

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Marciano Machado Saraiva

**PREVISÃO DA TEMPERATURA DO AR NO BRASIL UTILIZANDO ESTAÇÕES
METEOROLÓGICAS E MODELOS DE APRENDIZADO DE MÁQUINA**

Belo Horizonte
2020

Marciano Machado Saraiva

**PREVISÃO DA TEMPERATURA DO AR NO BRASIL UTILIZANDO ESTAÇÕES
METEOROLÓGICAS E MODELOS DE APRENDIZADO DE MÁQUINA**

Trabalho de Conclusão de Curso apresentado
ao Curso de Especialização em Ciência de Da-
dos e Big Data como requisito parcial à obtenção
do título de especialista.

Belo Horizonte
2020

SUMÁRIO

1. Introdução	4
1.1. Contextualização	4
1.2. O problema proposto	4
2. Coleta de Dados	5
2.1. Coleta dos Dados das Estações Meteorológicas Convencionais do INMET	6
2.2. Coleta dos Dados das Estações Meteorológicas Automáticas do INMET	7
2.3. Coleta dos Dados das Estações Meteorológicas Automáticas do LabMet	8
3. Processamento e Tratamento dos Dados	9
3.1. Convertendo todos os dados para um formato de arquivo estruturado	9
3.2. Transformando os dados das estações convencionais do INMET em registros diários	10
3.3. Transformando os dados das estações automáticas do INMET em registros diários	11
3.4. Registros diários das estações automáticas do LabMet	12
3.5. Eliminando valores inconsistentes	12
4. Análise e Exploração dos Dados	13
5. Criação dos Modelos de Aprendizado de Máquina	17
5.1. Criação do modelo ARIMA	18
5.1.1. Decomposição das séries temporais	18
5.1.2. Autocorrelação das séries temporais	20
5.1.3. Verificando a estacionariedade	22
5.1.4. Parametrizando o modelo ARIMA	26
5.2. Criação do modelo LSTM	27
5.2.1. Conjuntos de dados de treinamento, validação e teste	27
5.2.2. Normalização dos dados	28
5.2.3. Modelo LSTM desenvolvido	28
6. Apresentação dos Resultados	29
7. Links	33
REFERÊNCIAS	34

1. Introdução

1.1. Contextualização

As atividades de previsão desempenham um papel fundamental em nossas vidas. Todos os dias, a previsão do tempo nos informa como estará o tempo no dia seguinte, na semana seguinte, e até no mês seguinte. A temperatura, sendo um dos mais importantes parâmetros apresentados em previsões do tempo, tem um impacto direto na evaporação, derretimento de neve, geada e um impacto indireto nas condições atmosféricas e precipitação (Hansen et al. 2006). De acordo com recentes estudos sobre os impactos das mudanças climáticas, a agricultura, vegetação, recursos hídricos e o turismo são os setores mais afetados diretamente por mudanças de temperatura. Portanto, é necessário prever a temperatura com precisão para evitar perigos inesperados causados pela variação da temperatura, como geadas e secas que podem causar danos financeiros e perdas humanas (Kaymaz 2005).

1.2. O problema proposto

Diante desse contexto, este trabalho tem como objetivo prever o comportamento da temperatura média do ar para um intervalo de um ano utilizando séries temporais de temperatura obtidas de estações meteorológicas distribuídas por todo o território brasileiro.

Para facilitar o entendimento do problema e da solução proposta, utilizamos a técnica dos 5W's, que consiste em responder as seguintes perguntas:

Why?: Variações de temperatura podem ter impacto direto na produção agrícola, geração de energia, turismo e até na saúde da população, por isso, prever a temperatura com precisão é essencial para evitarmos esses riscos.

Who?: Os dados foram coletados das estações convencionais e automáticas do Instituto Nacional de Meteorologia do Brasil (INMET) e do Laboratório de Meteorologia da Universidade Federal do Vale do São Francisco (LabMet).

What?: Prever o comportamento da temperatura média do ar para um intervalo de um ano utilizando séries temporais de temperatura obtidas de estações meteorológicas espalhadas por todo o território brasileiro.

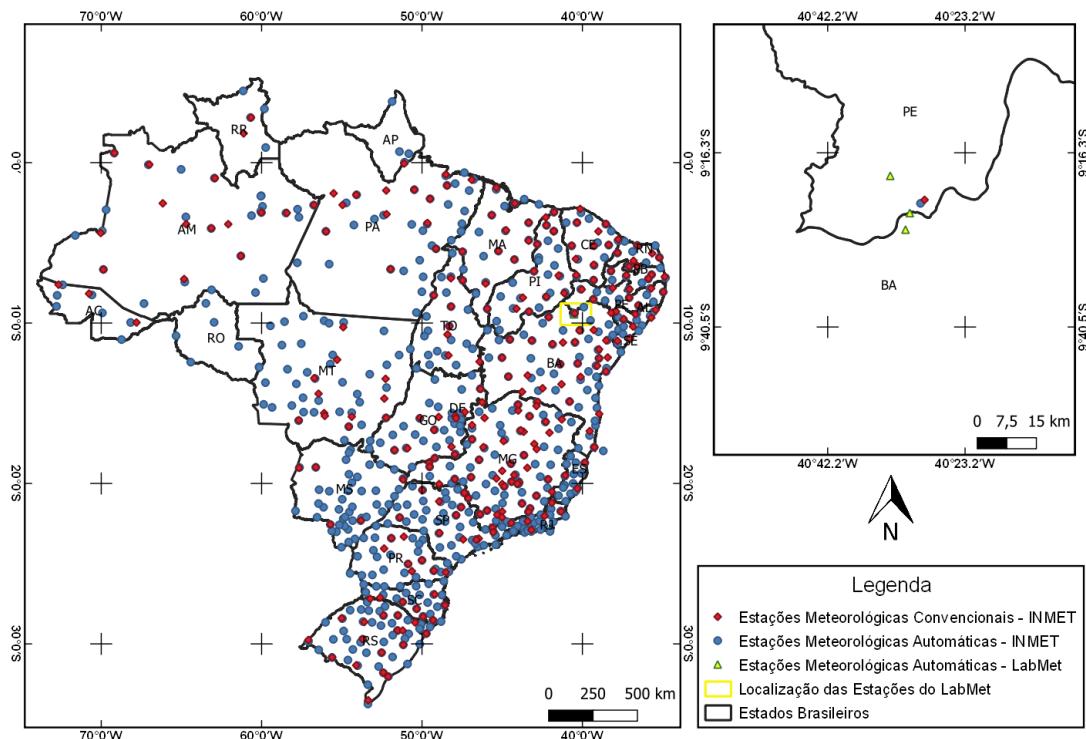
Where?: Estações meteorológicas espalhadas por todo o território brasileiro.

When?: O conjunto de dados das estações convencionais do INMET contém observações do período de 1961 à 2019. Já as estações automáticas, também do INMET, que começaram a ser implantadas no Brasil a partir no inicio deste século, possui dados de 2000 à 2019. Por último, as estações automáticas do LabMet possuem dados de 2007 à 2019.

2. Coleta de Dados

Neste projeto, utilizamos três diferentes conjuntos de dados meteorológicos. Os dois primeiros conjuntos, que representam maior parte dos dados, foram adquiridos de 265 estações meteorológicas convencionais e 610 estações automáticas vinculadas ao Instituto Nacional de Meteorologia (INMET). O terceiro conjunto de dados foi obtido de 3 estações meteorológicas automáticas administradas pelo Laboratório de Meteorologia da Universidade Federal do Vale do São Francisco (LabMet). Essas estações coletam dados tais como: precipitação, temperatura do ar, umidade relativa do ar, velocidade e direção do vento, radiação solar, dentre outras variáveis. A principal diferença entre esses dois tipos de estações, convencional e automática, é que, as convencionais requerem a presença diária do observador para a coleta dos dados, enquanto as automáticas operam por meio de sensores eletrônicos que alimentam o sistema de aquisição de dados, tendo como principal vantagem o registro contínuo de todas as variáveis. A Figura 1 ilustra a distribuição espacial no território nacional de todas as estações utilizadas neste trabalho. Nas próximas seções descreveremos a cobertura temporal, espacial e as demais informações de cada um dos conjuntos de dados.

Figura 1 – Distribuição espacial das estações meteorológicas utilizadas neste trabalho.



2.1. Coleta dos Dados das Estações Meteorológicas Convencionais do INMET

Distribuídas ao longo de todo o território brasileiro, as estações meteorológicas convencionais vinculadas ao INMET representam o conjunto de dados com maior série histórica dos três conjuntos, com observações que datam o ano de 1961. Ao todo, foram obtidas mais de 12 milhões de observações coletadas de todas as estações convencionais do INMET para o período de 1961 à 2019. Nas estações convencionais do INMET, cada observação corresponde à uma coleta realizada pelo observador em alguns períodos escolhidos do dia.

Os dados das estações convencionais do INMET foram baixados, de forma automatizada, em fevereiro de 2020, a partir do antigo portal do INMET¹ utilizando a biblioteca Selenium (Salunke 2014), versão 3.141.0. Desenvolvemos um script na linguagem de programação Python que percorreu cada uma das estações realizando

¹O antigo portal do INMET encontrava-se disponível, até a data do download dos dados, no endereço <http://www.inmet.gov.br>

o *download*, em formato HTML, de todas as variáveis disponíveis na plataforma. A Tabela 1 apresenta a lista das variáveis disponíveis nas estações convencionais.

Tabela 1 – Variáveis disponíveis no conjunto de dados das estações meteorológicas convencionais do INMET.

Nome da coluna	Descrição	Unidade de Medida
Estacao	Código da Estação	-
Data	Data da Coleta do Dado	DD/MM/YYYY
Hora	Hora da Coleta do Dado	HHMM
Precipitacao	Precipitação Acumulada	mm
TempBulboSeco	Temperatura do Bulbo Seco	°C
TempBulboUmido	Temperatura do Bulbo Úmido	°C
TempMaxima	Temperatura Máxima do Ar	°C
TempMinima	Temperatura Miníma do Ar	°C
UmidadeRelativa	Umidade Relativa do Ar	%
PressaoAtmEstacao	Pressão Atmosférico no Nível da Estação	mbar
PressaoAtmMar	Pressão Atmosférico no Nível do Mar	mbar
DirecaoVento	Direção do Vento	Código INMET
VelocidadeVento	Velocidade do Vento	m/s
Insolacao	Insolação	horas
Nebulosidade	Nebulosidade	décimos
Evaporacao Piche	Evaporação de Piche	mm
Temp Comp Media	Temperatura Compensada Média	°C
Umidade Relativa Media	Umidade Relativa Média do Ar	%
Velocidade do Vento Media	Velocidade Média do Vento	m/s

2.2. Coleta dos Dados das Estações Meteorológicas Automáticas do INMET

O segundo conjunto de dados utilizado neste trabalho foi obtido das estações meteorológicas automáticas vinculadas ao INMET, assim como as convencionais, distribuídas por todo o território nacional. Diferente das estações convencionais, que possuem três observações ao longo do dia, e dados desde 1961, as estações automáticas do INMET coletam dados horários desde o ano de 2000. Ao todo, foram baixadas quase 55 milhões de observações coletadas de todas as 610 estações meteorológicas automáticas do INMET, para o período de 2000 à 2019.

Os dados de cada uma das estações automáticas foram baixados, em formato CSV, em agosto de 2020 diretamente do novo portal do INMET². A Tabela 2 apresenta a lista das variáveis disponíveis nos dados baixados.

²O novo portal do INMET encontrava-se disponível, até a data do download dos dados, no endereço <https://portal.inmet.gov.br/>

Tabela 2 – Variáveis disponíveis no conjunto de dados das estações meteorológicas automáticas do INMET.

Nome da coluna	Descrição	Unidade de Medida
ESTACAO	Código da Estação	-
DATA (YYYY-MM-DD)	Data da observação	YYYY-MM-DD
HORA (UTC)	Hora da observação	HH
PRECIPITACAO TOTAL HORARIO (mm)	Precipitação Acumulada	mm
PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)	Pressão Atmosférica Instantânea no Nível da Estação	°C
PRESSAO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)	Pressão Atmosférica Máxima	mbar
PRESSAO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)	Pressão Atmosférica Mínima	mbar
RADIACAO GLOBAL (W/m2)	Radiação Sola Global	W/m2
TEMPERATURA DO AR - BULBO SECO, HORARIA (C)	Temperatura do Bulbo Seco	°C
TEMPERATURA DO PONTO DE ORVALHO (C)	Pressão Atmosférica no Nível da Estação	mbar
TEMPERATURA MAXIMA NA HORA ANT. (AUT) (C)	Pressão Atmosférica no Nível do Mar	mbar
TEMPERATURA MINIMA NA HORA ANT. (AUT) (C)	Direção do Vento	Código INMET
TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (C)	Velocidade do Vento	m/s
TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (C)	Insolação	horas
UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)	Nebulosidade	décimos
UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)	Evaporação de Piche	mm
UMIDADE RELATIVA DO AR, HORARIA (%)	Temperatura Compensada Média	°C
VENTO, DIRECAO HORARIA (gr)	Umidade Relativa Média do Ar	%
VENTO, RAJADA MAXIMA (m/s)	Velocidade Média do Vento	m/s
VENTO, VELOCIDADE HORARIA (m/s)	Velocidade Média do Vento	m/s

2.3. Coleta dos Dados das Estações Meteorológicas Automáticas do LabMet

O último conjunto de dados utilizado foi obtido de 3 estações meteorológicas automáticas localizadas no Vale do Rio São Francisco, entre os estados da Bahia e Pernambuco. Estas estações são administradas pelo Laboratório de Meteorologia da Universidade Federal do Vale do São Francisco (LabMet), que disponibilizam dados diário das estações através de seu portal³. Ao todo, baixamos, em formato XLS, mais de 9 mil registros disponibilizados para as três estações automáticas entre o período de 2007 à 2020. A Tabela 3 descreve as variáveis contidas no conjunto de dados.

Tabela 3 – Descrição dos campos/colunas dos datasets.

Nome da coluna	Descrição	Unidade de Medida
Estacao	Código gerado como identificar da estação	-
Data	Data da observação	YYYY-MM-DD
temperatura	Temperatura média do ar	°C
temperatura maxima	Temperatura máxima do ar	°C
temperatura minima	Temperatura mínima do ar	°C
umidade	Umidade relativa média do ar	%
umidade maxima	Umidade relativa máxima do ar	%
umidade minima	Umidade relativa mínima do ar	%
vento maxima 10m	Velocidade máxima do vento diária	m/s
rad. solar global	Radiação solar global	MJ/m ² /dia
evap. de referencia	Evapotranspiração de Referência	mm/dia

³O portal do LabMet encontrava-se disponível, até a data de agosto de 2020, no endereço <http://labmet.univasf.edu.br>

3. Processamento e Tratamento dos Dados

Nesta etapa, nosso principal objetivo foi compatibilizar os diferentes conjuntos de dados, de forma a construir um único conjunto de dados para facilitar o processo de análise e geração dos resultados.

3.1. Convertendo todos os dados para um formato de arquivo estruturado

Os dados obtidos para cada uma das estações meteorológicas convencionais do INMET, em formato HTML, foram convertidos para o formato CSV utilizando a biblioteca Pandas (McKinney et al. 2011) e o kit de ferramentas lxml (Behnel, Faassen e Bicking 2005). Em seguida, os dados de todas as estações convencionais foram combinados em um único arquivo estruturado em formato CSV.

Os dados das estações automáticas do INMET já foram obtidos em formato CSV, porém divididos em um arquivo por estação e por ano. Para este conjunto de dados, também utilizamos a biblioteca Pandas para combinar todos os dados em um único arquivo, também no formato CSV.

Os dados das estações automáticas do LabMet foram obtidos em formato XLS, um arquivo por variável para cada estação. Cada arquivo de uma determinada variável, para uma determinada estação, possuia múltiplas planilhas, cada planilha continha os dados para um determinado ano. Utilizamos a biblioteca Pandas para transformar e agrupar todos os dados das diferentes estações, variáveis e anos em um único arquivo estruturado em formato CSV.

Após este processo de transformação, obtivemos os conjuntos de dados descritos na Tabela 4.

Tabela 4 – Conjunto de dados utilizados neste trabalho.

Nome do Conjunto de Dados	Quantidade de Registros
Estações Meteorológicas Convencionais - INMET	12.251.335
Estações Meteorológicas Automáticas - INMET	54.840.384
Estações Meteorológicas Automáticas - LabMet	9.892

Antes de avançarmos no processo de tratamento e análise dos dados, prepa-

ramos um ambiente apropriado para o processamento desse conjunto de dados. Em um computador com 16 núcleos lógicos de processamento, 32GB de memória RAM e 256GB de SSD, configuramos um ambiente docker¹ com as plataformas Hadoop, versão 3.1.2, e Spark, versão 3.0.1. Todos os dados em formato CSV foram inseridos no sistema de arquivos HDFS do Hadoop para serem processados pelo Spark posteriormente.

O primeiro conjunto de dados, das estações meteorológicas convencionais do INMET, possui de 2 a 3 registros por dia, dependendo do ano observado, enquanto que as estações automáticas do INMET possui 24 registros por dia (horário), e as estações automáticas do LabMet possuem apenas um único registro por dia (diário). Pensando nisso, decidimos transformar todos os dados para a frequência diária. Esta transformação permitiu juntarmos todos os dados em um único grande conjunto com registros diários das variáveis analisadas. Nas próximas seções descreveremos os passos realizados para realizar essa transformação.

3.2. Transformando os dados das estações convencionais do INMET em registros diários

Ao analisar as variáveis de temperatura máxima, temperatura mínima e temperatura média das estações convencionais do INMET, identificamos em torno de 67% de registros nulos em cada uma das variáveis. Isso se deve, principalmente, pelo fato dessas variáveis estarem presentes em apenas um dos dois ou três registros diários disponibilizados pelas estações. A temperatura máxima e temperatura média estão presentes no registro das 00:00 horas (UTC) e a temperatura mínima no registro das 12:00 horas (UTC). Com isso, transformamos esse dado para a frequência diária utilizando como registro diário a primeira ocorrência do valor dessas variáveis no dia. Após o procedimento de transformação para dados diários, a quantidade de registro diminuiu de 12.251.335 para 4.224.027 e a porcentagem de dados nulos diminuiu de 67% para cerca de 6% nas temperaturas máximas e mínimas, e para próximo de 12% para a temperatura média. A Tabela 6 apresenta a quantidade de dados nulos antes

¹A configuração docker utilizada para criar o ambiente usado no processamento dos dados está disponível em: <https://github.com/saraivaufc/bigdata-docker>.

e após a transformação para dados diários.

Tabela 5 – Quantidade de registros nulos antes e após a conversão para dados diários das estações meteorológicas convencionais do INMET.

Variável	Registros nulos (Qtd.)	Registros nulos (%)	Registros nulos diários (Qtd.)	Registros nulos diários (%)
Temperatura máxima	8.300.413	67,75%	273.105	6,46%
Temperatura mínima	8.292.199	67,68%	264.891	6,27%
Temperatura média	8.528.574	69,61%	501.266	11,86%

3.3. Transformando os dados das estações automáticas do INMET em registros diários

Nas estações meteorológicas automáticas do INMET, os valores nulos estão indicados com o valor -9999, então o primeiro passo foi convertermos todos os valores -9999 para o valor "NULO". Em seguida, realizamos a contagem dos valores nulos dos dados horários das variáveis Temperatura Máxima, Temperatura Mínima e Temperatura do Bulbo seco. Utilizamos a temperatura do bulbo seco horária para estimar a temperatura média do ar diária. Esta primeira análise indicou que apenas aproximadamente 4% dos dados das variáveis analisadas estavam com valores nulos. Os dados estão disponíveis originalmente na periodicidade horária, para convertermos para dados diários de temperatura máxima, mínima e média, calculamos o máximo valor da temperatura máxima, o mínimo valor da temperatura mínima e a média dos 24 valores da temperatura do bulbo seco, respectivamente. Após a conversão dos dados horários para dados diários, a quantidade de registros passaram de 54.840.384 para 2.072.346, e a quantidade de dados nulos passaram de aproximadamente 4% para cerca de 6% nos dados diários. O aumento em proporção dos dados nulos na escala diária indica que a escala horária tinha muitos dados nulos agrupados em um mesmo período de um dia.

Tabela 6 – Quantidade de registros nulos antes e após a conversão para dados diários das estações meteorológicas automáticas do INMET.

Variável	Registros nulos (Qtd.)	Registros nulos (%)	Registros nulos diários (Qtd.)	Registros nulos diários (%)
Temperatura máxima	496.231	4,05%	125.226	6,04%
Temperatura mínima	496.396	4,05%	125.226	6,04%
Temperatura média	493.102	4,02%	125.014	6,03%

3.4. Registros diários das estações automáticas do LabMet

Por estarem na frequência diária, os dados das estações meteorológicas automáticas do LabMet não sofreram alterações. Verificando a ocorrência de valores nulos, não identificamos nenhum valor nulo na variável de temperatura mínima, um registro nulo na temperatura máxima, e dois registros nulos na variável da temperatura média.

3.5. Eliminando valores inconsistentes

Nesta etapa verificamos a ocorrência de dados inconsistentes. Geralmente, essas inconsistências são geradas por erros de leituras dos sensores, como problemas de calibração ou defeitos. Baseado em Baba, Vaz e Costa 2014, subdividimos as inconsistências em três grupos, inconsistências de limites, inconsistências lógicas, e inconsistências temporais.

- Inconsistência de limites: identifica valores que não respeitam os limites básicos da variável representada, ou seja, um valor fisicamente impossível de ser obtido. Nos dados de temperatura consideramos como inconsistência, para o Brasil, valores menores que -30°C e maiores que 50°C. Qualquer valor fora do intervalo estaria incorreto, pois representaria um valor impossível para a variável em questão;
- Inconsistências lógicas: quando temos dados de temperatura máxima, média e mínima referentes a um intervalo de tempo monitorado, estes dados devem respeitar sua condição lógica básica, ou seja, o valor de temperatura média para um período não pode ser maior que a temperatura máxima ou menor que a temperatura mínima deste mesmo período. Do contrário, esses valores serão considerados inconsistentes;
- Inconsistências temporais: identifica valores inconsistentes baseado em seus valores históricos. Para esta análise, utilizamos a metodologia elaborada por (Mateo e Leung 2008), em que consiste em determinar a consistência do dado

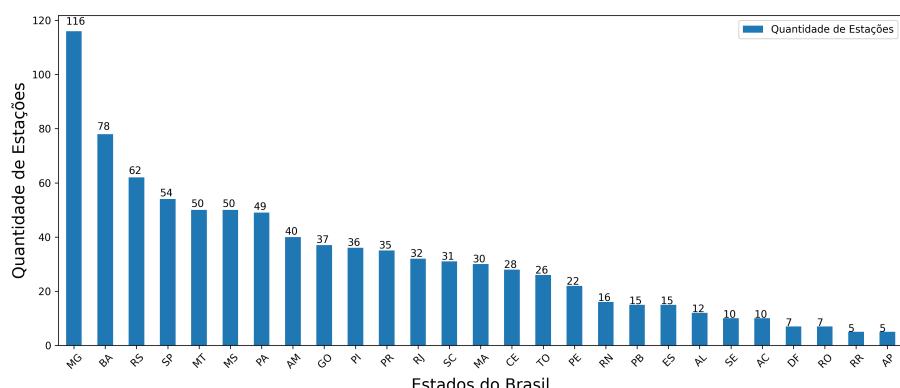
avaliado baseado no valor do dia anterior, no valor do dia seguinte e nos dados referentes aos três dias (anterior, corrente, posterior) do ano anterior. Se o valor do dado analisado se afastar da média dos valores analisados 3 vezes o valor do desvio padrão, este valor pode indicar problemas de leituras.

Ao final desta etapa, utilizamos os dados de temperatura máxima e mínima para validar o dado de temperatura média, mas após esta etapa eliminamos os dados de temperatura máxima e mínima e passamos a trabalhar apenas com a temperatura média. Antes de eliminar os valores inconsistentes, nosso conjunto de dados, resultado da combinação de todas as bases de dados, continha 6.300.381 registros no total, com 619.192 destes com valores nulos, ou seja, 9,82% de todos os registros. Após a eliminação dos valores inconsistentes, a quantidade de registros nulos subiu para 958.992, representando 15,22% do dado.

4. Análise e Exploração dos Dados

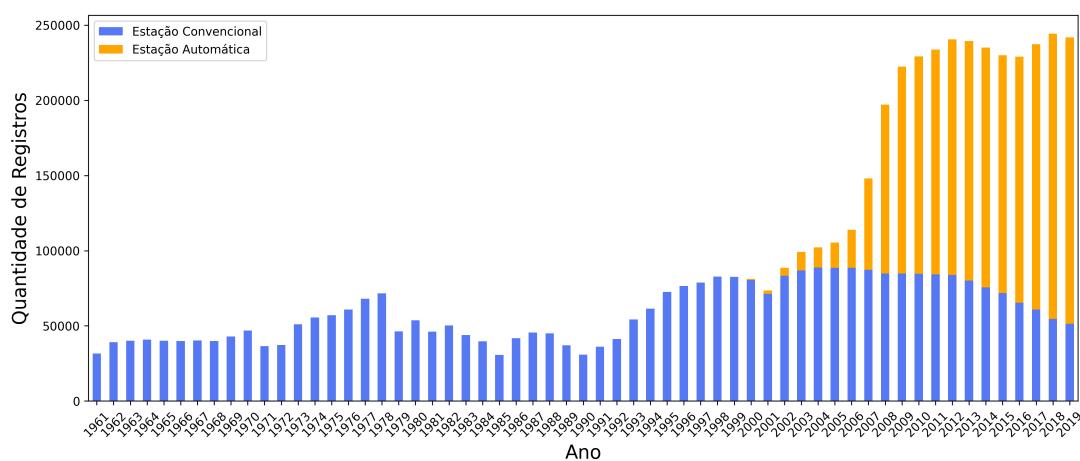
Nesta etapa da análise dos dados, uma das primeiras informações que levantamos foi a quantidade de estações por estado. Na Figura 2 podemos observar que os estados com maior quantidade de estações meteorológicas instaladas são, em primeiro lugar, Minas Gerais com 116 estações, seguido pelos estados da Bahia com 78 estações, Rio Grande do Sul com 62 estações e São Paulo com 54 estações. Em contrapartida, os estados de Rondônia, Roraima, e Amapá possuem apenas 7, 5 e 5 estações, respectivamente.

Figura 2 – Distribuição, por estado, das estações meteorológicas analisadas.



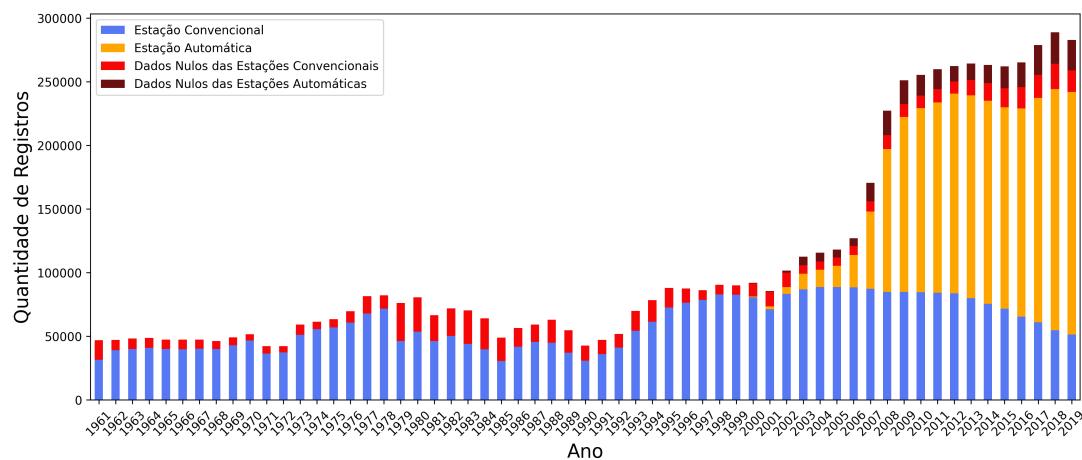
Dentre as estações analisadas, há estações convencionais com dados disponíveis desde o ano de 1961, até estações automáticas recentemente instaladas, com dados apenas para os últimos anos. Por isso, é importante entendermos como varia a disponibilidade de dados ao longo de todo o período afim de nos orientarmos melhor na criação dos nossos modelos. Para isto, apresentamos na Figura 3 a disponibilidade das observações das estações ao longo de todo o período analisado. Podemos observar que a partir do ano de 2002 há um crescimento na quantidade de dados disponíveis, com maiores crescimentos nos anos de 2007 e 2008. Isso se deve pela adoção ao processo de automação de dados meteorológicos através de monitoramento por estações automáticas que surgiu no Brasil no inicio deste século.

Figura 3 – Quantidade de registros disponíveis ao longo de todo o período analisado.



Ainda observando a Figura 3, é possível identificarmos um declínio na disponibilidade dos dados em alguns anos específicos. Por exemplo, de 1978 para 1979 houve uma diminuição na quantidade de dados disponíveis. Para entendermos melhor essa variação, e avaliarmos melhor a quantidade de dados que ficaram ausentes no conjunto final, ilustramos na Figura 4 a quantidade de registros que estão ausentes para os anos apresentados.

Figura 4 – Quantidade de registo disponíveis e nulos ao longo de todo o período analisado.



A ausência de dados em estações meteorológicas pode ser causada por várias razões, entre elas podemos citar a mudança geográfica da estação, mudança no instrumental, tempo de observação e práticas observacionais utilizadas (Oliveira e Souza 2019).

Uma informação que também é importante identificarmos nos dados, é se, valores extremos que ocorreram, de fato, na realidade, não foram removidos durante o processo de limpeza dos dados. Para isso, ilustramos nas Figuras 5 e 6 os recordes de temperatura máxima e mínima para as estações convencionais do INMET, pois estas são as que possuem maior série histórica e maior consistência dos valores.

Figura 5 – Recorde das maiores temperaturas já registradas no período de 1961 a 2019 nas estações convencionais do INMET.

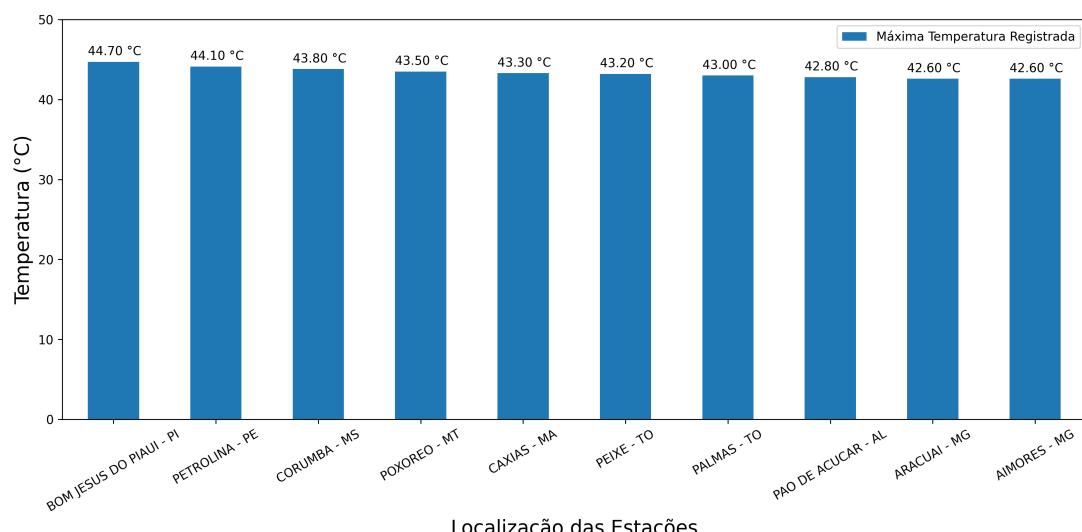
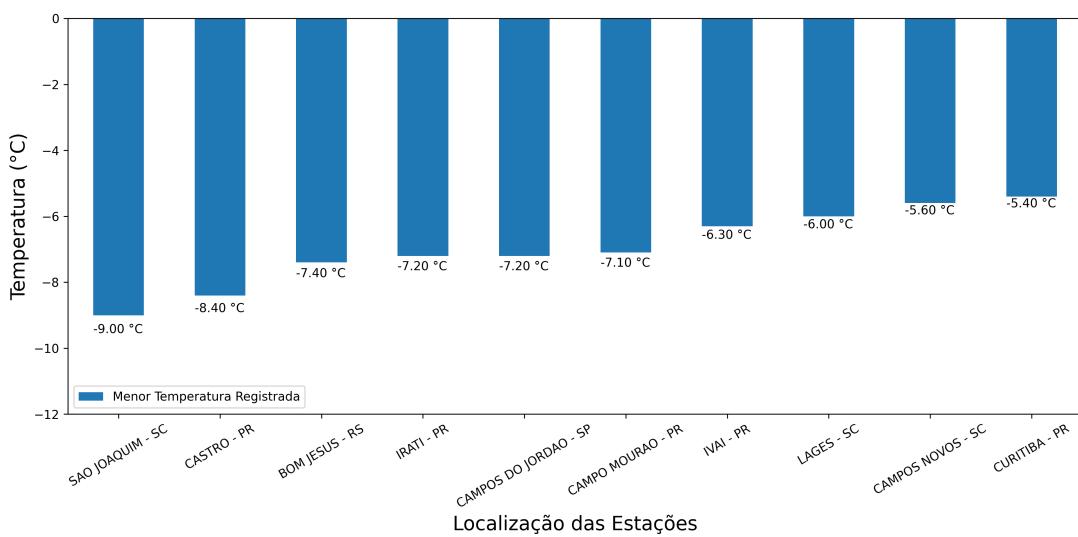


Figura 6 – Recorde das menores temperaturas já registradas no período de 1961 a 2019 nas estações convencionais do INMET.



Realizando uma pesquisa na internet sobre os recordes de temperatura no Brasil encontramos as seguintes informações:

"A maior temperatura registrada oficialmente no Brasil foi 44,7 °C em Bom Jesus, Piauí, em 21 de novembro de 2005, superando o recorde também oficial de Orleans, Santa Catarina, de 44,6 °C, de 6 de janeiro de 1963. Já a menor temperatura registrada foi de -17,8 °C no Morro da Igreja, em Urubici, Santa Catarina, em 29 de junho de 1996 (registro extraoficial). A menor temperatura registrada oficialmente no país foi de -14 °C, no município de Caçador, no mesmo estado, em 11 de junho de 1952." (Wikipedia 2020).

Essa referência aponta o município de Bom Jesus, no estado do Piauí, com a maior temperatura já registrada no Brasil, com a temperatura alcançando 44,70 °C em 21 de novembro de 2005, exatamente como os dados das estações convencionais apresentaram. Também segundo a referência, a menor temperatura já registrada após 1961 foi de -17,8 °C no Morro da Igreja, em Urubici, Santa Catarina, em 29 de junho de 1996. Analisando os dados, diferente do que apontou a informação encontrada, a menor temperatura registrada pelas estações convencionais foi no município de São Joaquim, em Santa Catarina. Cruzando o mapa de municípios de Santa Catarina com as estações que estamos analisando, observamos que o município de São Joaquim faz fronteira com o município de Urubici e que, dos dois municípios, apenas o de São

Joaquim possui, nos dados analisados, estações meteorológicas, o que indica um possível motivo das diferenças entre as informações.

5. Criação dos Modelos de Aprendizado de Máquina

Nesta seção, apresentaremos os modelos preditivos desenvolvidos em linguagem Python para a previsão da temperatura para o ano de 2019 utilizando modelos treinados com dados de anos anteriores. Neste trabalho, avaliamos dois modelos preditivos para a previsão da temperatura, o modelo Autoregressive Integrated Moving Average (ARIMA) (Whittle 1951) e um modelo baseado em redes neurais recorrentes utilizando a arquitetura Long Short-Term Memory (LSTM) (Hochreiter e Schmidhuber 1997).

O modelo ARIMA é um dos modelos lineares mais populares na previsão de séries temporais das últimas décadas, sua popularidade é devido, principalmente, às suas propriedades estatísticas, bem como à conhecida metodologia Box-Jenkins (Box, Jenkins e Reinsel 2011) no processo de construção do modelo (Zhang 2003).

Rede Neural Recorrente (RNN) é um tipo de Rede Neural onde a saída da etapa anterior é alimentada como entrada para a etapa atual, ou seja, ela possui uma “memória” que guarda todas as informações sobre o que foi calculado. A arquitetura de RNN que utilizamos foi a chamada Long Short-Term Memory - LSTM, uma arquitetura de rede neural recorrente específica que foi projetada para modelar sequências temporais e suas dependências de longo alcance com mais precisão do que RNNs convencionais (Sak, Senior e Beaufays 2014), característica importante quando trabalhamos com longas séries temporais como os dados histórico de temperatura no Brasil.

Para avaliarmos os modelos desenvolvidos, selecionamos, através de uma amostragem aleatória simples, 10% das estações convencionais do INMET, 10% das estações automáticas também do INMET e uma estação automática do LabMet. Após a amostragem, temos um conjunto de amostra de estações que guardam a mesma proporção do conjunto contendo todas as estações. Ao todo, utilizamos 88 estações para avaliar os modelos. A Tabela 7 apresenta a quantidade de estações que usamos para a avaliação dos modelos de previsão.

Tabela 7 – Proporção de estações utilizadas para avaliar os modelos desenvolvidos.

Fonte do Dado	Total de Estações	Estações da amostra de avaliação	%
Estações Convencionais do INMET	265	26	10%
Estações Automáticas do INMET	610	61	10%
Estações Automáticas do LabMet	3	1	33%
TOTAL	878	88	-

5.1. Criação do modelo ARIMA

5.1.1. Decomposição das séries temporais

Uma abstração útil para selecionar métodos de previsão é quebrar uma série temporal em componentes sistemáticos e não sistemáticos. Os componentes sistemáticos podemos considerar que são os que possuem consistência ou recorrência e podem ser descritos e modelados. Já os componentes não sistemáticos são os que não são possíveis de serem modelados (Box, Jenkins e Reinsel 2011).

Pensando nisso, podemos dividir os componentes das séries temporais em:

- nível: o valor médio da série;
- tendência: o valor crescente ou decrescente na série;
- sazonalidade: o ciclo repetitivo de curto prazo na série;
- ruído: a variação aleatória da série.

Para esta análise, apresentaremos apenas os resultados para três estações de exemplo que foram escolhidas a partir das estações de avaliação.

Decompondo as séries temporais das estações de exemplo, obtivemos os resultados apresentados nas figuras 7, 8 e 9.

Figura 7 – Decomposição da série temporal da temperatura do ar para a estação convencional localizada no município de Balsas, no estado do Maranhão.

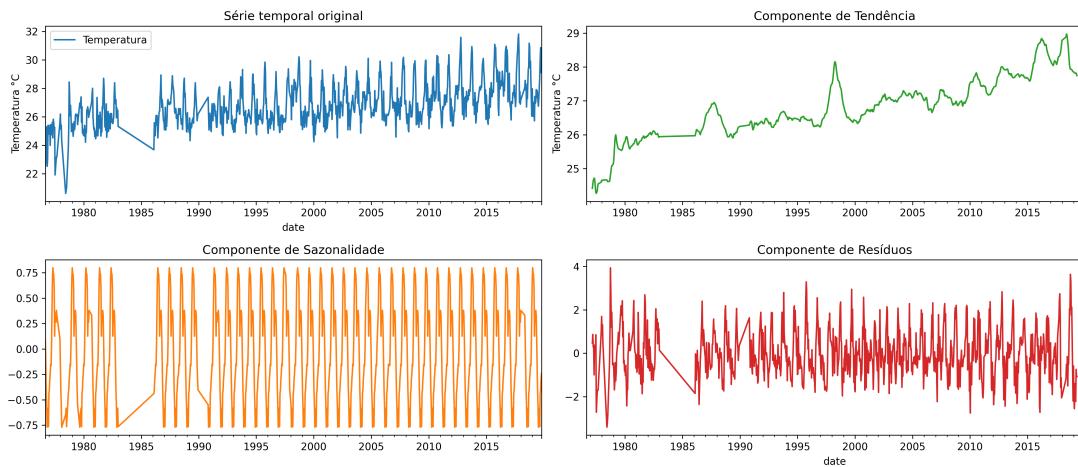


Figura 8 – Decomposição da série temporal da temperatura do ar para a estação automática localizada no município de Ariranha, no estado de São Paulo.

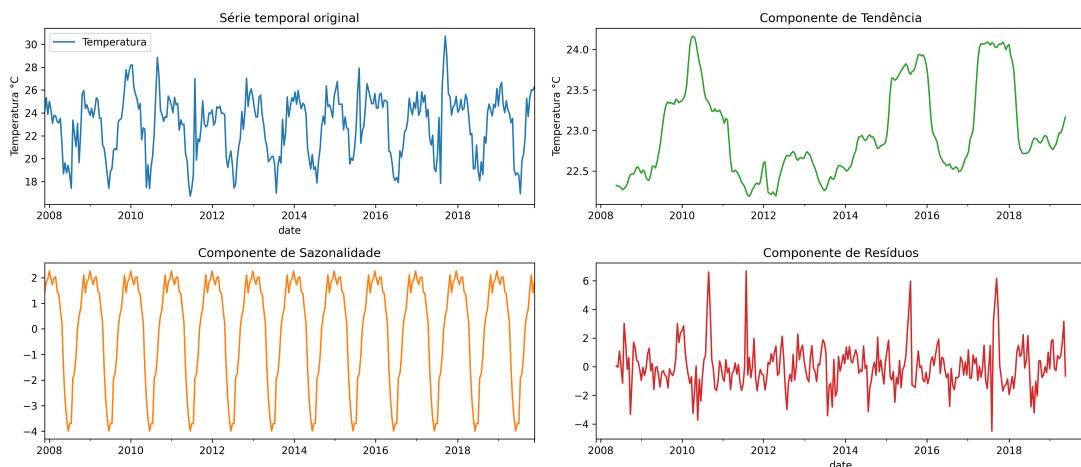
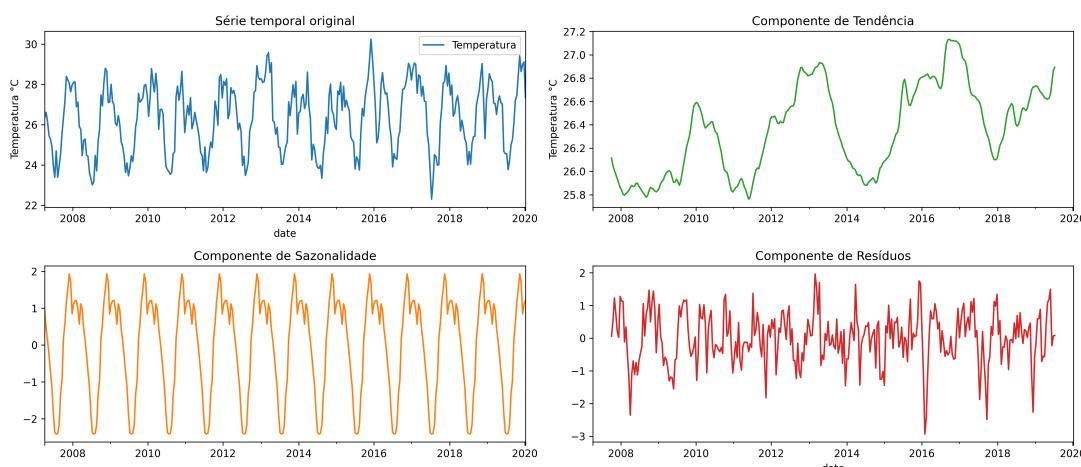


Figura 9 – Decomposição da série temporal da temperatura do ar para a estação automática localizada no município de Petrolina, no estado de Pernambuco.



Observando os componentes das séries temporais de exemplo, é possível identificarmos uma forte tendência anual nos dados de temperatura, confirmando a hipótese de que essa variável possui esta característica em seu comportamento.

5.1.2. Autocorrelação das séries temporais

Ao plotarmos a autocorrelação as estações de exemplo, temos as figuras 10, 11 e 12.

Figura 10 – Autocorrelação da série temporal da temperatura do ar para a estação convencional localizada no município de Balsas, no estado do Maranhão.

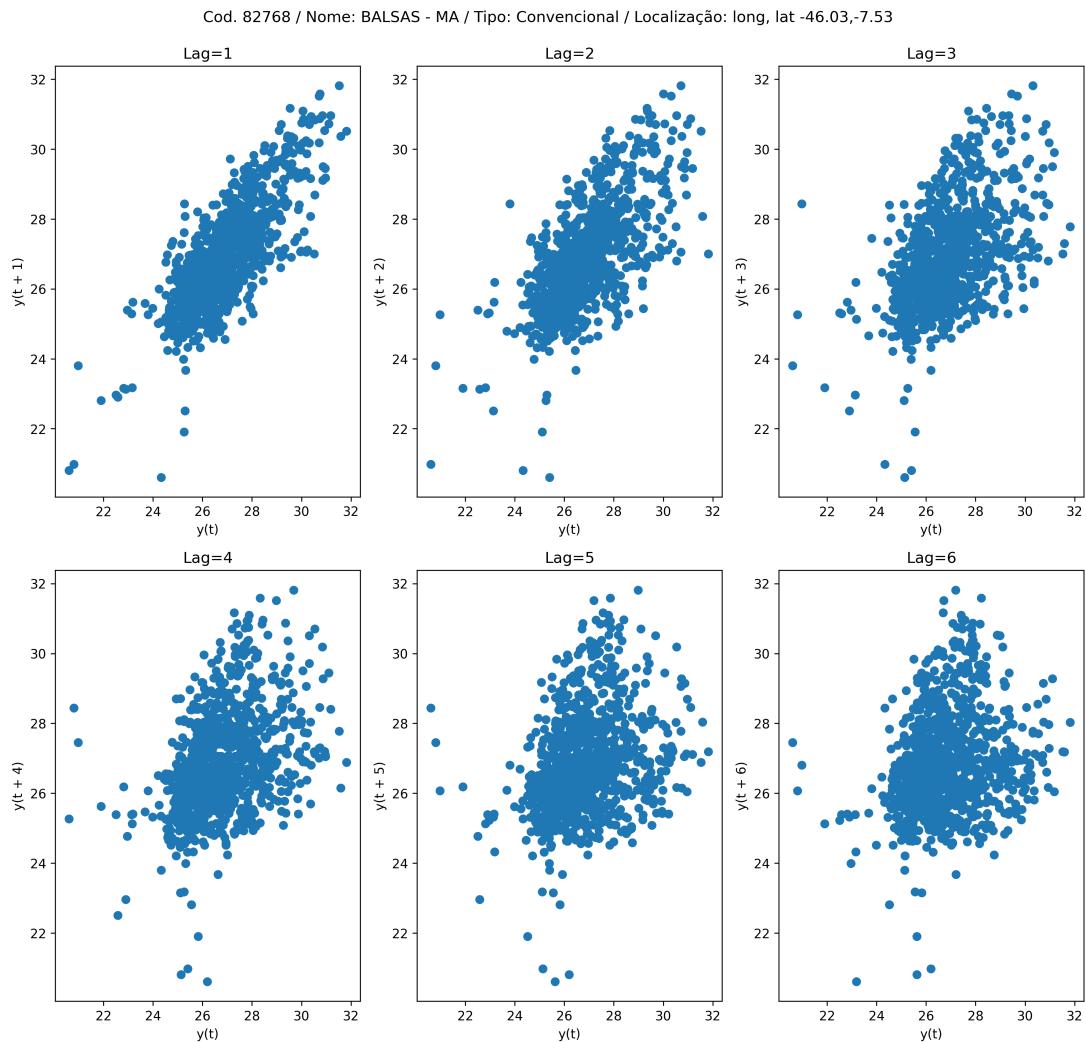


Figura 11 – Autocorrelação da série temporal da temperatura do ar para a estação automática localizada no município de Ariranha, no estado de São Paulo.

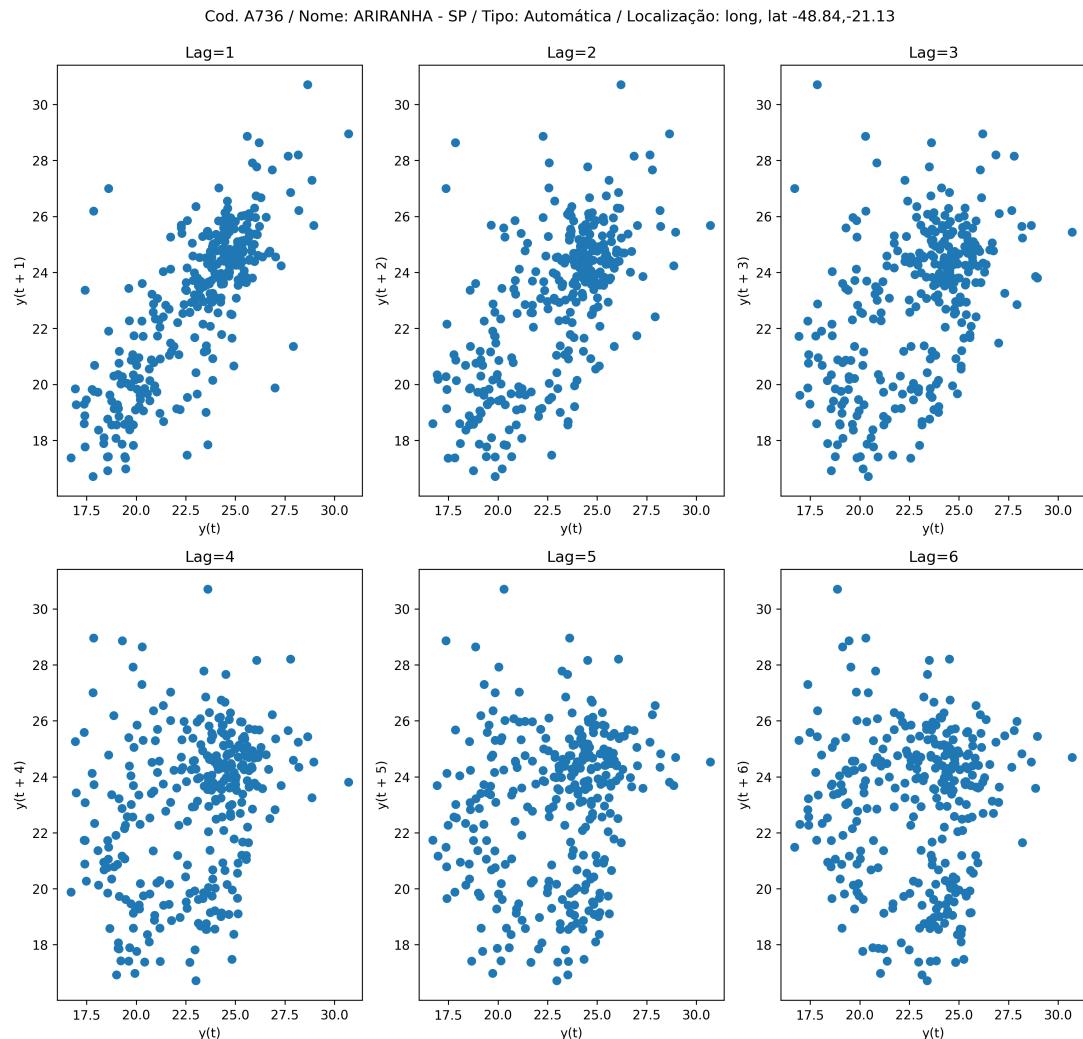
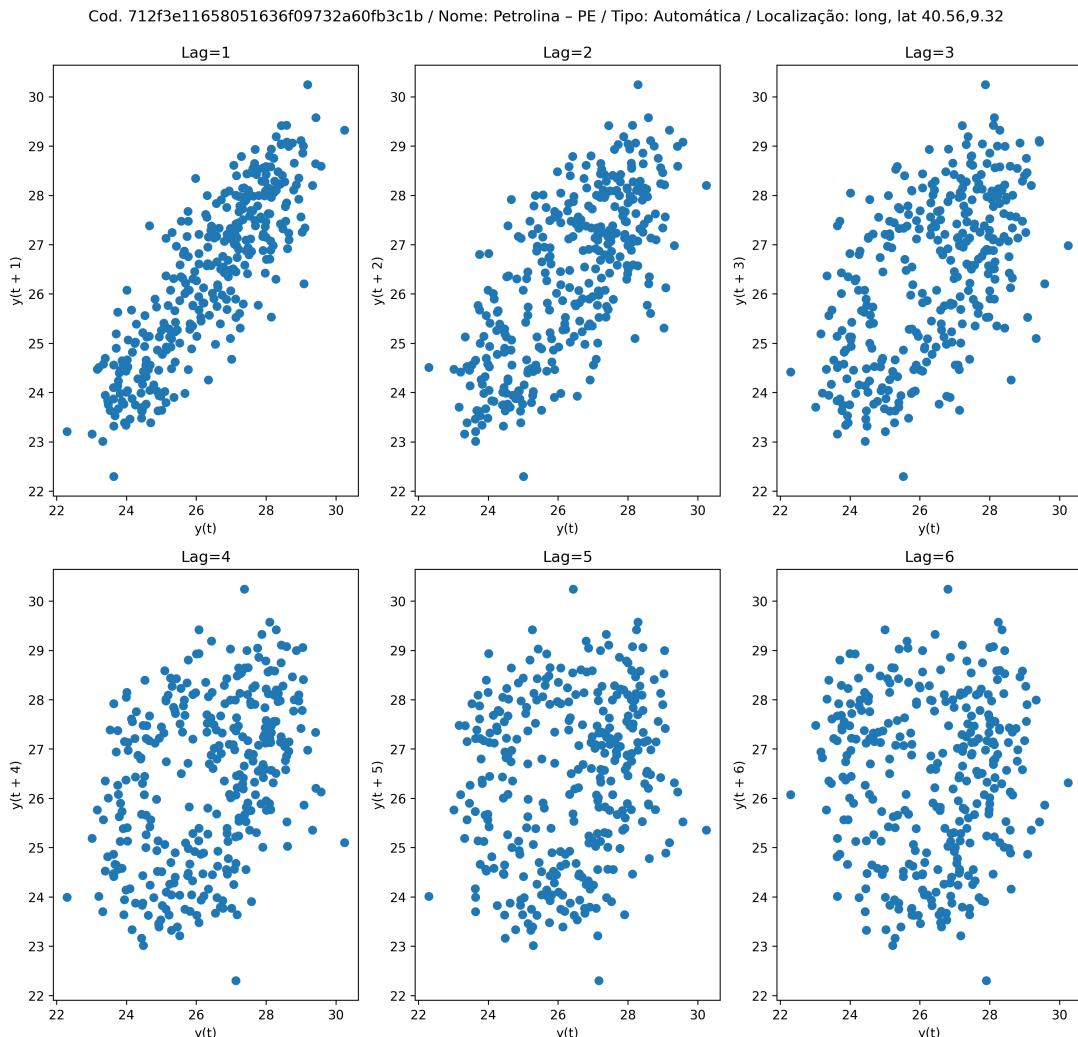


Figura 12 – Autocorrelação da série temporal da temperatura do ar para a estação automática localizada no município de Petrolina, no estado de Pernambuco.



Verificando os gráficos de autocorrelação, podemos identificar que não há uma alta autocorrelação nas estações de exemplo, por isso há a necessidade de se utilizar um modelo auto-arima para identificação do melhor modelo de forma automática.

5.1.3. Verificando a estacionariedade

A maioria dos modelos de previsão de séries temporais exigem que os dados sejam estacionários. Uma série temporal é considerada estacionária se suas propriedades estatísticas, como média, variância e covariância permanecerem constantes ao longo do tempo (Box, Jenkins e Reinsel 2011). Para verificarmos se as séries que

estamos trabalhando são estacionárias, utilizamos o teste de Dickey-Fuller aumentado (Said e Dickey 1984).

Apresentamos, nas figuras 13, 14 e 15, o resultado da verificação da estacionariedade das estações utilizadas como exemplo.

Figura 13 – Teste de estacionariedade da série temporal da temperatura do ar para a estação convencional localizada no município de Balsas, no estado do Maranhão.

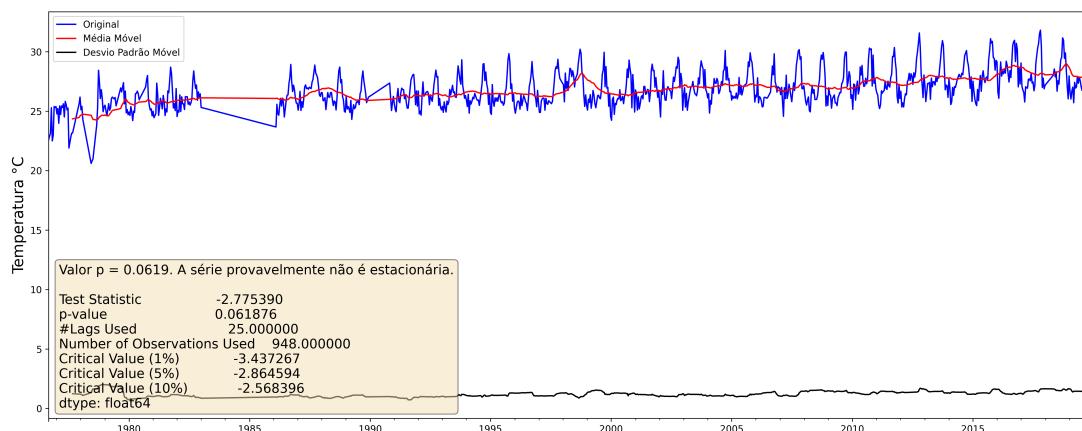


Figura 14 – Teste de estacionariedade da série temporal da temperatura do ar para a estação automática localizada no município de Ariranha, no estado de São Paulo.

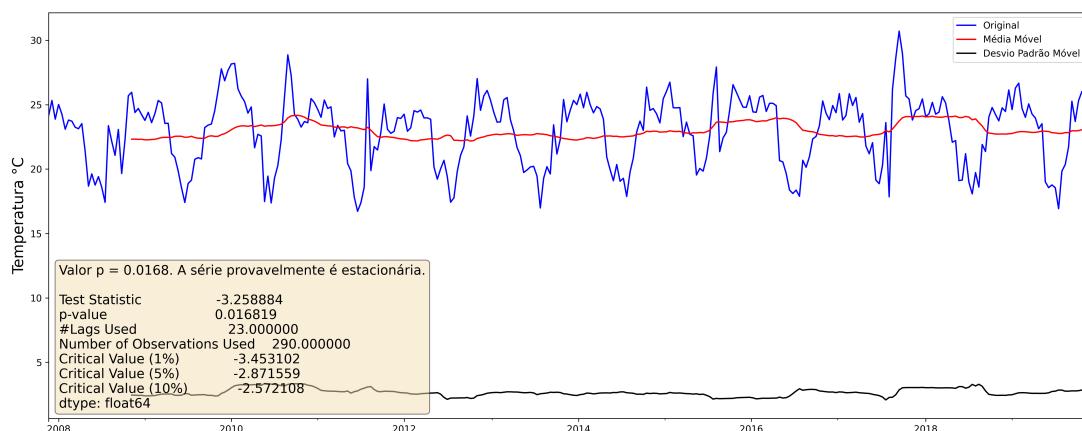


Figura 15 – Teste de estacionariedade da série temporal da temperatura do ar para a estação automática localizada no município de Petrolina, no estado de Pernambuco.



Pelo valor de p ser maior que 0,05, podemos afirmar que a primeira e a última série não são estacionárias. Já a segunda série, com o valor de p próximo de 0,01, indica que a série é estacionária. Para garantirmos que todas as séries utilizadas por nosso modelo sejam estacionárias, aplicamos uma diferenciação de primeira ordem com o objetivo de torna-las todas séries estacionárias.

Após o processo de diferenciação, realizamos novamente o teste de Dickey-Fuller aumentado e obtivemos os resultados apresentados das figuras 16, 17 e 18.

Figura 16 – Teste de estacionariedade da série temporal diferenciada da temperatura do ar para a estação convencional localizada no município de Balsas, no estado do Maranhão.

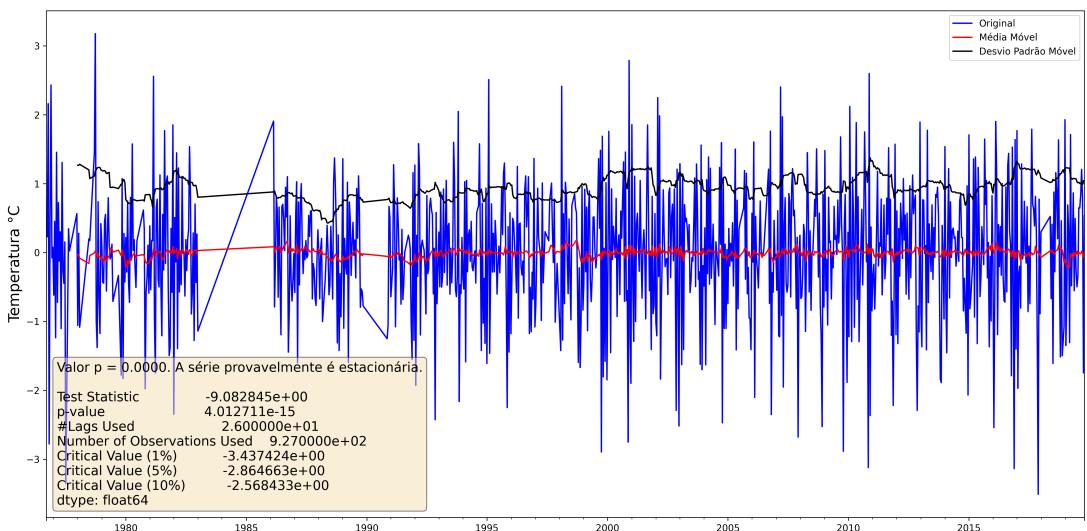


Figura 17 – Teste de estacionariedade da série temporal diferenciada da temperatura do ar para a estação automática localizada no município de Ariranha, no estado de São Paulo.

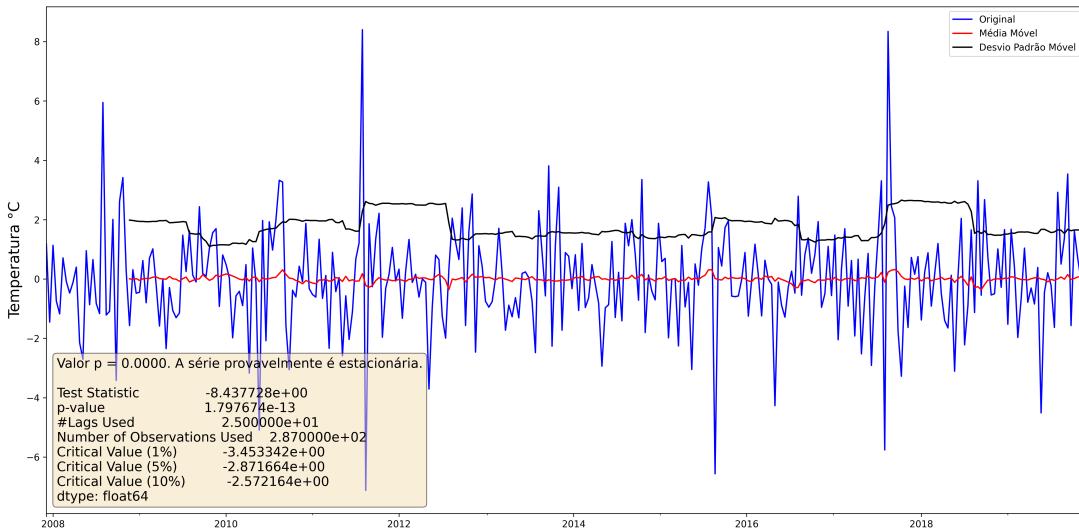
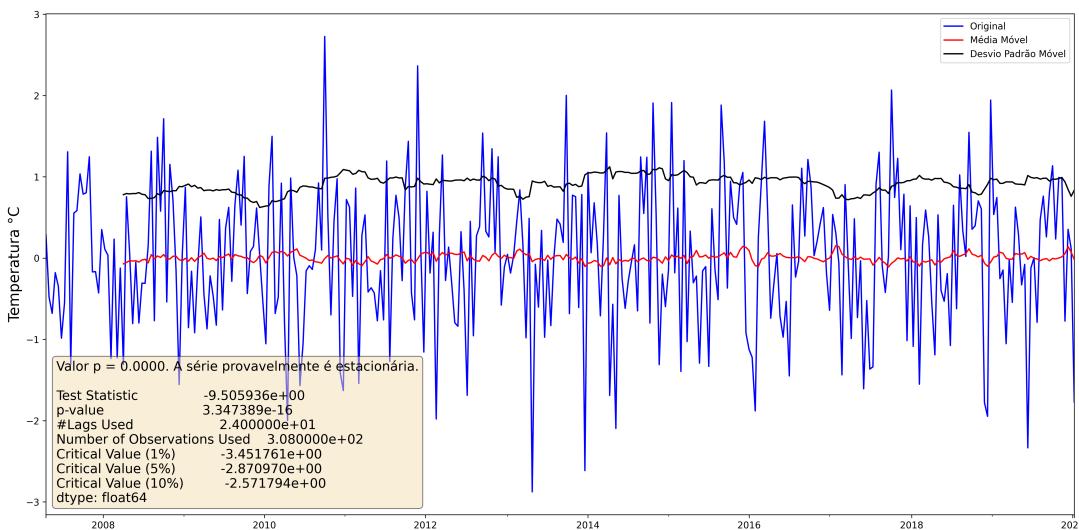


Figura 18 – Teste de estacionariedade da série temporal diferenciada da temperatura do ar para a estação automática localizada no município de Petrolina, no estado de Pernambuco.



Após a diferenciação, todas as séries obtiveram um valor de p abaixo de 0,05 e consequentemente passaram no teste da estacionariedade. Após este etapa, foi possível parametrizarmos e treinarmos o modelo de previsão ARIMA.

5.1.4. Parametrizando o modelo ARIMA

Para a parametrização do modelo, utilizamos a abordagem proposta por Hyndman, Khandakar et al. 2007. Essa abordagem testa iterativamente um conjunto de parâmetros buscando encontrar a combinação que gera um modelo com o menor valor para a métrica Akaike Information Criterion (AIC) (Sakamoto, Ishiguro e Kitagawa 1986), ou seja, ele busca o melhor modelo ajustado sem ter que testar exaustivamente todas as combinações possíveis. Apesar dessa abordagem buscar o modelo com a melhor combinação de parâmetros, é necessário informar antecipadamente o intervalo de possíveis valores para os parâmetros que ele utilizará para testar os modelos. Apresentamos na Tabela 8 a lista dos principais parâmetros que serão ajustados.

Tabela 8 – Parâmetros do modelo que serão ajustados iterativamente.

Nome do Parâmetro	Descrição
p	número de time lags do modelo auto-regressivo (AR)
q	ordem do modelo de média-móvel (MA)
d	grau de diferenciação
P	termo auto-regressivo para a parte sazonal
Q	termo da média-móvel para a parte sazonal
D	termo de diferenciação para a parte sazonal

Realizamos a busca pelo melhor modelo para cada uma das 88 estações selecionadas para análise. Para o processamento utilizamos a biblioteca escrita na linguagem Python pmdarima (Smith et al. 2017), versão 1.7.1. A Figura 19 apresenta todos parâmetros utilizados pela biblioteca pmdarima para buscar o melhor modelo para cada uma das estações avaliadas.

Figura 19 – Parâmetros utilizados na função `auto_arima`, da biblioteca `pmdarima`, para buscar o melhor modelo ajustado.

5.2. Criação do modelo LSTM

As redes neurais recorrente do tipo Long Short-Term Memory (LSTM) tem a promessa de aprender longas sequências de observações, o que é desejável em nosso caso, já que possuímos algumas estações convencionais com observações que datam do ano 1961, ou seja, 58 anos de observações. Mas para isso, elas exigem grandes conjuntos de dados de treinamento. Nesta seção, apresentaremos a arquitetura desenvolvida e os passos executados para realizar as previsões utilizando este modelo.

5.2.1. Conjuntos de dados de treinamento, validação e teste

Para o treinamento e análise da acurácia do modelo, os dados foram separados em três conjuntos distintos: treinamento, validação e teste. Segundo Hastie, Tibshirani e Friedman 2009, o conjunto de treinamento deve ser usado para ajustar o modelo; o conjunto de validação deve ser usado para estimar o erro da predição do modelo, e o conjunto de teste deve ser utilizado para a avaliação do erro da generalização do modelo final escolhido.

Para o conjunto de teste, assim como no modelo ARIMA, utilizamos as mesmas 88 estações que selecionamos na Figura 7 como amostras para a avaliação dos modelos. O conjunto de treinamento e validação geramos a partir das estações restantes, ou seja, de todas as 878 estações, utilizamos as mesmas 88 estações, que foram utilizadas no modelo ARIMA, para a geração dos dados de teste e as 788 estações restantes utilizamos para o treinamento e validação do modelo LSTM.

O nosso modelo receberá como dado de entrada, chamaremos de X, uma série de observações de temperatura e retornará como previsão, chamaremos aqui de Y, os valores de temperatura média previstos para o próximo ano. Para criarmos um grande conjunto de dados de treinamento, ao invés de dividirmos uma série temporal em um único par (X, Y) utilizando Y como o último ano e X como os dados dos anos anteriores, vamos utilizar aqui deslocamentos de um ano na série para obtermos, de uma única série, múltiplos pares (X, Y). Na Tabela 9 apresentamos todos os períodos utilizados para gerar os valores (X, Y) que foram utilizados para treinar o modelo LSTM.

Tabela 9 – Quantidade de registros por conjunto de dados obtidos.

Período para obter o dado de entrada (X)	Período para obter a saída esperada (Y)
todos os dados antes de 01/01/2019	01/01/2019 - 31/12/2019
todos os dados antes de 01/01/2018	01/01/2018 - 31/12/2018
todos os dados antes de 01/01/2017	01/01/2017 - 31/12/2017
todos os dados antes de 01/01/2016	01/01/2016 - 31/12/2016
todos os dados antes de 01/01/2015	01/01/2015 - 31/12/2015
todos os dados antes de 01/01/2014	01/01/2014 - 31/12/2014
todos os dados antes de 01/01/2013	01/01/2013 - 31/12/2013
todos os dados antes de 01/01/2012	01/01/2012 - 31/12/2012
todos os dados antes de 01/01/2011	01/01/2011 - 31/12/2011
todos os dados antes de 01/01/2010	01/01/2010 - 31/12/2010

Após processar todos os períodos, obtivemos 5.137 séries com os valores de entrada (X) e 5.137 séries com os valores esperados (Y). Destas, 75% dos pares(X, Y) foram alocados no conjunto de treinamento e 25% foram alocados no conjunto de validação.

5.2.2. Normalização dos dados

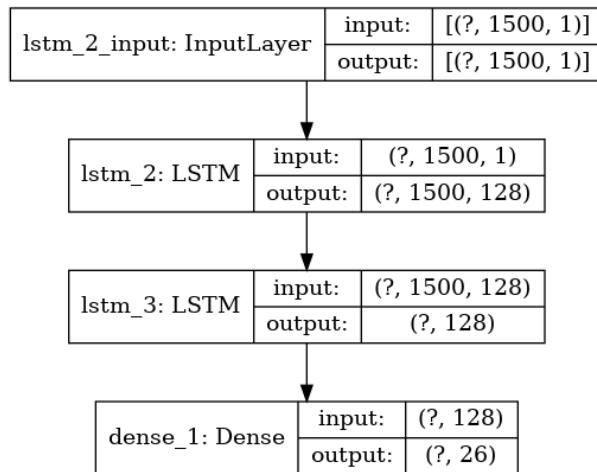
Para normalizar os dados de treinamento, utilizamos a função MinMaxScaler da biblioteca sklearn (Pedregosa et al. 2011), própria para aprendizado de máquina. Essa função tem como objetivo mapear os dados a serem normalizados no intervalo [0, 1]. Dessa forma, o maior valor que será normalizado receberá o valor de 1, e o menor, o valor de 0, os valores intermediários serão transformados em números dentro desse intervalo.

5.2.3. Modelo LSTM desenvolvido

Para a construção do modelo LSTM, utilizamos a biblioteca TensorFlow, versão 2.3.0. O modelo desenvolvido utiliza duas camadas LSTM, com 128 neurônios do tipo LSTM cada. Como camada de saída utilizamos uma camada densamente conectada com ativação linear e 26 neurônios, cada neurônio será responsável por prevê um dos 26 períodos de duas semanas que compõe o ano que será previsto pelo modelo.

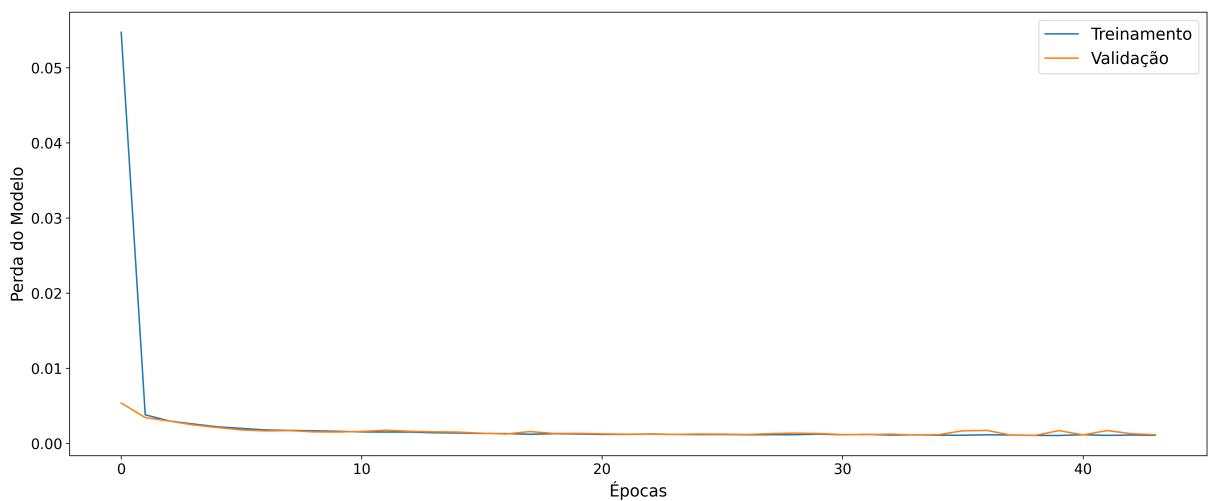
Configuramos o modelo para utilizar o otimizador Adam com taxa de aprendizado de 0,001. Como função de perda utilizamos o Mean Squared Error (MSE) e como métrica de avaliação o Mean Absolute Error (MAE). A Figura 20 ilustra o modelo completo desenvolvido.

Figura 20 – Arquitetura LSTM desenvolvida.



Para o treinamento do modelo, utilizamos lotes com 128 itens, treinando por 44 épocas. Utilizamos o *callback* EarlyStopping do Keras para realizar uma parada precoce no treinamento com o objetivo de evitar ajustes excessivos no modelo. A Figura 21 ilustra a perda do modelo ao longo do processo de treinamento.

Figura 21 – Perda do modelo durante o processo de treinamento.

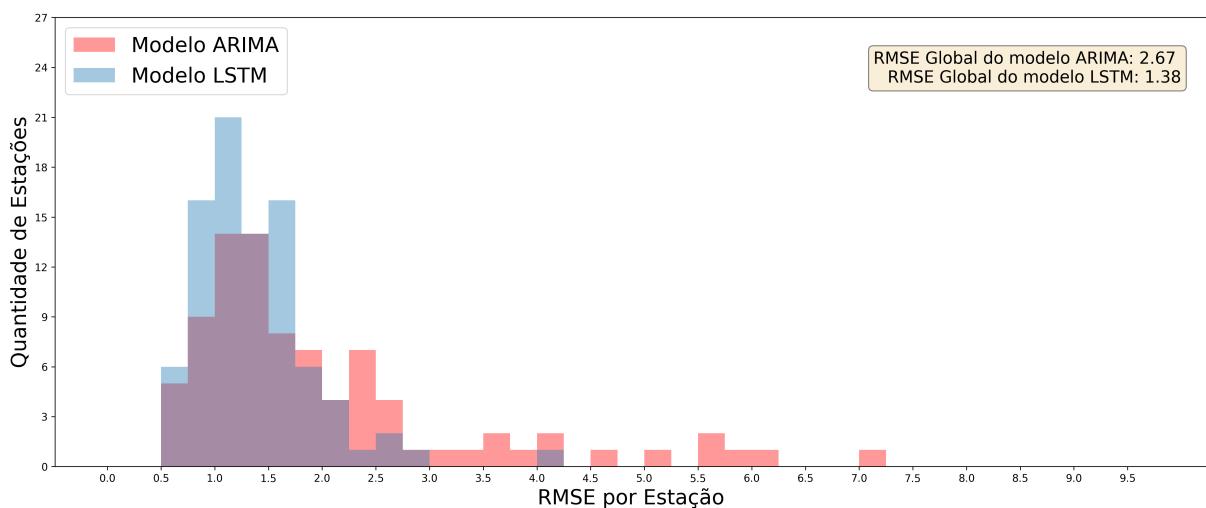


6. Apresentação dos Resultados

Nessa seção, apresentaremos os resultados obtidos através dos modelos preditivos ARIMA e LSTM para a previsão da temperatura do ano de 2019.

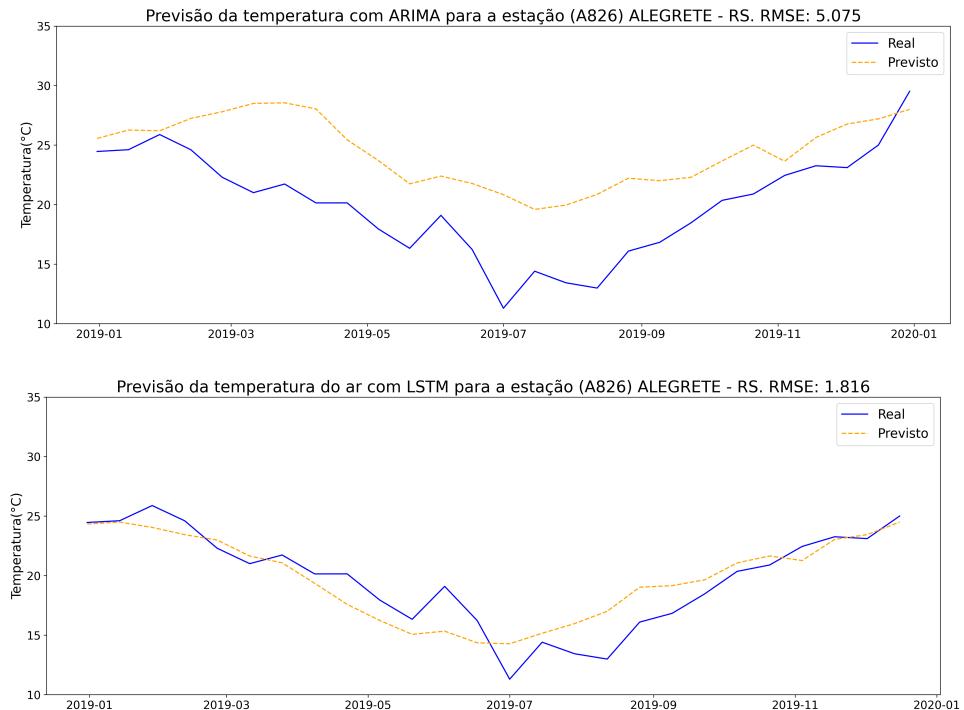
A Figura 22 apresenta um histograma de comparação entre o erro médio quadrático (RMSE) obtido em cada um dos modelos, ARIMA e LSTM, para a previsão da temperatura do ano de 2019 nas 88 estações avaliadas. Analisando o gráfico podemos afirmar que o modelo LSTM obteve mais estações com os menores erros, abaixo de 1,5, enquanto que o modelo ARIMA apresentou mais estações com os maiores erros, acima de 2,0. Para todas as 88 estações avaliadas, o RMSE Global, ou seja, a média de todos os erros, foi 2,67 para o modelo ARIMA e 1,38 para o modelo LSTM, indicando que o modelo LSTM obteve um desempenho, na previsão, melhor que o modelo ARIMA.

Figura 22 – Comparação entre os valores obtidos de RMSE para as 88 estações avaliadas.



Analizando as previsões individuais, ilustramos na Figura 23 o resultado da previsão para uma estação localizada no município de Alegrete, no estado do Rio Grande do Sul. Para esta estação em que o modelo ARIMA apresentou um maior erro, acima de 5, o modelo LSTM conseguiu uma boa performance alcançando um erro um pouco acima de 1,8. Esse comportamento se refletiu em diversas outras estações em que o modelo ARIMA obteve um resultado pior em relação ao seu próprio RMSE Global.

Figura 23 – Resultado da previsão da temperatura do ar para o ano de 2019 na estação meteorológica localizada no município de Alegrete, no estado do Rio Grande do Sul.



Avaliando as estações em que o modelo LSTM obteve um desempenho pior do que seu seu próprio RMSE Global, identificamos que, nessas estações, o modelo ARIMA não obteve grande vantagem, nas raras vezes em que foi melhor, em relação ao modelo LSTM. Na Figura 24 adicionamos um exemplo do resultado da previsão para uma estação em que o modelo LSTM não conseguiu bons resultados. Nas figuras 25 e 26 também apresentamos o resultados da previsão para outras estações.

Figura 24 – Resultado da previsão da temperatura do ar para o ano de 2019 na estação meteorológica localizada no município de Coxim, no estado do Mato Grosso do Sul.

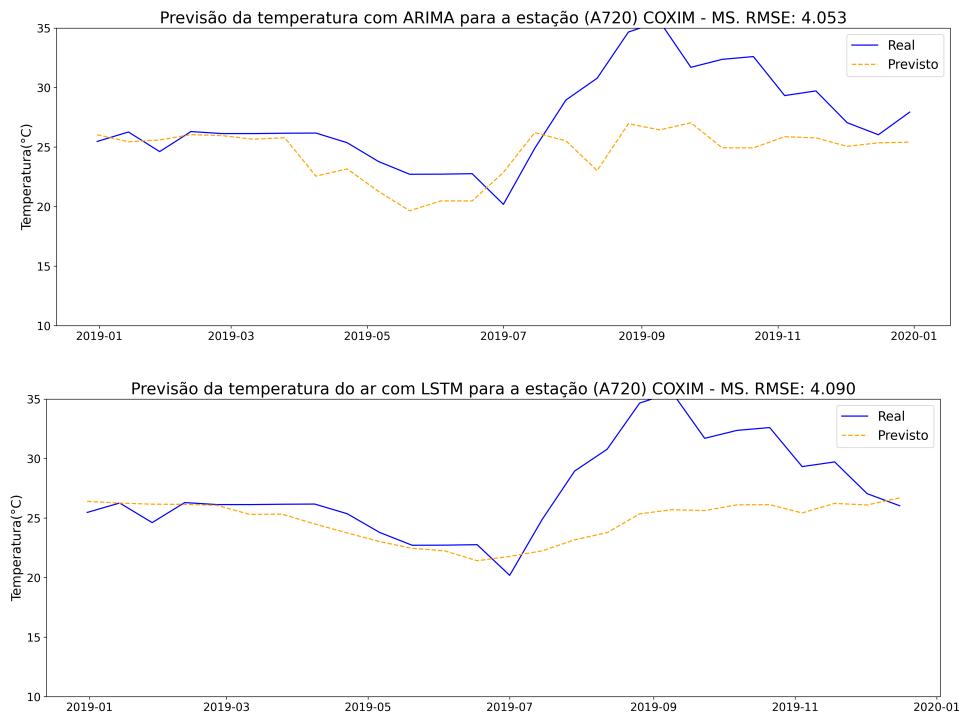


Figura 25 – Resultado da previsão da temperatura do ar para o ano de 2019 na estação meteorológica localizada no município de Jataí, no estado de Goiás.

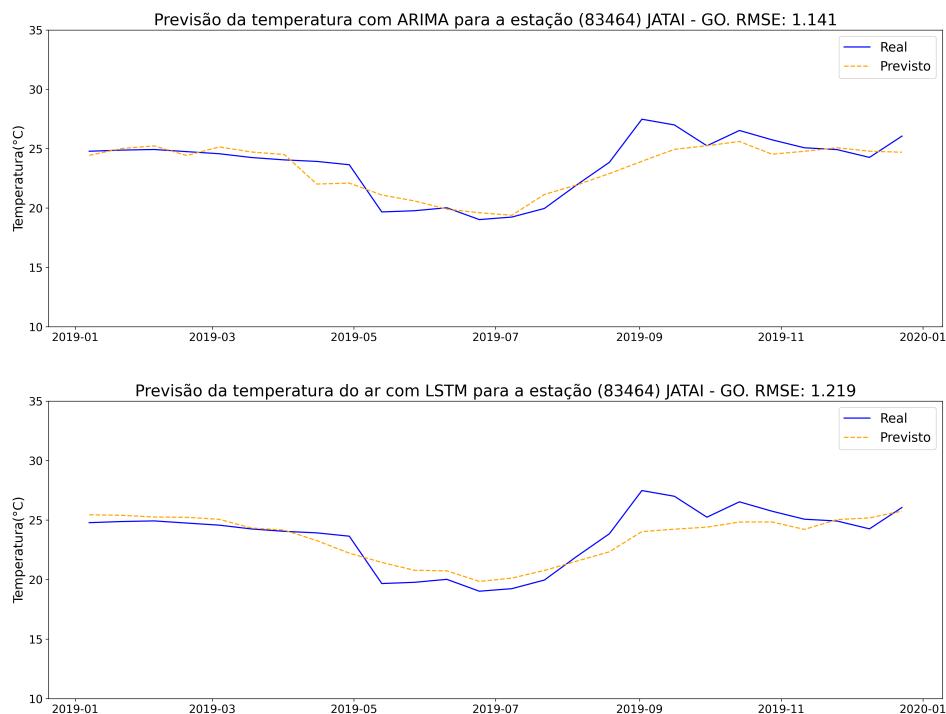
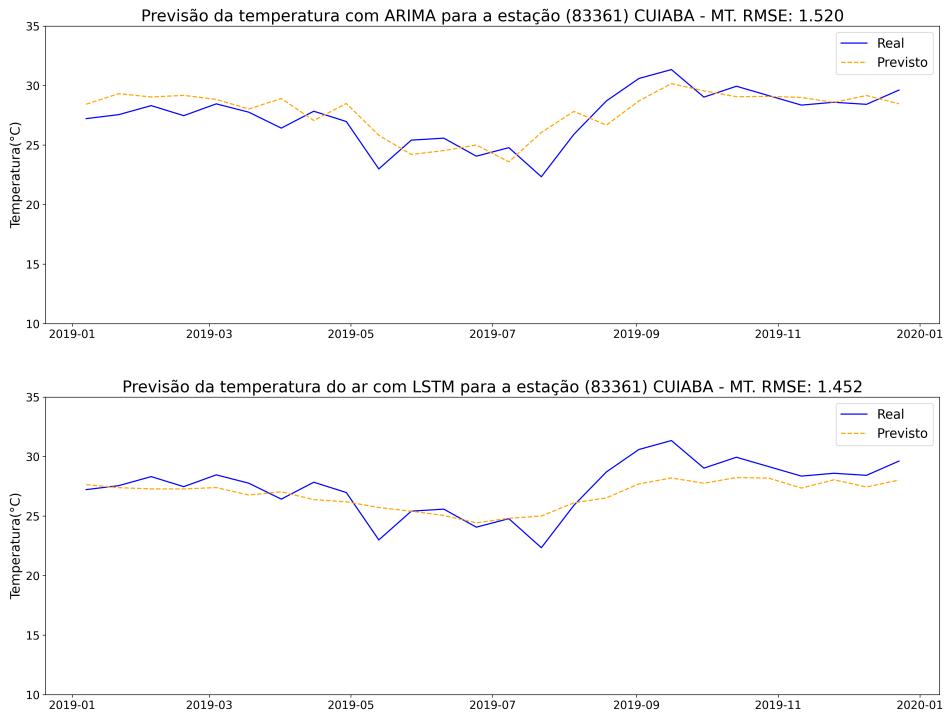


Figura 26 – Resultado da previsão da temperatura do ar para o ano de 2019 na estação meteorológica localizada no município de Cuiabá, no estado do Mato Grosso.



7. Links

Disponibilizamos os dados utilizados, scripts desenvolvidos e vídeo explicativo através dos links:

- Brazil Weather, Conventional Stations (1961-2019):
<https://www.kaggle.com/saraivaufc/conventional-weather-stations-brazil>
- Brazil Weather, Automatic Stations (2000-2019):
<https://www.kaggle.com/saraivaufc/automatic-weather-stations-brazil>
- LabMet - Automatic Weather Stations (2007-2019):
<https://www.kaggle.com/saraivaufc/automatic-weather-stations-labmet>
- Scripts desenvolvidos: <https://github.com/saraivaufc/TCC-Ciencia-de-Dados>
- Vídeo com explicação sobre este trabalho: <https://youtu.be/gYE2-oue06Y>

REFERÊNCIAS

- BABA, R. K.; VAZ, M. S. M. G.; COSTA, J. d. Correção de dados agrometeorológicos utilizando métodos estatísticos. *Revista Brasileira de Meteorologia*, SciELO Brasil, v. 29, n. 4, p. 515–526, 2014.
- BEHNEL, S.; FAASSEN, M.; BICKING, I. *lxml: XML and HTML with Python*. [S.I.]: Lxml, 2005.
- BOX, G. E.; JENKINS, G. M.; REINSEL, G. C. *Time series analysis: forecasting and control*. [S.I.]: John Wiley & Sons, 2011.
- HANSEN, J. et al. Global temperature change. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 103, n. 39, p. 14288–14293, 2006.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. [S.I.]: Springer Science & Business Media, 2009.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- HYNDMAN, R. J.; KHANDAKAR, Y. et al. *Automatic time series for forecasting: the forecast package for R*. [S.I.]: Monash University, Department of Econometrics and Business Statistics . . . , 2007.
- KAYMAZ, B. Hazards and their impact on human. In: *International Movement for Interdisciplinary Study of Estrangement Conference*. [S.I.: s.n.], 2005.
- MATEO, M. A. F.; LEUNG, C. K.-S. Design and development of a prototype system for detecting abnormal weather observations. In: *Proceedings of the 2008 C3S2E conference*. [S.I.: s.n.], 2008. p. 45–59.
- MCKINNEY, W. et al. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, Seattle, v. 14, n. 9, 2011.
- OLIVEIRA, A.; SOUZA, L. S. de. Estação meteorológica automática do cdtn: Tratamento das informações meteorológicas para reconhecimento da qualidade dos dados e apoio aos estudos de dispersão atmosférica. 2019.
- PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.
- SAID, S. E.; DICKEY, D. A. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, Oxford University Press, v. 71, n. 3, p. 599–607, 1984.