Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Práticas de Aprendizado de Máquina usando Previsão de Cimento como Estudo de Caso.

Pedro Fernandes

Monografia Final

mac 499 — Trabalho de Formatura Supervisionado

Supervisor: Prof. Dr. Marcelo Finger

São Paulo
2023

Esta seção é opcional e fica numa página separada;
ela pode ser usada para uma dedicatória ou epígrafe.

Agradecimentos

Do. Or do not. There is no try.
— Mestre Yoda

Texto texto. Texto opcional.

Resumo

Pedro Fernandes. **Práticas de Aprendizado de Máquina usando Previsão de Cimento como Estudo de Caso.**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Elemento obrigatório, constituído de uma sequência de frases concisas e objetivas, em forma de texto. Deve apresentar os objetivos, métodos empregados, resultados e conclusões. O resumo deve ser redigido em parágrafo único, conter no máximo 500 palavras e ser seguido dos termos representativos do conteúdo do trabalho (palavras-chave). Deve ser precedido da referência do documento. Texto texto. Texto texto.

Palavras-chave: Palavra-chave1. Palavra-chave2. Palavra-chave3.

Abstract

Pedro Fernandes. **Título em inglês não definido!**. Capstone Project Report (Bachelor).
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

Elemento obrigatório, elaborado com as mesmas características do resumo em língua portuguesa. De acordo com o Regimento da Pós-Graduação da USP (Artigo 99), deve ser redigido em inglês para fins de divulgação. É uma boa ideia usar o sítio www.grammarly.com na preparação de textos em inglês. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Keywords: Keyword1. Keyword2. Keyword3.

Lista de abreviaturas

ABNT	Associação Brasileira de Normas Técnicas
ARIMA	Média Móvel Integrada Autoregressiva (*AutoRegressive Integrated Moving Average*)
BCB	Banco Central do Brasil
CMA	Média Móvel Centrada (*Centrated Moving Average*)
EMBI	Índice de títulos de mercados emergentes (*Emerging Market Bond Index*)
FFNN	Rede Neural Feed Forward (*Feed Forward Neural Network*)
FGTS	Fundo de Garantia do Tempo de Serviço
FGV	Fundação Getulio Vargas
IA	Norma Brasileira
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IME	Instituto de Matemática e Estatística
INCC	Índice Nacional de Custo da Construção Civil
INMET	Instituto Nacional de Meteorologia
IPCA	Índice Nacional de Preços ao Consumidor Amplo
IPEA	Instituto de Pesquisa Econômica Aplicada
LSTM	Memória de longo e curto prazo (*Long-Short Term Memory*)
MAE	Média Percentual de Erro Absoluto (*Mean Absolute Error*)
MAPE	Média Percentual de Erro Absoluto (*Mean Absolute Percentage Error*)
ML	Machine Learning
NBR	Norma Brasileira
PIB	Produto Interno Bruto
RNN	Rede Neural Recorrente (*Recurrent Neural Network*)
SGD	Gradiente Descendente Estocástico (*Stochastic Gradient Descent*)
SNIC	Sindicato Nacional da Indústria do Cimento
RMSE	Média Raiz do Erro Quadrado (*Root Mean Squared Error*)
USP	Universidade de São Paulo

Lista de figuras

Lista de tabelas

Sumário

Introdução

Nas últimas décadas, os modelos baseados em aprendizado de máquina transcenderam as fronteiras da esfera acadêmica, tornando-se uma parte integral dos estudos de pesquisa e desenvolvimento em praticamente todos os setores da sociedade. Seu crescimento expressivo, tanto como campo de estudo quanto em aplicabilidade, deve-se ao fato de que esses modelos ultrapassam as expectativas comuns de um modelo matemático estatístico em termos de capacidade de previsão, classificação e geração.

A capacidade superior dos modelos baseados em aprendizado, especialmente os neurais, de capturar relações complexas entre os dados de entrada e representar bem às características do domínio ao qual o dado que queremos prever pertence tornou-se, em muitos casos, a primeira opção quando se trata de previsibilidade em domínios complexos cujos dados influenciadores são multivariados. Sua estrutura é versátil em relação à aceitação de entradas multidimensionais, ao contrário de modelos mais simples e univariados, como o ARIMA. Além disso, modelos mais sofisticados e profundos[1] como os das redes recorrentes, são capazes de mimetizar uma espécie de "memória", permitindo que o modelo capture variações sutis em séries temporais."

Sua aplicação se intensifica em diversas áreas, algumas delas como: a medicina, em que modelos treinados com grande quantidade de dados são capazes de gerar diagnósticos precisos, receitar medicamentos bem determinados com doses mais eficazes (Rajkomar *et al.*, 2019); na área de predição climática, onde o fornecimento de indicadores futuros dos processos da natureza leva a incentivos a projetos de mitigações na emissão de poluentes ou adaptações para condições inevitáveis inerentes as dinâmicas do planeta (Chatfield, 2000) e, finalmente, em áreas econômicas, em que é intrínsecos a análise de qualquer elemento por um prisma multivariado de parâmetros.

Todavia, uma simplificação desse uso acelerado, é muito bem delimitada por Mullainathan e Spiess (2017) que enxerga o sucesso do aprendizado de máquina como uma mudança de paradigma da busca por soluções de problemas. Iniciamos de um ponto de partida procedural, buscando deduzir todas as regras, que compõem o problema, para uma abordagem empírica, utilizando métodos indutivos para extrair relações de dados de forma automática; ou seja, os dados revelam suas informações constituintes e relacionais.

[1] Deep learning models

Estudo de Caso: Indústria de Cimento - Demanda

O cimento é um material de consumo tão intenso que pode ser considerado um dos pilares da sociedade moderna. Sendo o ingrediente chave na produção de concreto, está presente em todos os esforços construtivos relacionados à infraestrutura e moradia da civilização humana. Seu consumo mundial em 2020 foi equivalente a aproximadamente 4,16 bilhões de toneladas (SNIC, 2020), tornando-se, portanto, um dos materiais mais consumidos pela sociedade global. Sua importância é tão significativa que pode, por si só, servir como indicador de crescimento econômico de um país.

É de grande interesse para a indústria de cimento aumentar sua eficácia no atendimento da demanda, considerando que o cimento é um material com um prazo de validade relativamente curto, noventa dias, quando armazenado nas condições necessárias (NBR 16697, 2018). Após esse período, é possível reensaiá-lo, mas, caso não atenda às especificações necessárias, deve ser descartado. Isso impõe à indústria fornecedora a necessidade de um planejamento estratégico eficaz. Tanto a demanda quanto o posicionamento de suas plantas industriais e locais de armazenamento precisam ser cuidadosamente estudados.

No entanto, essa produção apresenta um aspecto negativo, pois gera uma das principais contribuições para as emissões antropogênicas de CO_2 e desperdício de resíduos sólidos, especialmente na ausência de políticas eficientes de coprocessamento[2] (Araujo, 2020).

O aprimoramento da eficácia na produção e distribuição do cimento é de vital importância, não apenas para a indústria fabricante, mas também para toda a sociedade. À medida que as indústrias aperfeiçoam suas cadeias de produção, melhorando sua logística e reduzindo o desperdício, temos, não somente benefícios da redução de custos, mas também contribuímos para a diminuição da poluição. Isso se torna ainda mais relevante em uma indústria que atende a um mercado com uma tendência de crescimento constante (Ighalo e Adeniyi, 2020).

Objetivo desse trabalho

O objetivo deste trabalho é buscar uma melhoria contínua de desempenho de um modelo, baseado em rede neural, de previsão do consumo de cimento por unidade da federação. Levando em consideração grande número de fatores que compõem o domínio do problema e técnicas e boas práticas na construção e treinamento do modelo preditivo.

Não é objetivo, desse trabalho, uma análise do estado da arte do uso de aprendizado de máquina e comparação entre diversos tipos de modelos preditivos, desde os menos sofisticados até os mais avançados. Já partimos do que a literatura atual considera como os modelos de melhor performance, dentre os neurais, para previsão de séries temporais, ou seja, modelos de rede neural recorrente como LSTM, stacked-LSTM e Bidirectional-LSTM (Sangiorgio e Dercole, 2020) (Barrera-Animas *et al.*, 2022) sendo a rede tradicional feed-forward somente usada devido sua compatibilidade com a biblioteca SHAP.

[2] O coprocessamento é uma técnica na qual os resíduos são empregados para substituir matérias-primas e/ou combustíveis na indústria de cimento.

Capítulo 1

Teoria

1.1 Séries temporais

1.1.1 Características inerentes

Séries temporais são dados coletados através de observações sequenciais em um espaço de tempo com intervalos pré-definidos e organizados de forma cronológica. Independentemente se a variável observada for de natureza contínua ou discreta, a essência da série temporal demanda algum processo de discretização.

Podemos selecionar três características como atributos principais das séries temporais:

- **Tendências**: Pode ser definida como a identificação de um comportamento de variação que ocorre quando analisada uma janela de eventos de tamanho considerável. Em uma análise unidimensional, como o consumo de cimento, esse comportamento, quando identificado, pode ser de queda, crescimento ou estagnação[1].
- **Sazonalidade**:Apesar do contexto mais amplo da palavra estar ligado a estações do globo terrestre, podemos estender esse conceito para uma série de padrões que são repetidos nos valores das séries temporais em intervalos cíclicos, esse padrão não necessita ser uniforme, contudo pode ser identificado com agregação no suposto ciclo, tendo seus padrões e variâncias evidenciados.
- **Ruído**:São flutuações irregulares cuja ocorrência não pode ser contextualizada pela tendência e sazonalidade. Dessa forma, dentro do conjunto de dados, elas podem ser descartadas; no entanto, sempre há a possibilidade de que façam parte integrante de um padrão não evidenciado ou de um conjunto de dados insuficiente para tal análise.

[1] Podendo ser identificado, ainda, o comportamento de queda ou crescimento como sendo linear, exponensial ou logarítmico.

1.1.2 Construções de séries temporais

Sendo dados coletados através de observações sequenciais em um espaço de tempo com intervalos pré-definidos, independentemente se a variável observada for de natureza contínua ou discreta, a essência da série temporal demanda algum processo de discretização. Esse processamento pode ser feito, de acordo com Chatfield (2000) dos seguintes modos:

i. Sendo suas amostras coletadas em intervalos de tempo arbitrários para o caso de variáveis de natureza contínua.

ii. Por agregação de dados coletados de forma irregular em um intervalo de tempo regular. Como, por exemplo, informações de consumo ou venda de um produto.

iii. Ou quando a variável observada já é naturalmente discreta e eventual. Por exemplo, pagamentos mensais parciais realizados sobre um montante a base de juros.

A grande maioria das variáveis observadas e delimitadas por uma série temporal apresentam um elevado grau de dependência entre passado e futuro, portanto, sua análise fundamentalmente deve levar em conta sua ordenação temporal.

1.1.3 Extrapolação como predição

Prever o futuro, seja com uma sofisticada rede neural ou uma simples presunção que o último valor observado se repetira, é uma forma de extrapolação de dados. Uma extrapolação gera, seguindo algum tipo de lógica, uma continuidade de informações na série temporal em um período cronológico não observado[2]

Independente de qual modelo seja usado sobre o dado presente, a extrapolação sempre assume uma hipótese: a de que o conjunto de padrões, independente dos mesmos terem sido identificados ou não, se repetirão nos dados extrapolados, isto é, que as nuancias intrínsecas dos dados observados e suas relações não mudaram durante o período cujo dado foi extrapolado. Essa particularidade sempre coloca a predição de dados em um território instável. Contudo, por termos o conhecimento dessa inevitabilidade, podemos também organizar o planejamento em torno dessa variabilidade considerando os erros de medidas e suas médias estatísticas.

A análise das séries temporais, segundo Chatfield (2000), tem quatro objetivos primordiais:

1. **Descrição**: Descrever os dados pelos seus atributos estatísticos ou visualmente através de gráficos.

2. **Modelagem**: Procurar uma representação matemática para o fenômeno observado que mimetize o seu comportamento esperado através de representações numéricas. Quando não conseguimos exprimir esse modelo deterministicamente por uma formula convencional temos, um caso muito adequado para o uso do recurso de aprendizado de máquina como ferramenta principal. (Abu-mostafa *et al.*, 2012).

[2] Isso é valido tanto para extrapolar o futuro quanto o passado.

3. **Previsão**: Com um modelo que represente bem o comportamento esperado do domínio estudado, podemos partir, não somente para a previsão de valores futuros na sequencial temporal, como também, a análise das variáveis que constituem a base problema.
4. **Controle**: Com um modelo de boa previsão e conhecimento dos elementos que determinam o resultado, torna-se possível adequar tomadas de decisões de acordo.[3]

1.2 Pré-processamento de dados

1.2.1 Substituição por último valor observado, sentinela ou moda

A mais simples substituição é quando repetimos o último valor de dado observado ou o próximo valor observado. Essas substituições apresentam custo computacional ínfimo, e são indicados a casos com poucos valores faltantes ou onde as questões temporais, de tendência e sazonalidade, não foram identificadas ou não se formalizam na granularidade de dados escolhida. Em um ambiente mais classificatório, onde os valores de dados não apresentam forte correlação com o passado, podemos optar por uma substituição por moda, ou valor mais comum dentro do ciclo ou estação, em um caso em que a sazonalidade é evidente.

No aprendizado de máquina, nos casos de modelos mais sofisticados, as vezes, podemos nos dar ao luxo de confiar na capacidade do modelo captar essas ausenciais sejam elas pontuais ou de janelas de dados faltantes. Esse tipo de imputação acrescenta um valor de fora do domínio (e.g.: -1) nos valores faltantes e é otimista em relação a capacidade do modelo reconhecer esses valores como ausentes de correlação e então ignorá-los nos pesos dos resultado de sua saída.

1.2.2 Interpolação linear

A interpolação linear consiste em imputar dados baseados em uma função linear entre dois pontos de dados observados em torno da sequência faltante e que estejam imediatamente antes e depois da mesma[4]. O Segmento de reta é delimitado segundo a fórmula:

$$y = y_{\text{anterior}} + (x - x_{\text{anterior}}) \cdot \left(\frac{y_{\text{posterior}} - y_{\text{anterior}}}{x_{\text{posterior}} - x_{\text{anterior}}} \right) \tag{1.1}$$

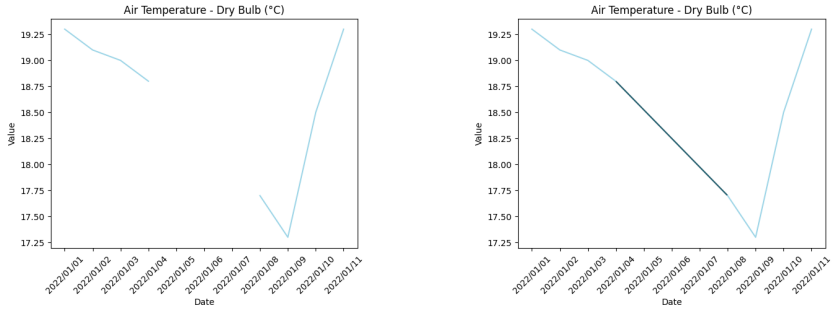
Onde:

- x é o ponto dentro da sequência cronológica que queremos estimar,

[3] Para o caso do cimento, isso permite a indústria adequar suas cadeias de produção de modo a aumentar eficiência e diminuir desperdícios. Permite também um planejamento mais assertivos sobre estratégias de implementação de fábricas ou estratégias de vendas.

[4] Quanto somente queremos interpolar um único ponto, com seu antecessor e sucessor $(x - 1)$ e $(x + 1)$ presentes a interpolação linear age como média de ambos pontos imediatos.

- y é o valor que queremos estimar,
- $x_{anterior}$ e $x_{posterior}$ são as posições dos pontos imediatos observados na ordem cronológica,
- $y_{anterior}$ e $y_{posterior}$ são os valores dos pontos imediatos, anterior e posterior.



(a) *Série com dados faltantes.* **(b)** *Dados imputados por interpolação linear.*

Figura 1.1: *Exemplo de interpolação linear*

A imputação por interpolação linear valoriza os dados observados ao passo que produz uma relação linear entre eles que age de forma similar a uma média ponderada. Esse método é eficiente computacionalmente e não gera ruídos com os dados novos. Contudo, a interpolação linear ignora completamente contextos mais amplos e próprios das séries temporais como as tendências e sazonalidades já discutidas.

1.2.3 Interpolação Polinomial

Outro tipo de interpolação por substituição de valores seguindo a regra de uma função é a interpolação polinomial. Esse tipo de imputação gera um polinômio:

$$P(x) = a_0 + a_1x + a_2x^2 + ... + a_dx^d \tag{1.2}$$

Utilizando parte dos pontos observados anteriores ao trecho faltante e parte dos pontos posteriores ao mesmo.

Existem vários métodos para interpolação polinomial, e um deles envolve a exploração da estrutura da matriz de Vandermonde para obter soluções numericamente estáveis em operações $O(n^2)$, em oposição à complexidade $O(n^3)$ da eliminação gaussiana. Esta abordagem é apenas uma das muitas técnicas usadas no ajuste polinomial. [5]

O método de criação desses polinômios, usado neste trabalho, através da biblioteca *NumPy: Numerical Python* (2022), envolve o método dos mínimos quadrados. Em que produramos reduzir o resultado da fórmula baixo:

[5] É importante mencionar que os métodos específicos empregados por bibliotecas como o `numpy.polyfit` do *NumPy: Numerical Python* (2022) para cálculo polinomial podem ser diferentes e muitas vezes são otimizados tanto para eficiência quanto para estabilidade numérica. Portanto, os detalhes precisos do método utilizado podem variar entre diferentes implementações de software.

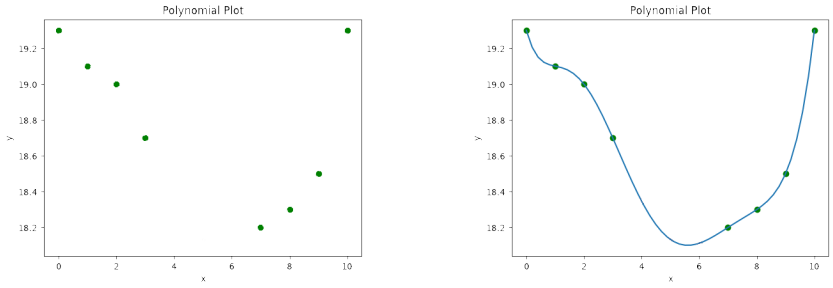
$$E = \sum_{i=1}^{n} [y_i - P(x_i)]^2 \tag{1.3}$$

- n é o número de dados disponibilizados.
- x_i e y_i são as coordenadas de cada um desses pontos no plano cartesiano.
- $P(x_i)$ é o valor previsto pelo polinômio a ser servido como equação base.
- E é, portanto, a média dos erros quadrados de cada ponto.

Isso gera os coeficientes de modo a diminuir o erro entre o polinômio e os pontos de dados observados oferecidos como base. Com esses coeficientes, geramos o polinômio da seguinte forma:

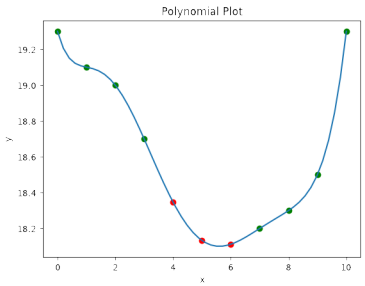
$$P(x) = (x - c_1)(x^2 - c_2)\dots(x^n - c_n) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n \tag{1.4}$$

Com essa função em mãos, simplesmente imputamos os valores de $P(x_f)$ para cada x_f faltante.



(a) *Série com dados faltantes.* **(b)** *Interpolação polinomial criada.*

Figura 1.2: *Exemplo de interpolação polinomial*



(a) *Série com dados faltantes.*

Figura 1.3: *Exemplo de interpolação polinomial.*

A grande vantagem da interpolação polinomial frente à linear é que ela considera o contexto, ou melhor, as tendências posteriores e anteriores dos dados presentes. Deste modo, temos uma interpolação que leva em conta tendências; no entanto, ela ainda ignora completamente sazonalidades.

1.2.4 Transformações

Transformações de dados estão mais ligadas a um pré-processamento com o intuito de melhorar o desempenho, ou seja, facilitar o modelo preditivo a ser treinado, a capturar as variações e relações entre os dados fornecidos e auxilia na de coordenação de atualizações em muitas camadas (Goodfellow *et al.*, 2017).

A normalização, também conhecida como escala mínimo-máximo[6], dimensiona uma série de dados para um intervalo específico, geralmente entre 0 e 1. Sua fórmula é:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1.5}$$

Onde:

- X_{norm} é o valor normalizado.
- X é o dado original da sequência selecionada.
- $\min(X)$ é o menor valor encontrado nessa sequência.
- $\max(X)$ é o maior valor encontrado nessa sequência.

A normalização mantém a relação entre os dados intacta, apenas os comprimindo em um intervalo bem definido, o que é útil quando o modelo precisa dessas relações proporcionais preservadas. No entanto, um grande inconveniente do uso da normalização é a distorção causada por ruídos (outliers) presentes nos dados. Essa distorção pode prejudicar o entendimento das variações entre a série de dados pelo modelo.

A padronização, também conhecida como escalonamento z-score, realiza uma transformação nos dados com o objetivo de tornar o conjunto resultante com média 0 e desvio padrão 1. Isso é alcançado com a seguinte operação:

$$Z = \frac{X - \mu}{\sigma} \tag{1.6}$$

Sendo:

- Z é o z-score.
- X é o dado original da sequência selecionada.
- μ é a média dos valores da sequência.
- σ é o desvio padrão do conjunto de valores da sequência.

A padronização é menos afetada por outliers em comparação a normalização e, portanto, é mais indicada para séries de dados mais ruidosas. No entanto, a relação entre os dados é distorcida, o que resulta em perda de interpretabilidade da sequência. A escolha entre essas transformações deve levar em consideração a natureza dos dados e suas relações. Em muitos

[6] Min-max scaling

casos, a escolha é feita de forma empírica, observando e selecionando a transformação que resulta no melhor desempenho preditivo do modelo (Qiao *et al.*, 2019).

1.2.5 Extrapolação por regressão linear

Vimos, anteriormente, que a previsão do futuro é um tipo de extrapolação. No entanto, em certas ocasiões, precisamos imputar dados relacionados ao passado. Uma decisão tomada para o treinamento é o intervalo temporal dos dados a serem fornecidos ao modelo para aprendizado. No entanto, nem todos os dados desse conjunto podem cobrir o intervalo requerido. Por exemplo, podemos decidir que o intervalo de treinamento será de dados referentes aos meses de 2003 até 2019. No entanto, um dos tipos de dados escolhidos para compor o conjunto de treinamento tem informações disponíveis somente no intervalo entre 2005 e 2019. Dessa forma, podemos decidir entre imputar esses dados com valores indicativos de ausência (1.2.1) ou podemos extrapolá-los usando métodos como o de regressão linear.

Cada situação pode requerer uma abordagem diferente. No entanto, por motivos práticos, não queremos gastar tempo extrapolando uma sequência que faz parte do conjunto de entradas, a fim de reservar mais tempo para a previsão do valor alvo em si (futuro). Uma simples regressão linear é, portanto, um método simples de extrapolação que pode ser eficaz para um curto período e possui um custo computacional relativamente baixo.

$$y = b_0 + b_1 x \tag{1.7}$$

Sendo:

- y é a variável de determinado ponto que desejamos imputar.
- x é o ponto da série temporal onde a previsão deve ocorrer.
- b_0 é o valor da variável onde a reta intercepta o eixo y.
- b_1 é a inclinação (slope) da equação linear.

A chave da regressão está em determinar b_0 e b_1 de forma a obter uma reta cujo erro quadrado em relação aos pontos seja o menor possível. A fórmula para o cálculo da inclinação é:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.8}$$

Onde:

- n é o número de dados presentes.
- x_i e y_i são as coordenadas de cada um desses pontos no plano cartesiano.
- $\bar{x}$ e $\bar{y}$ são as médias de posição na série e seu valor, respectivamente, de cada dado presente no conjunto.

O conjunto de dados usado para gerar a regressão não necessariamente precisa ser o conjunto total disponível. É possível escolher uma janela menor mais próxima ao intervalo a ser extrapolado para priorizar, desse modo, as relações entre os dados mais imediatos a esse intervalo.

1.2.6 Recuo de entradas e média móvel

Em muitos casos, desejamos reduzir o ruído na sequência de dados das séries temporais usada como entrada para o modelo de previsão. No entanto, também queremos preservar as variações dos dados de modo a não perder informações cíclicas e outras pequenas nuances. Um recurso interessante para isso é a média móvel. A média móvel substitui os dados por uma média de valores pré-estabelecida que naturalmente inclui o próprio valor.

O tipo mais comum de média móvel é a média móvel simples (SMA - Simple Moving Average). Ela leva em consideração, para cada dado, o próprio valor e uma janela anterior. O valor da média é imputado na posição, para cada posição, como a média de todos esses dados:

$$SMA(t) = \frac{X(t) + X(t-1) + X(t-2) + ... + X(t-n+1)}{n} \tag{1.9}$$

Sendo:

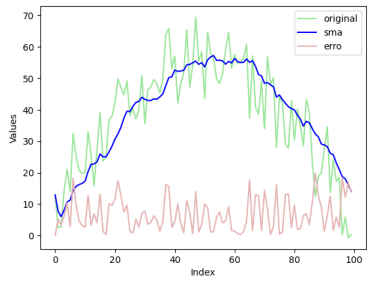
- $SMA(t)$ é o valor da média móvel simples para a posição t.
- $X(t)$ é o valor real para o dado na posição t na sequência.
- n é o número de pontos delimitados pela janela de dados escolhida.

A realização da média móvel causa um efeito colateral denominado recuo de entradas (lagged features). Esse recuo às vezes é desejado quando um tipo de entrada tem uma correlação maior com o dado alvo quando está atrasada. Quando esse atraso não é desejado, talvez, por outros procedimentos de recuo já terem sido realizados, podemos considerar o uso de uma média móvel centrada. Sua única diferença em relação à simples é que ela leva em consideração o passado e o futuro na mesma proporção, ou seja, o valor alvo encontra-se no meio da janela de sequência escolhida.

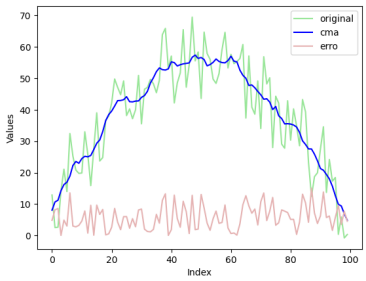
$$CMA(t) = \frac{nX(t-1) + X(t) + X(t+1) + ... + X(t+n-1)}{n} \tag{1.10}$$

Onde $CMA(t)$ é o valor da média móvel centrada para a posição t.

Essa operação tende a perder seu efeito à medida que o intervalo escolhido se torna indisponível, pois é maior do que os dados presentes, ou seja, perto do limite atual da série temporal.



(a) *Média móvel comum (SMA[a]).*



(b) *Média móvel centrada (CMA[a]).*

[a] Simple Moving Average

[a] Centered Moving Average

Figura 1.4: *Exemplo de médias móveis*

1.2.7 Imputação com sazonalidade

Como já vimos, características inerentes a séries temporais são as tendências, sazonalidade e ruídos. Ao identificarmos esses fatores podemos subdividi-los no que é denominado de decomposição de séries temporais (Chatfield, 2000). Os dados de séries temporais são modelados, simplificadamente, pela relação:

$$X_t = \mu_t \cdot i_t \cdot \varepsilon_t \tag{1.11}$$

Aqui estão as componentes deste modelo:

- μ_t representa a tendência ou componente de longo prazo.
- i_t é a componente de sazonalidade.
- ε_t denota a perturbação aleatória ou residual no tempo t.

Para facilitar a análise e modelagem, este modelo multiplicativo pode ser transformado em uma forma totalmente aditiva usando uma transformação logarítmica:

$$\log(X_t) = \log(\mu_t) + \log(i_t) + \log(\varepsilon_t) \tag{1.12}$$

Existem diversas formas de algoritmos que buscam separar e delimitar essas contribuições. Contudo, identificado um padrão cíclico, por exemplo o referente ao um ano, podemos então calcular a média de seus valores sazonais em escala de trimestre ou mês possibilitando uma imputação nesse sentido caso tenhamos algum valor-base confiável referente ao ciclo delimitado (o valor de consumo de cimento por ano, por exemplo).

1.3 Aprendizado de máquina

1.3.1 Definição e arcabouço para o aprendizado de máquina

Qualquer domínio cujo comportamento pode ser representado por um modelo de aprendizado de máquina (Abu-mostafa *et al.*, 2012) deve possuir um espaço X, denominado espaço de entrada, um conjunto D de pares entrada-saída no formato $(x_1, y_1), \dots, (x_n, y_n)$, e uma função desconhecida f no formato $f : X \to Y$. Modelos de aprendizado de máquina são aqueles que, com técnicas recursivas de treinamento, geram um modelo $g : X \to Y$ que se aproxima de f. O conjunto de todos os modelos, independentemente de quão bem se aproximam de f, é denominado conjunto de Hipóteses H.

Em relação ao conjunto D, quando temos conhecimento de todos os pares (x_i, y_i), ou seja, para cada x_i, conhecemos os valores de y_i, estamos em uma situação de aprendizado supervisionado. Nesse caso, o modelo é treinado tendo disponível para seu algoritmo de treinamento os alvos y_i correspondentes a cada entrada x_i.

No caso de um modelo de previsão de séries temporais, como neste trabalho, temos uma situação de aprendizado supervisionado para o treinamento. A previsão de valores futuros, por outro lado, não se encaixa nessa classificação. No entanto, desejamos usar x_j antes de termos os y_j correspondentes, em uma situação externa ao treinamento. Denominamos essa aplicação, que no nosso caso se refere à previsão de valores futuros com base em entradas do presente e do passado, de "teste fora da amostra"ou "teste no conjunto externo", cujos erros compõem o conjunto E_{out}.

Para buscar o melhor modelo no conjunto H, treinamos os modelos utilizando uma amostra conhecida, e os erros dentro da amostra são chamados de E_{in} (erros dentro da amostra). Os valores (x_i, y_i) são conhecidos, mas não são utilizados pelo algoritmo de treinamento para ajuste de parâmetros. O modelo que apresenta o melhor desempenho, de acordo com os resultados de E_{in}, é então escolhido para ser testado em E_{out}.

Independentemente dos resultados de treinamento em E_{in}, o modelo só demonstra sua qualidade se tiver um bom desempenho em E_{out}. Esse é o objetivo fundamental de praticamente todas as aplicações de aprendizado de máquina em situações de previsão de séries temporais.

1.3.2 Aprendizado vs. Design

Uma distinção pertinente explicada por Abu-mostafa *et al.* (2012), é a diferença de aprendizado e design. Quando consideramos regras de aplicação pré-determinadas, baseadas em particularidades, factuais ou não, do domínio estudado e a aplicamos no processo preditivo, seja na totalidade ou em alguma parcela de um fluxo procedural, temos então um design. Nesse design, ainda podem ser considerados parte do que hoje definimos como inteligência artificial que faz parte do aprendizado automático[7]. O aprendizado extrai suas relações dos dados oferecidos automaticamente, enquanto as regras de aplicação de design são delimitadas com algum conhecimento prévio.

[7] O aprendizado de máquina é amplamente considerado um subconjunto do que é conhecido como inteligência artificial (IA)

1.4 Alguns modelos de aprendizado de máquinas

1.4.1 Rede Neural Feed Forward - FFNN

Uma rede neural artificial, ou perceptron de multicamada, como o próprio nome sugere, possui uma rede conectada com mais de uma camada de perceptrons interconectados. As entradas são recebidas por uma camada de entrada que transmite o sinal de cada entrada para cada perceptron da camada seguinte, até chegar na camada final de saída, responsável por devolver o resultado da computação. Denominamos as camadas entre as camadas entrada e saída de camadas escondidas.

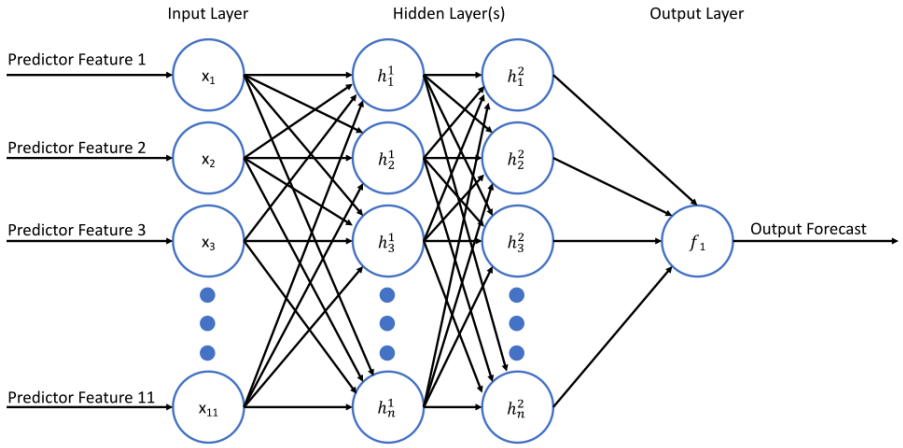


Figura 1.5: *Rede neural feed-forward*[a]*. Figura de* Barrera-Animas et al. *(2022)*

[a] Predictor Feature n (Tipo de dado de entrada); Input Layer (Camada de entrada); Hidden Layer (Camadas ocultas) e Output Layer (Camada de saída)

O nome "feed-forward"é aplicado a essa estrutura devido ao fato de que cada camada posterior se alimenta da saída da camada anterior em um fluxo, que apesar de ser ramificado, é direcionado. Esse aumento de complexidade permite ao modelo captar um maior número de relações entre os dados de entrada.Isso ocorre devido ao valor do dado de entrada ter uma trajetória muito mais variada, contribuindo de diferentes formas em diferentes fluxos de informação numérica.

$$a_i = \sigma(W_i \cdot a_{i-1} + b_i) \tag{1.13}$$

Onde:

- a_i é o vetor de ativação na camada i, representando a saída da camada.
- W_i é a matriz de pesos associada à camada i.
- a_{i-1} é o vetor de ativação da camada anterior (camada $i-1$).
- b_i é o vetor de viés da camada i.
- σ é a função de ativação.

O aprendizado automático ocorre através da alterações dos pesos da rede através da operação de gradiente descendente, que pode ser definido como um algoritmo que procura minimizar medidas de erro através de iterações contínuas que navegam pela superfície do espaço de erro em direção a um mínimo. Uma função de erro que pode ser usada como base é a do erro quadrado:

$$L_i = e_i^2 = (y_i - \hat{y}_i)^2 \tag{1.14}$$

Onde:

- L_i é o valor da função de perda para a predição i.
- e_i é o erro da predição i.
- y_i é o valor alvo ou valor observado.
- $\hat{y}_i$ é o valor previsto ou resultado do modelo.

Talvez, a forma mais comum de gradiente descendente seja o estocástico [8]. Importante não confundir com os otimizadores usados em bibliotecas de treinamento, muitos otimizadores, como o Adam, utilizam gradiente descendente estocástico combinando-o com diversas outras medidas. Esse gradiente descendente atualiza um parâmetro da rede neural na forma da regra:

$$w_{\text{novo}} = w_{\text{antigo}} - \eta \cdot \frac{\partial L}{\partial w} \tag{1.15}$$

Onde:

- w_{novo} é o novo peso atualizado.
- w_{antigo} é o peso antigo usado na atualização.
- η é a taxa de aprendizagem, que corresponde à proporção da atualização do peso em cada iteração.
- $\frac{\partial L}{\partial w}$ é o gradiente, que é computado usando a regra da cadeia sobre a função de perda (loss function).

O cálculo de redução do erro pode ser representado por uma superfície (para o caso multivariado) de erro onde há vales, cumes, pontos de cela e platôs. Dentre esses valores, alguns têm seus pontos de menor valor definidos como mínimos locais. Esses mínimos locais são extremamente indesejáveis para o processo de redução de erro, pois nos desviam do potencial de melhora que se configura como mínimo global.

[8] Stochastic Gradient Descent - SGD

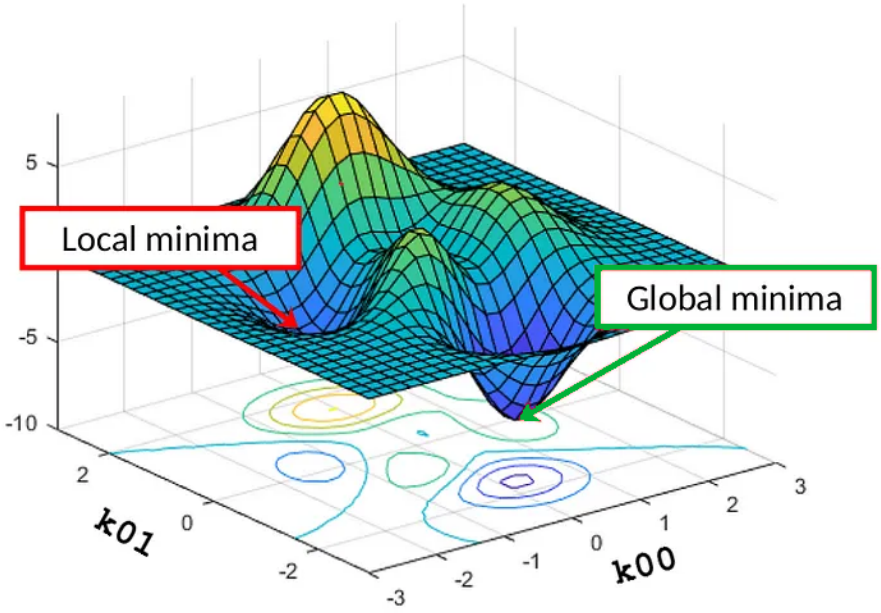


Figura 1.6: *Superfície da função de custo e exemplo de pontos de minimo local e global. Figura retirada de* Towards AI *(2023)*

O comportamento do SGD[9] é iterativo e envolve atualizar os valores até ser alcançada uma convergência. O nome "estocástico"é atribuído devido a não ser considerado todo conjunto de parâmetros em cada iteração. Esse procedimento origina ruído na atualização de parâmetros que pode ser favorável para variabilidade do processo, diminuindo as chances do valor convergir para um mínimo local.

1.4.2 Rede Neural Recorrente - RNN

Um tipo sofisticado de Rede Neural são as redes neurais recorrentes (RNN). Sua principal diferenciação se dá por meio de uma estrutura de fluxo de dados que promove a criação de uma espécie de memória. Esses fluxos funcionam como loops que registram valores cujos pesos são sensíveis ao contexto ou sequência dos dados de entrada.

Uma formalização simplificada de uma RNN pode ser definida da seguinte forma:

$$H_t = f(H_{t-1}, X_t) \tag{1.16}$$

$$\hat{Y}_{t+h|t} = g(H_t) \tag{1.17}$$

Onde:

- H_t é o estado denominado oculto (*hidden state*).
- f e g são funções passíveis de serem definidas.

[9] Stochastic Gradient Descent

- $\hat{Y}_{t+h|t}$ é a predição referente ao valor Y_{t+h}, levando em consideração o último estado t.

Essa formulação $\hat{Y}_{t+h|t}$ é a ideia chave da dependência sequencial que as redes recorrentes capturam.

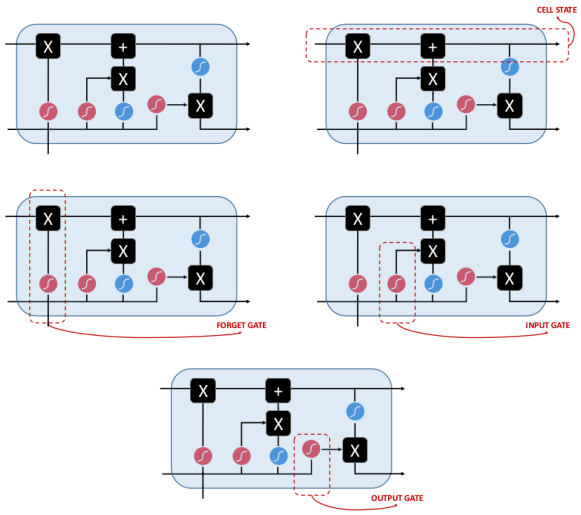


Figura 1.7: *Delimitações da Rede LSTM. Figura de* Masini *et al. (2023)*

A principal variante desse tipo de rede é denominada Rede LSTM (Long Short-Term Memory). Na LSTM, esse fluxo de loop fechado é controlado dinamicamente e é sensível à ordem das entradas.

Suas ramificações de fluxo são:

- *Input gate*: responsável por decidir quais informações serão adicionadas ao estado da célula atual C_t utilizando o estado anterior H_{t-1} e a nova entrada X_t através de uma função de ativação.
- *Forget gate*: Ele é responsável por delimitar quais informações do estado anterior C_{t-1} serão esquecidas e quais continuarão no fluxo de "memória".
- *Output gate*: Ele delimita quais informações serão passadas para o estado oculto (*hidden state*) e quais seguirão para o próximo estado de célula C_t.

1.5 Problemas comuns de treinamento de modelos

1.5.1 Overfitting

Segundo Abu-mostafa *et al.* (2012), overfitting é um fenômeno que ocorre quando encontramos um resultado adequado ou uma explicação para um problema que representa muito bem um cenário conhecido. Contudo, esse modelo perde a validade quando exposto a dados inéditos. Em outras palavras, o overfitting ocorre quando um modelo é escolhido no conjunto de hipóteses H devido à sua grande performance ou, melhor dizendo, baixo valor de erro E_{in}. No entanto, quando submetido a uma série de dados inéditos externos, temos uma baixa performance ou erro E_{out} alto. O problema é que muitas vezes recorremos

ao aprendizado de máquina quando o problema é muito complexo para o uso de técnicas mais triviais, e Abu-mostafa *et al.* (2012) afirma que existe um aumento probabilístico de overfitting quando há um aumento na complexidade do alvo.

Portanto, os três principais fatores que favorecem o overfitting são:

- Complexidade do alvo: um alvo complexo e com muitas nuances favorece o overfitting, pois, assim como um polinômio de alto grau é difícil de ser encontrado, é fácil que alguns dos seus pontos coincidam com outro polinômio.
- Conjunto reduzido de dados de treinamento: Um conjunto reduzido oferece pouca informação necessária para o modelo evitar o overfitting, ao passo que também é mais fácil "encaixar"um modelo em uma série de dados pequena.
- Ruído nos dados de entrada: Entradas incorretas de dados também contribuem para o overfitting devido à dificuldade do modelo em reconhecer esses *outliers*.

As redes neurais são muito eficientes em encontrar uma hipótese adequada para um conjunto de dados fornecidos para treinamento. Deste modo, o overfitting torna-se um dos principais problemas na busca de um modelo que tenha um bom desempenho em uma amostra inédita. Muitos dos outros problemas que veremos na sequência são definidos como tal devido a propiciarem a ocorrência de overfitting.

1.5.2 Dados insuficientes

Quando dispomos de poucos dados para treinar nossos modelos, enfrentamos sérios riscos de desenvolver um modelo hipotético suscetível ao overfitting. Isso ocorre porque não há amostras suficientes para que o modelo seja capaz de capturar as veradeiras relações entre os dados do domínio do problema. O modelo treinado nessas condições tenderá a gerar relações erradas ou correlações que, com a entrada de dados novos, acabem mostrando-se deturpadas. Por exemplo, um tipo de dado de entrada pode apresentar uma forte correlação com o dado alvo no trecho temporal delimitado para treinamento. Entretanto, somente com o avanço das observações na série cronológica, esse dado pode revelar-se não correlacionado no contexto real de causa e efeito. O modelo treinado sob essa limitação é propenso a capturar essa correlação inicial e tornar-se excessivamente dependente, em termos de valores de resultados, dessa entrada enganosa.

1.5.3 Dados imprecisos

Dados são a matéria-prima de todo modelo de aprendizado de máquina supervisionado. O modelo de aprendizado ajusta seus parâmetros de modo a melhor representar a relação $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ de entradas e saídas. Portanto, a quantidade, bem como a qualidade desses dados de entrada, é essencial para o treinamento do modelo. Contudo, não há total certeza se os dados de entrada e a relação que o modelo gera entre eles refletem bem o comportamento do domínio ao qual o modelo busca reproduzir. Uma variável de dado de entrada com forte correlação com o dado de saída pode contribuir para que o modelo tenha uma boa performance de treino. No entanto, somente em sua extensa aplicação em amostras externas, ela será julgada como fator determinante de causa-consequência em relação à variável alvo. Dados incorretos introduzem ruídos no modelo durante o

treinamento, comprometendo sua performance ao distorcer os parâmetros. Isso ocorre devido a uma série de dados imprecisos que não representam adequadamente nenhum aspecto do domínio.

1.6 Boas Práticas de Treinamento de Modelos de Aprendizado de Máquina

1.6.1 Entendendo o domínio do problema

O primeiro passo em qualquer empreendimento de treinamento de modelos de aprendizado de máquina é adquirir conhecimento sobre o domínio em que o alvo a ser previsto está inserido. Esse esforço é fundamental por diversos aspectos. A escolha de dados relevantes dentro do domínio estudado aumenta as chances do modelo captar relações interessantes entre eles, aproximando o resultado previsto da realidade (Aggarwal, 2018).

Ter um entendimento aprofundado do problema facilita a escolha das técnicas mais apropriadas para processar os dados. Por exemplo, em certos tipos de dados, a presença de ruído pode ser relevante, especialmente quando a variância é pequena. No entanto, o conhecimento prévio pode indicar que essas supostas variáveis de ruído podem interferir nos resultados.

Outro caso relevante, que é muito usado especialmente mercado financeiro (Krollner *et al.*, 2010), e que também relacionado ao conhecimento prévio, ocorre quando uma variável candidata a parâmetro de treinamento tem seu maior efeito, no resultado preditivo, quando a mesma é atrasada em algum intervalo conhecido. Isso se deve à natureza temporal das séries temporais e seus efeitos no domínio. Por exemplo, uma variável correspondente a investimentos em educação pode ser atrasada em alguns anos quando em um cenário relacionada a previsão de avanços tecnológicos ou consumo de livros.

1.6.2 Escolha de dados de entrada

A escolha dos tipos de dados a serem utilizados, como mencionado anteriormente, depende do conhecimento sobre o domínio do problema. No entanto, existem outros aspectos relacionados à escolha de dados específicos que devem ser cuidadosamente considerados para evitar problemas que possam resultar em distorções na predição do modelo.

Em primeiro lugar, os dados devem ser provenientes de fontes confiáveis, com seus processos de obtenção, processamento e publicação bem documentados. As distorções causadas por dados incorretos podem ser difíceis de mensurar devido à dificuldade em diagnosticar a causa do problema, e os dados incorretos podem não ser tão evidentes quanto dados ausentes.

Outras considerações relevantes para a escolha dos dados estão relacionadas ao seu processamento. A granularidade dos intervalos de amostragem dos dados, no caso de séries temporais, geralmente envolve unidades de medida temporais, como horas, dias, semanas, meses e anos. É importante considerar a granularidade mais apropriada de acordo

com o objetivo da predição. No entanto, uma granularidade mais fina sempre oferece mais oportunidades para modelos de aprendizado automático, pois eles se beneficiam do aumento da quantidade de dados fornecidos para treinamento [10].

1.6.3 Métodos de regularização

Métodos de regularização são procedimentos técnicos e sistemáticos que auxiliam modelos de aprendizado a evitar o overfitting, promovendo melhor desempenho de generalização. A maioria dos métodos envolve a modificação do comportamento do algoritmo de treinamento para penalizar certas condições ou impor restrições que o forçam a se comportar de maneira mais benéfica.

> "Regularização é qualquer modificação que fazemos em um algoritmo de aprendizagem que visa reduzir seu erro de generalização, mas não seu erro de treinamento."(Goodfellow *et al.*, 2017)

Dentre os procedimentos, destacam-se os que penalizam as normas dos parâmetros. Esses métodos adicionam uma penalidade de norma na forma de um parâmetro $\Omega(\theta)$. A fórmula geral é a seguinte:

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta) \tag{1.18}$$

Onde:

- $\tilde{J}$ é a função objetivo regularizada com seus parâmetros ajustados.
- α, com $\alpha \in [0, \infty)$, é um hiperparâmetro que atua como o coeficiente de contribuição do termo de penalidade de norma Ω na função objetivo.
- θ é o tamanho dos parâmetros.

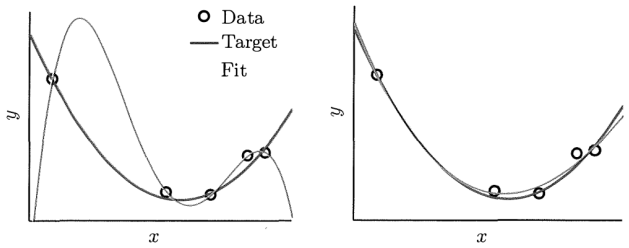


Figura 1.8: *Exemplo de correção de overfitting por regularização*[a]*. Figura retirada de Abu-mostafa et al. (2012)*

[a] O Gráfico a esquerda representa um polinômio desenvolvido que, apesar de adequar-se perfeitamente aos dados (Data), claramente não representa bem o alvo real (Target). A figura a direita, por sua vez, aproxima muito melhor o resultado desejado, devido a regularização.

[10] Aggarwal (2018) considera, como os dois fatores mais proeminentes para o avanço do uso de ML o aumento do poder computacional e o aumento ao acesso de dados.

É importante ressaltar que esses procedimentos de penalização ocorrem na maioria dos casos sobre os pesos W das redes neurais e não sobre seus bias. Dois regularizadores de parâmetros que trabalham com penalidades são a Regularização de Parâmetros L2, também conhecida como decaimento de peso [11], e a Regularização de Parâmetros L1, que tende a propiciar uma solução considerada mais esparsa (Goodfellow *et al.*, 2017), com alguns parâmetros tendendo a zero e, portanto, atua como uma selecionadora de entradas.

Outras formas de regularização impõem restrições aos modelos durante sua execução. Um exemplo é a técnica chamada dropout, que aleatoriamente redefine o valor de um conjunto preestabelecido de variáveis para zero, promovendo diversificação do modelo, pois o modelo não pode contar com a mesma variável de entrada em cada iteração[12] de aprendizado.

1.6.4 Validação

Para termos um maior grau de certeza de que o modelo está performando bem em um ambiente de testes que não foi usado no treinamento, ou seja, está generalizando bem, podemos realizar esse tipo de teste durante o treinamento na chamada validação cruzada[13]. Nesse método, os dados são divididos em partes não sobrepostas, de modo que a cada iteração do treinamento, o modelo seja testado com uma parte dos dados separados para teste. Esse algoritmo favorece um monitoramento da performance do modelo, que é mais valioso do que um cálculo de perda sobre uma entrada que foi usada para o treinamento, evitando, portanto, o overfitting.

A base da validação consiste em inserir dados como entrada em um modelo treinado, ou em processo de treinamento, e medir o erro de sua saida frente ao valor real[14] com alguma métrica. Existem diversas formas de medirmos esse erro, algumas delas são:

- MAE (Mean Absolute Error): Calcula a média dos erros absolutos registrados em cada previsão. Sua fórmula é:

$$MAE = \frac{1}{n} \sum |Y - \hat{Y}|, \tag{1.19}$$

  Onde n é o número de amostras, Y é o dado real e $\hat{Y}$ é o dado previsto.

- MAPE (Mean Absolute Percentage Error): Similar ao MAE, contudo, seu erro medido em cada predição é percentual. Pode ter seu valor final multiplicado por 100 para ser exibido como porcentagem ou como fração. Seu cálculo é:

$$MAPE = \frac{1}{n} \sum \left| \frac{(\hat{Y} - Y)}{\hat{Y}} \right| \times 100\%. \tag{1.20}$$

[11] Também referida como regressão de crista ou regularização Tikhonov

[12] Denominadas em ML como epochs (épocas)

[13] k-fold cross-validation

[14] Em um caso de aprendizado supervisionado.

- RMSE (Root Mean Squared Error): É uma variante dos cálculos de erro que tende a penalizar maiores variações no cálculo da média de erros, uma vez que eleva ao quadrado a diferença do erro antes de aplicar uma raiz quadrada sobre sua média. RMSE oferece uma informação interessante para os casos em que grandes erros ou outliers não podem ser tolerados, e sua ocorrência acaba sendo diluída em um cálculo de MAE com muitas previsões. Ele pode ser obtido através do cálculo abaixo:

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{Y} - Y)^2}. \tag{1.21}$$

1.7 Explicabilidade

1.7.1 Interpretando Modelos Preditivos

Os modelos neurais estão entre os modelos preditivos de maior complexidade. Entender como eles funcionam em um cenário multivariado pode se tornar uma tarefa muito difícil devido à grande quantidade de permutações do tipo verdadeiro ou falso que teríamos que aplicar sobre suas variáveis de entrada, bem como à sua presença ou ausência. Os benefícios de um modelo interpretável são valiosos e diversificados, pois fornece indicadores sobre a importância de determinadas variáveis em relação a outras, auxiliando nos processos de decisões, tanto para os mantenedores do modelo quanto, principalmente, para aqueles que trabalham no domínio ao qual o modelo foi projetado.

1.7.2 Valores SHAP

Valores SHAP atuam como um conjunto de medidas que buscam avaliar a importância atribuída a cada dado de entrada.

Para um determinado modo de previsão, o valor SHAP para uma entrada i é calculado da seguinte forma:

$$\phi(i) = \sum_{z' \subseteq N} \frac{|z'|!(M - |z'| - 1)!}{M!}[f_x(z') - f_x(z' \setminus \{i\})] \tag{1.22}$$

Onde:

- $\phi(i)$ é o valor SHAP para a variável de entrada i.
- $f_x(z')$ é a previsão do modelo considerando um subconjunto de entradas z'.
- $f_x(z' \setminus \{i\})$ é a previsão do modelo para um subconjunto de entradas z' sem a entrada i.
- z' representa um subconjunto do conjunto de tipos de entrada N. No contexto dos valores SHAP, z' é um conjunto de recursos ou variáveis que está sendo considerado. A fórmula calcula o valor de Shapley para o recurso i somando as diferenças sobre todos os subconjuntos possíveis z' que contêm o recurso i.

- M representa o número total de recursos no conjunto de tipos de entrada N.

Não é objetivo deste trabalho aprofundar nas propriedades e teoremas que levaram a essa expressão. Entretanto, caso haja interesse, o trabalho de Lundberg e Lee (2017) as delimita detalhadamente. A figura abaixo ilustra como esses valores se relacionam:

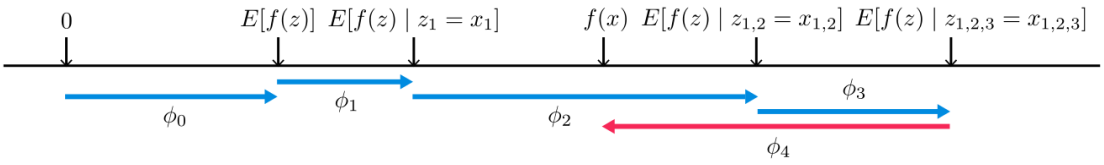


Figura 1.9: *Figura explicativa de valores SHAP (Lundberg e Lee, 2017).*

Na figura acima, ϕ de 0 a 4 são os valores SHAP, $f(x)$ é o resultado da previsão do modelo, e $E[f(z)]$ é o valor-base, ou seja, o valor do resultado caso todas as variáveis estivessem presentes e com valores médios. Então $E[f(z)|z_i = x_i]$ é o valor aproximado do modelo quando a variável z_i é definida como entrada com o valor x_i. Dessa forma, o valor SHAP ϕ_i é a diferença $E[f(z)|z_i = x_i] - E[f(z)]$. Essas esperanças são sensíveis à ordem em que os tipos de entrada são adicionados, e assim os valores SHAP surgem da média dos valores de ϕ_i em todas as ordenações possíveis.

Capítulo 2

Metodologia

2.1 Entendimento e decisões baseadas na dinâmica do consumo de cimento

Como já foi introduzido, a boa previsão de demanda de cimento por localidade estratégica é essencial para uma produção e distribuição eficaz da indústria de cimento. As decisões de projeto deste trabalho foram tomadas levando em consideração tanto as condições da indústria de cimento quanto a necessidade de predição, a disponibilidade de dados públicos e os limites de tempo inerentes a um trabalho de pouco menos de um ano.

O consumo de cimento é um evento cuja pontualidade é imprevisível, deste modo o Sindicato Nacional da Indústria de Cimento (SNIC, 2020) reúne as informações disponibilizadas pelos seus participantes e utiliza agregação[1] para gerar uma série temporal de consumo de cimento por mês por estado da e federação brasileira.

Para a previsão, foram escolhidos os modelos de redes neurais do tipo perceptron multicamadas [2] e as do tipo LSTM. Como as dinâmicas dos estados não são as mesmas, a previsão de consumo de cimento para cada estado contou com sua própria rede neural. Essa escolha foi tomada com o intuito de facilitar o modelo a captar relações próprias de cada estado, partindo do pressuposto de que determinada variável pode não apresentar o mesmo efeito no valor alvo em um modelo de previsão para um determinado estado federativo em comparação com outro. [3]

[1] A agregação já foi definida em 1.1.2

[2] Outro modo de referir-se a Rede Neural Feed-Forward

[3] modelos complexos de redes neurais usados nesse trabalho são capazes de "aprender"essas nuances, contudo ao separarmos os casos manualmente, garantimos um esforço a menos, diminuindo a complexidade do problema

2.1.1 Divisão estadual

Como vimos, a indústria carece de previsões precisas e bem definidas quanto ao seu espaço geográfico, sendo a sua setorização essencial para o planejamento de implantações de plantas industriais e ajustes logísticos de transporte. Neste trabalho, foi escolhida a delimitação em unidades federativas da união, ou seja, estados. Isso se deve ao fato de que o SNIC já fornece dados de consumo de cimento por unidade da federação com granularidade mensal e à disponibilidade da maioria dos dados públicos relacionados a índices socioeconômicos pertinentes à questão estudada, que apresentam seus valores divididos por estados como a divisão mínima[4].

2.1.2 Recuo de um ano

Uma previsão de valores futuros não pode contar com dados do futuro para ser formulada. Precisamos prever o futuro com base nas informações disponíveis no presente e no passado. Portanto, todos os dados usados para a previsão são defasados em um intervalo de um ano. Esse procedimento também nos aproxima de um ambiente real de treinamento contínuo e previsão de intervalos de tempo futuros. Ao termos informações de dados referentes ao ano x, podemos iniciar a previsão de valores para o ano $x + 1$.

2.2 Preparação dos dados e organização

Este trabalho aproveitou os dados de entrada definidos no trabalho anterior feito pela aluna Julia Leite [5], também sob a supervisão do Prof. Dr. Marcelo Finger. Portanto, existem 15 colunas de dados de entrada relacionados a índices de desenvolvimento, índices socioeconômicos e valores quantitativos relacionados à população e produção de cimento.

Além disso, procuramos adicionar uma série de novos dados de entrada para fins experimentais. O consumo de cimento é um fenômeno complexo, e muitas vezes não sabemos quem são seus fatores influentes e como eles afetam o resultado.

Portanto, foram adicionados dois tipos de séries de dados. Primeiramente, incluímos dados meteorológicos por estado. Isso foi feito por duas razões principais. Primeiro, suspeitamos que variações nas condições dos fenômenos atmosféricos podem influenciar o consumo de cimento, afetando sua durabilidade em situações em que os materiais não são armazenados adequadamente [6] ou porque foram vendidos no varejo e os compradores não seguiram as precauções de armazenamento adequadas. Em segundo lugar, um modelo treinado com esses parâmetros se torna mais sensível às variações sazonais associadas aos fenômenos atmosféricos, aumentando a probabilidade de o modelo capturar as variações sazonais no consumo de cimento em si.

---

[4] Alguns dados do IBGE podem ser encontrados em divisões de municípios, mas o processamento desses dados em escalas menores resultaria em um esforço desproporcional ao escopo da disciplina à qual este trabalho está inserido

[5] "Previsão de consumo de cimento nos estados do Brasil usando métodos de aprendizado automático- https://linux.ime.usp.br/~jleite/mac0499/

[6] Por exemplo, quando um depósito não atende às especificações das normas brasileiras

Outra série de dados adicionada refere-se a dados econômicos relacionados ao poder de compra e endividamento da população. Essa inclusão de dados econômicos se baseia na análise dos relatórios anuais sobre o consumo e produção de cimento elaborados pelo SNIC. Atualmente, é amplamente reconhecido que o consumo de cimento está intrinsecamente ligado às condições econômicas da nação, e que o poder de investimento, até mesmo em nível individual, é um fator determinante nesse contexto.

2.2.1 Dados e fontes

Abaixo, segue a tabela de dados iniciais:

Dado	Unidade	Divisão	Granularidade	Período Original
Consumo de cimento	mil ton.	Estadual	Mensal	2003-2022
Produção de cimento	mil ton.	Estadual-	Mensal	2003-2022
Valor do Cimento[a]	R$/Kg	Estadual	Mensal	1994-2022
Desemprego	%	Estadual	Trimestral	2002-2016 e 2014-2022
IDH	-	Estadual	Anual	1970-2021
PIB - Estadual	R$ (mil.)	Estadual	Anual	1939-2020
PIB - Construção Civil	R$ (mil.)	Estadual	Anual	1939-2020
PIB - Estadual per capita	R$ (mil)[b]	Estadual	Anual	1985-2022
PIB - Preços de Mercado	R$ (mil.)	Estadual	Anual	1939-2020
NFSP	% PIB	Nacional	Mensal	1999-2022
ELCF	R$ de 2010 (mi.)	Nacional	Anual	1947-2022
IGP-DI	%	Nacional	Mensal	1944-2022
INCC	%	Nacional	Mensal	1944-2022
IPCA	-[c]	Nacional	Mensal	1994-2022
População	Habitante	Nacional	Anual	1872-2022
SELIC	%	Nacional	Diário	1986-2022

Tabela 2.1: *Tabela de informações dos dados iniciais*

[a] Portland 32

[b] Preços do ano 2010

[c] Número-Indice (1993=100)

Esses dados iniciais são os mesmos utilizados no trabalho anterior. Optamos por reutilizá-los devido ser uma escolha adequada visto que eles representam um conjunto tradicional de indicadores socioeconômicos de acesso público muito utilizados em diversas áreas de pesquisa.

Dado	Método[a] I	Método II	Fonte	URL
Consumo de cimento	IS	-	SNIC	I
Produção de cimento	-	-	SNIC	I
Valor do Cimento	IP	-	SNIC	I
Desemprego	IP	-	IBGE + BCB	II + III
IDH	IP	IL	IPEA	II
PIB - Estadual	-	-	IBGE	II
PIB - Construção Civil	-	-	IBGE	II
PIB - Estadual per capita	-	-	IBGE	II
PIB - Preços de Mercado	-	-	IBGE	II
NFSP	-	-	BCB	IV
ELCF	-	-	BCB	IV
IGP-DI	-	-	FGV	II
INCC	-	-	FGV	II
IPCA	IP	-	IBGE	V
População	IP	-	IBGE	II
SELIC	IL	-	IBGE	VI

IS	Imputação com Sazonalidade
IL	Interpolação Linear
IP	Interpolação Polinomial
CMA	Média Móvel Centrada[b]

I	http://www.cbicdados.com.br/
II	http://www.ipeadata.gov.br/Default.aspx
III	https://www.bcb.gov.br/estatisticas
IV	https://dadosabertos.bcb.gov.br/
V	https://www.ibge.gov.br/estatisticas/
VI	https://www.debit.com.br/tabelas/

Tabela 2.2: *Tabela de informação de métodos de pré-processamento dos dados iniciais*

[a] Método de pré-processamento usado para imputação sobre os dados de entrada

[b] Centered moving average - OBS.: Todos os dados, exceto o alvo consumo de cimento, foram processadas por média móvel centrada

Os dados meteorológicos foram todos retirados do INMET [7], abaixo a relação:

Dado	Unidade	Divisão	Granularidade	Período Original
Precipitação	mm/dia	Estadual	Diário	(2001|2008)-2022
Pressão atmosférica	mB	Estadual	Diário	(2001|2008)-2022
Radiação global	kj/m²	Estadual	Diário	(2001|2008)-2022
Temperatura do ar	°C	Estadual	Diário	(2001|2008)-2022
Temp. ponto de orvalho	°C	Estadual	Diário	(2001|2008)-2022
Precipitação	%	Estadual	Diário	(2001|2008)-2022
Precipitação	m/s	Estadual	Diário	(2001|2008)-2022

Tabela 2.3: *Tabela de informações dos dados do INMET*

Todos os dados meteorológicos foram retidados das bases públicas do INMET[8] e foram processadas por uma mistural de regressão linear, imputação por interpolação polinomial e, finalmente, média móvel centrada.

Os dados econômicos adicionais, ou séries expandida, foram os da tabela seguinte:

Dado	Unidade	Divisão	Granularidade	Período Original
Custo m²	R$	Estadual	Mensal	1999-2022
Depósito Poupança	R$	Estadual	Anual	2003-2022
IDH - Educação	-	Estadual	Anual	1970-2021
IDH - Longevidade	-	Estadual	Anual	1970-2021
IDH - Renda	-	Estadual	Anual	1970-2021
NFSP - Fluxo Mensal	R$ (milhões)	Nacional	Mensal	1999-2022
Operações de crédito[a]	R$ (milhões)	Nacional	Mensal	1990-2022
EMBI Risco-Brasil	-	Nacional	Diário	1994-2022
FGTS	R$ (mil)	Nacional	Mensal	1988-2022
Operações de crédito[b]	-[c]	Nacional	Mensal	2010-2022
PIB[d]	R$ (mil)	Nacional	Mensal	1996-2022
PPC[e]	R$ (mil)	Nacional	Anual	1940-2022

Tabela 2.4: *Tabela de informações dos dados iniciais*

[a] Saldo da carteira de crédito

[b] Inadimplência da carteira de crédito

[c] Número-Indice (1995=100)

[d] Atividades imobiliárias

[e] Salário mínimo - paridade do poder de compra (PPC)

[7] Instituto Nacional de Meteorologia

[8] https://portal.inmet.gov.br/

Dado	Método I	Método II	Fonte	URL
Custo m²	-	-	SIDRA	VII
Depósito Poupança	IL	-	BCB	II
IDH - Educação	IP	IL	IPEA	II
IDH - Longevidade	IP	IL	IPEA	II
IDH - Renda	IP	IL	IPEA	II
NFSP - Fluxo Mensal	-	-	BCB	IV
Operações de crédito	-	-	IBGE	II
EMBI Risco-Brasil	-	-	J.P. Morgan[a]	II
FGTS	-	-	IBGE	II
Operações de crédito	IP	IL	BCB	II
PIB[b]	-	-	IBGE	II
PPC[c]	-	-	IPEA	II

IL	Interpolação Linear
IP	Interpolação Polinomial
CMA	Média Móvel Centrada[d]

II	http://www.ipeadata.gov.br/Default.aspx
IV	https://dadosabertos.bcb.gov.br/
VII	https://sidra.ibge.gov.br/

Tabela 2.5: *Tabela de informação de métodos de pré-processamento dos dados iniciais*

[a] Banco de investimento multinacional americano e empresa de serviços financeiros - https://www.jpmorgan.com.br/pt/about-us

[b] Atividades imobiliárias

[c] Salário mínimo - paridade do poder de compra (PPC)

[d] Centered moving average - OBS.: Todos as entradas foram processadas por médica móvel centrada

2.2.2 Escolha de granularidade e janela de ação

A escolha da granularidade decorreu da granularidade mínima oferecida pela fonte de dados de consumo de cimento, ou seja, a fornecida pelo SNIC. Desta forma, todas as séries temporais dos dados de entrada usados para a predição forma postos dessa forma ou processados para isto.

A janela de ação, ou janela escolhida para treinamento e predição, foi a a de início do ano de 2003 até o final do ano de 2019. Essa escolha se estabeleceu devido a maior presença de dados confiáveis durante esse período e pelo mesmo ser relativamente mais estável quando comparado a décadas, que foram marcadas por elevada inflação e adoção de politicas econômicas expressivas como a do plano real.

2.3 Pré-processamento do alvo de predição

A coluna de dados de consumo de cimento é o nosso alvo de predição. Portanto, é essencial não alterá-la. No entanto, durante os anos de 2014, 2015 e 2016, não possuímos

informações detalhadas na granularidade mensal. A única informação disponível é o consumo total anual.

Para lidar com essa falta de dados mensais, desenvolvemos um procedimento de imputação que mantém os atributos essenciais da série temporal, preservando a precisão do consumo anual. Como mencionado anteriormente, a sazonalidade é um atributo crucial das séries temporais. Dada a informação anual de consumo, nosso objetivo foi reconstruir uma distribuição mensal que fosse consistente com o comportamento sazonal dos dados observados anteriormente.

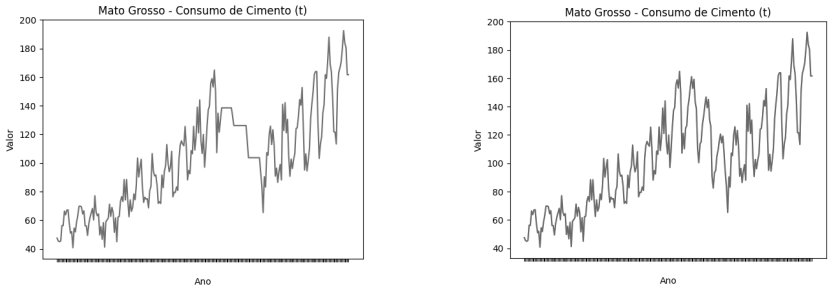
O procedimento seguiu as etapas a seguir:

Primeiramente, calculamos a média dos valores de consumo de cimento para o conjunto de dados, agrupados por mês. Isso nos forneceu uma espécie de sazonalidade média, considerando que o de cimento geralmente segue um ciclo anual. O cálculo foi o seguinte:

$$C_{\text{input}}(m, a) = C_{\text{anual}}(a) \times \frac{M_{\text{sazonal}}(m)}{\sum_{i=0}^{11} M_{\text{sazonal}}(i)} \tag{2.1}$$

Onde:

- $C_{\text{input}}(m, a)$ é o valor a ser imputado para o mês referente ao consumo do ano observado.
- $C_{\text{anual}}(a)$ é o consumo de cimento referente ao ano "a"disponível.
- $M_{\text{sazonal}}(m)$ é a média de consumo de cimento relativa ao mês "m"calculada dentro do intervalo observado.
- $\sum_{i=0}^{11} M_{\text{sazonal}}(i)$ é a soma de todas as médias mensais para cada um dos doze meses "i".



(a) *Séria temporal com dados imputados por média.* **(b)** *Dados imputados com sazonalidade média.*

Figura 2.1: *Exemplo de interpolação linear*

Esse gráfico mostra "a"mostra a imputação por média para o consumo de cimento do estado do Mato Grosso havendo ausência dos dados na granularidade de mês entre os anos de 2014 e 2016. A média do consumo anual é dividia pelos meses criando platôs nas séries. Em contrapartida, a imputação "b"leva em conta a média de sazonalidade e, devido sua fórmula, mantêm o valor anual intacto.

2.4 Análise de resultados e adição de dados

Para cada experimento realizado, foram coletados as amostras das previsões em intervalos de tempo pré-determinados sendo os seus valores de entradas nunca utilizadas pelo modelo durante o treinamento. Para cada unidade federativa, foram realizadas previsões mensais sendo seus erros absolutos calculados. Esses erros, por sua vez, foram agrupados nos intervalos estabelecidos de maneira a podermos gerar as estatísticas de MAE, MAPE e RSME sobre o conjunto de valores previstos por estado.

2.5 Tecnologias utilizadas

Neste trabalho[9] foi utilizada uma coleção de software bem comuns no campo da ciência de dados e modelagem baseada em aprendizado de máquina. Abaixo uma lista com as principais:

[9] Os arquivos de código que constituíram este trabalho estão disponíveis em https://github.com/Pedro84skynet/TCC-MAC0499.

Tecnologia	Descrição
Python:	Uma das linguagens de alto nível mais utilizadas devido sua elevada versatilidade e acessibilidade. Sua gama de bibliotecas auxiliares a coloca entre as principais linguagem para trabalhos com ciência de dados. (*Python Programming Language* 2023)
NumPy:	Biblioteca veloz, por ser escrita em C, e fundamental para computações numéricas complexas em Python. Oferece suporte a arrays multidimensionais e diversas funções matemáticas. (*NumPy: Numerical Python* 2022)
Pandas:	Principal biblioteca para manipulação e análise de dados, facilitando operações com seus dataframes que são estruturas de dados tabulares. (*Pandas: Python Data Analysis Library* 2023)
Jupyter Notebook:	Ambiente interativo para desenvolvimento de código, muito adequada para quando necessitamos de feedback visual ágil e experimentações. (*Jupyter Notebook* 2023)
TensorFlow:	Um dos principais softwares de código aberto usado para o desenvolvimento de modelos de aprendizado de máquina e redes neurais. Muito versátil, com uma ampla coleção de métodos para diversos tipos de modelos e casos. (*TensorFlow: An Open Source Machine Learning Framework for Everyone* 2023)
SHAP:	Ferramenta relativamente nova desenvolvida para o estudo de explicabilidade de modelos de aprendizado de máquina. Seu principal uso se formaliza nas estimativas de valor Shapley, calculando a contribuição de cada tipo de entrada para a predição. (*SHAP (SHapley Additive exPlanations)* 2023)
Matplotlib:	Biblioteca gráfica de visualização de dados em Python, oferecendo diversas construções visuais como gráficos e plots. (*Matplotlib: Visualization with Python* 2023)
Statsmodels:	Biblioteca estatística em Python muito usada para estimar erros de modelos, incluindo regressões, e análises de séries temporais. (*Statsmodels: Statistical Models in Python* 2023)

Tabela 2.6: *Descrição das Tecnologias Utilizadas*

Capítulo 3

Experimentos e Resultados

3.1 Pré-processamento por mês

3.1.1 Treinamento com pré-processamento interpolação e média móvel

3.2 Explicabilidade usando métodos SHAP

3.3 Adição de dados meteorológicos

3.4 Adição de dados extras

3.5 Interpretação final dos resultados

Capítulo 4

Conclusão

4.1 Avaliação pessoal e próximos passos

Referências

[Abu-mostafa *et al.* 2012] Yaser S. Abu-mostafa, Malik Magdon-ismail e Hsuan-tien Lin. *Learning from Data: A Short Course*. AMLBook.com, 2012 (citado nas pgs. viii, 4, 12, 16, 17, 19).

[Aggarwal 2018] Charu C. Aggarwal. *Neural networks and deep learning*. Springer, 2018 (citado nas pgs. 18, 19).

[Araujo 2020] Geraldo Jose Ferraresi de Araujo. "O coprocessamento na indústria de cimento: definição, oportunidades e vantagem competitiva". *Revista Nacional de Gerenciamento de Cidades* 8.57 (2020), pp. 52–61 (citado na pg. 2).

[Barrera-Animas *et al.* 2022] Ari Yair Barrera-Animas, Lukumon Oyedele, Juan Manuel Davila Delgado e Lukman Adewale Akanbi. "A comparative analysis of modern machine learning algorithms for time-series forecasting". *Machine Learning with Applications* 7 (2022) (citado nas pgs. viii, 2, 13).

[Chatfield 2000] Chris Chatfield. *Time-series forecasting*. CRC press, 2000 (citado nas pgs. 1, 4, 11).

[Goodfellow *et al.* 2017] Ian Goodfellow, Yoshua Bengio e Aaron Courville. *Deep Learning*. 22ª ed. The MIT Press, 2017 (citado nas pgs. 8, 19, 20).

[Ighalo e Adeniyi 2020] Joshua O. Ighalo e Adewale George Adeniyi. "A perspective on environmental sustainability in the cement industry". *Waste Disposal and Sustainable Energy* 2.3 (jan. de 2020), pp. 161–164. doi: 10.1007/s42768-020-00043-y (citado na pg. 2).

[*Jupyter Notebook* 2023] *Jupyter notebook*. Versão 6.5.4. 2023. url: https://jupyter.org (citado na pg. 31).

[Krollner *et al.* 2010] Bjoern Krollner, Bruce Vanstone e Gavin Finnie. "Financial time series forecasting with machine learning techniques: a survey" (2010), pp. 25–30 (citado na pg. 18).

[Lundberg e Lee 2017] Scott M. Lundberg e Su-In Lee. "A unified approach to interpreting model predictions". *Advances in neural information processing systems* 30 (2017) (citado nas pgs. viii, 22).

[Masini *et al.* 2023] Ricardo P. Masini, Marcelo C. Medeiros e Eduardo F. Mendes. "Machine learning advances for time series forecasting". *Journal of economic surveys* 37.1 (2023), pp. 76–111 (citado nas pgs. viii, 16).

[*Matplotlib: Visualization with Python* 2023] *Matplotlib: visualization with python*. Versão 3.8.0. 2023. URL: https://matplotlib.org (citado na pg. 31).

[Mullainathan e Spiess 2017] Sendhil Mullainathan e Jann Spiess. "Machine learning: an applied econometric approach". *Journal of Economic Perspectives* 31.2 (2017), pp. 87–106. DOI: 10.1257/jep.31.2.87 (citado na pg. 1).

[NBR 16697 2018] NBR 16697. *Cimento Portland - Requisitos*. 2018 (citado na pg. 2).

[*NumPy: Numerical Python* 2022] *Numpy: numerical python*. Versão 1.21.3. 2022. URL: https://numpy.org (citado nas pgs. 6, 31).

[*Pandas: Python Data Analysis Library* 2023] *Pandas: python data analysis library*. Versão 2.0.3. 2023. URL: https://pandas.pydata.org (citado na pg. 31).

[*Python Programming Language* 2023] *Python programming language*. Versão 3.9.17. 2023. URL: https://www.python.org (citado na pg. 31).

[Qiao *et al.* 2019] Siyuan Qiao, Haoran Wang, Chao Liu, Wei Shen e Alan Yuille. "Micro-batch training with batch-channel normalization and weight standardization". *arXiv preprint arXiv:1903.10520* (2019) (citado na pg. 9).

[Rajkomar *et al.* 2019] Alvin Rajkomar, Jeffrey Dean M.D. e Isaac Kohane. "Machine learning in medicine". *New England Journal of Medicine* 380.14 (mai. de 2019), pp. 1347–1358. DOI: 10.1056/NEJMra1814259 (citado na pg. 1).

[Sangiorgio e Dercole 2020] Matteo Sangiorgio e Fabio Dercole. "Robustness of lstm neural networks for multi-step forecasting of chaotic time series". *Chaos, Solitons & Fractals* 139.6 (out. de 2020), pp. 726–728. DOI: 10.1016/j.chaos.2020.110045 (citado na pg. 2).

[*SHAP (SHapley Additive exPlanations)* 2023] *Shap (shapley additive explanations)*. Versão 0.42.1. 2023. URL: https://github.com/slundberg/shap (citado na pg. 31).

[SNIC 2020] SNIC. *Relatório Anual Sindicato Nacional da Indústria do Cimento*. http://snic.org.br/numeros-relatorio-anual.php. Accessed: 20/11/23. 2020 (citado nas pgs. 2, 23).

[*Statsmodels: Statistical Models in Python* 2023] *Statsmodels: statistical models in python*. Versão 0.14.0. 2023. URL: https://www.statsmodels.org (citado na pg. 31).

[*TensorFlow: An Open Source Machine Learning Framework for Everyone* 2023] *Tensorflow: an open source machine learning framework for everyone*. Versão 2.13.0. 2023. URL: https://www.tensorflow.org (citado na pg. 31).

[Towards AI 2023] Towards AI. *Deep Learning from Scratch in Modern C: Gradient Descent*. Accessed on 10/November/2023. 2023. URL: https://pub.towardsai.net/deep-learning-from-scratch-in-modern-c-gradient-descent-670bc5889112 (citado nas pgs. viii, 15).