UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

OTÁVIO PIGOZZO MARTELLI

MODELOS LSTM PARA PREDIÇÃO DO PREÇO DA SOJA COM BASE EM DADOS CLIMÁTICOS BRASILEIROS

TRABALHO DE CONCLUSÃO DE CURSO

PATO BRANCO
2021

OTÁVIO PIGOZZO MARTELLI

MODELOS LSTM PARA PREDIÇÃO DO PREÇO DA SOJA COM BASE EM DADOS CLIMÁTICOS BRASILEIROS

Trabalho de Conclusão de Curso apresentado ao Curso de engenharia de computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Jefferson Tales Oliva
Universidade Tecnológica Federal do Paraná

Coorientador: Ives Rene Venturini Pola
Universidade Tecnológica Federal do Paraná

PATO BRANCO
2021

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEP. ACADEMICO DE INFORMATICA DAINF-PB

TERMO DE APROVAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO - TCC

MODELOS LSTM PARA PREDIÇÃO DO PREÇO DA SOJA COM BASE EM DADOS CLIMÁTICOS BRASILEIROS

Por

Otavio Pigozzo Martelli

Monografia apresentada às 14 horas 00 min. do dia 17 de agosto de 2021 como requisito parcial, para conclusão do Curso de Engenharia da Computação da Universidade Tecnológica Federal do Paraná, Campus Pato Branco. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação e conferidas, bem como achadas conforme, as alterações indicadas pela Banca Examinadora, o trabalho de conclusão de curso foi considerado APROVADO.

Banca examinadora:

Prof. Dr. Dalcimar Casanova	Membro
Profa. Dra. Viviane Dal Molin de Souza	Membro
Prof. Dr. Jefferson Tales Oliva	Orientador
Profa. Dra. Viviane Dal Molin de Souza	Professor(a) responsável TCCII

Documento assinado eletronicamente por (Document electronically signed by) **JEFFERSON TALES OLIVA**, **PROFESSOR(A) ORIENTADOR(A)**, em (at) 17/08/2021, às 19:47, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 4º, § 3º, do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por (Document electronically signed by) **DALCIMAR CASANOVA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 17/08/2021, às 20:12, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 4º, § 3º, do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por (Document electronically signed by) **VIVIANE DAL MOLIN DE SOUZA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 17/08/2021, às 22:35, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 4º, § 3º, do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site (The authenticity of this document can be checked on the website) https://sei.utfpr.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador (informing the verification code) **2201968** e o código CRC (and the CRC code) **C681A048**.

Referência: Processo nº 23064.034853/2021-39 SEI nº 2201968

Criado por vivianesouza, versão 2 por vivianesouza em 17/08/2021 19:28:28.

Dedico este trabalho a minha família e amigos,
em especial a minha avó.

RESUMO

PIGOZZO MARTELLI, Otávio. Modelos LSTM para predição do preço da soja com base em dados climáticos brasileiros. 2021. 39 f. Trabalho de Conclusão de Curso – Curso de engenharia de computação, Universidade Tecnológica Federal do Paraná. Pato Branco, 2021.

A importância do Brasil no cenário agrícola é inquestionável, sendo um dos maiores produtores de soja do mundo. Com uma crescente produtividade, o Brasil tem alcançado cada vez mais destaque no âmbito mundial. O clima influencia de forma ativa na produção de soja, entretanto, vale lembrar que o Brasil é um pais continental. Por essa razão, os dados climáticos se diferenciam de uma região a outra. O objetivo deste trabalho é realizar uma análise preditiva do preço da soja usando dados climáticos brasileiro e identificar a influência que o clima do Brasil tem sobre o preço internacional de soja. Para isso, foi obtida uma base de dados climáticos e dados complementares, tais como, o valor do dólar, a inflação e a produção anual de soja. Esses dados foram coletados no período de 2000 a 2020 de todo o território brasileiro. Para a avaliação experimental, foram escolhidos os dados climáticos dos estados com maior produção de soja, sendo eles o Mato Grosso, Mato Grosso do Sul, Goiás, Paraná e Rio Grande do Sul. Após o pré-processamento desses dados, modelos preditivos foram construídos utilizando o método *Long Short Term Memory* (LSTM). De forma que, o método foi aplicado para cada estado e duas configurações diferentes (uma com dados climáticos e outra sem dados climáticos), ao total foram construídos dez modelos. Na avaliação de desempenho, foram calculados o *Root Mean Squared Error* (RMSE) e o *Mean Absolute Error* (MAE) para cada modelo. Neste caso, os modelos construídos utilizando os dados do Paraná atingiu os melhores resultados para ambas configurações nas duas métricas. Após isso, testes estatísticos de hipótese foram aplicados e nenhum deles demonstrou diferença estatisticamente significativa entre os modelos, ou seja, a utilização de dados climáticos brasileiros não influenciam significativamente no preço internacional da soja. Por mais que os resultados tenham demonstrado que dados climáticos brasileiros não tiveram influência no preço do soja na avaliação experimental, este trabalho teve diversas contribuições, tais como a criação de uma base de dados climática para utilização em outros trabalhos e a análise preditiva do preço da soja, possibilitando auxílio aos agricultores em tomadas de decisão.
Palavras-chave: Inteligência Artificial. Aprendizado de Máquina. Clima Brasileiro. Modelos Preditivos. Long Short Term Memory. LSTM. Redes Neurais Artificiais. Séries Temporais.

ABSTRACT

PIGOZZO MARTELLI, Otávio. LSTM models for prediction of soybeans price, based on brasilian climate data. 2021. 39 f. Trabalho de Conclusão de Curso – Curso de engenharia de computação, Universidade Tecnológica Federal do Paraná. Pato Branco, 2021.

The importance of Brazil in the agricultural scenario is unquestionable, being one of the largest soy producers in the world. With a growing productivity, Brazil has achieved more and more prominence in the world scope. The climate actively influences soy production, however, it is worth remembering that Brazil is a continental country. For this reason, climate data differ from one region to another. The objective of this work is to carry out a predictive analysis of soybean prices using Brazilian climate data and to identify the influence that Brazil's climate has on the international price of soybeans. For this, a climate database and complementary data were obtained, such as the value of the dollar, inflation and the annual production of soy. These data were collected from 2000 to 2020 from the entire Brazilian territory. For the experimental evaluation, climate data from the states with the highest soy production were chosen, namely Mato Grosso, Mato Grosso do Sul, Goiás, Paraná and Rio Grande do Sul. After the pre-processing of these data, predictive models were built using the *Long Short Term Memory* (LSTM) method. So, the method was applied to each state and two different configurations (one with climate data and the other without climate data), in total ten models were built. In the performance evaluation, the *Root Mean Squared Error* (RMSE) and the *Mean Absolute Error* (MAE) were calculated for each model. In this case, the models built using Paraná data achieved the best results for both configurations in the two metrics. After that, statistical hypothesis tests were applied and none of them showed a statistically significant difference between the models, that is, the use of Brazilian climate data does not significantly influence the international price of soy. As much as the results have shown that Brazilian climate data had no influence on the price of soybeans in the experimental evaluation, this work had several contributions, such as the creation of a climate database for use in other studies and the predictive analysis of the price of soybeans, enabling help for producers in decision-making.

Keywords: Artificial Intelligence. Machine Learning. Brazilian Climate. Predictive Models. Long Short Term Memory. LSTM. Artificial Neural Networks. Time Series.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EST	Estação Meteorológica
IT	Insolação Total
PT	Preciptacão Total
TCM	Temperatura Compensada Média
TMM	Temperatura Miníma Média
URM	Umidade Relativa Média
PS	Preço da Soja
IPCA	Índice Nacional de Preços ao Consumidor Amplo
EL	El Niño
LA	La Niña
CO	Colheita
RNA	Redes Neurais Artificiais
KDD	Knowledge Discovery in Databases
IoT	Internet of Things
PCA	Principal Component Analysis
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
RAE	Relative Absolute Error
MSE	Mean Square Error
RBF	Radial Basis Function
GA	Genetic Algorithm
GD	Gradient Descent
GDGA	Gradient Descent and Genetic Algorithms
LSTNet	Long-and Short-term Temporal Patterns with neuralNetworks

SUMÁRIO

1 INTRODUÇÃO

O Brasil está entre os maiores produtores de soja no mundo, com uma produtividade crescente desse grão. No entanto, em alguns períodos climáticos severos, como secas ou geadas (EBC, 2020), a produção de grãos pode sofrer impactos negativos, como a redução da sua produção (BRASILAGRO, 2019). Todavia, sabendo que o Brasil possui uma extensa área territorial, o clima é diversificado em sua extensão, favorecendo a produtividade da soja em algumas regiões em que o clima é mais adequado (THOMPSON, 1988). Por exemplo, regiões em que temperaturas oscilam entre 20ºC e 30ºC faz com que o desenvolvimento da planta seja de forma mais rápida e uniforme. Porém, para solos com temperaturas acima de 40ºC, o plantio sofre danos que provocam a diminuição das vagens, prejudicando a produção e consequentemente a produtividade (AGEITEC, 2007).

A falta de informações referentes à influência do clima do Brasil na cotação da soja no mercado internacional, das consequências que o clima adverso causa sobre a safra (G1, 2019) e da variação da produtividade devido à instabilidade climática (REUTERS, 2012), tem colaborado para possíveis reduções nos lucros dos produtores, Nesse cenário, este trabalho, além da possibilidade de auxiliar agricultores em processos de tomadas de decisão, propõe quantificar a influência do clima brasileiro na cotação internacional da soja. Sendo assim, qual é o impacto direto que o clima brasileiro tem sobre o preço internacional da *commoditie*? Para responder essa pergunta é necessário o levantamento dos dados climáticos brasileiros de múltiplas estações meteorológicas.

Para obter os dados climáticos brasileiros é necessário a utilização de métodos e ferramentas proveniente dos avanços na área da Tecnologia da Informação, que têm viabilizado o registros de grandes volumes de dados (GOLDSCHMIDT; PASSOS, 2005). Nas bases de dados, a observação manual desses dados é inviável pelos humanos, pois além de demandar muito tempo, também é suscetível a falhas (GOLDSCHMIDT; PASSOS, 2005) Nesse contexto, diversos recursos podem ser utilizados para a obtenção de informações relevantes, como a mineração de dados, que é um processo que explora grandes quantidades de dados e extrai informações relevantes, cujos padrões podem ser utilizados para a geração de modelos preditivos (ZAKI; MEIRA, 2014).

Outra dificuldade é a realização do levantamento das bases de dados de fontes distintas, sendo necessário o uso de *Data Warehouse*, que é um depósito de dados de informações coletadas de diversas fontes e armazenadas em um banco de dados de forma consolidada (HAN; PEI; KAMBER, 2011). Utilizada para sanar alguns eventuais problemas, como o armazenamento de informações dessas fontes distintas em tabelas separadas, das quais, os dados de cada uma devem passar por um pré-processamento dos dados.

Com o *Data Warehouse* estabelecido, os seus dados devem ser pré-processados, para ser utilizada como entrada em diversas aplicações computacionais, como nos modelos preditivos

criados através dos métodos de aprendizado de máquina, os quais, serão aplicados em uma base de teste para a avaliação de desempenho (FACELI et al., 2011). Após, nos resultados da avaliação podem ser aplicados testes estatísticos para a verificação de diferenças significativas entre os modelos.

1.1 Objetivos

1.1.1 Objetivo Geral

Investigar a influência do clima do Brasil sobre o preço da soja internacional, por meio de modelos, construídos utilizando métodos de aprendizado de máquina, capazes de realizar análises preditivas da variação da cotação da soja internacional, baseado em dados climáticos brasileiros.

1.1.2 Objetivos Específicos

- Coleta de dados necessários para a criação de uma base de dados unificada.
- Aplicação de pré-processamento e métodos de aprendizado de máquina para construção de modelos preditivos.
- Avaliar os modelos preditivos por meio de medidas de desempenho e de testes estatísticos.
- Realizar a comparação de desempenho entre os modelos preditivos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

Devido aos avanços tecnológicos, a produção e a capacidade de armazenamento dos dados tem aumentando consideravelmente nos últimos anos. Com os avanços da internet das coisas (*internet of things* – IoT) (PRIESNITZ FILHO et al., 2019), a produção de dados aumentam de forma ainda mais acelerada. Entretanto, a forma como lidamos com os dados tem mudado, pois muitas técnicas e métodos foram propostos para facilitar essa tarefa (GOLDSCHMIDT; PASSOS, 2005). A Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD), que se refere ao processo geral de descoberta de conhecimento útil a partir de dados, a mineração de dados se refere a uma etapa específica desse processo de descoberta (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O KDD é composto por métodos e ferramentas apropriadas capazes de auxiliar na extração e transformação de padrões significativos que podem ser úteis, como em processos de tomadas de decisão (GOLDSCHMIDT; PASSOS, 2005).

Mantendo em vista a natureza dos dados utilizados para extração de padrões significativos e o comportamento das suas observações ordenadas no tempo, atribui-se o rótulo de séries temporais, sendo um conjunto de observações feitas sequencialmente no tempo. Esse tipo de dados são utilizados em diversas áreas do conhecimento para descrever e observar e explicar o comportamento temporal dos dados observados. Além disso, a análise de séries temporais tem a finalidade de realizar predição com base nas observações anteriores (CHATFIELD, 2016). Dessa forma, esse capitulo é organizado da seguinte forma: na Seção 2.2 são apresentados os fundamentos teóricos sobre séries temporais; na Seção 2.3 são apresentados os fundamentos teóricos sobre as etapas de Pré-Processamento de dados; na Seção 2.4 é apresentado alguns modelos de previsão univariados e multivariados; na Seção 2.6 é abordado os métodos de avaliação de desempenho nos modelos de previsão utilizados e a apresentação dos testes estatísticos necessários para a validação dos modelos de previsão.

2.2 Fundamentos de Séries Temporais

Uma série temporal é o conjunto de observações de forma ordenada no tempo (MORETTIN; TOLOI, 2006), diversos fenômenos podem ser representados por séries temporais, tais como Sinais Biológicos, Sensoriamentos, Mercado Financeiro e entre outros.

As série temporais podem ser classificada da seguinte forma (BROCKWELL et al., 2016):

- Contínua: observações coletadas continuamente sem interrupções no tempo.
- Discreta: observações coletadas em intervalos de tempo discreto.

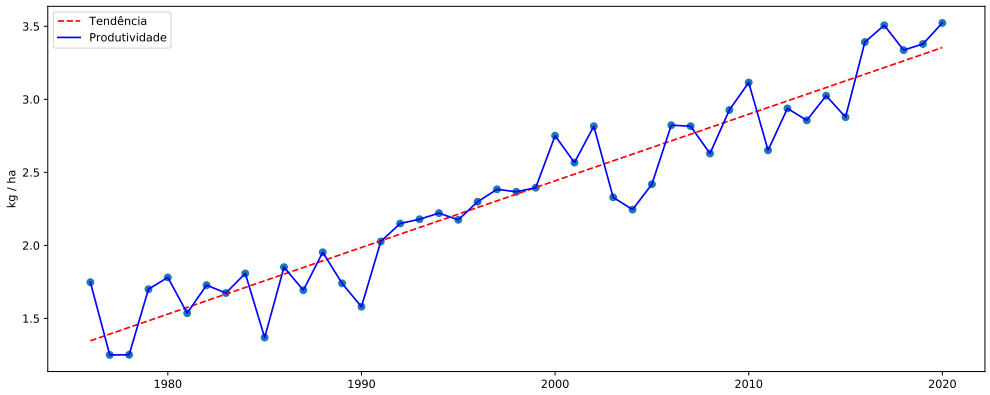
A Série Temporal pode ser composta pelos seguintes componentes (MORETTIN; TOLOI, 2006):

- X_t – Tendência: movimentos regulares com comportamento crescente ou decrescente.
- Y_t – Sazonalidade: movimentos periódicos com ocorrência regulares, conhecidos como Padrões.
- Z_t – Resíduo: movimentos com comportamento aleatório ou irregular, considerados como Ruídos.

Desse modo, a série temporal no instante t pode ser representada pela Equação 1.

$$S_t = X_t + Y_t + Z_t \tag{1}$$

Um exemplo de série temporal pode ser vista na Figura 1 [1], note que além da série temporal, existe a linha de tendência, que por sua vez é uma dos componentes que formam uma série temporal.



Fonte: Elaborada pelo Autor.

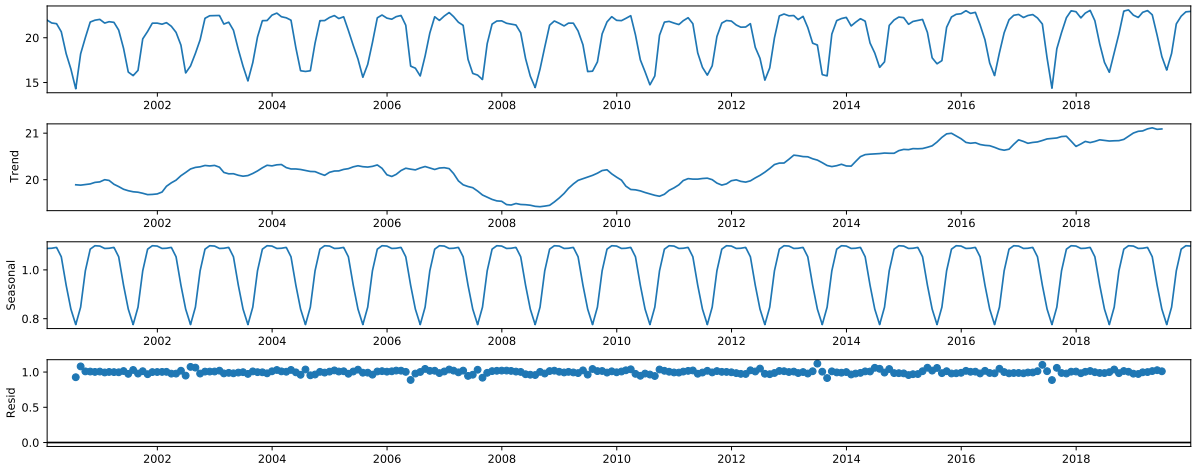
Figura 1 – Série Histórica da Produtividade da Soja no Brasil.

Para exemplificar a constituição de uma série temporal e seus componentes, observe na Figura 2 [2], pode ser observados os seus componentes para as observações da Temperatura Mínima Média do estado do Mato Grosso.

As séries temporais podem ser estacionária, ou seja, os valores no decorrer do tempo, variam em volta de uma média constante, mantendo estabilidade. Entretanto, grande parte das séries temporais são não estacionárias, ou seja, ela possui tendência de movimento, tanto positivo quanto negativo, de forma que, sua variação seja ao redor de uma reta inclinada (MORETTIN; TOLOI, 2006).

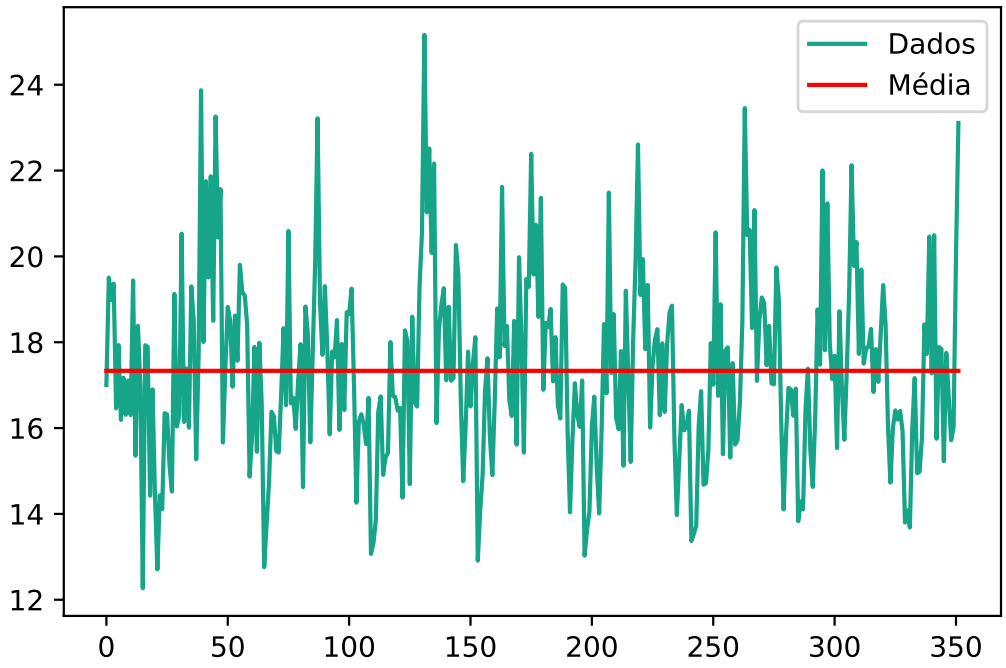
Na Figura 1 é ilustrado o comportamento de uma série temporal não estacionária. Na Figura 3 é ilustrado um exemplo de uma série temporal estacionária.

[1] Figura 1 criada a partir dos dados da Companhia Nacional de Abastecimento – (Conab).
[2] Figura 2, criada com dados do Instituto Nacional de Meteorologia – (INMET)



Fonte: Elaborada pelo Autor.

Figura 2 – Temperatura Mínima Média do Estado do Mato Grosso.



Fonte: Elaborada pelo Autor.

Figura 3 – Série Temporal Estacionária.

As séries temporais podem ser univariadas, ou seja, possui um atributo observado no tempo, por exemplo, na Figura 1 é ilustrada um exemplo série temporal univariada, pois possui uma observações únicas da Produtividade registrada sequencialmente no tempo. As séries temporais também podem ser multivariadas, quando possuem múltiplos atributos observados

no tempo, por exemplo uma série temporal de Mercado de Ações que possuem atributos como Abertura, Alta, Baixa e Fechamento dos valores das ações no decorrer do tempo (CHATFIELD, 2016).

2.3 Pré-Processamento de Dados

As séries temporais podem ser utilizadas em métodos de aprendizado de máquina. No entanto, grande parte dos algoritmos de aprendizado de máquina são sensíveis à qualidade da entrada de dados. Nesse sentido, pode ser necessário o pré-processamento da base de dados (entrada) para a preparação e estruturação das informações em um formato apropriado para o modelo de aprendizagem (HAN; PEI; KAMBER, 2011). Para isso, diversas abordagens podem ser aplicadas para o pré-processamento, tais como: limpeza (Seção 2.3.1), transformação (Seção 2.3.2) e redução de dados (Seção 2.3.3).

2.3.1 Limpeza dos Dados

Muitas vezes os dados podem estar incompletos, com ruídos ou inconsistentes. Assim, o papel da limpeza dos dados é melhorar a qualidade dos mesmos (GOLDSCHMIDT; PASSOS, 2005). Essa limpeza pode ser realizada por meio dos seguintes procedimentos:

- Exclusão de casos: dependendo da qualidade na base de dados, pode haver alguns atributos que possuem vários valores nulos ou redundantes. Também, nesses conjuntos de dados podem haver atributos irrelevantes. Uma das estratégias para solucionar esses problemas seria a exclusão de linhas e/ou colunas das bases de dados (TAN; STEINBACH; KUMAR, 2016).
- Preenchimento de dados faltantes: alguns campos de dados podem possuir valores nulos. Assim, algumas técnicas ou medidas estatísticas podem ser utilizadas para o preenchimento dos dados, tais como média, mediana, moda, entre outras (TAN; STEINBACH; KUMAR, 2016).
- Correção de Erros: muitas informações são obtidas por sensores, dos quais é possível que alguns deles tenham mal funcionamento, acarretando em dados com ruídos. Nesse caso, dependendo do atributo, podem ser aplicados métodos de filtragem para a remoção ou redução de ruídos, como o filtro passa alta (GOLDSCHMIDT; PASSOS, 2005).
- *Outliers*: diferentemente de ruídos, *outliers* podem ser valores legítimos, mas atípicos, podendo acarretar anomalias durante o treinamento de modelos regressivos. Para a solução desse problema, uma das abordagens comumente aplicada é a remoção de exemplos considerados fora da normalidade.

2.3.2 Transformação de Dados

Normalmente, os dados originais não estão preparados para o uso imediato (*e.g.* algoritmos de aprendizado de máquina), sendo necessário algum tipo de tratamento nos

mesmos, o que envolve os seguinte métodos:

- Normalização de Dados: Os dados podem estar representados em unidades de medidas diferentes, como quilograma e libra. Para a normalização de dados, diversos métodos podem ser aplicados, como o *Min-Max* A normalização pode ser realizada por meio da aplicação da (Equação 2), que reescala um atributo para valores entre 0 e 1 (HAN; PEI; KAMBER, 2011) e (GOLDSCHMIDT; PASSOS, 2005).

$$A^{'} = \frac{(A - Min)}{(Max - Min)} \tag{2}$$

  Por exemplo, Na Equação 2, $A^{'}$ é o valor normalizado; A é o valor do atributo, Min é o Valor mínimo do atributo e Max é o valor máximo do atributo.
- Padronização: em inglês *Standard Score*, também conhecido como *Z–Score*, que por sua vez, é o número de desvios padrão que está acima ou abaixo do valor médio da população. A Padronização pode ser realizada por meio da aplicação do da Equação 3, onde $\hat{\mu}$ é a média das amostras de observações dos dados que serão substituído pelo *z-score* e o σ é a variância dos mesmos, após essa transformação, a média $\mu = 0$ e o desvio padrão $\sigma = 1$ (ZAKI; MEIRA, 2014)

$$x_i^{'} = \frac{x_i - \hat{\mu}}{\sigma} \tag{3}$$

2.3.3 Redução de Dados

Dependendo do tamanho do conjunto de entrada, torna-se muito custoso computacionalmente a utilização de todos os dados, podendo acarretar em custo de processamento muito alto. Para lidar com isso, o uso de algumas ferramentas de mineração de dados, tais como, Seleção de Atributos, Redução de Dimensionalidade e Extração de Características, as quais, são aliadas poderosas na redução do custo computacional (GOLDSCHMIDT; PASSOS, 2005).

- Seleção de Atributos: em alguns casos, o conjunto de entrada contém centenas de atributos, dos quais, vários podem ser irrelevantes. Nesse caso, é recomendado a remoção dos mesmos, pois eles podem prejudicar a descoberta de padrões e retardar o processo de mineração de dados. Para determinar quais são os melhores e os piores atributos podem ser realizados teste estatísticos para a verificação do nível de significância. Existem outras abordagens para a seleção de atributos, por exemplo, baseando-se na correlação (HAN; PEI; KAMBER, 2011; ZHENG; CASARI, 2018).
- Redução de Dimensionalidade: um dos problemas de termos uma alta dimensionalidade é a quantidade excessiva de espaço de armazenamento e o aumento do custo computacional (HAN; PEI; KAMBER, 2011). Para tratar esse problema, uma das soluções é a redução do tamanho do conjunto de dados, a qual pode ser feita através das técnicas de redução de dimensionalidade. Uma das técnicas comumente aplicada para redução de dimensionalidade é a Análise de Componentes Principais (PCA), que busca por

vetores ortonormais de k atributos, que tem a capacidade de representar os dados em uma dimensão menor se comparada a dimensão da base de dados original (HAN; PEI; KAMBER, 2011).

- Extração de Características: é de fundamental importância, ainda mais quando aplicado para processos de aprendizado de máquinas, os quais necessitam de dados para realizar a identificação de padrões e fazer predições. O propósito da extração de características é obter os recursos certos para reduzir a dificuldade de aprendizado e melhorar seus resultados. Para isso, existem duas abordagens, como a engenharia de recurso e o aprendizado de características. A engenharia de características requer o uso de técnicas para a extração de atributos a partir dos dados brutos. O aprendizado de características utiliza técnicas que permite o algoritmo aprender características de maneira automatizada (ZHENG; CASARI, 2018).

2.4 Predição de Séries Temporais

Predição é uma das aplicações comumente utilizadas em séries temporais, tendo em vista que os modelos de predição de séries temporais, podem ser propostos para dados temporais univariados e/ou multivariados. A escolha do modelo de previsão para análise de uma serie temporal em particular, depende muito das características da série temporal, da natureza dos dados e do comportamento de suas observações. Os modelos de predição para séries temporais, podem ser separados em Modelos Paramétricos e Modelos Não Paramétricos (MORETTIN; TOLOI, 2006).

2.4.1 Modelos Paramétricos

Os método paramétrico deve ser aplicado em séries temporais que estejam dentro de qualquer distribuição. Os modelos mais utilizados dentro do grupo de modelos paramétricos são conhecidos como Modelos Autorregressivos de Médias Moveis (ARMA) e Modelos Autorregressivos Integrados de Médias Moveis (ARIMA), o qual, pode ser aplicado em séries temporais não estacionária (BROCKWELL et al., 2016).

- Modelos Autorregressivos de Médias Moveis (ARMA – *autorregressive-moving-average*): É a fusão dos modelos Autorregressivo (AR – *Autoregressive*) e Médias Móveis (MA – *Moving Average*). Os Modelos ARMA são usados para descrever séries temporais estacionárias (LAZZERI, 2020).
- Modelos Autorregressivos Integrados de Médias Moveis (ARIMA – *autoregressive integrated moving average*): é o aprimoramento do ARMA que, além de incluir AR e MA, também utiliza a integração (I) para a predição de séries temporais. No ARIMA, a integração tem a finalidade de tornar séries não estacionária em estacionária. Esse modelo é recomendado para séries temporais que possuam componentes de tendência não sazonais. Para séries temporais sazonais, é sugerido o uso do SARIMA, uma versão

do ARIMA para dados temporais com sazonalidade (LAZZERI, 2020).

2.4.2 Modelos Não Paramétricos

Os Modelos Não Paramétricos ao contrário dos Modelos Paramétricos, não faz suposição quanto à distribuição de probabilidade da população em estudo (MORETTIN; TOLOI, 2006). Alguns dos modelos mais populares não paramétricos são o K-ésimo Vizinho mais Próximo e Árvore de Decisão.

- *K-ésimo Vizinho mais Próximo*: Adapta a quantidade de suavização da densidade de dados em um determinado local. O nível da suavização depende de K, que é o número de vizinhos mais próximos, de forma que K é muito menor que N, o tamanho total das amostras (ALPAYDIN, 2020).
- *Árvore de Decisão*: um estrutura hierárquica que é gerada utilizando a estratégia de dividir para conquistar em termos de seus atributos, até atingir o nível mais simples para ser rotulado. A árvore de decisão é um método não paramétrico eficiente, usando tanto para regressão quanto para classificação. Além disso, é possível aprender diretamente a base de regras utilizadas que levaram ao resultado. (ALPAYDIN, 2020).

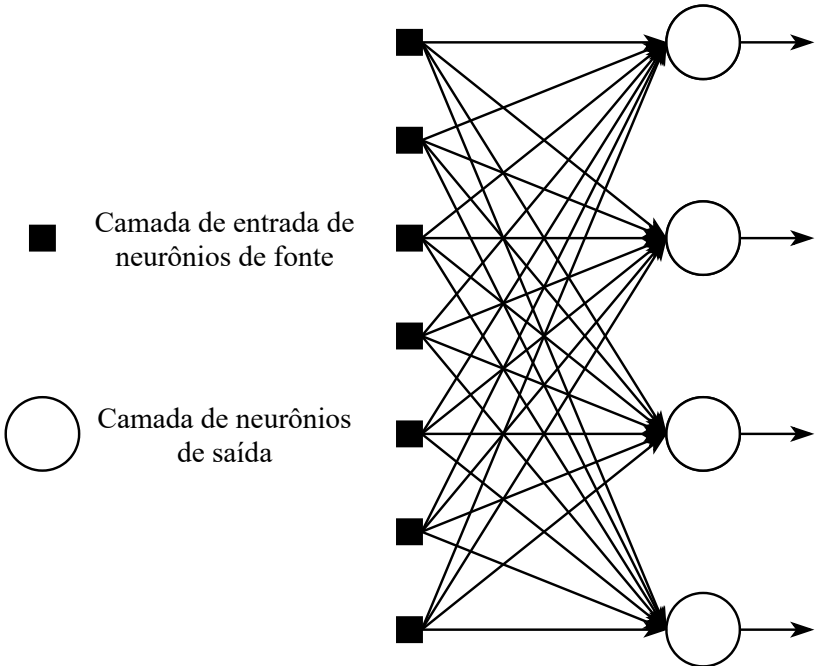
2.5 Redes Neurais Artificiais

O cérebro humano processa informações de forma complexa, não-linear e paralelamente, motivando estudos com o propósito de criar métodos de processamento de informação inspirado no sistema neural biológico. As Redes Neurais Artificiais são baseadas no comportamento do sistema nervoso humano, na captação de informações pelos receptores, os quais as convertem em estímulos elétricos, para alimentar a rede neural com o propósito de identificar padrões relevantes. Após isso, esses padrões são entregues aos Atuadores que tem como objetivo realizar a conversão dos estímulos elétricos em respostas discerníveis. As Redes Neurais Artificiais são compostas por camadas, que por sua vez, são constituídas de unidades de processamento de informações conhecidas como neurônio, possuindo três elementos básicos(HAYKIN, 2007).

- Sinapse: Cada unidade possui um peso ou força própria, podendo variar entre valores positivos e negativos.
- Somador: Um neurônio artificial que realiza a soma de todos os sinais de entrada, calculados pelo peso ou força de cada sinapse.
- Função de Ativação: Limita os valores de saída em intervalos de amplitude de valores finitos, os valores de saída após a Função de Ativação, possuem valores 0 e 1.

As Redes Neurais que possuem o comportamento direcional dos neurônios de entrada para a camada de saída sem realimentação, são conhecidas como *Feedforward neural network* (HAYKIN, 2007). Possuindo apenas a camada de saída sem a camada oculta de neurônios. Na Figura 4 é ilustrado um exemplo de rede de camada única, e Na Figura 5 é ilustrado um exemplo de múltiplas camadas de nós computacionais, também denominados como neurônios

artificiais.



Fonte: Elaborada pelo Autor.

Figura 4 – Rede de Camada Única.

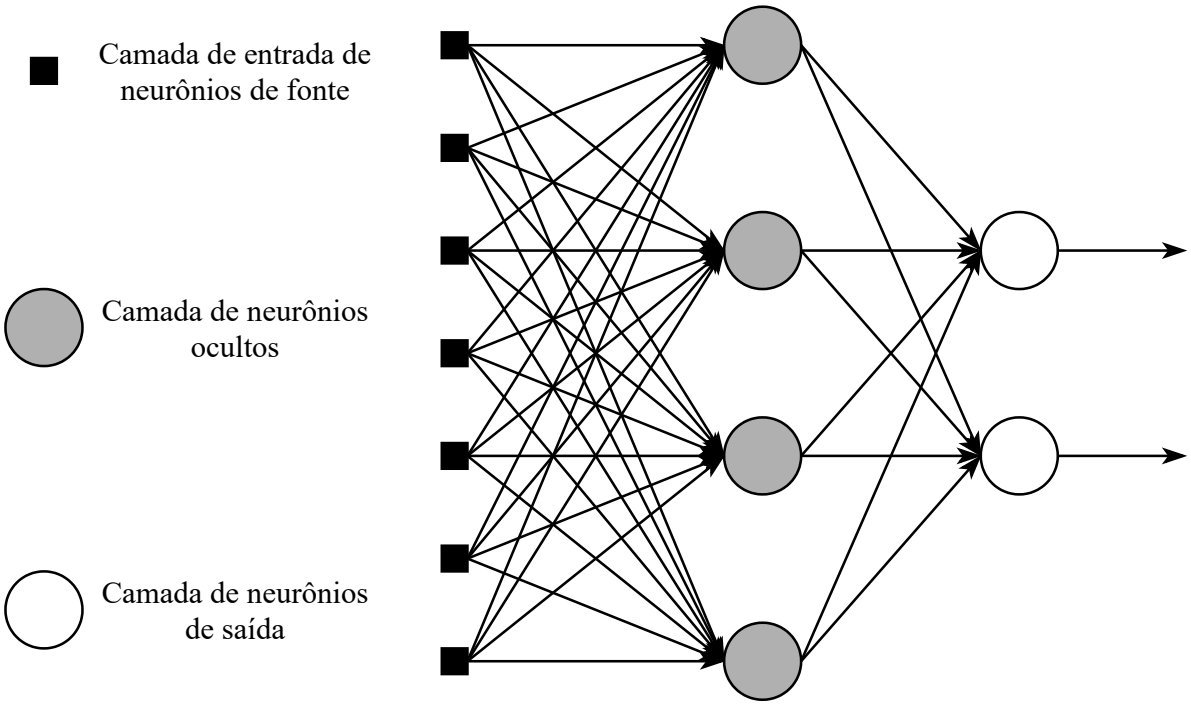
Note que na Figura 4, a camada de entrada de neurônios de fonte não possuem nós computacionais. Portanto, a única camada que possui nós computacionais é a camada de neurônios de saída, formando uma rede neural de camada única.

Para a Figura 5, existe a camada de neurônios ocultos, que possuem nós computacionais. Sendo assim, essa rede neural é composta por múltiplas camadas.

2.5.1 Rede Neural Recorrente

As Redes Neurais Artificiais – (RNA) com alimentação adiante assumem que as entradas das informações são independentes umas das outras. Entretanto, isso não acontece para todos os dados, como os sequenciais, tais como, palavras em um texto, preço da soja ao longo do tempo, reconhecimento de fala, dentre outros. Para o processamento desses dados, foi proposta as Redes Neurais Recorrentes (RNN – *Recurrent Neural Networks*), que são apropriadas para dados sequenciais devido à habilidade de cada neurônio armazenar a informação da entrada anterior em sua memória interna. No entanto, o RNN simples não é capaz de obter as dependências de longo prazo em dados sequenciais (GULLI; KAPOOR; PAL, 2019)

A variação de Rede Neural Recorrente mais utilizada atualmente é a *Long Short-Term Memory* (LSTM), desenvolvida por (HOCHREITER; SCHMIDHUBER, 1997), que tem a capacidade de aprender dependências de longo prazo, possibilitando o processamento de sequências



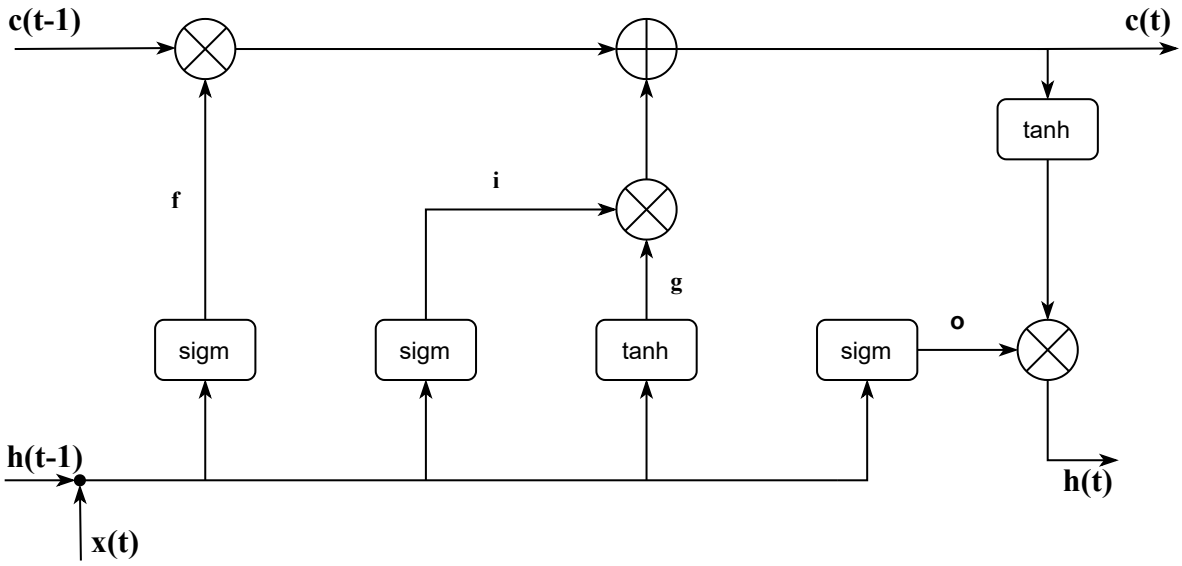
Fonte: Elaborada pelo Autor.

Figura 5 – Rede de Múltiplas Camadas.

longas de entradas. LSTM é utilizada em diversas aplicações, tais como reconhecimento de voz, aprendizado de escrita, identificação de padrões em séries temporais, entre outras (GULLI; KAPOOR; PAL, 2019).

A LSTM implementa a recorrência com a interação de quatro camadas de forma específica, A Figura 6 ilustra um exemplo de célula LSTM, onde i, f, o, são respectivamente as portas de entrada (*input*), esquecimento (*forget*) e saída (*output*), eles são calculados da mesma maneira, mas com matrizes de parâmetros diferentes W_i, U_i, W_f, U_f, W_o, U_o, sendo que W é a matriz de parâmetros de entrada de $x(t)$ e U a matriz de parâmetros de entrada de h_{t-1}.

O primeiro passo é definir o que descartar, para isso, a função sigmoide, que é a mais amplamente usada para Função de Ativação, modula a saída entre 0 e 1, de forma que ele permite multiplicar elemento por elemento do vetor de saída produzido e definir quais elementos serão descartados.



Fonte: Elaborada pelo Autor.

Figura 6 – Estrutura da Célula LSTM.

Os componentes que representam um LSTM são apresentados a seguir:

- A porta de esquecimento f, apresentada pela Equação 4, define o quanto do estado anterior h_{t-1} vai ser mantido.

$$f = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{4}$$

- A Porta de Entrada i, apresentada pela Equação 5, define o quanto do estado calculado da entrada atual x_t será mantido.

$$i = \sigma\left(W_i h_{t-1} + U_i x_t + b_i\right) \tag{5}$$

- A Porta de Saída o, apresentada pela Equação 6, é responsável por definir o quanto do estado interno é exposto para a próxima camada.

$$o = \sigma\left(W_o h_{t-1} + U_o x_t + b_o\right) \tag{6}$$

- O estado interno oculto g, apresentado pela Equação 7, é calculado baseado na entrada atual x_t e no estado anterior h_{t-1}.

$$g = tanh\left(W_g h_{t-1} + U_g x_t + b_g\right) \tag{7}$$

- O estado c_t, apresentado na Equação 8, é o combinador de memória anterior com a nova entrada. Caso a porta de esquecimento f seja definida como 0, é ignorado a memória antiga. Se a porta de entrada i for definida como 0, é ignorado o estado recém calculado.

$$c_t = (f * c_{t-1}) + (g * i) \tag{8}$$

- O estado oculto h_t, no instante t, é calculado como a memória c_t no instante t juntamente com a porta de saída o (GULLI; KAPOOR; PAL, 2019).

$$h_t = tanh\,(c_t) * o \tag{9}$$

2.6 Avaliação dos Modelos

Com os resultados dos modelos preditivos disponíveis, é necessário uma avaliação do desempenho dos mesmos, dessa forma esse capitulo apresenta algumas das ferramentas de avaliação de desempenho, tais como: Validação Cruzada (Seção 2.6.1), Métricas de Desempenho (Seção 2.6.2) e Testes Estatísticos (Seção 2.6.3).

2.6.1 Validação Cruzada

A Validação Cruzada particiona o conjunto de dados em subconjuntos, divididos entre conjunto de teste e de treinamento. Um dos métodos de particionamento mais utilizados é o *K-fold*, que faz o particionamento dos dados em k subconjuntos (ou *folds*) de tamanhos iguais, sendo o i-ésimo subconjunto utilizado para teste e os $k-1$ subconjuntos restantes são utilizados para o treinamento de modelos. Os dados de teste são aplicados no modelo construído a partir dos dados de treinamento. O treinamento e o teste são repetidos k vezes, sendo cada um com um subconjunto de teste diferente (ALPAYDIN, 2020).

No entanto, a validação cruzada não leva em consideração que os dados possam ser sequenciais, ou seja, quando a ordem importa, como é o caso para séries temporais. Nesse sentido, para realizar o particionamento da base de dados entre conjuntos de treino e de teste para dados temporais é utilizado o Split Sequencial, cujo funcionamento é ilustrado na Figura 7.

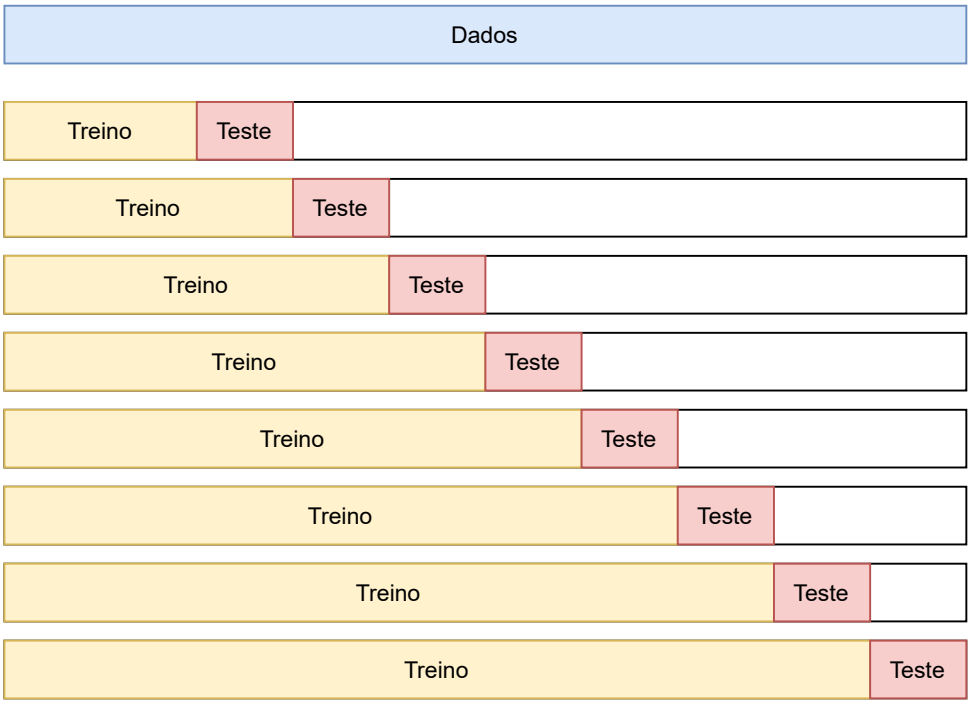
Na Figura 7 foi realizado 8 split sequenciais, sempre levando em consideração e mantendo a sequência das observações, mesmo que isso signifique aumentar o conjunto de treino. Entretanto, o conjunto de teste continua com o mesmo tamanho.

2.6.2 Métricas de Desempenho

O Erro absoluto médio (MAE – *Mean Absolute Error*): é a média dos valores absolutos dos erros de forma individual, supondo que $\hat{y}_i$ é o valor previsto do i-ésima amostra e y_i é o valor verdadeiro correspondente, então o MAE estimado para $n_{samples}$ é definido como:

$$MAE_{(y,\hat{y})} = \frac{1}{n_samples} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \tag{10}$$

O Erro quadrático médio da raiz (RMSE – *Root Mean Square Error*): é a raiz quadrada das médias ao quadrado. Sendo assim, os efeitos de cada erros é proporcional ao erro elevado



Fonte: Elaborada pelo Autor.

Figura 7 – Particionamento de Dados Temporais para Validação Cruzada.

ao quadrado.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{11}$$

y_i são os valores observados, $\hat{y}_i$ são os valores previstos e n é o numero de observações.

2.6.3 Testes Estatísticos

Para a comparação de modelos preditivos, podem ser aplicados testes estatísticos de hipótese. Tais testes devem ser escolhidos de acordo com as características dos conjuntos de dados, tais como pareamento e distribuição. Se os conjuntos de dados estão possuem alguma ligação (*e.g.* originados da mesma fonte), os mesmos são pareados e, caso contrário, são não pareados. Na análise de distribuição dos dados, geralmente é aplicado um teste de normalidade para verificar se os conjuntos de dados estão dentro da distribuição normal (LEHMANN; ROMANO, 2006).

O Teste de Normalidade é utilizado para descobrir se os erros estimados estão normalmente distribuídos ou não (GUJARATI; PORTER, 2009). A obtenção dessa informação é útil para a escolha de quais teste estatísticos de hipótese poderá ser empregado. Por exemplo, para a escolha de testes estatísticos de hipóteses paramétricos, as amostras precisam ser retiradas de forma aleatória de uma população normalmente distribuída, caso contrário, utiliza-se teste não paramétricos (CORDER; FOREMAN, 2014).

O teste de hipótese é composto por duas hipóteses, uma é a hipótese a ser testada, conhecida como hipótese nula, e a outra é a a hipótese alternativa. Realizando o teste de hipótese, pode-de decidir qual hipótese será rejeitada. Se o resultado do teste de hipótese resultar em $P_{valor} <=$ nível de significância, a hipótese nula é considerada verdadeira. Se $P_{valor} >$ nível de significância, rejeita-se a hipótese nula. Dentre os testes de hipótese, existem os testes paramétricos, os quais devem ser utilizados para quando as amostras de uma população estão normalmente distribuída, caso contrário, é utilizado os não paramétricos (WEISS, 2011), alguns exemplos são apresentados a seguir:

1. Testes Não Paramétricos:
    - *Wilcoxon's signed rank test:* método de distribuição usado para testar a diferença entre duas populações com amostras emparelhadas. O teste é baseado nas diferenças absolutas de observações entre as duas amostras (WILCOXON, 1945).
    - *Mann–Whitney U-test:* teste de distribuição usado para avaliar se duas populações têm a mesma mediana (MANN; WHITNEY, 1947).
2. Teste Paramétricos:
    - t-Student: teste de significância para avaliar hipóteses sobre médias populacionais. Existem duas versões desse teste, uma para dados emparelhados e a outra, para dados não emparelhados (FREEDMAN; PISANI; PURVES, 2007).
    - Análise de Variância *(ANOVA – Analysis of variance)*: método para comparar médias de três ou mais populações (FISHER, 1956).

3 TRABALHOS RELACIONADOS

3.1 Considerações Iniciais

Neste capítulo são apresentados alguns trabalhos relacionados, onde a influência do clima sobre as commodities é avaliada por meio de diversas técnicas de aprendizado de máquina, tais como, modelos de redes neurais artificiais, regressão multivariada e regressão interativa (VU, 2007).

3.2 Análises Preditivas

Li, Xu e Li (2010) Constata a dificuldade da análise preditiva dos preços nos mercados de curto prazo. Além disso, é apresentado um comparativo entre o modelo Auto-regressivo integrado de médias móveis (*Autoregressive Integrated Moving Average* – ARIMA), e uma Redes Neurais Artificias – (RNA) com alimentação avante (*feed-forward*). Esses modelos foram aplicados em uma série temporal para a predição dos preços do tomate a partir dos dados coletados entre o período de 1996 e 2010. Como resultado do comparativo, o modelo utilizando RNA superou em desempenho do ARIMA, se comparado o erro percentual entre os valores observados e os valores previstos de ambos os métodos.

Wang e Gao (2018) apresenta uma abordagem diferente ao realizar um análise preditiva com o objetivo de obter os valores futuros dos preços mais altos e mais baixos da soja para auxiliar a previsão do preço de fechamento da commoditie. Para isso, foi utilizado um conjunto de dados da *Dalian Commodity Exchange* – (DCE) como entrada em uma Rede Neural Recorrente do tipo *Long Short-Term Memory* – (LSTM). Feito isso, utilizando MAE como métrica de desempenho, o modelo de predição foi capaz de obter 80% de precisão quando os preços mínimos e máximos possuem alta volatilidade.

Em Zhang et al. (2018) é apresentado um modelo de rede neural de função de base radial (*radial basis function* – RBF) de regressão quantílica. Em primeiro momento, esse modelo foi usado para descrever a distribuição do preço da soja. Após isso, foi utilizada uma rede neural RBF que mostra ter capacidade de aproximar a função contínua (DE FREITAS et al., 2001). Sendo assim, para aproximar os componentes não lineares do preço da soja foi utilizado o modelo apresentado anteriormente. Para realizar a otimização dos parâmetros do modelo, foi proposto a utilização de um algoritmo híbrido utilizando o método Algoritmos Genéticos (*Genetic Algorithm* – GA). Foi utilizado juntamente o algoritmo de otimização Gradiente Descendente (GD – *Gradient Descent*), sendo capaz de realizar otimização global, mas a convergência para o ótimo local é lenta. Assim, foi proposto o algoritmo híbrido conhecido como *Gradient Descent and Genetic Algorithms*–(GDGA) (AHMAD et al., 2010). Feito isso, os resultados constataram que o algoritmo híbrido forneceu resultados melhores do que o método GA puro.

Ouyang, Wei e Wu (2019) utilizam séries temporais multivariadas para construção de modelos capazes de prever preços futuros das commodities. Como os preços futuros das commodities sofrem influência de informações de curto e de longo prazo, constata-se que uma abordagem usando o modelo auto regressivo de médias móveis pode falhar. Entretanto, foi proposta um alternativa para solução desse problema por meio de um modelo de uma rede de série temporal de longo e de curto prazo (*Long-and Short-term Temporal Patterns with neural Networks* – LSTNet) Lai et al. (2018), que utiliza a rede neural convolucional e a rede neural recorrente para obter padrões de dependência local em curto prazo. Com base nos resultados empíricos, o LSTNet atingiu maiores valores de acurácia em relação a diversos trabalhos estado-da-arte.

Prashantha et al. (2020) ressalta a importância da produção agrícola e o impacto que a falta de previsão dos preços das commodities tem sobre os agricultores. Muitas vezes, os produtores não têm obtido um preço adequado referente aos seus plantios. Tendo em vista que o rendimento das safras tem como parâmetros, os componentes do solo, precipitação e umidade do solo, a solução foi desenvolver um sistema mais eficiente para predição de preços de commodities agrícolas. Para isso foi usado métodos de aprendizado de máquina com o propósito de reduzir o risco para o agricultor, tais como Modelos de Regressão Linear, *Decision Tree* e *K-Nearest Neighbor* – (KNN), tendo como resultado a média dos preços futuros das commodities.

Reis Filho et al. (2020) destaca a importância dos preços das commodities no mercado global e apresenta o desafio de análise preditiva dos preços futuros das commodities, tendo em vista eventos políticos e crises econômicas. É apresentada a hipótese de que notícias relacionadas ao agronegócio possuem informações que auxiliariam nas análises preditivas. Sendo assim, técnicas de mineração de textos foram aplicadas com o propósito de incrementar a base de dados e incorporá-los nas séries temporais das commodities agrícolas que, por sua vez, foram obtidas a partir do *World Agricultural Supply and Demand Estimates* – (WASDE). Para isso, foram utilizados métodos de aprendizado de máquina, como RBF e Regressão Polinomial. Para a validação dos modelos, foram consideradas as seguintes medidas: (MAE) – *Mean Absolute Error*, (MAPE) – *Mean Absolute Percentage Error*, (MSE) – *Mean Square Error* e (RMSE) – *Root Mean Square Error*.

Mahto et al. (2021) ressalta a importância da previsão de informações de mercado bem fundamentadas, como a previsão de curto prazo dos preços das commodities. Foram considerados os dados do preço do grão de soja no mercado local de Maharashtra, India, que foram coletados no período de janeiro de 2014 a dezembro de 2018. Além disso, também foram considerados os dados referente ao preço das sementes de girassol no mercado local de Andhra Pradesh, India, cujos dados foram coletados no período de janeiro de 2011 a dezembro de 2016. Para a predição, foi utilizado o modelo RNA, cujos resultados foram comparados com os resultados do modelo ARIMA tradicional. As métricas de desempenho utilizadas foram a *Mean Absolute Percentage Error* (MAPE) e a *Root Mean Squared Percentage Error*

(RMSPE). Considerando as métricas de desempenho, o modelo RNA atingiu melhores valores de desempenho em relação ao modelo ARIMA.

3.3 Rendimento de Produção

Kaul, Hill e Walthall (2005) investigaram modelos de RNA para predição do rendimento do milho e da soja no estado de Maryland, Estados Unidos. Além disso, foi realizada uma comparação da capacidade de previsão dos modelos em nível estadual, regional e local. Para isso, foi utilizado uma base de dados sobre históricos do rendimento de milho e soja do estado de Maryland entre os períodos de 1978 a 1998. Após, o resultados constataram que a utilização dos modelos de RNA, têm potencial para ser uma ferramenta capaz de auxiliar os especialistas no desenvolvimento, na revisão ou na atualização do plano de gestão de nutrientes.

Alves et al. (2018) propõem a utilização de RNA para estimar a produtividade da soja, baseado-se nas características agronômicas e em seus hábitos de crescimento. Além das características das sementes, para a construção de modelos foi feito uso dos dados agronômicos da safra de 2013/2014 de soja em Anápolis – GO Alves et al. (2018). Após isso, o método RNA foi utilizado para o treinamento de modelos, como resultado, o modelo mais acurado obteve 98% de taxa de acerto sobre os dados de treinamentos. Para os dados de teste, foi obtida uma taxa de acerto de 72%. Sendo assim, foi concluído que os modelos RNA são capazes de estimar, com uma acurácia aceitável, a produtividade da soja baseada em suas características agronômicas e hábitos de crescimento.

Elavarasan et al. (2018) ressalta a influência dos avanços tecnológicos sobre os dados referente a agricultura. Dessa forma, pode-se realizar as investigações desses dados com o propósito de alcançar melhorias nas safras, ser capaz de prever os rendimentos e identificar os estresses hídricos causado pelo plantio. Para isso, técnicas de aprendizado de máquina foram utilizadas no decorrer desse trabalho, tais como Árvore de Decisão, Floresta Aleatória, RNA, Redes Bayesianas, Máquina de vetores de suporte, modelo de Markov, *K-means clustering* e Algoritmo de maximização de expectativa. Comparando os modelos entre eles, utilizando métodos de validação de desempenho, tais como (RMSE) e (MAE), além disso, o modelo de aprendizado supervisionado que alcançou o menor valor de MAE foi o RNA e o modelo não supervisionado que alcançou o menor MAE foi o Algoritmo de maximização de expectativa.

Sun et al. (2019) utiliza uma base de dados de sensoriamento remoto para auxiliar na previsão do rendimento de safra, por meio de Redes Neurais Convolucionais (*Convolutional Neural Network* – CNN) e LSTM. Os dois métodos de redes neurais foram utilizados para a geração de um modelo híbrido CNN-LSTM alimentado com dados ambientais, amostras de temperatura terrestre e informações históricas de produção de soja, e comparar seu desempenhos com os demais modelos anteriores, utilizando RMSE e Erro percentual como métrica de desempenho, que por sua vez o modelo híbrido alcançou valores menores de RMSE se comparado aos outros modelos.

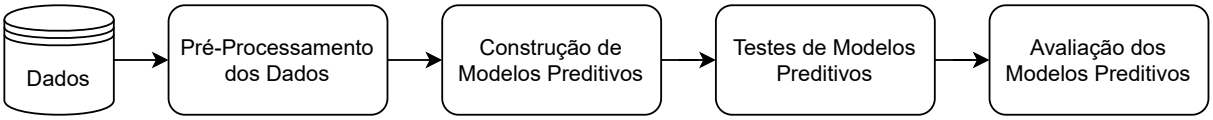
3.4 Considerações Finais

Nesse capitulo foram apresentados pesquisas relacionadas ao tema desse trabalho, demonstrando os diferentes métodos de análises preditivas sobre a produção de commodities, a predição de preços futuros e a influência climática nesse cenário. Nesse contexto, é importante ressaltar que, em nenhum dos trabalhos relacionados apresentados foi desenvolvido um modelo de predição baseado em dados climáticos brasileiros. Adicionalmente também não foi feito um modelo para a verificação da influência que o clima do Brasil possa ter sobre o preço internacional da soja.

4 CONFIGURAÇÃO EXPERIMENTAL

4.1 Considerações Iniciais

A mineração de dados pode ser resumida em alguns passos, com métodos e propósitos variados. Nesse trabalho foi utilizado métodos de aprendizado de máquina para construção de modelos para a predição do preço da soja a partir dos dados climáticos. Na Figura 8 é apresentado o fluxograma implementado para alcançar o objetivo deste trabalho.



Fonte: Elaborada pelo autor.

Figura 8 – Fluxograma Implementado.

Dessa forma, este capitulo tem como propósito apresentar a metodologia utilizada em cada etapa do processo, sendo organizado da seguinte forma: na Seção 4.2 é apresentado as ferramentas e as tecnologias utilizadas no desenvolvimento desse trabalho; na Seção 4.3 será apresentado os passos e as fontes de onde os dados foram obtidos; na Seção 4.4 é abordado os tratamentos de dados utilizado no decorrer desse trabalho; na Seção 4.5 é demonstrado qual foi o método utilizado para executar o treinamento e como foram separados a base de teste e treino; na Seção 4.6 é apresentado o método utilizado para a avaliação de desempenhos dos modelos treinados.

4.2 Ferramentas e Tecnologias

Nesse trabalho foram utilizadas as seguintes ferramentas e tecnologias:

- Anaconda Navigator é uma interface gráfica de usuário que permite iniciar aplicativos e gerenciar facilmente pacotes conda.
- Spyder é um ambiente de desenvolvimento integrado de código aberto para programação em Python.
- Python 3.8 é a linguagem mais utilizada para aprendizado de máquina e possui grande quantidade de bibliotecas para auxiliar no treinamento de modelos preditivos.
- Pacotes: NumPy [1], Pandas [2], Seaborn [3], Pyplot [4], Scipy [5], Datetime [6], Sklearn [7]

[1] https://numpy.org/
[2] https://pandas.pydata.org/
[3] https://seaborn.pydata.org/
[4] https://matplotlib.org/2.0.2/api/pyplot_api.html
[5] https://www.scipy.org/
[6] https://docs.python.org/3/library/datetime.html
[7] https://scikit-learn.org/stable/

4.3 Aquisição de Dados

Os dados foram divididos em duas categorias, a primeira sendo os dados climáticos, e a segunda, como dados complementares, as quais foram agrupadas em uma única tabela. Sendo assim, essa Seção será organizado da seguinte forma: na Subseção 4.3.1 é abordado a aquisição dos dados climáticos; na Subseção 4.3.2 são apresentados os dados complementares.

4.3.1 Base de Dados Climática

Os dados climáticos brasileiros foram obtidos por meio do Instituto Nacional de Meteorologia (Inmet) Meteorologia (2020), cujos dados se resumem basicamente entre os seguintes atributos:

- Estação: é o numero de identificação da estação meteorológica.
- Data: é data em que os dados foram observados.
- Velocidade do vento média: é o valor médio, em metros por segundo, da velocidade do vento.
- Velocidade vento máxima média: é a média do valor máximo, em metros por segundo, da velocidade do vento.
- Evapotranspiração potencial: é o valor, em milímetros, que representa o potencial processo de perda de água para atmosfera.
- Evapotranspiração real: é o valor, em milímetros, que representa a quantidade de água que foi transferida para atmosfera por evaporação e transpiração em condições reais.
- Insolação total: é o valor do intervalo total de tempo em horas, entre o nascimento e o por do sol, em que o mesmo não foi obstruído por nuvens.
- Nebulosidade média: é o valor, em décimos, do céu encoberto. Por exemplo, se o valor for 0.5, significa a metade da abóboda celeste estava encoberta por nuvens. Caso o valor indicar zero, indica que nenhuma nuvem foi identificada.
- Temperatura máxima média: é a média do valor máximo da temperatura em Graus Celsius.
- Temperatura compensada média: é o valor da média de três leituras de temperatura mais a temperatura máxima e miníma. A média deses 5 valores é conhecida como Temperatura Compensada Média.
- Temperatura miníma média: é a média das temperaturas minímas em Graus Celsius.
- Umidade relativa média: é a umidade relativa média do ar em porcentagem.
- Precipitação total: é a quantidade de água em milímetros da atmosfera que foi depositada na superfície terrestre.

Esses dados foram coletados, mensalmente, entre 2000 e 2020, a partir de cada estação meteorológica de todo o território brasileiro, foram ao todo, 264 estações meteorológicas cujos dados foram processados e separados por cada estados, levando em consideração que alguns estados possuíam mais estações meteorológicas que outros.

4.3.2 Base de Dados Complementares

Somente os dados climáticos podem não ser suficientes para alcançar os objetivos desde trabalho. Sendo assim, foram necessárias bases de dados complementares, as quais estão apresentadas a seguir:

- **Série Histórica dos Valores do Índice Nacional de Preços ao Consumidor Amplo (IPCA):** obtida em Geografia e Estatística (2020), seus dados consistem em valores mensais, em porcentagem, da inflação no Brasil entre os períodos de 2000 a 2020.
- **Série Histórica da Cotação do Dólar:** obtida em Investing (2020), são dados mensais referentes à cotação do dólar em reais no período de 2000 a 2020.
- **Série Histórica da Safra de Soja:** obtida em Companhia Nacional de Abastecimento (2020), são dados anuais referentes à à produção, à produtividade e à área plantada de soja no Brasil durante o período de 2000 a 2020.
- **Série Histórica do Preço da Soja Internacional:** obtida em International Monetary Fund (2020), são dados mensais que resumem no preço da soja em dólares por toneladas métricas no período de 2000 a 2020.

4.4 Pré-Processamento

Após os dados de fontes diferentes serem coletados e armazenados em uma única tabela, é necessário o tratamento dos dados. Para isso, primeiramente, foi realizado uma verificação por dados nulos, e foi constatado que dois atributos possuíam 30% de seus campos nulos. Nesse caso, esses atributos foram removidos da base de dados, tais como *'evapobhpotencial' e 'evapobhreal'*.

Após, os atributos restantes que ainda possuíam dados nulos, foram tratados da seguinte forma:

1. Verificação e remoção de *outliers*
2. Separação dos dados climáticos por estações meteorológicas.
3. Com as estações separadas, os dados foram agrupados por meses, de janeiro a dezembro, de forma que todos os dados referente ao mês de janeiro, independente do ano ou do dia, eram agrupados e o mesmo foi feito para os outros meses.
4. Para cada mês, nos dados de cada estação meteorológica, a média dos valores dos atributos foram atribuídas nos campos de valores nulos dos seus respectivos atributos.

Os atributos restantes, que não são dados climáticos, foram distribuídos da seguinte forma na base de dados:

1. Para os dados da série histórica das safras de soja, foi utilizado a seguinte abordagem: como os dados climáticos eram separados por meses e os dados de produção eram anuais, foi decidido por substituir o valor da produção da soja entre os meses de colheita daquele estado, dividindo-os entre os números de estações meteorológicas dos seus respectivos estados, pois alguns estados possuem mais estações meteorológicas que outros.

2. Os valores da cotação do dólar foram distribuídos de forma parelha ao período dos dados climáticos.
3. Os dados referente aos IPCA foram alocados por suas data equivalente ao mesmo período dos dados climáticos.

Após esses tratamentos, os dados foram separados entre os cinco estados mais produtivos do Brasil: Mato Grosso (MT), Mato Grosso do Sul (MS), Goiás (GO), Paraná (PR), Rio Grande do Sul (RS). Esses estados foram responsáveis pela produção de 74.9% do soja da safra de 2019/2020 [8]. Feito isso, foi calculada a média entre os atributos no mesmo período para cada estação meteorológica pertencente ao mesmo estado. Dessa maneira, foi possível encontrar a média dos dados climáticos de cada estado. E, seguida, foi removido o atributo *'estação'* para cada base de dados correspondente ao seu estado.

4.5 Construção de Modelos de Predição

Após o preprocessamento, foi gerada uma base de dados unificada para cada estado considerado nessa avaliação experimental, de forma que, os atributos observados na Tabela 11, os quais, também estão presentes na base de dados dos outros estados considerados neste trabalhos, são os seguintes:

- PS – Preço da Soja
- PT – Precipitação Total
- TCM – Temperatura Compensada Média
- TMM – Temperatura Mínima Média
- URM – Umidade Relativa Média
- IT – Insolação Total
- IPCA – Índice Nacional de Preços ao Consumidor Amplo
- Dolar – Dólar Americano.
- CO – Colheita.

Tabela 1 – Base de Dados do Paraná.

Data	PS	PT	TCM	TMM	URM	IT	IPCA	Dolar	CO
2000-01-31	*180.38*	*195.443*	*23.048*	*18.808*	*77.433*	*185.886*	*0.62*	*1.784*	*222.95*
2000-02-29	*185.75*	*254.529*	*22.380*	*18.672*	*81.252*	*150.334*	*0.13*	*1.768*	*222.95*
2000-03-31	*190.97*	*130.557*	*21.515*	*17.838*	*80.921*	*158.101*	*0.22*	*1.739*	*222.95*
2000-04-30	*197.10*	*17.643*	*20.006*	*15.220*	*73.445*	*202.179*	*0.42*	*1.805*	*222.95*
2000-05-31	*200.87*	*34.571*	*16.176*	*11.707*	*76.768*	*164.616*	*0.01*	*1.824*	*0.000*

Fonte: Autoria Própria.

Em cada base de dados pré-processada, foi aplicada a normalização com o propósito de que o algoritmo de aprendizagem associe os atributos no tempo passado da base de dados

[8] Cálculo baseado nas informações da série histórica da safra de soja em https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras

com o preço da soja observado no tempo seguinte, em outras palavras, o algoritmo vai levar em consideração todos os atributos do mês anterior e atual para realizar a predição do preço da soja.

Por fim, cada base de dados foi dividida entre conjuntos de treino e treinamento. Em seguida, modelos preditivos foram construídos utilizando o método *Long Short-Term Memory* (LSTM) para predição, por ser capaz de aprender as dependências de longo prazo dos dados de entrada. Foi construído um modelo para cada estado com duas configurações diferentes, uma delas com dados climáticos e outro sem os dados climáticos.

4.6 Avaliação de Desempenho dos Modelos

Após a construção dos modelos preditivos, os mesmos foram avaliados para comparar o desempenho entre eles. Para isso, foi usado o método split sequencial, uma versão de validação da cruzada para dados temporais. Para cada modelo de predição, foi obtido valores de *Root Mean Squared Error* (RMSE) e *Mean Absolut Error* (MAE) amostrais, os quais são os valores obtidos de cada *fold* da validação cruzada. As medidas de RMSE e MAE globais, são as métricas de desempenho de todos os *folds*. Para o cálculo do RMSE global, foi utilizado os valores amostrais como pode ser visto na Equação 12 de forma que, G representa o RMSE Global, A representa os valores amostrais do RMSE, o m é os meses do ano e o f é o numero de split sequencial realizado.

$$G = \sqrt{\frac{\sum_i^n (A_i^2.m)}{f.m}} \tag{12}$$

Por fim, os resultados da validação cruzada foram utilizados para a realização de testes estatísticos de hipótese utilizando o nível de significância de 5%. As análises estatísticas foram conduzidas nas seguintes configurações:

- Por estado: comparação entre modelos que utilizaram dados climáticos e os que não utilizaram. Essa análise é feita separada por estado (*e.g.* PR com dados climáticos *vs.* PR sem dados climáticos). Nessa configuração, foram realizados, no total, cinco testes.
- Por base de dados: comparação entre modelos para cada base de dados. Em outras palavras, uma comparação é realizada entre modelos construídos utilizando dados climáticos e outra, entre os que não consideraram dados climáticos para o seu treinamento. Desse modo, foram realizados dois testes estatísticos para essa configuração.

Nesse contexto, os testes de hipótese devem ser conduzidos para dados não pareados.

Para a escolha do teste de hipótese adequado, foi computado o teste de normalidade Shapiro-Wilk para verificar se os testes devem ser paramétricos ou não paramétricos. Esse teste de normalidade é um dos mais utilizados por ser considerado o mais eficiente para pequenos conjuntos de dados, ou seja, os que contêm menos que 50 amostras Royston (1992).

4.7 Considerações Finais

Nesse capitulo, foram apresentados os passos utilizados para o desenvolvimento desse trabalho, como o pré-processamento, a construção dos modelos preditivos e a avaliação de desempenho dos modelos criados. Para o Capitulo 5, são apresentados os resultados e discussões.

5 RESULTADOS E DISCUSSÕES

Após a aquisição e tratamento dos dados climáticos e com a inclusão dos dados referente produção anual de soja, foi necessária a transformação para dados mensal. Para isso, foi dividido a produção da safra anual, entre os meses de colheita de cada estados (COMPANHIA NACIONAL DE ABASTECIMENTO, 2019) e o numero de estação meteorológica que cada estado possuía.

Ao realizar a união dos dados das estações meteorológicas juntamente com os dados complementares, os dados de algumas estações tiveram que passar por mais etapas do pré-processamento nos períodos de tempos observados, além disso, alguns estados possuíam mais estações meteorológicas que outros, justificando a diferença entre o valores de colheita, conforme apresentado nos Apêndice A.

Com a etapa do pré processamento concluída, os modelos preditivos (LSTM) – *Long Short Term Memory* foram criados para cada estado com duas configurações, uma com dados climáticos e outra sem dados climáticos. Após, para avaliar o desempenho de cada modelo, foi utilizada a Validação Cruzada para dados temporais nos modelos, usando 10 folds. Em seguida, foram aplicadas métricas de desempenho sobre os resultados da validação para cada *fold*. Dessa forma, os valores de (RMSE) – *Root Mean Squared Error* e os valores de (MAE) – *Mean Absolute Error*, obtidos para cada *fold* são medidas amostrais. As medidas amostrais são apresentadas no Apêndice B. É importante ressaltar que as medidas amostrais foram mensuradas para cada *fold* da validação cruzada. Sendo assim, essas medidas são utilizadas para o cálculo da RMSE e da MAE globais, cujos resultados são apresentados nas Tabelas 2 e 3.

Tabela 2 – Medidas de Desempenho Globais do Modelo LSTM com dados climáticos.

Estados	RMSE Global	MAE Global
MT	23,98	19,29
MS	23,68	18,93
GO	23,41	19,03
PR	22,11	17,51
RS	24,06	18,99

Fonte: Autoria Própria.

Na Tabela 2, é apresentado os valores globais de RMSE e MAE para cada modelo preditivo com a configuração utilizando dados climáticos. Na Tabela 3, são apresentados os RMSE e MAE globais para os modelos que foram construídos sem dados climáticos.

Como foi demonstrado pelos resultados da Tabela 2, é possível notar que o estado do Paraná, foi o estado que alcançou um RMSE menor, isso foi possível pela qualidade dos dados coletados pelos sensores das estações meteorológicas do Paraná. Para os resultados apresentados na Tabela 3 o estado do Rio Grande do Sul (RS) foi o que obteve o menor valor

Tabela 3 – Medidas de Desempenho Globais do Modelo LSTM sem dados climáticos.

Estados	RMSE Global	MAE Global
MT	22,62	18,26
MS	23,01	18,61
GO	23,68	18,98
PR	22,24	17,84
RS	22,18	17,91

Fonte: Autoria Própria.

para o RMSE. Nesse caso, o fator determinante para esse desempenho foi os valores da colheita do estado do RS que foram atribuídas aos seus respectivos meses de colheita.

Neste trabalho foi proposta uma investigação a respeito da influência do clima no Brasil sobre o preço internacional da soja. Para responder a essa pergunta, apenas o uso de medidas de RMSE e MAE, não são o suficiente para verificar a hipótese deste trabalho, portanto, são necessárias algumas análises estatísticas em cima dos resultados obtidos entre os modelos treinados.

Para a escolha de um teste estatístico apropriado, foi necessário verificar se as amostras estavam normalmente distribuídas. Para isso, foi aplicado o teste de normalidade Shapiro-Wilk, considerando um nível de significância de 5% (0,05). Os resultados do teste Shapiro-Wilk são apresentados na Tabela 4.

Tabela 4 – Valor de P do Teste de Shapiro-Wilk para as Medidas de Desempenho.

	RMSE	
	Com Dados Climáticos	**Sem Dados Climáticos**
MT	0.59608	0.68518
MS	0.70378	0.25664
GO	0.43360	0.13044
PR	0.20909	0.41385
RS	0.15277	0.48572
	MAE	
	Com Dados Climáticos	**Sem Dados Climáticos**
MT	0.63050	0.44692
MS	0.28124	0.59656
GO	0.40691	0.21749
PR	0.21878	0.53817
RS	0.16500	0.77340

Fonte: Autoria Própria.

Sendo assim, de acordo os p-valores apresentados na Tabela 4, a hipótese de que os dados estão normalmente distribuídas não pode ser rejeitada, ou seja, os modelos foram aprovados no teste de normalidade. Dessa forma, foi escolhido um teste estatístico de hipótese paramétrico para dados não pareados com o proposito de comparar os modelos com dados

Tabela 5 – Valor de P dos Testes T-Student entre os Resultados dos Modelos Preditivos.

	RMSE	**MAE**
MT	0.70611	0.76832
MS	0.89802	0.89988
GO	0.94585	0.93904
PR	0.98245	0.97506
RS	0.62191	0.66803

Fonte: Autoria Própria.

Tabela 6 – Valor de P do Teste ANOVA Para Cada Base de Dados.

	Com Dados Climáticos	**Sem Dados Climáticos**
RMSE	0.98907	0.99934
MAE	0.98799	0.99923

Fonte: Autoria Própria.

climáticos com os sem dados climáticos. Além desse teste, foi escolhido outro teste não pareado para comparar os resultados entre os estados para cada configuração.

Desse modo, foi aplicado o teste *t-Student* não pareado, comparando os modelos preditivos que utilizaram dados climáticos, com os modelos preditivos sem os dados climáticos. Vale ressaltar que essa comparação foi feita por estado (e.g. MT com dados climáticos vs. MT sem dados climáticos). Para a aplicação desse teste, foi considerado um nível de significância de 5% (0,05), cujos resultados podem ser observados na Tabela 5.

Conforme apresentado na Tabela 5, não foi constatada diferença estatisticamente significativa em nenhuma comparação por pares entre modelos. Desse modo, podemos afirmar, com 95% de certeza, que os dados climáticos brasileiros não influenciaram no preço do soja.

Além disso, duas análises, por meio do teste ANOVA, foram realizadas, sendo uma para comparação entre modelos construídos considerando dados climáticos e a outra, para modelos em que não foram considerados dados climáticos. Os resultados do teste Anova são apresentados na Tabela 6.

De acordo com os resultados apresentados na Tabela 6, não foi encontrada estatisticamente significativa entre os modelos preditivos com dados climáticos e sem os dados climáticos.

A falta de isonomia nos dados climáticos entre os estados, tem contribuído para a diferença das métricas de desempenho. Em outras palavras, nos resultados dos modelos com dados climáticos, o estado do Paraná foi o estado com o menor valor de RMSE devido à qualidade dos dados adquiridos. Em contra partida, o estado do Rio Grande do Sul, o qual possui mais estações meteorológicas, necessitando de mais tratamento nos dados, teve o pior desempenho de RMSE comparado aos modelos com dados climáticos. Porém, em comparação com os modelos sem dados climáticos, o estado do RS é o que tem se saído melhor considerando a medida RMSE. Apesar disso, as diferenças entre os modelos não são

estatisticamente significantes.

6 CONCLUSÃO E TRABALHOS FUTUROS

Os avanços tecnológicos na agricultura, tem viabilizado análises preditivas dos preços ou produtividade das commodities. Nesse contexto, esse trabalho foi desenvolvimento com o propósito de realizar análises preditivas do preço da soja por meio de modelos (LSTM) – *Long Short Term Memory* com duas configurações, sendo um com dados climáticos e outro sem dados climáticos.

Com as comparações das métricas de desempenho feitas entre as duas configurações e os testes estatísticos aplicados, foi possível constatar que não houve diferença significativa entre os modelos, evidenciando que o uso dos dados climáticos não acarretou em melhora preditiva de modelos. Em outras palavras, o clima no Brasil não demonstrou influência significativa no preço da soja internacional.

6.0.1 Limitações

Durante o desenvolvimento deste trabalho, foram constatadas as seguintes limitações:

- Falta de Isonomia entre os dados das estações meteorológicas.
- Diferença de quantidade de estações meteorológicas em cada estado.
- A predição do preço internacional da soja foi feita levando em consideração apenas os dados climáticos do Brasil.

6.0.2 Principais Contribuições

Este trabalho forneceu as seguintes contribuições:

- Verificação da influência climática do Brasil sobre o preço internacional da soja.
- Predição do preço da soja, utilizando modelos (LSTM – *Long Short Term Memory*) considerando dados climáticos brasileiros.
- Possibilidade de auxílio aos agricultores em tomadas de decisão.
- Criação de uma base de dados climática.
- Aplicação de teses estatísticos.
- Engenharia de dados na base de dados climáticos.

6.0.3 Trabalhos Futuros

Como sugestão de aprimoramento e consolidação dos resultados apresentados nesse trabalho, atividades futuras incluem:

- Utilização de outros métodos para a construção de modelos.
- Utilizar dados climáticos de outros países com maior produção de soja do mundo.
- Acrescentar informações referentes à oferta e demanda.

Referências

AGEITEC. **Temperatura**. 2007. Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/soja/arvore/CONT000fzr67cri02wx5ok0cpoo6aeh331my.html>. Acesso em: 05/05/2021. Citado na página 1.

AHMAD, F. et al. Performance comparison of gradient descent and genetic algorithm based artificial neural networks training. In: **10th International Conference on Intelligent Systems Design and Applications**. Cairo, Egito: IEEE, 2010. p. 604–609. Citado na página 16.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge, EUA: MIT press, 2020. Citado 2 vezes nas páginas 9 e 13.

ALVES, G. R. et al. Estimating soybean yields with artificial neural networks. **Acta Scientiarum. Agronomy**, SciELO Brasil, v. 40, 2018. Citado na página 18.

BRASILAGRO. **Clima afeta soja, e produção deverá ser inferior à do ano passado**. 2019. Disponível em: <https://www.brasilagro.com.br/conteudo/clima-afeta-soja-e-producao-devera-ser-inferior-a-do-ano-passado.html>. Acesso em: 20/11/2020. Citado na página 1.

BROCKWELL, P. J. et al. **Introduction to time series and forecasting**. Nova Iorque, EUA: Springer, 2016. Citado 2 vezes nas páginas 3 e 8.

CHATFIELD, C. **The Analysis of Time Series: An Introduction**. Boca Raton, EUA: CRC Press, 2016. Citado 2 vezes nas páginas 3 e 6.

COMPANHIA NACIONAL DE ABASTECIMENTO. **CALENDÁRIO AGRÍCOLA (PLANTIO E COLHEITA)**. 2019. Disponível em: <https://www.conab.gov.br/institucional/publicacoes/outras-publicacoes/item/download/28424_34d371f808b23d9bd37b9101c8ed5094>. Citado na página 26.

COMPANHIA NACIONAL DE ABASTECIMENTO. **SÉRIE HISTÓRICA DAS SAFRAS**. 2020. Disponível em: <https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras/item/download/34187_6be3627158aa3be10359cb18490ad520>. Acesso em: 18/11/2020. Citado na página 22.

CORDER, G. W.; FOREMAN, D. I. **Nonparametric statistics: A step-by-step approach**. Indianápolis, EUA: John Wiley & Sons, 2014. Citado na página 14.

DE FREITAS, N. et al. Sequential monte carlo methods for neural networks. In: **Sequential Monte Carlo methods in practice**. Nova Iorque: Springer, 2001. p. 359–379. Citado na página 16.

EBC. **Seca no Sul ameça produção de soja no estado**. 2020. Disponível em: <https://radios.ebc.com.br/brasil-rural/2020/03/seca-no-sul-ameca-producao-de-soja-no-estado>. Acesso em: 20/11/2020. Citado na página 1.

ELAVARASAN, D. et al. Forecasting yield by integrating agrarian factors and machine learning models: A survey. **Computers and Electronics in Agriculture**, Elsevier, v. 155, p. 257–282, 2018. Citado na página 18.

FACELI, K. et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro, Brasil: LTC, 2011. Citado na página 2.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996. Citado na página 3.

FISHER, R. A. Statistical methods and scientific inference. Hafner Publishing Co., 1956. Citado na página 15.

FREEDMAN, D.; PISANI, R.; PURVES, R. **Statistics**. Nova Iorque, EUA: Norton, 2007. Citado na página 15.

G1. **Clima adverso indica replantio para soja do Brasil e eleva atenção sobre 2ª safra**. 2019. Disponível em: <https://g1.globo.com/economia/agronegocios/noticia/2019/10/10/clima-adverso-indica-replantio-para-soja-do-brasil-e-eleva-atencao-sobre-2a-safra.ghtml>. Acesso em: 20/11/2020. Citado na página 1.

GEOGRAFIA E ESTATíSTICA, I. **Índice Nacional de Preços ao Consumidor Amplo - IPCA**. 2020. Disponível em: <ftp://ftp.ibge.gov.br/Precos_Indices_de_Precos_ao_Consumidor/IPCA/Serie_Historica/ipca_SerieHist.zip>. Acesso em: 18/11/2020. Citado na página 22.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. São Paulo, Brasil: Gulf Professional Publishing, 2005. Citado 4 vezes nas páginas 1, 3, 6 e 7.

GUJARATI, D.; PORTER, D. **Basic Econometrics**. Pensilvania, EUA: McGraw-Hill Irwin, 2009. Citado na página 14.

GULLI, A.; KAPOOR, A.; PAL, S. **Deep learning with TensorFlow 2 and Keras: regression, ConvNets, GANs, RNNs, NLP, and more with TensorFlow 2 and the Keras API**. Birmingham, Reino Unido: Packt Publishing Ltd, 2019. Citado 3 vezes nas páginas 10, 11 e 13.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Waltham, EUA: Elsevier, 2011. Citado 4 vezes nas páginas 1, 6, 7 e 8.

HAYKIN, S. **Redes neurais: princípios e prática**. São Paulo, Brasil: Bookman Editora, 2007. Citado na página 9.

HOCHREITER, S.; SCHMIDHUBER, J. Lstm can solve hard long time lag problems. **Advances in neural information processing systems**, MORGAN KAUFMANN PUBLISHERS, p. 473–479, 1997. Citado na página 10.

INTERNATIONAL MONETARY FUND. **IMF PRIMARY COMMODITY PRICES**. 2020. Disponível em: <https://www.imf.org/-/media/Files/Research/CommodityPrices/Monthly/ExternalData.ashx>. Acesso em: 18/11/2020. Citado na página 22.

INVESTING. **USD/BRL - Dólar Americano Real Brasileiro**. 2020. Disponível em: <https://br.investing.com/currencies/usd-brl-historical-data>. Acesso em: 18/11/2020. Citado na página 22.

KAUL, M.; HILL, R. L.; WALTHALL, C. Artificial neural networks for corn and soybean yield prediction. **Agricultural Systems**, Elsevier, v. 85, n. 1, p. 1–18, 2005. Citado na página 18.

LAI, G. et al. Modeling long-and short-term temporal patterns with deep neural networks. In: ACM, 41. **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**. Ann Arbor, EUA, 2018. p. 95–104. Citado na página 17.

LAZZERI, F. **Machine Learning for Time Series Forecasting with Python**. Indianápolis, EUA: John Wiley & Sons, 2020. Citado 2 vezes nas páginas 8 e 9.

LEHMANN, E. L.; ROMANO, J. P. **Testing statistical hypotheses**. Nova Iorque, EUA: Springer Science & Business Media, 2006. Citado na página 14.

LI, G.-q.; XU, S.-w.; LI, Z.-m. Short-term price forecasting for agro-products using artificial neural networks. **Agriculture and Agricultural Science Procedia**, Elsevier, v. 1, p. 278–287, 2010. Citado na página 16.

MAHTO, A. K. et al. Short-term forecasting of agriculture commodities in context of indian market for sustainable agriculture by using the artificial neural network. **Journal of Food Quality**, Hindawi, v. 2021, 2021. Citado na página 17.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, JSTOR, p. 50–60, 1947. Citado na página 15.

METEOROLOGIA, I. N. de. **Instituto Nacional de Meteorologia**. 2020. Disponível em: <https://bdmep.inmet.gov.br/>. Acesso em: 18/11/2020. Citado na página 21.

MORETTIN, P.; TOLOI, C. de C. **Análise de séries temporais**. São Paulo, Brasil: Edgard Blucher, 2006. Citado 4 vezes nas páginas 3, 4, 8 e 9.

OUYANG, H.; WEI, X.; WU, Q. Agricultural commodity futures prices prediction via long-and short-term time series network. **Journal of Applied Economics**, Taylor & Francis, v. 22, n. 1, p. 468–483, 2019. Citado na página 17.

PRASHANTHA, S. et al. A survey on crop analysis & agriculture commodities price prediction using machine learning techniques. 2020. Citado na página 17.

PRIESNITZ FILHO, W. et al. Uma visão dos avanços tecnológicos da iot em aplicações domésticas. In: **10th International Symposium on Technological Innovation**. Aracaju, Brasil: API, 2019. Citado na página 3.

REIS FILHO, I. J. R. et al. A integração de séries temporais e dados de textos para a previsão de preços futuros de milho e soja. 2020. Citado na página 17.

REUTERS. **Weather erratic for Brazil soy as harvesting starts**. 2012. Disponível em: <https://www.reuters.com/article/us-soy-brazil-weather/weather-erratic-for-brazil-soy-as-harvesting-starts-idUSTRE80217020120103>. Acesso em: 20/11/2020. Citado na página 1.

ROYSTON, P. Approximating the shapiro-wilk w-test for non-normality. **Statistics and computing**, Springer, v. 2, n. 3, p. 117–119, 1992. Citado na página 24.

SUN, J. et al. County-level soybean yield prediction using deep cnn-lstm model. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 19, n. 20, p. 4363, 2019. Citado na página 18.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Delhi, Índia: Pearson Education India, 2016. Citado na página 6.

THOMPSON, L. M. Effects of changes in climate and weather variability on the yields of corn and soybeans. **Journal of Production Agriculture**, Wiley Online Library, v. 1, n. 1, p. 20–27, 1988. Citado na página 1.

VU, K. M. **The ARIMA and VARIMA time series: their modelings, Analyses and Applications**. Ottawa, Canadá: AuLac Technologies Inc., 2007. Citado na página 16.

WANG, C.; GAO, Q. High and low prices prediction of soybean futures with lstm neural network. In: IEEE. **IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)**. Pequim, China, 2018. p. 140–143. Citado na página 16.

WEISS, N. **Introductory Statistics**. 9th. ed. Boston, EUA: Pearson, 2011. Citado na página 15.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics Bulletin**, JSTOR, v. 1, n. 6, p. 80–83, 1945. Citado na página 15.

ZAKI, M. J.; MEIRA, W. **Data mining and analysis: fundamental concepts and algorithms**. Nova Iorque, EUA: Cambridge University Press, 2014. Citado 2 vezes nas páginas 1 e 7.

ZHANG, D. et al. Prediction of soybean price in China using QR-RBF neural network model. **Computers and Electronics in Agriculture**, Elsevier, v. 154, p. 10–17, 2018. Citado na página 16.

ZHENG, A.; CASARI, A. **Feature engineering for machine learning: principles and techniques for data scientists**. Sebastopol, EUA: O'Reilly Media, 2018. Citado 2 vezes nas páginas 7 e 8.

APÊNDICE A – Prévia das Bases de Dados de Cada Estado

Tabela 7 – Base de Dados de Mato Grosso.

Data	PS	PT	TCM	TMM	URM	IT	IPCA	Dolar	CO
2000-01-31	*180.384*	*212.418*	*25.923*	*21.968*	*84.192*	*159.203*	*0.62*	*1.784*	*235.969*
2000-02-29	*185.758*	*222.718*	*25.416*	*21.648*	*87.707*	*127.782*	*0.13*	*1.768*	*181.338*
2000-03-31	*190.971*	*249.300*	*25.475*	*21.568*	*87.519*	*144.554*	*0.22*	*1.739*	*181.338*
2000-04-30	*197.104*	*111.200*	*25.406*	*20.665*	*84.311*	*204.950*	*0.42*	*1.805*	*130.712*
2000-05-31	*200.870*	*4.363*	*24.246*	*18.219*	*78.523*	*256.452*	*0.01*	*1.824*	*34.942*

Fonte: Autoria Própria.

Tabela 8 – Base de Dados de Mato Grosso do Sul.

Data	PS	PT	TCM	TMM	URM	IT	IPCA	Dolar	CO
2000-01-31	*180.384*	*126.40*	*27.363*	*22.775*	*72.331*	*241.220*	*0.62*	*1.784*	*99.753*
2000-02-29	*185.758*	*201.86*	*26.093*	*22.217*	*78.816*	*188.548*	*0.13*	*1.768*	*99.753*
2000-03-31	*190.971*	*207.48*	*25.051*	*21.441*	*83.369*	*170.240*	*0.22*	*1.739*	*71.812*
2000-04-30	*197.104*	*89.06*	*24.160*	*19.134*	*71.702*	*250.520*	*0.42*	*1.805*	*66.378*
2000-05-31	*200.870*	*41.78*	*20.616*	*15.726*	*73.076*	*203.800*	*0.01*	*1.824*	*57.685*

Fonte: Autoria Própria.

Tabela 9 – Base de Dados de Goiás.

Data	PS	PT	TCM	TMM	URM	IT	IPCA	Dolar	CO
2000-01-31	*180.38*	*275.727*	*24.243*	*20.316*	*78.390*	*158.237*	*0.62*	*1.784*	*28.585*
2000-02-29	*185.75*	*292.436*	*23.962*	*20.143*	*79.917*	*153.503*	*0.13*	*1.768*	*0.000*
2000-03-31	*190.97*	*279.255*	*23.955*	*20.169*	*79.297*	*172.064*	*0.22*	*1.739*	*0.000*
2000-04-30	*197.10*	*60.182*	*23.633*	*18.540*	*70.313*	*246.498*	*0.42*	*1.805*	*52.361*
2000-05-31	*200.87*	*4.536*	*22.236*	*16.274*	*62.974*	*269.313*	*0.01*	*1.824*	*52.361*

Fonte: Autoria Própria.

Tabela 10 – Base de Dados de Rio Grande do Sul.

Data	PS	PT	TCM	TMM	URM	IT	IPCA	Dolar	CO
2000-01-31	*180.384*	*128.744*	*23.391*	*18.512*	*71.435*	*242.090*	*0.62*	*1.784*	*0.000*
2000-02-29	*185.758*	*111.405*	*22.439*	*18.025*	*73.829*	*195.426*	*0.13*	*1.768*	*68.958*
2000-03-31	*190.971*	*182.355*	*20.678*	*16.371*	*77.931*	*199.169*	*0.22*	*1.739*	*68.958*
2000-04-30	*197.104*	*109.977*	*19.294*	*15.206*	*78.624*	*172.389*	*0.42*	*1.805*	*68.958*
2000-05-31	*200.870*	*161.477*	*14.656*	*10.889*	*80.285*	*162.085*	*0.01*	*1.824*	*68.958*

Fonte: Autoria Própria.

Tabela 11 – Base de Dados do Paraná.

Data	PS	PT	TCM	TMM	URM	IT	IPCA	Dolar	CO
2000-01-31	*180.38*	*195.443*	*23.048*	*18.808*	*77.433*	*185.886*	*0.62*	*1.784*	*222.95*
2000-02-29	*185.75*	*254.529*	*22.380*	*18.672*	*81.252*	*150.334*	*0.13*	*1.768*	*222.95*
2000-03-31	*190.97*	*130.557*	*21.515*	*17.838*	*80.921*	*158.101*	*0.22*	*1.739*	*222.95*
2000-04-30	*197.10*	*17.643*	*20.006*	*15.220*	*73.445*	*202.179*	*0.42*	*1.805*	*222.95*
2000-05-31	*200.87*	*34.571*	*16.176*	*11.707*	*76.768*	*164.616*	*0.01*	*1.824*	*0.000*

Fonte: Autoria Própria.

APÊNDICE B – Resultados dos Modelos Preditivos LSTM

Tabela 12 – Modelo Preditivo LSTM com Dados Climáticos.

Estados	Treino	Teste	RMSE	MAE
MT	2000-2009	2010	29,12	22,97
	2000-2010	2011	24,69	21,02
	2000-2011	2012	38,20	31,7
	2000-2012	2013	24,13	18,93
	2000-2013	2014	29,40	23,15
	2000-2014	2015	13,87	10,76
	2000-2015	2016	22,02	16,55
	2000-2016	2017	18,67	14,83
	2000-2017	2018	12,63	9,83
	2000-2018	2019	14,28	12,18
MS	2000-2009	2010	27,61	23,37
	2000-2010	2011	23,45	18,69
	2000-2011	2012	38,51	28,78
	2000-2012	2013	19,71	15,24
	2000-2013	2014	33,61	28,39
	2000-2014	2015	10,33	8,96
	2000-2015	2016	22,88	17,29
	2000-2016	2017	16,15	12,65
	2000-2017	2018	12,92	11,21
	2000-2018	2019	15,46	12,63
GO	2000-2009	2010	23,96	20,16
	2000-2010	2011	27,41	23,33
	2000-2011	2012	40,09	32,01
	2000-2012	2013	22,81	19,17
	2000-2013	2014	29,68	23,79
	2000-2014	2015	11,81	9,06
	2000-2015	2016	19,61	14,46
	2000-2016	2017	15,85	12,23
	2000-2017	2018	13,29	11,38
	2000-2018	2019	14,10	12,03
PR	2000-2009	2010	22,62	17,71
	2000-2010	2011	19,38	15,77
	2000-2011	2012	37,18	28,54
	2000-2012	2013	21,59	18,67
	2000-2013	2014	28,08	21,44
	2000-2014	2015	13,06	11,57
	2000-2015	2016	23,68	18,08
	2000-2016	2017	15,05	11,63
	2000-2017	2018	14,98	12,93
	2000-2018	2019	13,43	10,87
RS	2000-2009	2010	27,90	23,47
	2000-2010	2011	24,29	21,24
	2000-2011	2012	41,50	29,93
	2000-2012	2013	18,70	15,28
	2000-2013	2014	29,90	23,46
	2000-2014	2015	17,02	13,73
	2000-2015	2016	19,83	14,9
	2000-2016	2017	18,83	14,96
	2000-2017	2018	13,50	11,13
	2000-2018	2019	15,42	13,04

Fonte: Autoria Própria.

Tabela 13 – Modelo Preditivo LSTM sem Dados Climáticos.

Estados	Treino	Teste	RMSE	MAE
MT	2000-2009	2010	23,85	20,13
	2000-2010	2011	26,52	22,94
	2000-2011	2012	35,94	26,37
	2000-2012	2013	22,38	18,29
	2000-2013	2014	28,36	22,32
	2000-2014	2015	12,84	11,7
	2000-2015	2016	23,18	20
	2000-2016	2017	13,85	10,65
	2000-2017	2018	15,7	12,39
	2000-2018	2019	10,54	8,52
MS	2000-2009	2010	22,34	18,25
	2000-2010	2011	23,92	20,89
	2000-2011	2012	34,52	26,24
	2000-2012	2013	21,78	17,61
	2000-2013	2014	35,94	29,94
	2000-2014	2015	15,01	11,13
	2000-2015	2016	22,62	19,24
	2000-2016	2017	13,09	10,51
	2000-2017	2018	15,42	11,57
	2000-2018	2019	10,88	7,77
GO	2000-2009	2010	16,38	13,72
	2000-2010	2011	30,31	23,65
	2000-2011	2012	43,71	33,76
	2000-2012	2013	20,81	18,18
	2000-2013	2014	29,88	23,76
	2000-2014	2015	11,63	10,14
	2000-2015	2016	21,53	18,58
	2000-2016	2017	13,38	10,65
	2000-2017	2018	16,62	13,22
	2000-2018	2019	11,4	9,36
PR	2000-2009	2010	23,36	19,91
	2000-2010	2011	22,32	17,98
	2000-2011	2012	37,71	28,75
	2000-2012	2013	20,19	16,54
	2000-2013	2014	29,25	23,06
	2000-2014	2015	12,92	11,01
	2000-2015	2016	22,77	19,62
	2000-2016	2017	13,72	10,31
	2000-2017	2018	15,07	11,83
	2000-2018	2019	10,95	9,04
RS	2000-2009	2010	23,11	19,41
	2000-2010	2011	23,2	19,22
	2000-2011	2012	37,23	28,4
	2000-2012	2013	18,2	15,92
	2000-2013	2014	29,78	23,9
	2000-2014	2015	12,59	10,46
	2000-2015	2016	21,26	18,46
	2000-2016	2017	14,94	11,65
	2000-2017	2018	16,82	13,09
	2000-2018	2019	11,28	8,68

Fonte: Autoria Própria.