número de neurônios	32,64,128 e 256
Função de ativação	ReLU e <i>swish</i>
Epochs	50, 100 e 150
Pré-processamento de dados	normalização, <i>min-max scaler</i> e <i>power transformer</i>

Tabela 3.3: Parâmetros testados nas redes *feed forward*

² Cada *epoch* corresponde a uma passagem pelos dados de entrada ([PATTERSON e GIBSON, 2017](#))

Cada experimento corresponde a uma combinação dos parâmetros presentes na tabela 3.3, por exemplo: um modelo de uma camada de 32 neurônios, com função de ativação ReLU e 50 *epochs*. A partir do que se falou, os resultados dos experimentos foram separados de acordo com o número de camadas na rede para melhor visualização, de tal forma que o desempenho dos melhores modelos – em relação a outros com o mesmo número de camadas – é dado na tabela 3.4:

Número de camadas	MAE	RMSE	MAPE
1	51317	87398	0.50
2	33028	60233	0.34
3	26880	50830	0.27
4	25977	42925	0.31

Tabela 3.4: Experimentos com melhor desempenho segundo o número de camadas das redes *feed forward*

É interessante notar que ao priorizar os erros absolutos (MAE e RMSE), o modelo com quatro camadas apresentou um desempenho ligeiramente melhor que o de três camadas, contudo, possui maior erro percentual. Assim, por mais que os erros absolutos sejam menores, ao levar em consideração o valor que o modelo pretende prever, os erros desse modelo são maiores que os gerados pelo modelo de três camadas.

Neste trabalho, priorizou-se modelos com menor erro percentual (MAPE), então o modelo com melhor desempenho entre as redes *multilayer feed forward* foi o de três camadas. Esse modelo utiliza a função de ativação *swish*, o método de normalização para pré-processar os dados e foi treinado com 100 *epochs*, além disso a primeira camada oculta possui 256 neurônios, a segunda, 64 e a terceira, 32.

3.3 Redes recorrentes

Neste trabalho, foram utilizadas redes recorrentes LSTM e GRU, para ambas as redes, realizaram-se experimentos alterando os mesmos parâmetros, a saber, número de camadas, número de neurônios, função de ativação, número de *epochs* e método de pré-processamento de dados. Na tabela 3.5, encontra-se a relação dos valores testados para cada parâmetro:

Parâmetro	Valores
Número de camadas	1 e 2
Número de neurônios	32,64,128 e 256
Função de ativação	tangente hiperbólica ^a , ReLU e <i>swish</i>
<i>Epochs</i>	50, 100 e 150
Pré-processamento de dados <i>time step</i> ^b	normalização, <i>min-max scaler</i> e <i>power transformer</i> 3 e 5

Tabela 3.5: *Parâmetros testados nas redes recorrentes*

^a Por padrão, as redes LSTM e GRU utilizam a tangente hiperbólica (*tanh*) como função de ativação.

^b O *time step* é o tamanho da sequência que a rede recebe como entrada.

Foram realizados os mesmos experimentos para ambas as redes, assim como para as redes neurais tradicionais, os resultados foram agrupados de acordo com o número de camadas da rede. O desempenho dos modelos com melhores resultados estão na tabela 3.6:

Número de camadas	Rede	MAE	RMSE	MAPE
1	LSTM	22694	43918	0.19
1	GRU	26751	52281	0.24
2	LSTM	25002	47565	0.19
2	GRU	24992	50602	0.20

Tabela 3.6: *Experimentos com melhor desempenho segundo o número de camadas das redes recorrentes*

Observa-se que as redes LSTM obtiveram um melhor desempenho, em geral, que as redes GRU, além disso a rede LSTM mais simples, com uma camada, fez as previsões mais precisas. A rede recorrente com melhor desempenho, então, é uma rede LSTM de uma camada com 32 neurônios, que utiliza a ReLU como função de ativação, com 150 *epochs* e normalização como método de processamento de dados.

3.4 Redes bidirecionais

As redes bidirecionais utilizam duas redes recorrentes para realizar as previsões. Neste estudo, realizaram-se experimentos com redes LSTM e GRU, alterou-se também o número de camadas, a quantidade de neurônios nas camadas, a função de ativação utilizada nas redes, o método de pré-processamento de dados e a quantidade de *epochs*. A relação dos valores testados para cada parâmetro está na tabela 3.7:

Parâmetro	Valores
Número de camadas	1 e 2
Número de neurônios	32, 64, 128 e 256
Função de ativação	tangente hiperbólica ^a , ReLU e <i>swish</i>
<i>Epochs</i>	50, 100 e 150
Pré-processamento de dados	normalização, <i>min-max scaler</i> e <i>power transformer</i>
<i>time step</i> ^b	3 e 5

Tabela 3.7: *Parâmetros testados nas redes bidirecionais*

^a Por padrão, as redes LSTM e GRU utilizam a tangente hiperbólica (*tanh*) como função de ativação.

^b O *time step* é o tamanho da sequência que a rede recebe como entrada.

O resultado dos modelos foi agrupado de acordo com o número de camadas da rede e com a categoria de rede recorrente escolhida (LSTM ou GRU). A relação dos modelos com melhor desempenho é dada na tabela 3.8:

Número de camadas	Rede	MAE	RMSE	MAPE
1	LSTM	19185	33942	0.17
1	GRU	19952	33979	0.18
2	LSTM	22964	38688	0.20
2	GRU	20993	35817	0.20

Tabela 3.8: *Experimentos com melhor desempenho das redes bidirecionais*

Observa-se que o modelo com melhor desempenho das redes bidirecionais utiliza uma camada de rede LSTM de dois neurônios para fazer as previsões, além disso, o experimento foi realizado com 150 *epochs* e com o método de pré-processamento de dados *min-max scaler*.

Capítulo 4

Resultados

No capítulo 3, foram realizados experimentos para encontrar a combinação de parâmetros que proporcionava melhor desempenho para as categorias de modelos testadas neste estudo, regressão linear, redes *feed forward*, redes recorrentes e redes bidirecionais. Neste capítulo, são apresentadas as previsões dos diferentes modelos.

É possível comparar o desempenho dos modelos a partir da tabela 4.1:

Modelo	MAE	RMSE	MAPE
Regressão linear	26435	38970	0.32
Rede <i>feed forward</i>	26880	50830	0.27
Rede recorrente	22694	43918	0.19
Rede bidirecional	19185	33942	0.17

Tabela 4.1: Comparação do desempenho dos modelos

Na análise da tabela acima, nota-se que o desempenho nas previsões melhora à medida que se aumenta a robustez do modelo. A regressão linear é utilizada como base para avaliar o desempenho das redes neurais por ser um modelo mais simples.

É possível perceber que houve uma ligeira melhora das previsões com as redes *feed forward* em relação à regressão linear que se refletiu na redução do erro médio percentual em 5%. Essa melhora, contudo, não é perceptível nos indicadores de erro absoluto, enquanto a MAE apresenta um valor similar ao da regressão, a RMSE teve um aumento significativo, o que indica a presença de *outliers*.

Observa-se uma significativa melhora, de 8%, das redes *feed forward* para os modelos de redes recorrentes, reflexo da capacidade de reconhecer sequências e contexto. A melhora também se reflete nos erros absolutos calculados pela MAE, contudo, a RMSE apresentou piora em relação à regressão linear, provavelmente por conta de *outliers*, uma vez que a RMSE é mais sensível a valores discrepantes.

As redes bidirecionais apresentaram melhor desempenho em comparação com os outros modelos, isso em virtude da arquitetura mais complexa com duas redes recorrentes.

Percebeu-se também uma redução de 2% da MAPE em relação às redes recorrentes e de 15% em relação à regressão linear. Esse comportamento também se reflete nos indicadores de erro absoluto que são os menores entre as classes de algoritmos testados.

No presente trabalho, foi adotado o estado de São Paulo como objeto de observação mais detalhado para comparar as previsões e o valor real do consumo, essa escolha foi tomada por se tratar do maior consumidor de cimento no Brasil¹, de tal forma que a evolução do consumo mensal de cimento em São Paulo pode ser vista na figura 4.1:

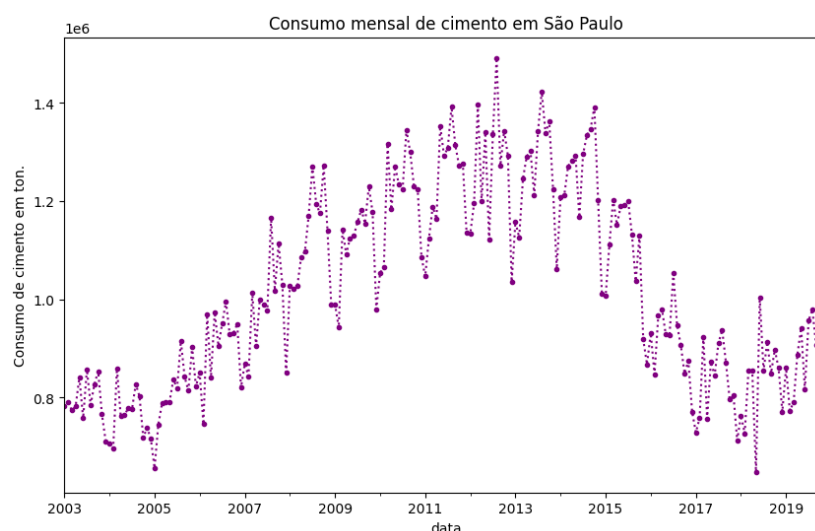


Figura 4.1: *Evolução do consumo mensal de cimento em São Paulo*

Na imagem acima, os círculos representam as medições mensais do consumo de cimento em São Paulo. Observa-se que a demanda por cimento no estado vinha em tendência de queda desde 2013 e passou a apresentar alta de 2018 em diante, além disso, há certa sazonalidade da demanda.

Os modelos foram utilizados para prever o consumo mensal de cimento a partir de julho de 2017. Os gráficos do consumo real e da previsão realizada pelo modelo são sobrepostos para melhor visualização. Nos gráficos das previsões apresentados neste capítulo, o consumo real está representado em azul claro e as previsões do modelo, em laranja, além disso, a linha tracejada em cinza marca julho de 2017, início das previsões.

4.1 Regressão linear

Comparou-se as previsões realizadas pela regressão linear com o valor real que foi demandado de cimento, tendo isso em mente, na imagem 4.2, representa-se distribuição do erro percentual das previsões:

¹ Em 2019, o estado de São Paulo foi responsável por 21.3% da demanda por cimento no Brasil, 8.6% à frente do segundo maior consumidor, Minas Gerais.

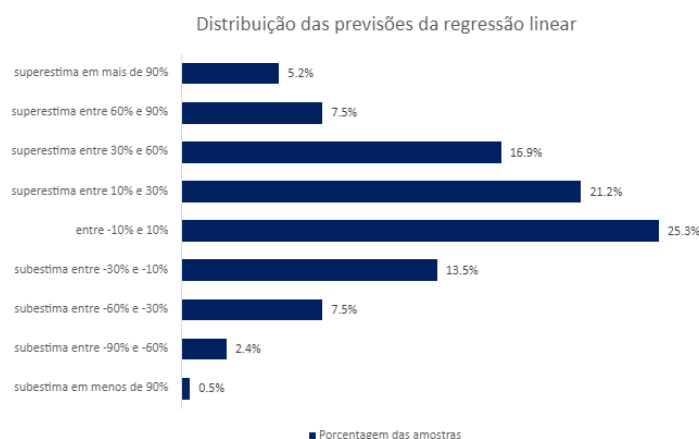


Figura 4.2: Distribuição da variação percentual das previsões da regressão linear

No gráfico acima, é possível ver que o modelo concentra 60% das previsões com erro percentual menor ou igual a 30%. Além disso, observa-se uma tendência de superestimar o valor do consumo, uma vez que 50.8% das previsões superestimam o valor real em mais de 10%.

As previsões realizadas pela regressão linear para o estado de São Paulo podem ser visualizadas na imagem 4.3:

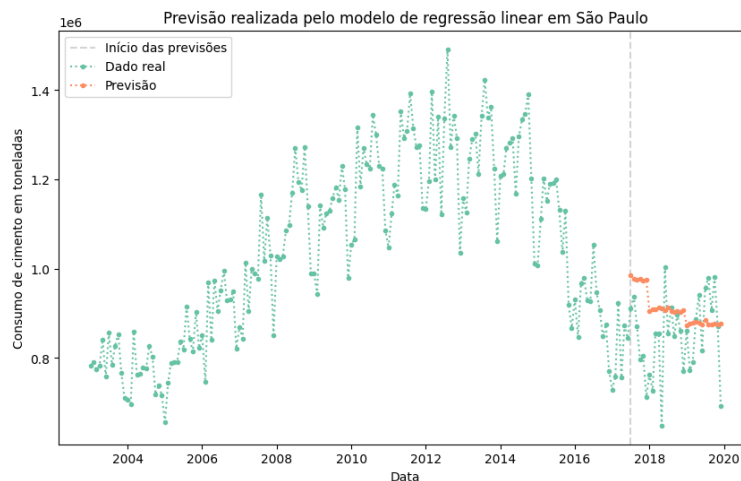


Figura 4.3: Previsão realizada pela regressão linear para o consumo de cimento em São Paulo

Destaca-se na figura 4.3 a proximidade das previsões realizadas ao longo dos meses de um ano, são perceptíveis variações apenas na passagem de um ano para o seguinte. Uma possível explicação para esse comportamento são as variáveis que originalmente possuíam granularidade anuais utilizadas como atributos.

Para o estado de São Paulo, as previsões estão com valores próximos aos reais. Apesar disso, o modelo previu que a tendência de queda, presente nos dados até 2018, continuaria em 2018 e 2019, contudo o consumo real passou a apresentar alta.

4.2 Redes *feed forward*

As redes *feed forward* são modelos mais robustos que a regressão linear. A distribuição da diferença percentual entre as previsões e os valores reais do consumo é dada na figura 4.4:

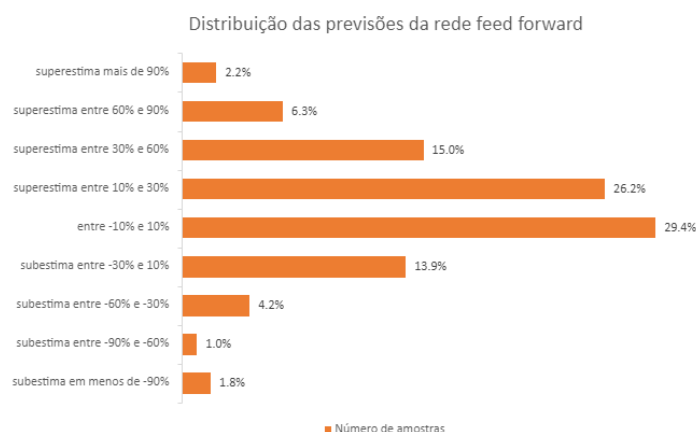


Figura 4.4: Distribuição da variação percentual das previsões das redes *feed forward*

Ao analisar o gráfico acima, observa-se que as redes *feed forward* concentram 69.5% das previsões com erro percentual de até 30%, dessa forma, há melhora de 9.5% em comparação com a regressão linear. Essas redes ainda apresentam tendência a superestimar o consumo, contudo, em menor grau que a regressão linear, já que 49.7% das previsões sobrestimam o valor real em mais de 10%.

Na imagem 4.5, são dadas as previsões do modelo para o estado de São Paulo:

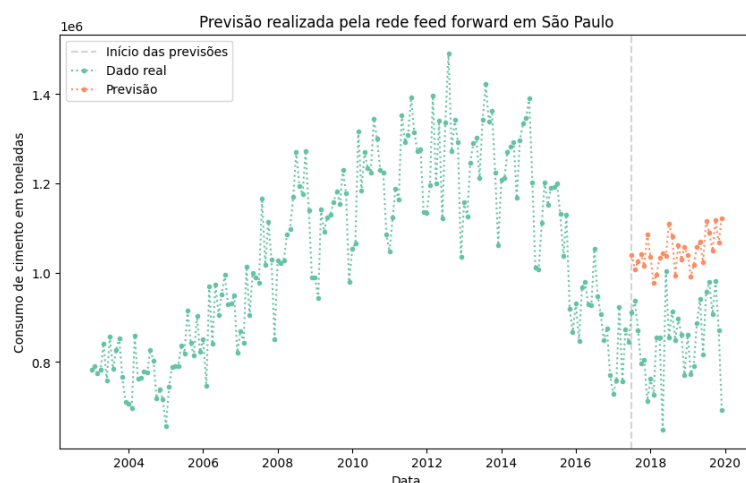


Figura 4.5: Previsão realizada pelas redes *feed forward* em São Paulo

Observa-se na imagem que o modelo capturou a leve tendência de aumento do consumo a partir de 2018, contudo as previsões superestimam o valor real do consumo.

4.3 Redes recorrentes

A distribuição do erro percentual das previsões realizadas pelas redes recorrentes pode ser observada na figura 4.6:

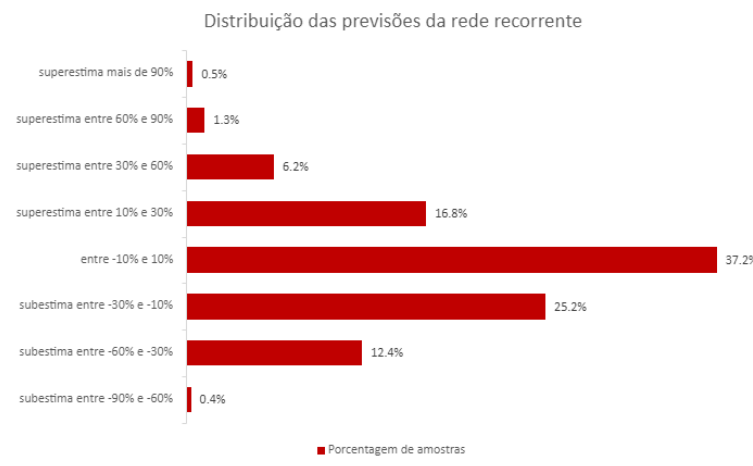


Figura 4.6: Distribuição da variação percentual das previsões das redes recorrentes

Na análise do gráfico acima, é perceptível o impacto da análise de contexto histórico nas previsões realizadas pelas redes recorrentes. As previsões com erro percentual de até 30% correspondem a 79.2% do conjunto de teste, um aumento de 9.7% em relação à rede *feed forward* e de 19.2% em relação à regressão linear. Além disso, destaca-se que 37.2% das previsões apresentaram erro percentual de até 10%. Inversamente à regressão linear e das redes *feed forward*, as redes recorrentes apresentam tendência de subestimar o consumo, com 38% das previsões com valores subestimando o valor real do consumo em mais de 30%.

Na figura 4.7, é possível observar as previsões realizadas pelas redes recorrentes no estado de São Paulo:

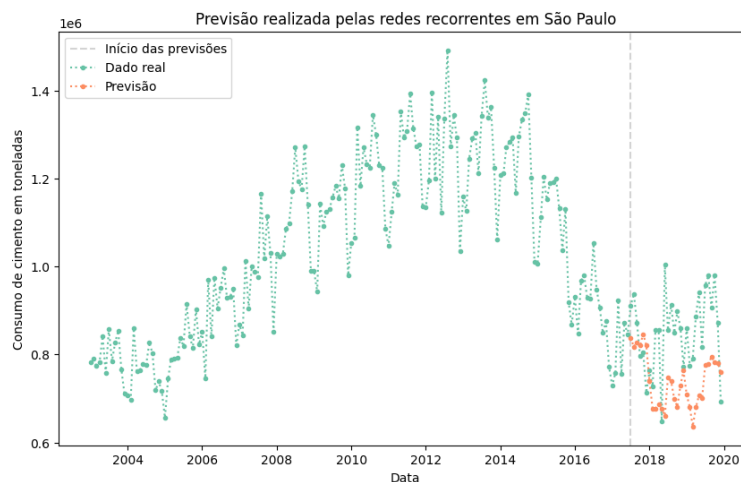


Figura 4.7: Previsão realizada pelas redes recorrentes em São Paulo

Por meio do gráfico, é possível evidenciar que o modelo capturou com precisão a tendência do consumo, primeiro de queda até 2018 e aumento após essa data. Destaca-se, também, a proximidade das previsões realizadas pelas redes recorrentes e o valor real do consumo, embora o modelo tenha tendência de subestimar o valor.

4.4 Redes bidirecionais

Na imagem 4.8, está a distribuição do erro percentual das previsões realizadas ao utilizar redes bidirecionais:

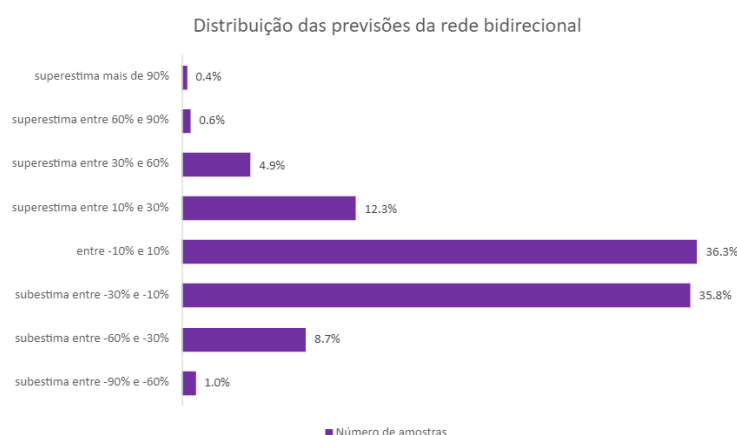


Figura 4.8: Distribuição da variação percentual das previsões das redes bidirecionais

Observa-se no gráfico acima, a melhora no desempenho ao utilizar a arquitetura robusta das redes bidirecionais: 84.4% das previsões desse modelo apresentam erro de até 30%, um crescimento de 5.2% ao comparar com as redes recorrentes e de 24.4% ao comparar com a regressão linear. Destaca-se, também, uma tendência maior de subestimar o valor real, 45.5% subestimam o consumo real em mais de 30%,

Na imagem 4.9, encontram-se as previsões realizadas pelo modelo para São Paulo:

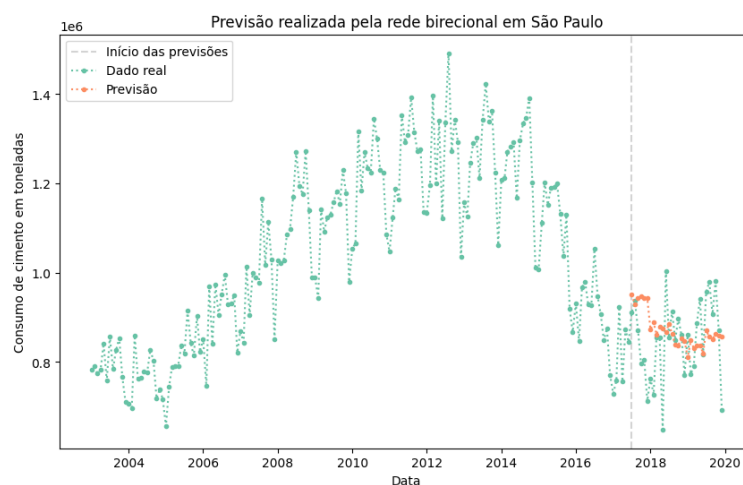


Figura 4.9: Previsão realizada pelas redes bidirecionais em São Paulo

Destaca-se na figura acima a assertividade das previsões realizadas pelas redes bidirecionais. O modelo não apenas captura de forma certa a tendência do consumo, como prevê valor muito próximos ao valor real. Por isso mesmo, em comparação com as previsões realizadas pelos outros modelos testados nesta trabalho, trata-se da previsão mais precisa.

Capítulo 5

Discussão

Nesta dissertação, a diferença entre 1 e o valor da MAPE de um modelo foi empregada como métrica de assertividade, com o propósito de mensurar o grau de acerto das previsões.

Durante a análise dos experimentos realizados, foi possível observar que as previsões se tornaram mais assertivas à medida que se elevou a complexidade do modelo. Essa melhora nas previsões indicou que não houve sintomas de falta de dados e era esperada, visto que as redes *feed forward* acrescentaram à regressão linear a capacidade de reconhecer padrões não lineares nos dados por meio do processamento matemático realizado nas camadas de neurônios. As redes recorrentes, de maneira análoga, adicionaram às redes *feed forward* a capacidade de identificar contexto histórico e sequências e, por fim, as redes bidirecionais agregaram às recorrentes a habilidade de avaliar as entradas nos dois sentidos das sequências.

A regressão linear é o modelo mais simples empregado neste estudo e, apesar de apresentar as previsões menos exatas entre os modelos, obteve 68% de assertividade. Seu erro absoluto médio foi de 26.4 mil toneladas, uma taxa de erro aceitável ao levar em consideração que a média de consumo mensal de cimento no Brasil é de 151.2 mil toneladas.

O processamento mais robusto das redes *feed forward*, conferido pelas camadas de neurônios, resultou em uma assertividade de 73% em média, um aumento de 5% em relação à regressão linear. Essa melhora, contudo, não se refletiu no erro absoluto que foi de 26.9 mil toneladas e houve um provável aumento nos *outliers* – amostras cuja previsão foi distante do valor real – indicado pelo aumento da RMSE.

A capacidade de analisar contexto histórico garantida às redes recorrentes pela arquitetura complexa teve como consequência uma assertividade de 81%, uma melhora de 8% em comparação com as redes *feed forward* e de 13% em comparação com a regressão linear. Do mesmo modo, o impacto se refletiu na redução do erro absoluto para 22.7 mil toneladas, ou seja, a rede apresentou um melhor desempenho tanto ao considerar o valor absoluto do erro, quanto ao comparar a proporção do erro em relação ao resultado que o modelo pretendia prever.

Por fim, as duas redes combinadas conferiram às redes bidirecionais uma assertividade de 83% em média, um aumento de 2% em relação às redes recorrentes e de 15% em relação à regressão linear. Além disso, o erro absoluto foi o menor entre os modelos testados no estudo – 19.1 mil toneladas –, e houve menor incidência de previsões muito discrepantes do valor real, uma vez houve queda na RMSE. Esse modelo também apresentou uma boa capacidade de generalização, haja vista que apresentou 12.6% de erro no conjunto de treino e 17.4% no de teste, ou seja, um baixo *overfitting*.

Assim, a capacidade de analisar contexto histórico nos dois sentidos das redes neurais bidirecionais configurou-se como diferencial e garantiu ao modelo o melhor desempenho entre os testados. Ainda que as previsões desse modelo não coincidam exatamente com os dados reais de consumo, esse comportamento é esperado, uma vez que esse indicador é sujeito a influências externas não mensuráveis pelo modelo, como políticas governamentais em um estado específico, ocorrências climáticas, grandes obras da indústria privada, entre outros fatores.

Capítulo 6

Conclusão

Este trabalho tinha o objetivo de propor uma solução para a demanda da indústria cimenteira por um modo bem fundamentado para prever a demanda por cimento nos estados brasileiros. Por conta disso, foi investigado o desempenho das previsões de diferentes modelos de *machine learning*, que utilizavam dados econômicos, sociais, monetários e da construção civil para prever o consumo mensal de cimento; os modelos testados, em ordem crescente de complexidade, foram regressão linear, redes neurais *feed forward*, recorrentes e bidirecionais.

Ao longo desta monografia, foram apresentados os conceitos de *machine learning*, regressão linear, redes neurais *feed forward*, recorrentes e bidirecionais. Além disso, descreveu-se a metodologia aplicada, o tratamento realizado nos dados, os métodos para mensurar o desempenho dos modelos e os experimentos realizados. Foram detalhadas, também, as previsões realizadas pelos modelos e a avaliação dessas.

Evidenciou-se ao longo do trabalho o aumento na assertividade das previsões à medida que se aumentava a robustez dos modelos utilizados, em geral, a melhora também foi perceptível nas métricas de erro absoluto. Partindo do que foi dito, observou-se que as redes bidirecionais são o modelo com melhor desempenho dentre os testados no estudo, com previsões assertivas que capturaram com precisão a tendência do consumo. Apesar do erro associado às previsões — já esperado, uma vez que o fenômeno que se deseja prever é influenciado por fatores além da capacidade do modelo de capturar —, foi alcançada uma assertividade relevante nas previsões, além de um baixo erro absoluto.

Dessa forma, infere-se que o uso de modelos de *machine learning* para prever a demanda por cimento é promissor e capaz de produzir relevantes resultados e, por isso mesmo, se configura como uma possível solução para a deficiência da indústria cimenteira, de um modo embasado para estimar o consumo futuro de cimento.

Espera-se que os experimentos, resultados e conclusões realizados neste trabalho de conclusão de curso auxiliem futuros pesquisadores a retomar a pesquisa deste projeto.

Referências

- [DSA 2022] Data Science ACADEMY. *Deep Learning Book*. 2022. URL: <https://www.deeplearningbook.com.br> (acesso em 27/12/2022) (citado nas pgs. 4, 7–10, 13–15).
- [ALLA 2021] Samhita ALLA. *Advanced Recurrent Neural Networks: Bidirectional RNNs*. 2021. URL: <https://blog.paperspace.com/bidirectional-rnn-keras/> (acesso em 28/01/2023) (citado nas pgs. 15, 16).
- [IBRAM 2021] IBRAM Mineração do BRASIL. *Votorantim Cimentos inicia a operação de nova fábrica em Pecém com lançamento de produto no mercado cearense*. 2021. URL: <https://ibram.org.br/noticia/votorantim-cimentos-inicia-a-operacao-de-nova-fabrica-em-pecem-com-lancamento-de-produto-no-mercado-cearense/> (citado na pg. 1).
- [CARVALHO 2008] Maria Beatriz Maury de CARVALHO. “Impactos e Conflitos da Produção de Cimento no Distrito Federal”. Diss. de mestr. Universidade de Brasília, 2008 (citado na pg. 2).
- [CHO *et al.* 2014] Kyunghyun CHO *et al.* *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. DOI: 10.48550/ARXIV.1406.1078. URL: <https://arxiv.org/abs/1406.1078> (citado na pg. 15).
- [ABCP 2002] Associação Brasileira de CIMENTO PORTLAND. “Guia básico de utilização do cimento portland”. Em: 2002, pg. 5 (citado na pg. 1).
- [DESAI 2019] Rashmi DESAI. *Top 10 Python Libraries for Data Science*. 2019. URL: <https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266> (acesso em 28/01/2023).
- [ESSELMAN 2022] Amy ESSELMAN. “What is a boxplot?” 2022. URL: <https://www.storytellingwithdata.com/blog/what-is-a-boxplot> (citado na pg. 19).
- [GOODFELLOW *et al.* 2016] Ian GOODFELLOW, Yoshua BENGIO e Aaron COURVILLE. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016 (citado nas pgs. 4, 5, 17).

- [HEWAMALAGE *et al.* 2022] Hansika HEWAMALAGE, Klaus ACKERMANN e Christoph BERGMEIR. *Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices*. 2022. DOI: [10.48550/ARXIV.2203.10716](https://arxiv.org/abs/2203.10716). URL: [5Curl%7Bhttps://arxiv.org/abs/2203.10716%7D](https://arxiv.org/abs/2203.10716) (citado nas pgs. 22, 23).
- [HOAGLIN 2013] David C. HOAGLIN. “Volume 16: how to detect and handle outliers”. Em: 2013 (citado na pg. 18).
- [HOCHREITER e SCHMIDHUBER 1997] Sepp HOCHREITER e Jürgen SCHMIDHUBER. “Long short-term memory”. Em: *Neural computation* 9 (dez. de 1997), pgs. 1735–80. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735) (citado na pg. 13).
- [HYNDMAN e ATHANASOPOULOS 2021] Rob J. HYNDMAN e George ATHANASOPOULOS. *Forecasting: Principles and practice*. 3ª ed. Accessed on 17 December 2022. OTexts, 2021. URL: <https://otexts.com/fpp3/> (citado nas pgs. viii, 6, 7).
- [JIANG *et al.* 2022] Hao-sheng JIANG, Jia-li CHEN e Chong-qi ZHANG. *Constructing K-optimal designs for different Scheffé models*. 2022. DOI: [10.48550/ARXIV.2210.07922](https://arxiv.org/abs/2210.07922). URL: <https://arxiv.org/abs/2210.07922> (citado na pg. 26).
- [JULIO DA MOTTA SINGER 2022] Pedro Alberto Morettin e JULIO DA MOTTA SINGER. *Estatística e ciência de dados*. 2022.
- [MCCARTHY 2007] John MCCARTHY. “What is artificial intelligence?” Em: (nov. de 2007), pg. 2. URL: <http://www-formal.stanford.edu/jmc/whatisai.pdf> (citado na pg. 4).
- [NICHOLAS LACASCIO 2017] Nikhil Buduma e NICHOLAS LACASCIO. *Fundamentals of Deep Learning*. O’Reilly Media, Inc, 2017 (citado na pg. 8).
- [NIELSEN 2015] Michael A. NIELSEN. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: <http://neuralnetworksanddeeplearning.com/> (citado nas pgs. viii, 10, 12).
- [NORMAS TÉCNICAS 2018] Associação Brasileira de NORMAS TÉCNICAS. “Cimento Portland — requisitos”. Em: 2018 (citado na pg. 1).
- [PATTERSON e GIBSON 2017] Josh PATTERSON e Adam GIBSON. *Deep Learning: A practitioner’s approach*. 2017, pg. 4 (citado nas pgs. viii, 4, 5, 7–11, 13, 26).
- [POOLE *et al.* 1998] David POOLE, Alan MACKWORTH e Randy GOEBEL. *Computational Intelligence: A Logical Approach*. Oxford University Press, 1998.
- [RAMACHANDRAN *et al.* 2017] Prajit RAMACHANDRAN, Barret ZOPH e Quoc V. LE. *Searching for Activation Functions*. 2017. DOI: [10.48550/ARXIV.1710.05941](https://arxiv.org/abs/1710.05941). URL: <https://arxiv.org/abs/1710.05941> (citado nas pgs. 11, 12).
- [S. RASCHKA 2014] Sebastian RASCHKA. “About feature scaling and normalization - and the effect of standardization for machine learning algorithms”. 2014. URL: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html.

- [RASCHKA 2014] Sebastian RASCHKA. *About Feature Scaling and Normalization – and the effect of standardization for machine learning algorithms*. Jul. de 2014. URL: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html (acesso em 27/01/2023) (citado nas pgs. 21, 22).
- [RUSSELL e NORVIG 1995] Stuart J. RUSSELL e Peter NORVIG. *Artificial Intelligence: A Modern Approach, third edition*. 1995 (citado na pg. 3).
- [SAXENA 2021] Shipra SAXENA. *Introduction to Gated Recurrent Unit (GRU)*. 2021. URL: <https://www.analyticsvidhya.com/blog/2021/03/introduction-to-gated-recurrent-unit-gru/> (acesso em 27/01/2023).
- [SCHUSTER e PALIWAL 1997] M. SCHUSTER e K.K. PALIWAL. “Bidirectional recurrent neural networks”. Em: *IEEE Transactions on Signal Processing* 45.11 (1997), pgs. 2673–2681. DOI: [10.1109/78.650093](https://doi.org/10.1109/78.650093) (citado na pg. 15).
- [SILVA BARBOZA 2016] Lucas da SILVA BARBOZA. “Estudo sobre o impacto da redução do consumo de cimento no comportamento mecânico do concreto autoadensável”. Diss. de mestr. Universidade Federal de São Carlos, 2016.
- [SIQUEIRA 2021] Daniel SIQUEIRA. *Melhorando a análise com o Boxplot*. 2021. URL: <https://www.alura.com.br/artigos/melhorando-a-analise-com-o-boxplot> (acesso em 29/01/2023).
- [SNIC 2021] SNIC. “Relatório nacional 2021 sindicato nacional da indústria do cimento”. 2021.
- [STECK *et al.* 2021] Harald STECK *et al.* “Deep learning for recommender systems: a netflix case study”. Em: *AI Magazine* 42.3 (nov. de 2021), pgs. 7–18. DOI: [10.1609/aimag.v42i3.18140](https://doi.org/10.1609/aimag.v42i3.18140). URL: <https://ojs.aaai.org/index.php/aimagazine/article/view/18140> (citado na pg. 3).
- [TUKEY 1977] John W. TUKEY. *Exploratory Data Analysis*. Addison-Wesley, 1977 (citado na pg. 18).
- [TURING 1950] A. M. TURING. “Computing machinery and intelligence”. Em: *Mind* 59.236 (1950) (citado na pg. 4).
- [VASCONCELOS 2022] Ieda VASCONCELOS. *Informativo Econômico PIB*. 2022. URL: <https://cbic.org.br/wp-content/uploads/2022/03/informativo-economico-pib-04-marco-2022.pdf> (citado na pg. 1).
- [WALTER PITTS 1943] Warren McCulloch e WALTER PITTS. “A logical calculus of the ideas immanent in nervous activity”. Em: *Bulletin of Mathematical Biology* 5 (1943) (citado na pg. 8).

- [WANG 2019] Pei WANG. “On defining artificial intelligence”. Em: *Journal of Artificial General Intelligence* 10.2 (2019), pgs. 1–37. DOI: [doi:10.2478/jagi-2019-0002](https://doi.org/10.2478/jagi-2019-0002). URL: <https://doi.org/10.2478/jagi-2019-0002>.
- [WILLIAMSON *et al.* 1989] D WILLIAMSON, RA PARKER e Juliette KENDRICK. “The box plot: a simple visual method to interpret data”. Em: *Annals of internal medicine* 110 (jul. de 1989), pgs. 916–21. DOI: [10.1059/0003-4819-110-11-916](https://doi.org/10.1059/0003-4819-110-11-916) (citado na pg. 19).
- [YEO e JOHNSON 2000] In-Kwon YEO e Richard A. JOHNSON. “A new family of power transformations to improve normality or symmetry”. Em: *Biometrika* 87.4 (dez. de 2000), pgs. 954–959. ISSN: 0006-3444. DOI: [10.1093/biomet/87.4.954](https://doi.org/10.1093/biomet/87.4.954). eprint: <https://academic.oup.com/biomet/article-pdf/87/4/954/633221/870954.pdf>. URL: <https://doi.org/10.1093/biomet/87.4.954> (citado na pg. 22).
- [ZHANG *et al.* 2021a] Aston ZHANG, Zachary C. LIPTON, Mu LI e Alexander J. SMOLA. “Dive into deep learning”. Em: *arXiv preprint arXiv:2106.11342* (2021) (citado na pg. 15).
- [ZHANG *et al.* 2021b] Aston ZHANG, Zachary C. LIPTON, Mu LI e Alexander J. SMOLA. “Dive into deep learning”. Em: *arXiv preprint arXiv:2106.11342* (2021) (citado na pg. 10).