**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Marcelo Pereira Cunha**

**APLICAÇÃO DE TÉCNICAS DE DEEP LEARNING EM PREVISÃO DE VARIÁVEIS CLIMÁTICAS**

Belo Horizonte

2021

**Marcelo Pereira Cunha**

**APLICAÇÃO DE TÉCNICAS DE DEEP LEARNING EM PREVISÃO DE VARIÁVEIS CLIMÁTICAS**

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

Belo Horizonte

2021**SUMÁRIO**

[1. Introdução 4](#_Toc78280095)

[1.1. Contextualização 4](#_Toc78280096)

[1.2. O problema proposto 8](#_Toc78280097)

[2. Coleta de Dados 9](#_Toc78280098)

[3. Processamento/Tratamento de Dados 13](#_Toc78280099)

[4. Análise e Exploração dos Dados 17](#_Toc78280100)

[5. Criação de Modelos de Machine Learning 27](#_Toc78280101)

[6. Apresentação dos Resultados 36](#_Toc78280102)

[7. Links 39](#_Toc78280103)

[REFERÊNCIAS 40](#_Toc78280104)

# 1. Introdução

## Contextualização

A previsão do tempo através de métodos numéricos possui Vilhelm Bjerknes (1904) como um importante pesquisador, pois o mesmo identificou que o estado futuro da atmosfera poderia, ser obtido pela integração das equações diferenciais que governam o seu comportamento, sendo a atmosfera modelada como um sistema dinâmico (Ynoue, 2011). As condições iniciais das equações diferenciais utilizadas, descreveriam um estado observado da atmosfera. O pesquisador britânico Lewis Fry Richardson foi o primeiro a realizara integração numérica compreensível dessas equações.

Richardson, partindo de observações dos dados meteorológicos das 7 UTC do dia 20 de maio de 1910, calculou a derivada da pressão na Alemanha em função do tempo. A variação prevista pelo seu modelo numérico para a variável pressão no período de 6 horas foi de 146 hPa, valor muito maior do que aquele realmente observado. Todavia, mesmo com um resultado aquém do esperado, o trabalho de Richardson mostrou quais seriam os obstáculos a sobrepujar no campo de previsões climáticas através de modelos numéricos.

Para realizar a previsão de uma variável atmosférica para o intervalo de apenas um dia, mesmo nos dias atuais, se faz necessário uma exorbitante quantidade de cálculos, sendo demandada considerável rapidez no processamento destes cálculos. Além disso, os dados utilizados para representar o estado inicial da atmosfera não eram suficientes na experiência realizada por Richardson. Através dos experimentos realizados, pode-se observar também que, caso as técnicas numéricas utilizadas não fossem devidamente implementadas, poderiam ser gerados erros que se propagariam ao longo de toda a computação do modelo numérico proposto, o que reduziria a precisão e acurácia da previsão.

Ao longo da década de 1960, iniciou-se a utilização dos modelos numéricos globais de tempo em previsões climáticas, possibilitando realizar a previsão de como se faria a evolução da atmosfera a longo prazo. Considerando que o tempo atmosférico e o clima possuem escalas temporais distintas, ajustes nos modelos se tornaram necessários para a representação dos processos físicos que regem cada escala de tempo. A previsão de tempo e de clima é um ramo altamente especializado e em constante evolução.

Um aspecto importante a se ressaltar é a importância em diferenciar a previsão de tempo da previsão de clima. Conforme citado anteriormente, a previsão de tempo é realizada para poucos dias consecutivos, fato este que se deve ao escasso conhecimento das condições iniciais (estado observado da atmosfera) fornecidas aos modelos numéricos utilizados.

Na previsão climática não se tem interesse em prever com exatidão o local e o momento da ocorrência de um sistema atmosférico como na previsão do tempo e, sim, que seja realizada uma simulação pelo modelo. Caso o objetivo seja prever como será a precipitação no outono austral, uma simulação é feita para essa estação nesta região; após isto é calculado o total acumulado de precipitação e, como último passo, este é comparado com o valor climatológico (que é proveniente da média de um longo período de precipitação observada). Desta maneira, é possível descobrir se tal estação será mais úmida ou mais seca do que a climatologia observada previamente.

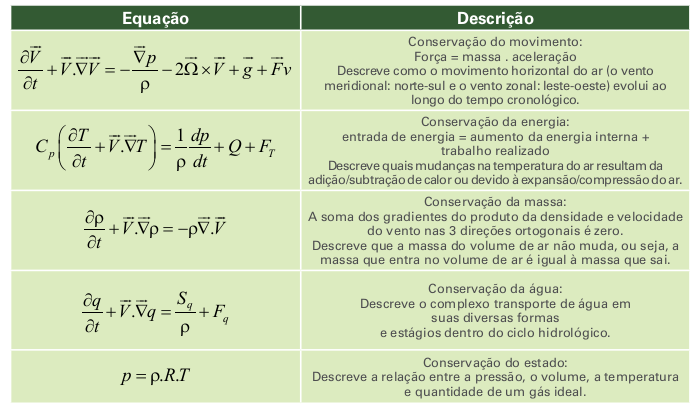
A principal ferramenta utilizada atualmente no campo das previsões de tempo e de clima são os modelos numéricos conhecidos como Modelos Numéricos de Circulação Geral – MCG, sendo um conjunto de equações físicas descritas em forma numérica e resolvidas com o auxílio de computadores.

Os procedimentos gerais para uma previsão de tempo ou de clima incluem 3 etapas (análise, previsão e pós-processamento).

Na fase da análise, as observações meteorológicas são fornecidas a programas computacionais, que preparam os dados para os modelos de previsão. Uma vez que a rede de observações global de dados não cobre regularmente a superfície da Terra, as observações meteorológicas são submetidas a métodos matemáticos para se tornarem uniformes, isto é, valores iniciais com espaçamento horizontal regular. Embora isso seja somente um passo preparatório, a tarefa é difícil, pois há milhões de dados provenientes de diferentes fontes (satélites, navios, estações meteorológicas de superfície etc.) e que não necessariamente foram medidos no mesmo horário. Além disso, nenhuma das medidas é completamente livre de erros. Assim, é necessário o máximo possível dos erros para produzir campos atmosféricos consistentes. O procedimento da análise também inclui a junção de condições iniciais observadas e previsões do modelo numérico para a elaboração de condições iniciais para esse modelo. Em síntese, a fase da análise tem como objetivo criar um conjunto de dados com espaçamento horizontal uniforme para ser fornecido aos modelos de previsão como condição inicial. Terminada a etapa da análise, o modelo pode ser executado para produzir as previsões.

Ao longo da etapa de previsão, utiliza-se um modelo numérico cujo objetivo é resolver as equações básicas que descrevem o comportamento da atmosfera. Abaixo são exibidas as equações. Vale ressaltar que o vento possui 3 componentes nas direções ortogonais x, y e z, ou seja, V = ( u , v , w ).

Figura : Equações do modelo numérico de previsão do tempo

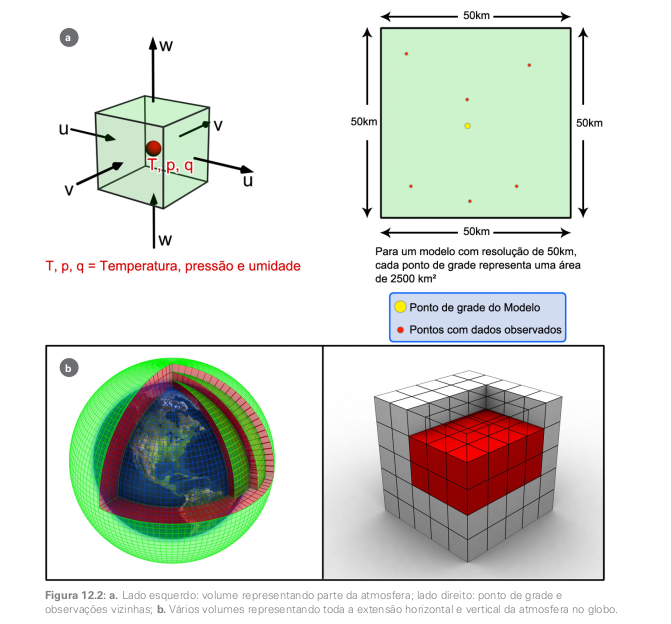


Fonte: USP (Universidade de São Paulo) Departamento de Geografia

As equações utilizadas nos modelos de previsão de tempo e clima são resolvidas em pontos de grade, ou seja, a superfície do planeta é dividida em latitudes e longitudes e a intersecção delas fornece os pontos de grade. Já a distância entre 2 pontos de grade vizinhos fornece a resolução horizontal do modelo. A atmosfera também é dividida em níveis verticais, sendo esta divisão conhecida como resolução vertical.

Para uma previsão futura, o modelo é iniciado com as condições iniciais observadas e a previsão para aquele momento. Também é informado no intervalo de tempo (resolução temporal previamente determinada), que ele vai fazer a integração das equações até chegar ao final do tempo desejado (por exemplo, 24 h, 48 h, 72 h). As condições da atmosfera observadas são utilizadas apenas na primeira integração, pois nas demais o modelo utilizará a previsão feita anteriormente como condição inicial. Com isso, é elaborado um conjunto de previsões: para algumas horas, um dia ou alguns dias etc.

Figura : Modelo de Temperatura, Pressão e Umidade



Fonte: USP (Universidade de São Paulo) Departamento de Geografia

Durante a etapa de pós processamento, embora os modelos numéricos façam cálculos para intervalos de tempo de minutos, os arquivos de saída com os resultados dos cálculos são reduzidos para intervalos regulares de horas (por exemplo, a cada 12 horas) e em pontos de grade. Com os arquivos de saída dos modelos, os meteorologistas elaboram mapas de diferentes variáveis atmosféricas e, após o estudo desses mapas, elaboram os boletins da previsão de tempo e clima, dependendo do modelo utilizado. Essas atividades são chamadas de pós-processamento. Os boletins são importantes para o planejamento de várias atividades, como por exemplo o setor agrícola, hidrelétricas, comercial, turismo, lazer, prevenção de desastres, entre outras.

Os modelos utilizados para a previsão de tempo e clima podem ser globais ou regionais. Os modelos globais simulam as condições atmosféricas globalmente. Em sua maioria tem grade com resolução horizontal em torno de 200 km, 28 níveis verticais e são inicializados com condições iniciais de todo o planeta. Esses modelos possuem grande eficiência em simular as características gerais da circulação atmosférica de larga escala, mas não as características locais como brisas de vale e de montanha. Atualmente, é inviável aumentar a resolução horizontal de tais modelos, pois isso envolve elevado custo computacional devido ao maior tempo de processamento das previsões que envolvem um número massivo de dados e mais espaço físico de disco necessário para armazená-las. Por isso, para obtenção de informações mais detalhadas da atmosfera, são utilizados os modelos regionais (ou de área limitada), que simulam as condições atmosféricas em pequenas porções do planeta e possuem grade com resolução horizontal, em torno de 20 km. Esses modelos, por terem resolução horizontal maior do que os globais, representam melhor a superfície e os fenômenos mais regionais, tornando os resultados mais precisos. Os modelos de área limitada, além das condições iniciais, precisam de condições de fronteira lateral, já que abrangem pequenas porções do globo. As fronteiras laterais têm possibilidade de serem provenientes das saídas de modelos de circulação geral da atmosfera.

## 1.2. O problema proposto

Considerando a elevada complexidade em resolver equações diferenciais parciais computacionalmente, tendo em vista também o massivo volume de dados referentes à condições iniciais que são demandados para uma resolução eficaz das equações que compõem os modelos climáticos, bem como os grandes erros provocados por possíveis falhas e/ou inconsistências nos dados que compõe as condições iniciais supracitadas, o objetivo do presente trabalho é propor um modelo simplificado, que possua um número reduzido de variáveis de entrada para realizar a predição do valor da variável temperatura do ar, simplificando o processo de coleta armazenamento, análise e processamento das diversas variáveis utilizadas nos modelos numéricos de equações diferenciais usadas no processo de previsão do tempo.

Para a realização de tal intento, serão utilizados os ítens listados abaixo:

* Realizar análise exploratória e descritiva dos dados fornecidos pelo Instituto Nacional de Meteorologia, doravante INMET (2020 e 2021) e Climatempo (2021).
* Desenvolver um modelo de machine learning utilizando o algoritmo LSTM (Long Short Term Memory).
* Desenvolver um modelo preditivo utilizando o modelo estatístico ARIMA (Auto Regressive Moving Average).
* Comparar os valores de predição obtidos pelo LTSM e ARIMA com os dados previstos pelo Clima Tempo, obtidos através da API de previsão do tempo do Clima Tempo.
* Os dados analisados e coletados compreendem a região de Belo Horizonte, Minas Gerais, entre 01/01/2020 e 14/07/2021. (INMET) e 14/07/2021 e 16/07/2021 (Clima Tempo).

# 2. Coleta de Dados

A tabela a seguir exibe o detalhamento dos dados climáticos do INMET para 2020, obtidos através do link abaixo, obtidos em 07/07/2021:

<https://portal.inmet.gov.br/dadoshistoricos>

Tabela : Dados Meteorológicos INMET 2020

|  |  |  |
| --- | --- | --- |
| Nome da coluna/campo | Descrição | Tipo |
| Data | Data. | OBJECT |
| Hora UTC | Hora. | OBJECT |
| PRECIPITACAO TOTAL. HORARIO (mm) | Valor de precipitação. | float64 |
| PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO. HORARIA (mB) | Pressão atmosférica medida do nível da estação meteorológica. | float64 |
| PRESSAO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB) | Pressão atmosférica máxima medida do nível da estação meteorológica. | float64 |
| PRESSAO ATMOSFERICA MIN.NA HORA ANT. (AUT) (mB) | Pressão atmosférica mínima medida do nível da estação meteorológica. | float64 |
| RADIACAO GLOBAL (Kj/m3) | Valor de radiação mensurado pela estação meteorológica. | float64 |
| TEMPERATURA DO AR - BULBO SECO. HORARIA (Celsius) | Média horária da temperatura do ar. | float64 |
| TEMPERATURA DO PONTO DE ORVALHO (Celsius) | Temperatura na qual o vapor de água que está em suspensão no ar começa a se condensar | float64 |
| TEMPERATURA MINIMA NA HORA ANT. (AUT) (Celsius) | Temperatura mínima medida na hora anterior. | float64 |
| TEMPERATURA MAXIMA NA HORA ANT. (AUT) (Celsius).1 | Temperatura máxima medida na hora anterior. | float64 |
| TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (Celsius) | Temperatura máxima do ponto de orvalho na última hora. | float64 |
| TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (Celsius) | Temperatura mínima do ponto de orvalho na última hora. | float64 |
| UMIDADE REL. MAX. NA HORA ANT. (AUT) | Umidade relativa do ar máxima na hora anterior. | float64 |
| UMIDADE REL. MIN. NA HORA ANT. (AUT) | Umidade relativa do ar mínima na hora anterior. | float64 |
| UMIDADE RELATIVA DO AR. HORARIA (%) | Porcentagem horária de umidade relativa do ar. | float64 |
| VENTO. DIRECAO HORARIA (gr) ((gr)) | Direção do vento horária. | float64 |
| VENTO. RAJADA MAXIMA (m/s) | Velocidade máxima do vento. | float64 |
| VENTO. VELOCIDADE HORARIA (m/s) | Média horária de velocidade do vento. | float64 |

A tabela a seguir exibe o detalhamento dos dados climáticos do INMET para 2021, obtidos através do link abaixo, obtidos em 14/07/2021:

<https://bdmep.inmet.gov.br/$2a$10$HxsOeG1EWoLlEoMx3gi6eSJdzuuYD7iOA8qVyfKi.tjgYmx3QE..zip>

Tabela : Dados Climáticos INMET 2021

|  |  |  |
| --- | --- | --- |
| Nome da coluna/campo | Descrição | Tipo |
| Data Medicao | Data da medição. | object |
| Hora Medicao | Hora da medição. | int64 |
| PRECIPITACAO TOTAL. HORARIO(mm) | Precipitação total horária em milímetros. | int64 |
| PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO. HORARIA(mB) | Pressão ATM horária. | float64 |
| PRESSAO ATMOSFERICA REDUZIDA NIVEL DO MAR. AUT(mB) | Pressão Atmosférica considerando o nível do mar. | float64 |
| PRESSAO ATMOSFERICA MAX.NA HORA ANT. (AUT)(mB) | Valor máximo de pressão atmosférica obtido na hora anterior | float64 |
| PRESSAO ATMOSFERICA MIN.NA HORA ANT. (AUT)(mB) | Valor mínimo de pressão atmosférica obtido na hora anterior. | float64 |
| TEMPERATURA DA CPU DA ESTACAO(°C) | Temperatura méria horária medida na CPU da estação meteorológica. | int64 |
| TEMPERATURA DO AR - BULBO SECO. HORARIA(°C) | Temperatura média do ar. | float64 |
| TEMPERATURA DO PONTO DE ORVALHO(°C) | Temperatura na qual o vapor de água que está em suspensão no ar começa a se condensar. | float64 |
| TEMPERATURA MINIMA NA HORA ANT. (AUT)(°C) | Valor mínimo de temperatura ambiente obtido na hora anterior. | float64 |
| TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT)(°C) | Temperatura máxima do ponto de orvalho na última hora. | float64 |
| TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT)(°C) | Temperatura mínima do ponto de orvalho na última hora. | float64 |
| TENSAO DA BATERIA DA ESTACAO(V) | Valot de tensão elétrica na bateria da estação meteorológica em (V). | float64 |
| UMIDADE REL. MAX. NA HORA ANT. (AUT)(%) | Umidade relativa do ar máxima na hora anterior. | int64 |
| UMIDADE REL. MIN. NA HORA ANT. (AUT)(%) | Umidade relativa do ar mínima na hora anterior. | int64 |
| UMIDADE RELATIVA DO AR. HORARIA(%) | Porcentagem horária de umidade relativa do ar. | int64 |
| VENTO DIRECAO HORARIA (gr)(° (gr)) | Direção do vento horária. | int64 |
| VENTO. RAJADA MAXIMA(m/s) | Velocidade máxima do vento. | float64 |
| VENTO. VELOCIDADE HORARIA(m/s); | Média horária de velocidade do vento. | float64 |

Os dados do Climatempo foram obtidos via chamadas HTTP em código python. O cadastro para uso dessa API pode ser feito no site, obtidos em 14/07/2021:

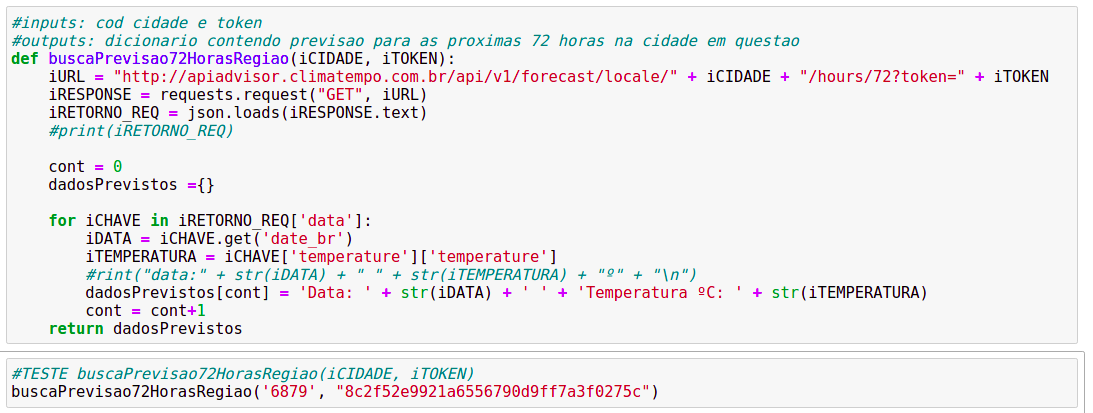
<https://advisor.climatempo.com.br/>

Tabela : Dados Obtidos Utilizando a API Clima Tempo

|  |  |  |
| --- | --- | --- |
| Nome da coluna/campo | Descrição | Tipo |
| Data | Data e Hora da medição | object |
| Temperatura | Temperatura do ar (Ambiente) | int64 |

A leitura e os médotos disponibilizados por essa API podem ser visualizados no notebook “apiClimaTempo.ipynb”.

Figura : Código Python de Consulta da API Clima Tempo



Fonte: Autor

Para o desenvolvimento do projeto, utilizou-se a plataforma Anaconda Navigator 1.9.12. O Anaconda é um projeto open-source que já contém a linguagem Python e centenas de bibliotecas “embutidas”.

Utilizando o Anaconda, todas as bibliotecas “obrigatórias” para fazer ciência de dados. Além das diversas bibliotecas disponíveis, o software é Anaconda é multiplataforma, ou seja, funciona em Windows, Linux e MacOS.

No presente trabalho, utilizou-se o sistema operacional Linux UBUNTU 19.04. Utilizou-se a ferramenta Jupyter Notebook através do ambiente Anaconda Navigator. O Jupyter Notebook é uma evolução do bloco de notas. De forma prática, é possível compilar trechos de códigos de diversas linguagens de programação. No trabalho em questão, utilizou-se a linguagem python na versão 3.7.6.

# 3. Processamento/Tratamento de Dados

Ao longo da etapa de processamento e tratamento dos dados, utilizou-se as bibliotecas pandas e numpy. Alguns ajustes foram neessários usando o editor de textos "MS Visual Studio Code". Os arquivos “DadosBHCercadinho2020.csv” e “DadosBHCercadinho2021.csv” passaram pelos ajustes.

* Substituição de ";" por “,” para fazer a separação das colunas, para evitar falhas de leitura ao longo do código.
* Casas decimais eram separadas por vírgulas (seguindo o padrão utilizado no Brasil) sendo convertido para ponto, para evitar falhas de leitura ao longo do código.

Na etapa de processamento e tratamento dos dados, foi realizada usando o notebook “INMETDATACLEANING.ipynb” Primeiramente, realizou-se o carregamento dos dados, como segue na imagem abaixo:

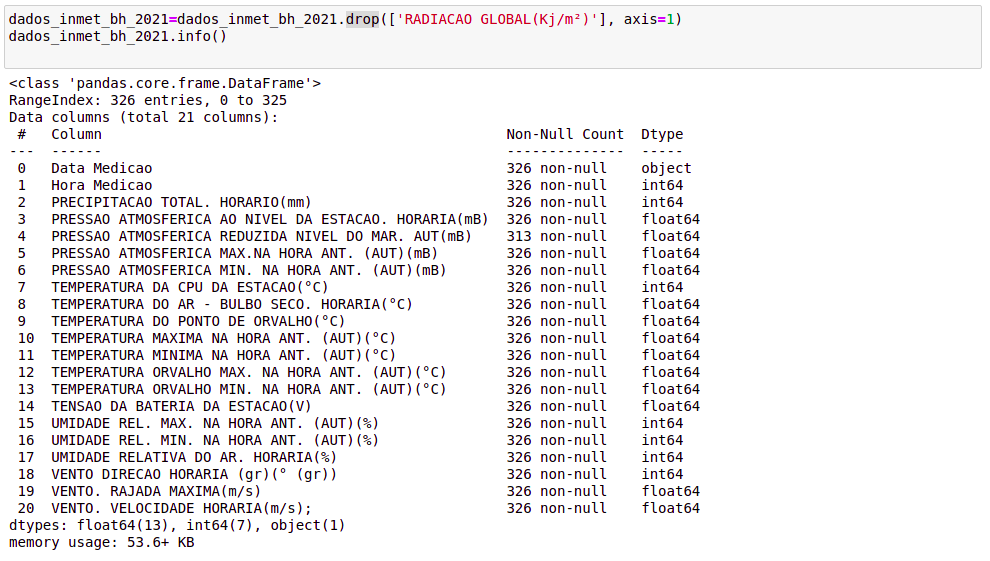
Figura : Carregamento de Dados do INMET



Fonte: Autor

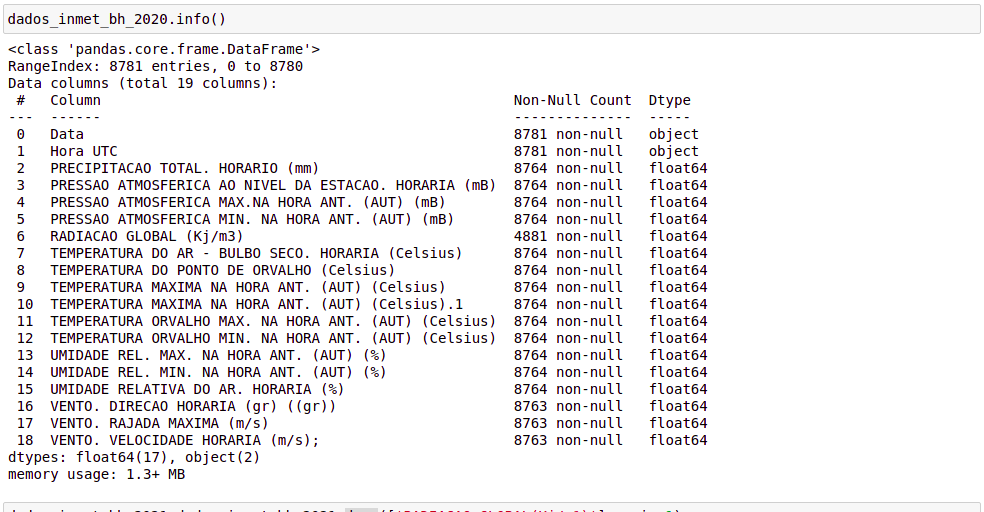
Visualizou-se que os dados das variáveis “Data” e “Hora UTC” estavam completos ao longo de todos os registros para o dataset de 2020, porém, as demais variáveis, apresentavam um valor equivalente a 0,204% dos dados faltantes, exceto pela variável “RADIACAO GLOBAL (Kj/m3)”, possuindo 44,48% dos dados nulos. Optou-se por deletar a variável “RADIACAO GLOBAL (Kj/m3)”, como segue na imagem abaixo. Essa variável foi eliminada do datasets de 2020 e 2021, devido ao grande percentual de dados nulos que esta possuia. No dataset de 2021 apenas a variável “PRESSAO ATMOSFERICA REDUZIDA NIVEL DO MAR. AUT(mB)” apresentou 3,09% de dados nulos.

Figura : Detalhamento dos dados do INMET de 2021 durante tratamento de dados



FONTE: Autor

Figura Detalhamento dos dados do INMET de 2020 durante tratamento de dados



Fonte: Autor

Para as demais variáveis do arquivo de 2020, exceto “RADIACAO GLOBAL (Kj/m3)” que foi eliminada, utilizou-se a substituição dos valores nulos pela média, levando em consideração os pontos:

Positivos:

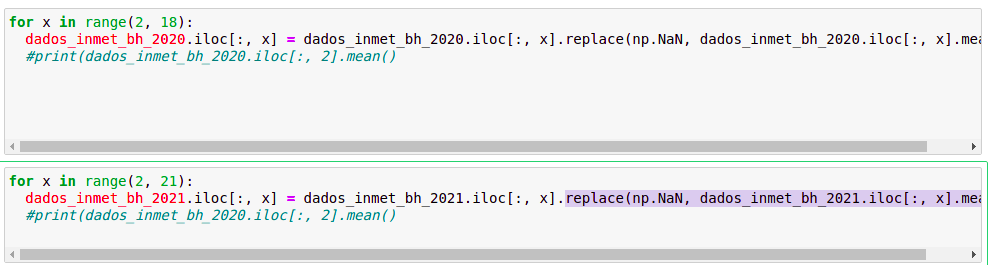
1. Resultado satisfatório quando o volume de dados não é massivo.
2. Melhor tratativa do que eliminar todas as ocorrências onde as valores estão nulos, pois geraria considerável perda de dados.

Negativos:

1. Inserir aproximações adiciona BIAS e variância aos dados.
2. Por se tratar de um método de implementação mais simples e rudimentar, tem o resultado inferior ao de outros métodos mais modernos, como por exemplo utilizar regressão linear para prever os dados faltantes.

Foi realizada a tratativa de substituição dos dados nulos pelos valores de média das variáveis usando o código exibido na imagem abaixo.

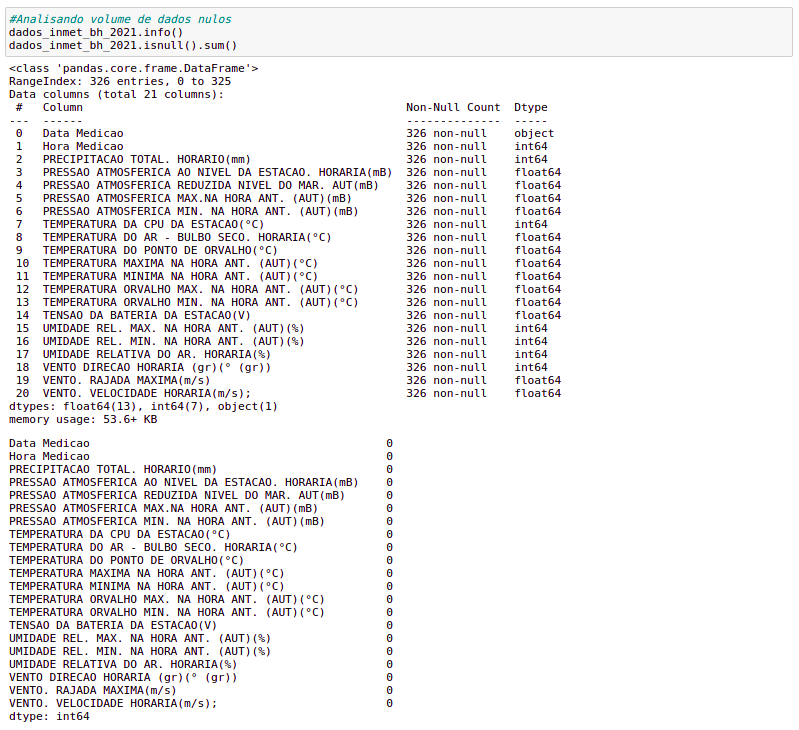
Figura : Tratamento de dados nulos



Fonte: Autor

Após a tratativa proposta, observou-se que todos os dados nulos foram substituídos pelos devidos valores médios das variáveis, como segue abaixo.

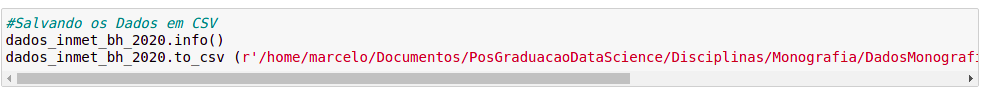
Figura : Dados INMET após tratamento



Fonte: Autor

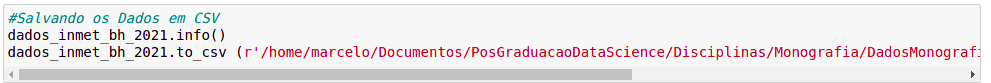
Após realizada a tratativa proposta, os dados foram salvos nos arquivos “dados\_inmet\_filtrados\_2020.csv” e “dados\_inmet\_filtrados\_2021.csv” para serem utilizados nas etapas posteriores.

Figura : Salvando os dados no formato CSV



Fonte: Autor

Figura : Salvando os dados no formato CSV



Fonte: Autor

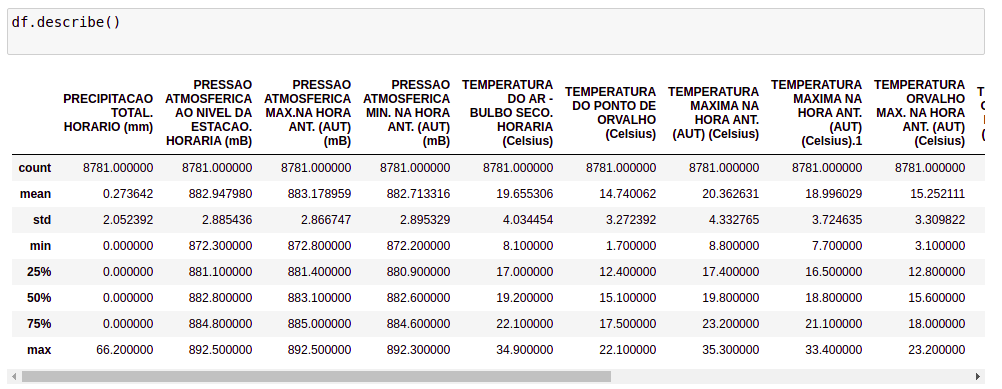
Os dados obtidos pela API Climatempo vieram a presença de dados nulos e não demandaram tratamento nesta etapa.

# 4. Análise e Exploração dos Dados

Ao longo da etapa de análise, processamento e exploração de dados, utilizou-se as bibliotecas pandas, numpy, matplotlib e seaborn.

Foi utilizado o método dataframe.describe() do pandas para analisar as métricas estatísticas média, desvio padrão, mediana e etc.

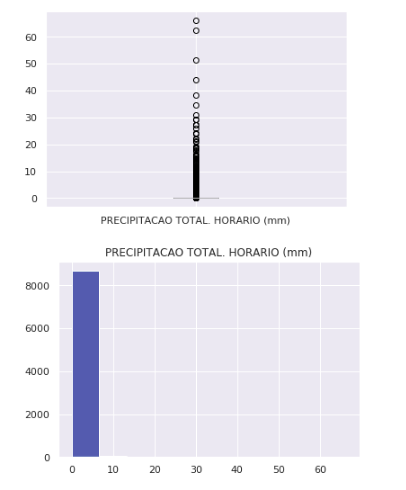
Figura : Análise estatística dos dados



Fonte: Autor

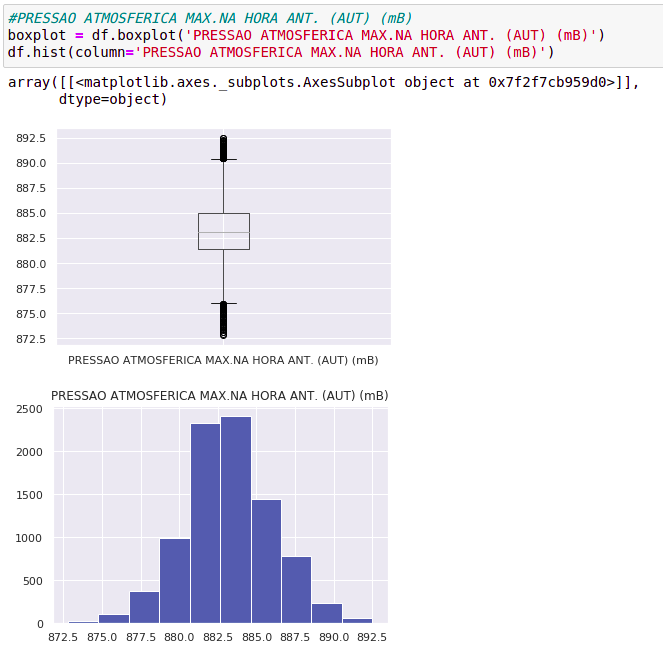
Nessa etapa foi observou-se que as variáveis apresentavam um baixo nível de dispersão, possuindo um desvio padrão entre 2 e 5, de maneira geral. Exceto pelas variáveis “Vento Direção Horária", “Umidade Relativa Horaria”, “Umidade Rel Min na hora anterior” e “Umidade Rel Max na hora anterior”, que apresentavam os valores 85.580479, 17.612537, 18.415922 e 16.635963, respectivamente. Uma análise de box plot e distribuição dos dados se faz mais detalhadamente, á seguir.

Figura : Análise da Precipitação Total

  
Fonte: Autor

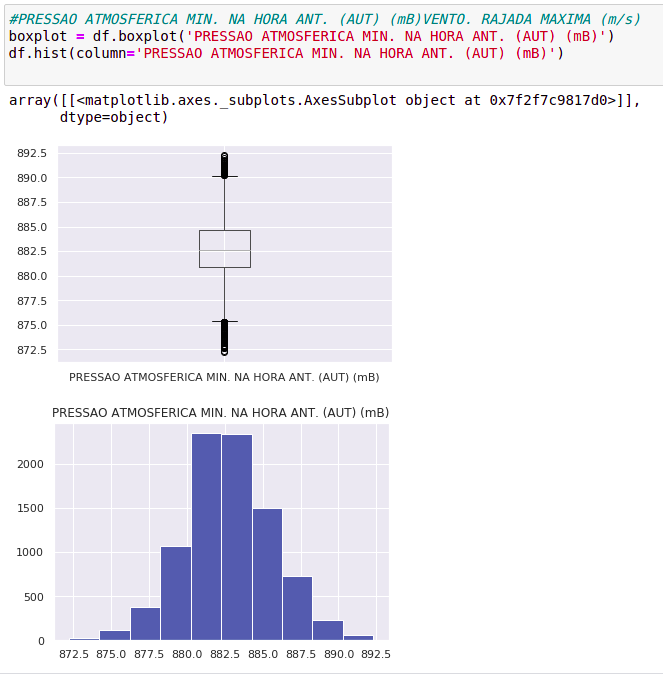
A variável relacionada à precipitação total horária, apresentou distribuição não-normal e a grande maioria dos dados entre 0 e 10 mm, sendo os demais valores acima de 10 considerados outliers. Isso se deve ao fato do índice pluviométrico no período de tempo que compreende o volume de dados analisado ter sido baixo.

Figura : Análise da Pressão Atmosférica Máxima na Hora Anterior



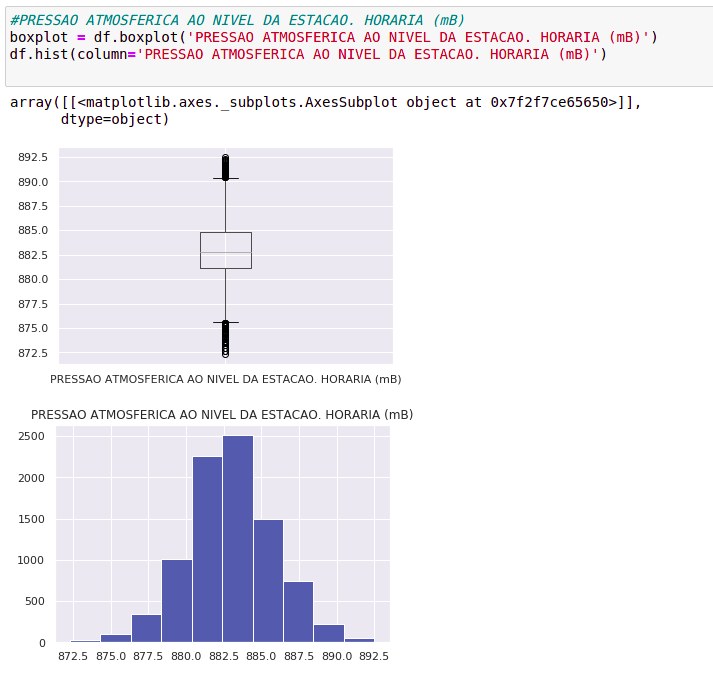
Fonte: Autor

Figura : Análise da Pressão Atmosférica Mínima na Hora Anterior



Fonte: Autor

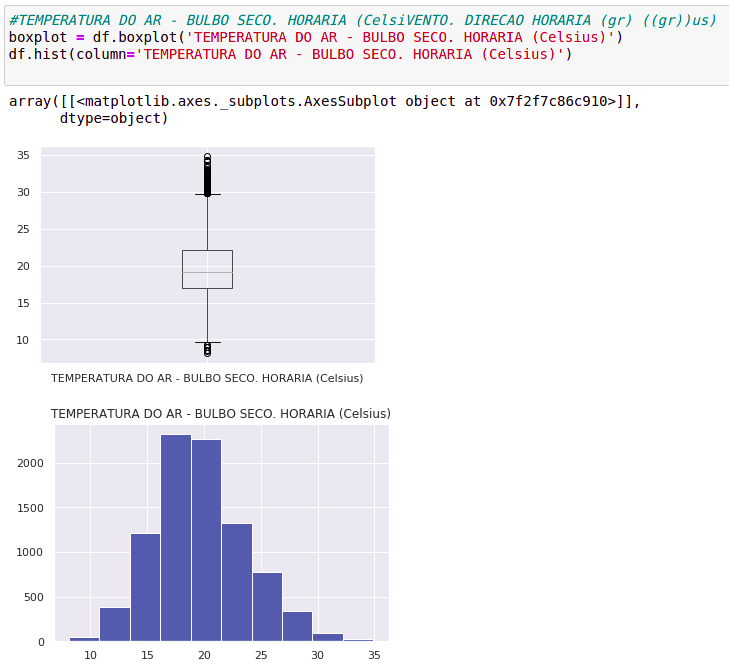
Figura : Análise da Pressão Atmosférica Média



Fonte: Autor

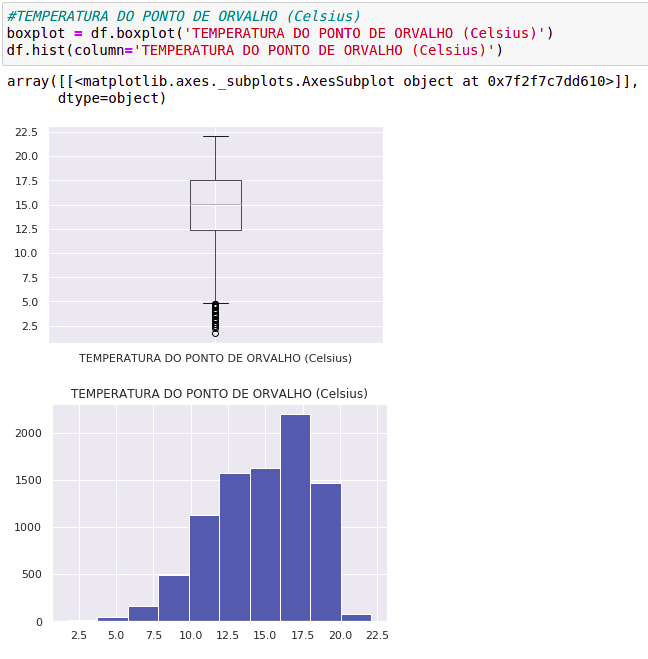
As variáveis relacionadas à pressão atmosférica, apresentaram comportamento bastante semelhante, tendo sua distribuição aproximadamente normal e o número de outliers acima e abaixo do valor máximo semelhantes.

Figura : Análise da Temperatura do Ar



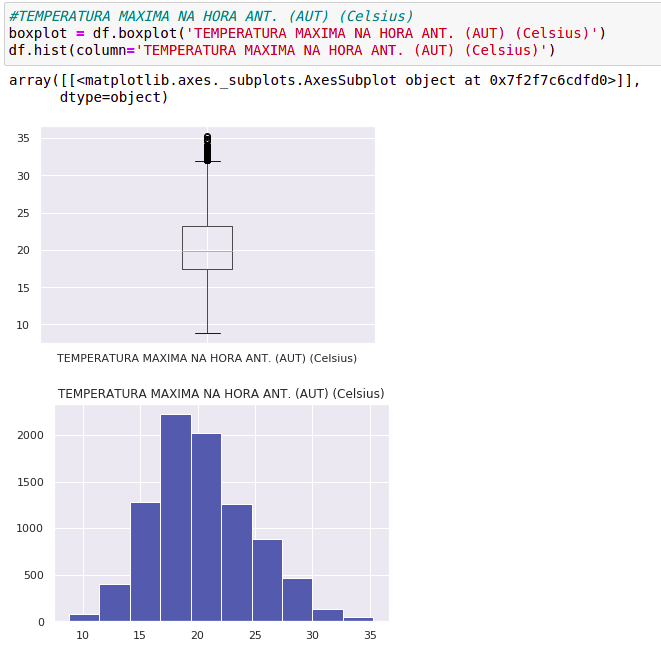
Fonte: Autor

Figura : Análise da Temperatura do Ponto de Orvalho



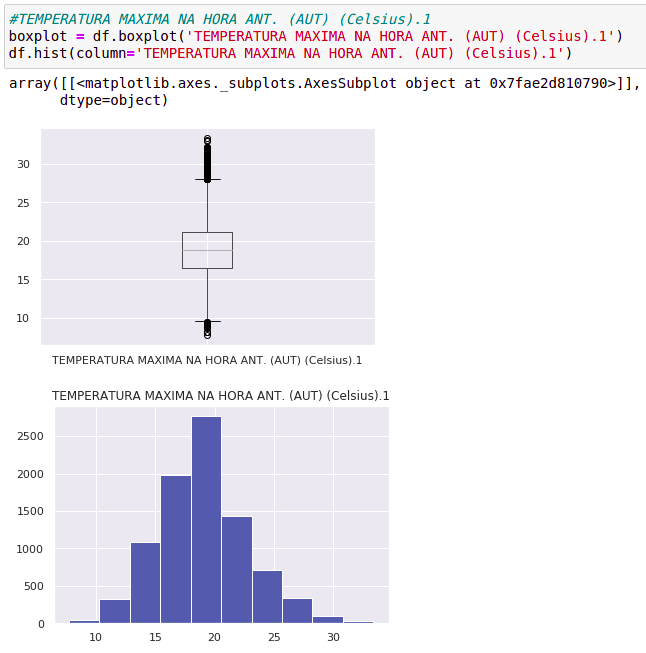
Fonte: Autor

Figura : Análise da Temperatura Máxima do Ar na Hora Anterior



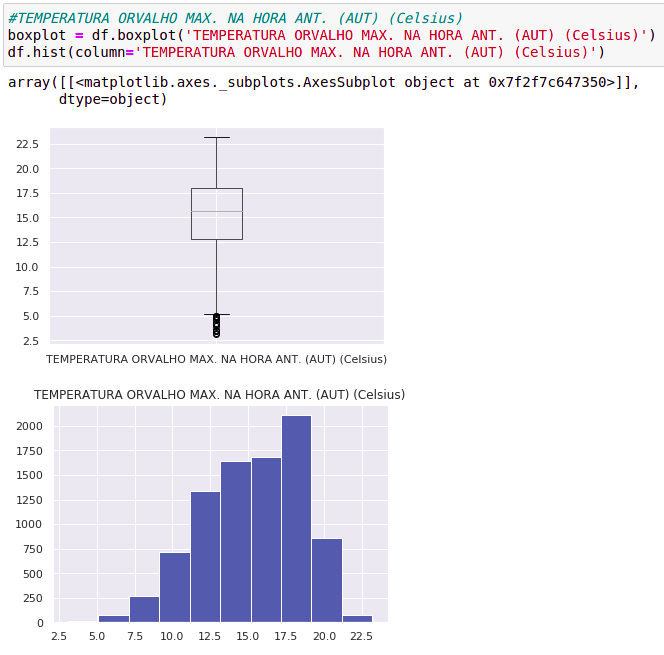
Fonte: Autor

Figura : Análise da Temperatura Mázima do Ar na Hora Anterior



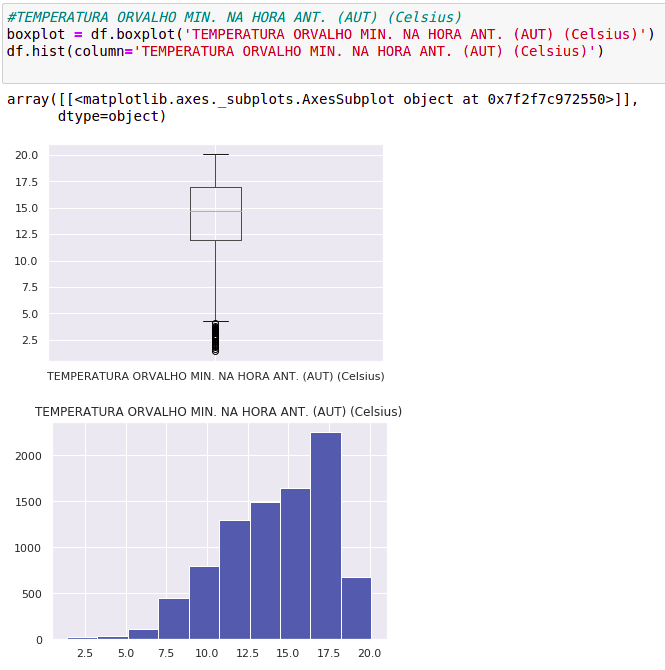
Fonte: Autor

Figura : Análise da Temperatura de Orvalho Máxima



Fonte: Autor

Figura : Análise Temperatura de Orvalho Mínima Na Hora Anterior

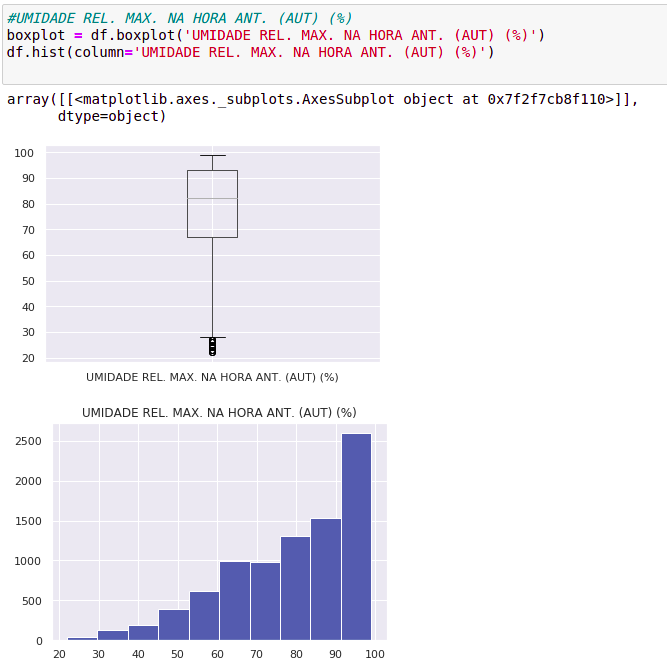


Fonte: Autor

Analisou-se que as variáveis relacionadas à temperatura e observou-se que não possuíam as mesmas características de distribuição de probabilidades e outliers. Todas possuem uma distribuição que se aproxima da distribuição normal e em relação à outliers, segue análise obtida.

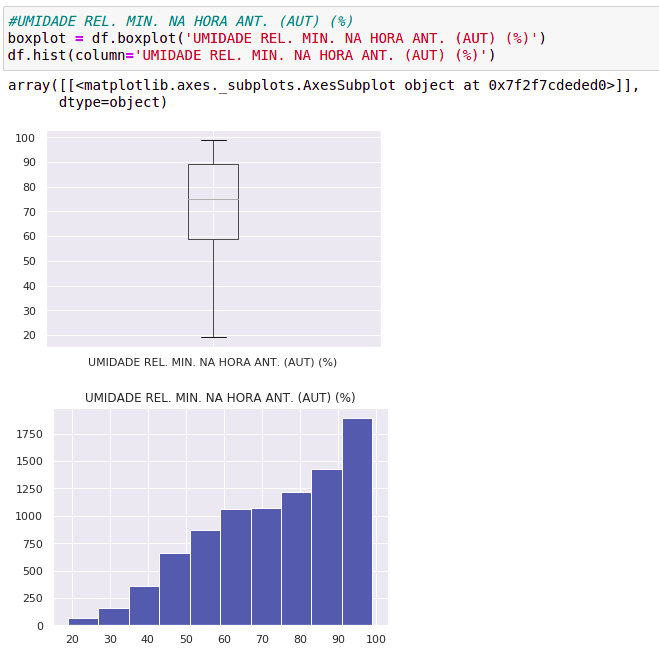
* Temperatura do Ar – Bulbo seco apresentou um volume considerável de outliers acima do valor máximo e poucos outliers abaixo do valor mínimo.
* Temperatura do ponto de Orvalho apresentou um grande volume de outliers abaixo do valor mínimo.
* Temperatura máxima na Hora anterior apresentou outliers acima do valor máximo.
* Temperatura Orvalho Máximo na hora anterior e Temperatura Orvalho Mínimo na hora anterior apresentaram um valor considerável de outliers abaixo do valor mínimo.

Figura : Análise Umidade Relativa Máxima Na Hora Anterior



Fonte: Autor

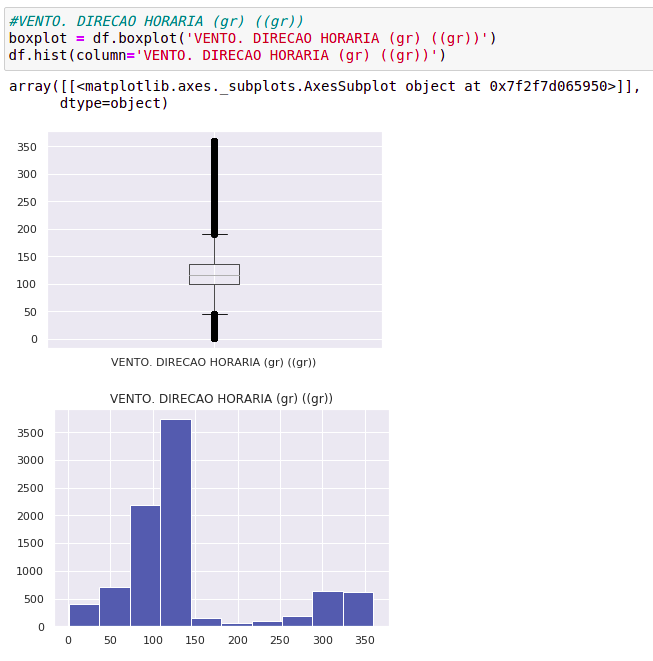
Figura : Análise Umidade Mínima na Hora Anterior



Fonte: Autor

As variáveis relacionadas a Umidade relativa do ar, apresentaram comportamento parecido. Todas as 3 variáveis apresentaram distribuição não normal e apenas a variável referente à umidade relativa máxima na hora anterior apresentou outliers abaixo do valor mínimo. As demais não apresentaram outliers.

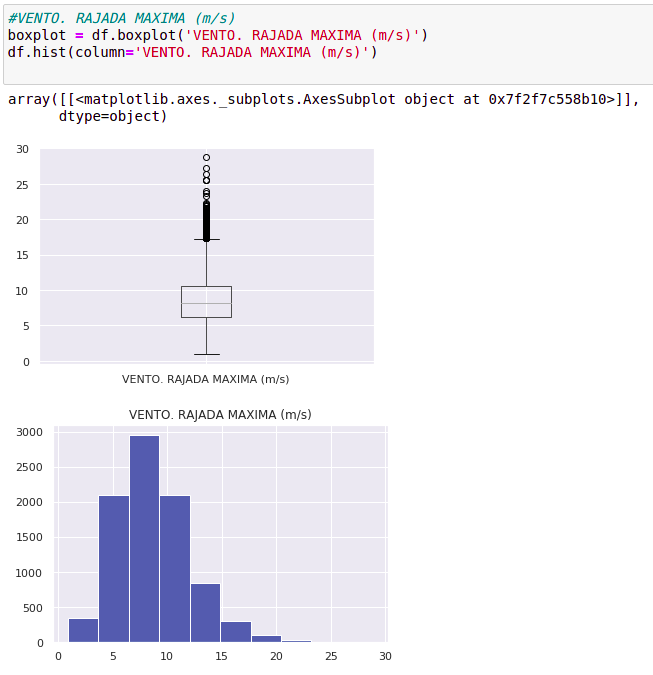
Figura : Análise Vento Direção Horária



Fonte: Autor

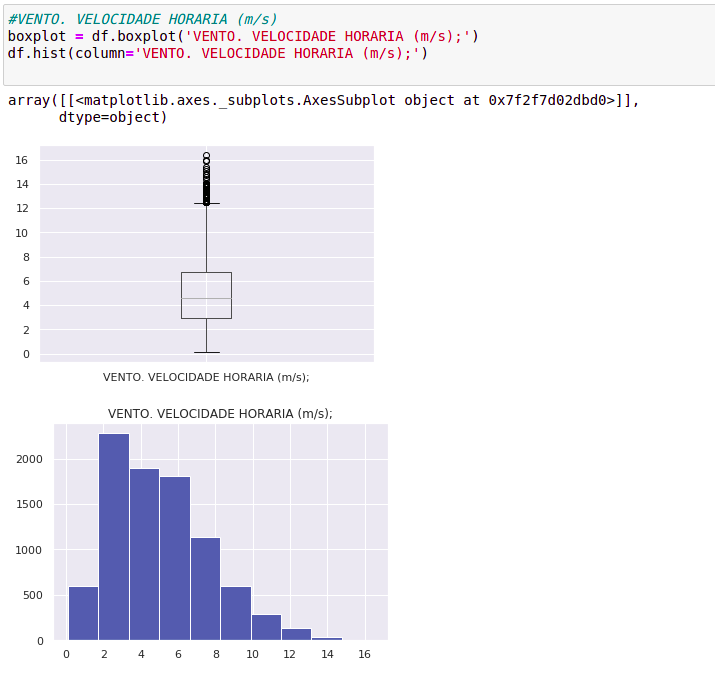
A variável relacionada a direção horária do vento apresentou um grande volume de outliers acima e abaixo dos valores máximo e mínimo, respectivamente.

Figura : Análise de Velocidade Máxima do Vento



Fonte: Autor

Figura : Análise da Velocidade Média do Vento

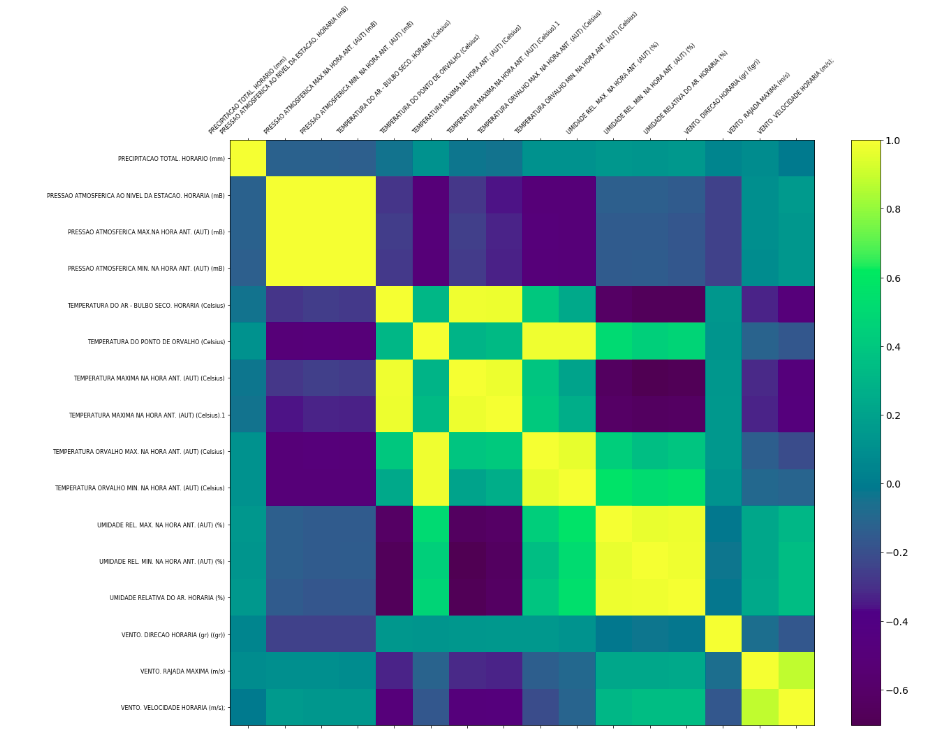


Fonte: Autor

Observou-se que as variáveis relacionadas à velocidade média horária e velocidade máxima horária apresentaram um certo número de outliers acima do valor máximo e uma distribuição estatística bastante semelhante. Os valores médios, máximo e mínimo foram diferentes pois uma das variáveis considera o valor médio horário do vento, enquanto outra considera o valor máximo horário.

Após as análises de distribuição e boxplot, realizou-se a análise de correlação dos dados e, afim de reduzir o número de variáveis à serem consideradas no processo de modelagem utilizando machine learning e ferramentas estatísticas, decidiu-se utilizar a análise de correlação e eliminar as variáveis que possuíssem correlação maior que 0,95 entre si.

Figura : Gráfico de Correlação Entre as Variáveis



Fonte: Autor

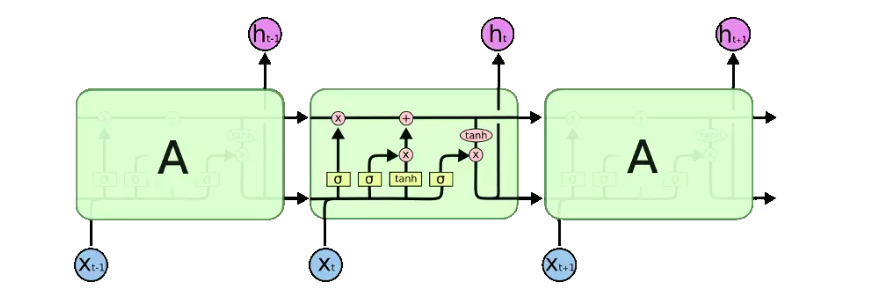
Através da análise realizada, pode-se observar correlação entre às variáveis de temperatura do Ar e Temperatura Máxima na Hora anterior. Na etapa de criação do modelo de machine learning, optou-se por utilizar a Temperatura Máxima na hora anterior e umidade relativa máxima na hora anterior como variáveis independentes, devido à correlação que apresentaram.

# 5. Criação de Modelos de Machine Learning

O modelo de rede neural recorrente utilizada para o processo de predição de temperatura do ar foi a arquitetura LSTM (“Long Short Term Memory”), também chamada de “redes de memória de longo prazo”. Essa arquitetura foi projetada para evitar o problema de dependência de longo prazo. Afinal, recordar a informação por longos períodos de tempo é praticamente o seu comportamento padrão.

Possui diversas linhas, onde cada linha carrega um vetor inteiro (desde a saída de um nó até as entradas dos outros); os círculos cor-de-rosa representam operações pontuais (como a adição de vetores); as caixas amarelas são camadas de rede neural aprendidas; as linhas que se fundem significam concatenação, enquanto uma linha de bifurcação denota que o seu conteúdo é copiado e enviado para locais diferentes.

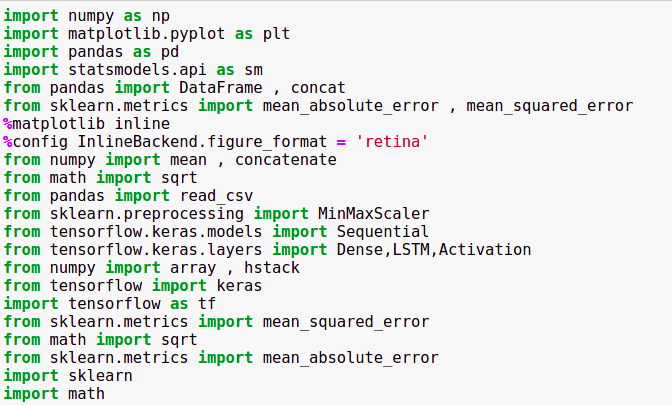
Figura : Representação no Modelo LSTM (Long Short Term Memory)



Fonte: Autor

Segue abaixo as bibliotecas python utilizadas para a etapa de desenvolvimento do modelo de machine learning.

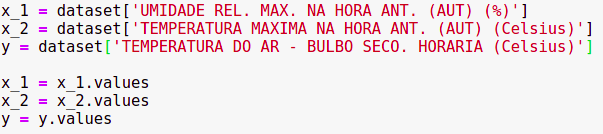
Figura : Bibliotecas Python Utilizadas na Criação do Modelo



Fonte: Autor

Levou-se em consideração, a fim de maior simplicidade do modelo, o menor número possível de variáveis de entrada para se obter a variável dependente, que no caso proposto seria a temperatura do ar. Considerando a análise de correlação e a premissa supracitada, optou-se por utilizar as variáveis abaixo como entrada do modelo(x\_1 e x\_2) pois de acordo com estudos da área, as variáveis relacionadas à umidade estão fortemente conectadas com mudanças de temperatura atmosférica, chuvas e etc. Optou-se por utilizar a temperatura máxima na hora anterior por esta variável possuir forte correlação com a variável dependente. Foram carregados 5000 linhas do dataset “dados\_inmet\_filtrados\_2020.csv” para as etapas de treinamento e testes. 300 linhas carregadas para a etapa de validação do modelo.

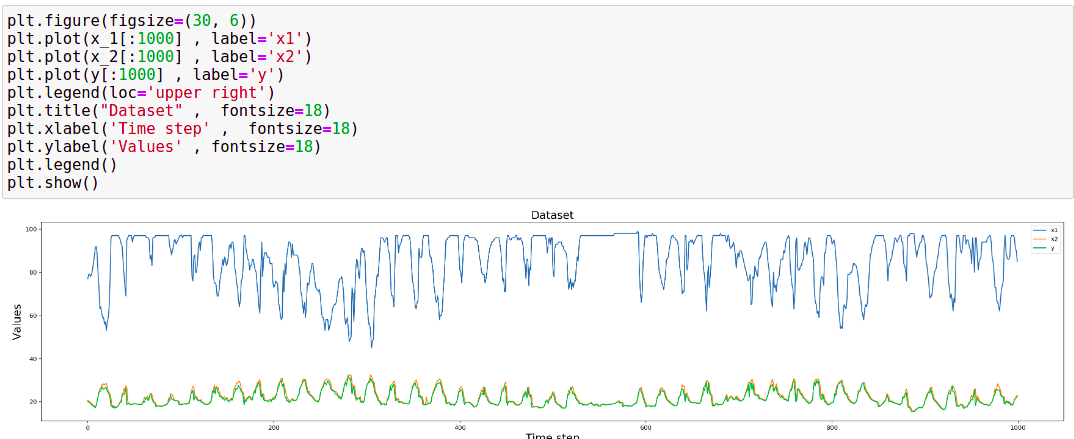
Figura : Variáveis Utilizadas no Modelo



Fonte: Autor

Segue abaixo o gráfico das tendẽncias das 3 variáveis em questão.Sendo a de Temperatura do Ar à variável à ser predita pelo modelo.

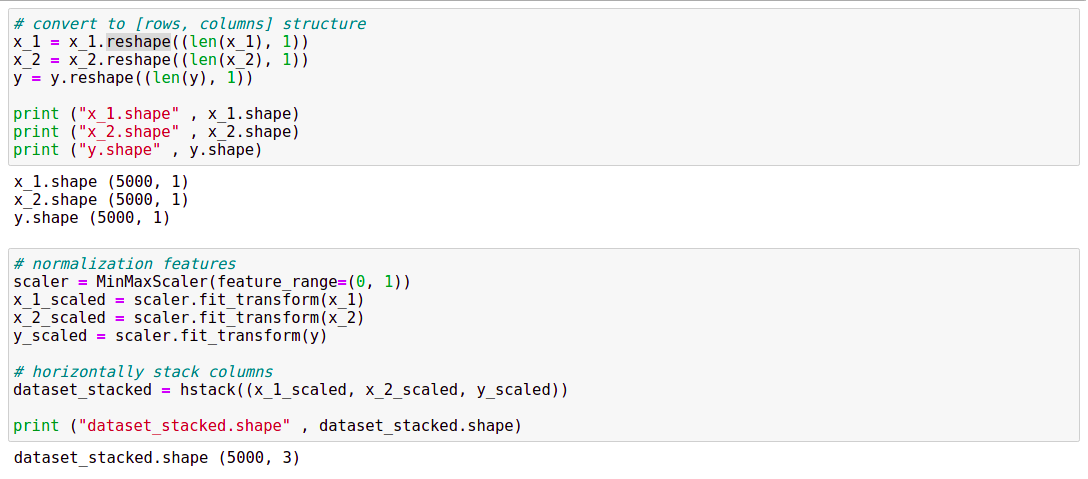
Figura : Visualização dos Variáveis Dependentes e Independente



Fonte: Autor

Os próximos passos foram a formatação dos dados e definição do dataset padronizado para ser utilizado pelo LTSM, tendo sido o método “MinMaxScaler” utilizado para normalização dos dados entre 0 e 1.

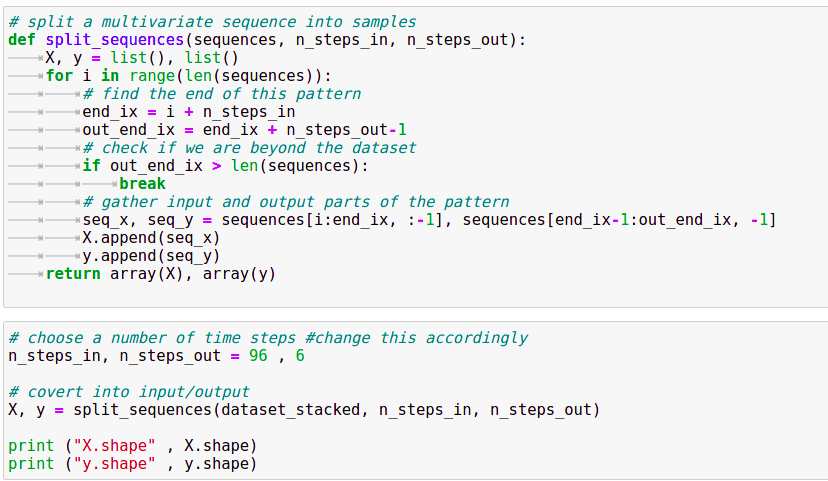
Figura : Normalização das Variáveis Independentes e Dependente



Fonte: Autor

A função split\_sequences() recebe os dados da sequência, o valor de quantos dados do passado deve ser usado para prever cada step à frente e o número de steps à ser previsto.

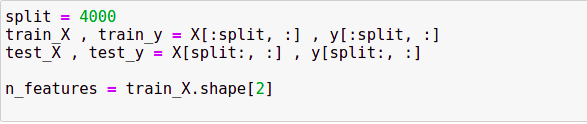
Figura : Preparação de Dados para Inserí-los no Modelo LTSM



Fonte: Autor

Em seguida, foi realizada a divisão entre dados de treinamento do modelo e dados de teste, usando a variável split com o valor atribuído de 4000, que indica 4000 dados para treinamento e 1 para teste.

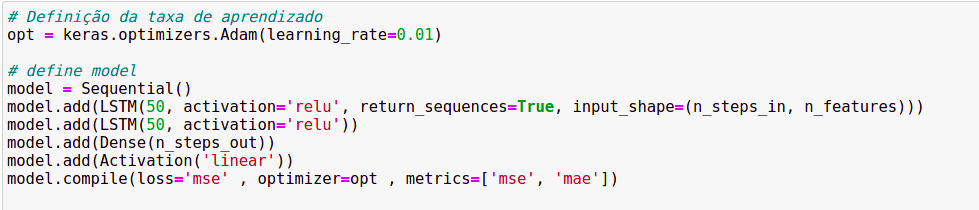
Figura : Separação dos Dados em Treinamento e Teste



Fonte: Autor

Definiu-se a taxa de aprendizado como 0,01. Este parâmetro representa a velocidade com o qual a rede neural aprenderá, sendo valores próximos à zero um aprendizado mais lento e gradual e valores próximos a um mais acelerado, mas também com grandes probabilidades de aumentar o erro no modelo gerado.

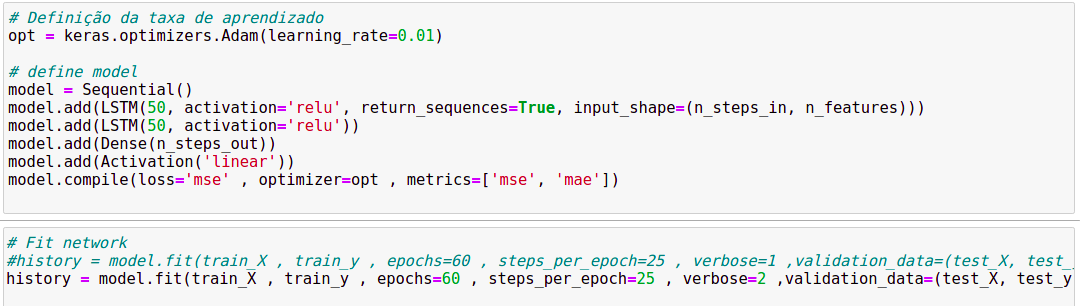
Figura : Definição e Configuração do Modelo



Fonte: Autor

Utilizou-se o modelo Sequential(), que é uma pilha linear e sequencial de camadas formando a rede neural recorrente, definiu-se a função de ativação da rede como linear, a função de perda e usou as métricas mean square error e mean average error para validação do modelo em treinamento.

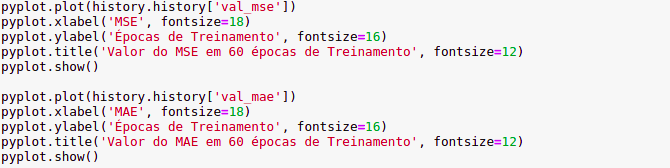
Figura : Configuração do Modelo e Inicio da Etapa de Treinamento



Fonte: Autor

Configurou-se o modelo a treinar com 60 épocas de treinamento e 25 passos de treinamento por época. Verificou-se que pelo método de análise Elbow, o valor de 60 épocas para o número de épocas como adequado, tendo em vista o decaimento exponencial do erro e um valor mínimo obtido próximo à 60 épocas de treino. O treinamento da rede LSTM é feito com a função model.fit().

Figura : Geração dos Gráficos de Erros de Treinamento



Fonte: Autor

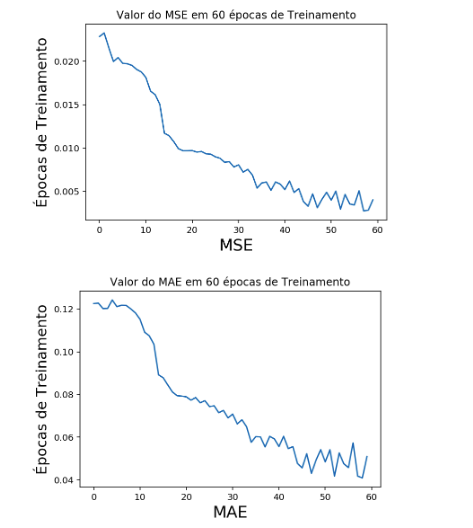
A biblioteca matplotlib foi utilizada para plotar os gráficos de MSE e MAE. Seguem abaixo os valores das métricas calculados para a última época de treinamento.

Figura : Erros Obtidos ao Fim das 60 Épocas de Treinamento



Fonte: Autor

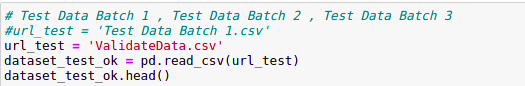
Figura : Análise de Erros de Treinamento



Fonte: Autor

Em seguida, os dados de validação foram carregados e parametrizados à mesma maneira que os dados de treinamento da rede, utilizando-se dos mesmos métodos e funções de parametrização.

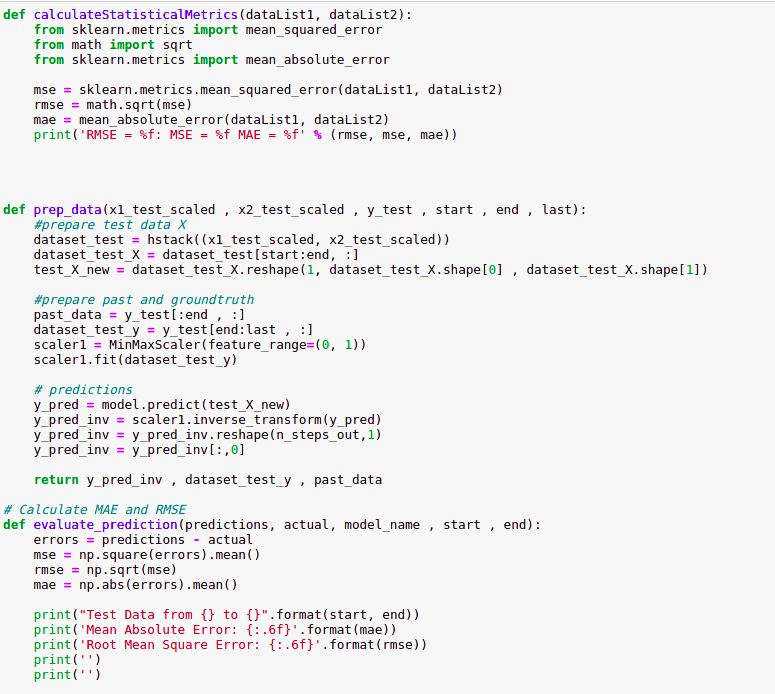
Figura : Carregando Dados de Validação



Fonte: Autor

Os cálculos das métricas estatísticas se fez usando funções da biblioteca sklearn. As funções descritas na imagem abaixo foram utilizadas para a preparação dos dados de validação para serem inseridos no modelo previamente treinado e também para cálculo das métricas para validar a acurácia das previsões realizadas.

Figura : Funções Para Cálculo de Métricas Estatísticas e Preparação de Dados

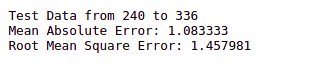


Fonte: Autor

Os valores de previsão obtidos através dos dados de validação foram retornados pela função model.predict(test\_X\_new), que recebe como parâmetro os dados de validação das variáveis independentes parametrizadas previamente.

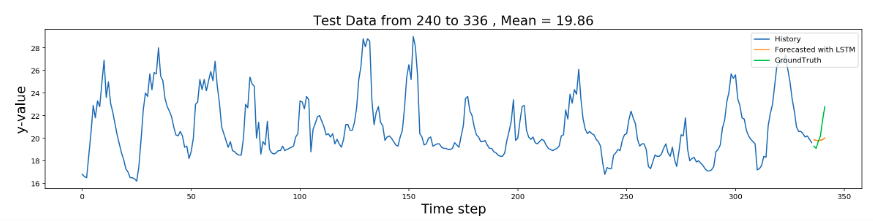
Na imagem abaixo pode-se visualizar o gráfico de valores históricos, comparados aos valores extrapolados pela previsão, juntamente às métricas estatísticas para validação do modelo.

Figura : Valores de Erro Obtido



Fonte: Autor

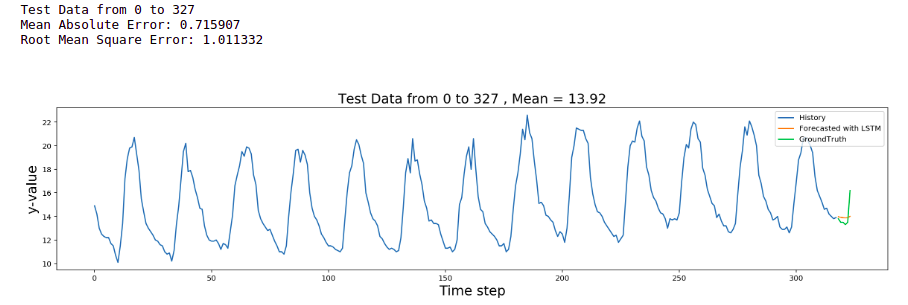
Figura : Gráfico Exibindo Dados Históricos e Valores das Previsões



Fonte: Autor

Para o processo de previsão da variável Temperatura do Ar para 6 horas no futuro, utilizou-se como dados de entrada o dataset de dados dados\_inmet\_filtrados\_2021.csv, composto por dados entre 01/07/2021 e 14/07/2021 e possuindo 327 registros históricos. Os métodos de preparação e inserção de dados no modelo foram os mesmos utilizados nas etapas de treinamento e validação dos dados já mencionadas e destrinchadas previamente no presente trabalho. Os valores das métricas estatísticas bem como o gráfico contendo os dados históricos e os valores previstos pelo modelo LTSM segue abaixo.

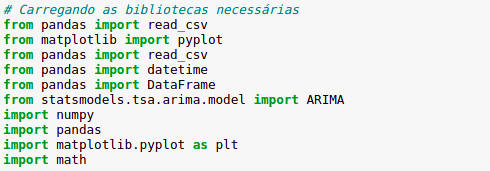
Figura : Valores Históricos e Predição Realizada pelo LSTM



Fonte: Autor

Paralelamente, para comparação com os resultados obtidos pelo LSTM, utilizou-se o ARIMA (Auto Regressive Model Moving Average), fortemente indicado em modelos preditivos. As bibliotecas abaixo foram utilizadas.

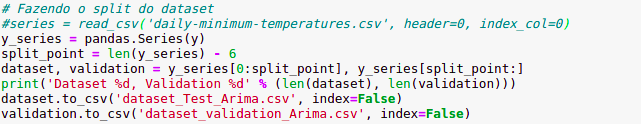
Figura : Bibliotecas Python Utilizadas



Fonte: Autor

O dataset utilizado foi o “dados\_inmet\_filtrados\_2021.csv” já descrito anteriormente no presente documento. Realizou-se a separação entre dados de treinamento e validação, como segue.

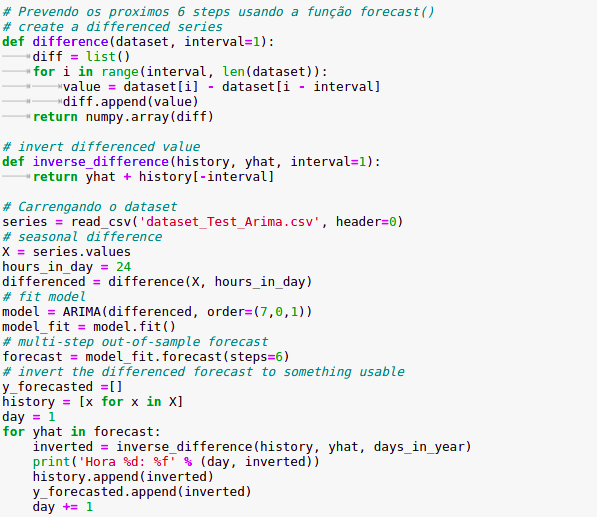
Figura : Preparação dos Dados de Validação



Fonte: Autor

Utilizando a função Forecast(), utilizando os dados de teste foi previsto para as próximas 6 horas no futuro o valor da temperatura do ar, considerando apenas esta variável no processo de modelagem.

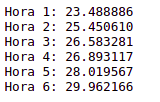
Figura : Inicialização do Modelo ARIMA



Fonte: Autor

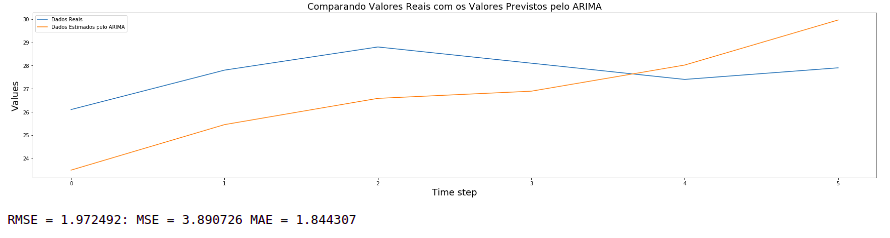
O resultado obtido foi:

Figura : Valores de Temperatura Preditos pelo ARIMA entre 14:00 e 19:00



Fonte: Autor

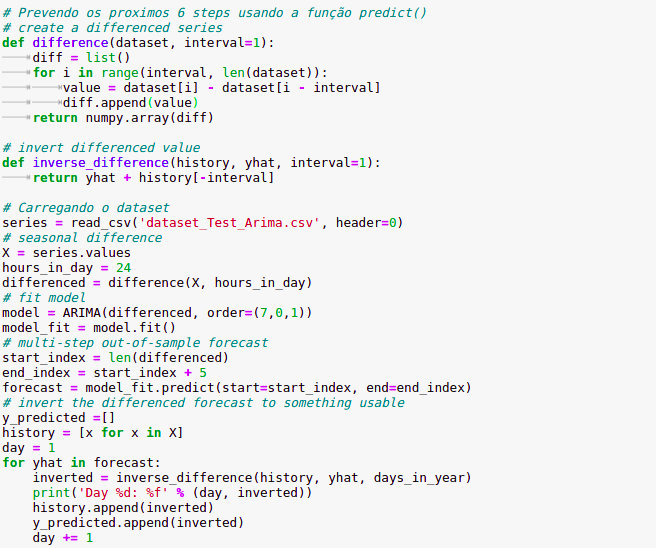
Figura : Gráfico de Comparação dos Valores



Fonte: Autor

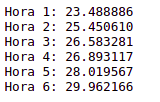
Utilizou-se o método Predict() para gerar as previsões, porém os resultados obtidos foram consideravelmente similares.

Figura : Configuração e Parametrização do ARIMA



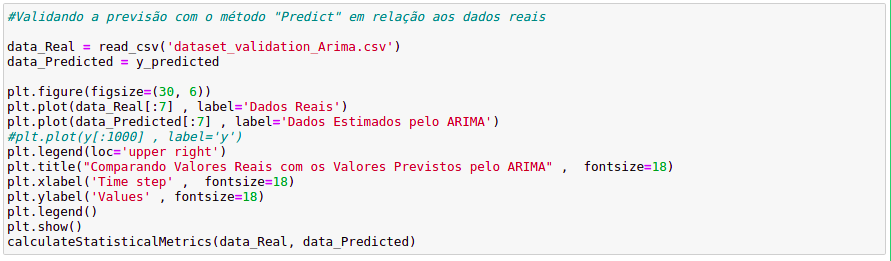
Fonte: Autor

Figura : Valores de Temperatura Preditos pelo ARIMA entre 14:00 e 19:00

  
Fonte: Autor

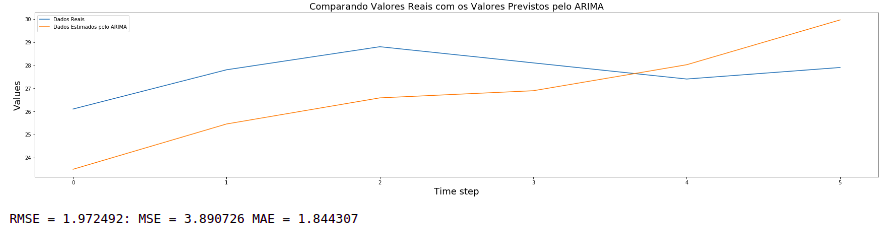
Realizando a validação do modelo ARIMA frente aos dados reais presentes no arquivo “dataset\_validation\_Arima.csv”, obteve-se o resultado abaixo:

Figura : Código Gerado Para Exibir os Resultados do Modelo



Fonte: Autor

Figura : Gráfico Contendo Dados Reais e Preditos pelo ARIMA

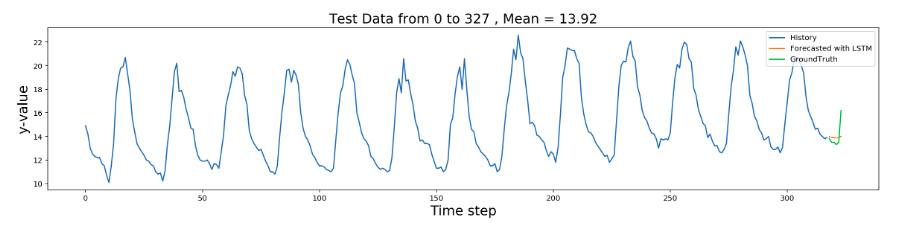


Fonte: Autor

# 6. Apresentação dos Resultados

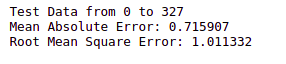
O resultado obtido utilizando o LSTM apresentou menores erros nas métricas estatísticas utilizadas (mean absolute error, root mean square error e mean square error) frente ao ARIMA, para realizar a previsão dos dados. Os dados previstos por ambos os modelos foram do dia 14/07/2021 das 14:00 às 19:00.

Figura : Resultados Obtidos pelo LSTM



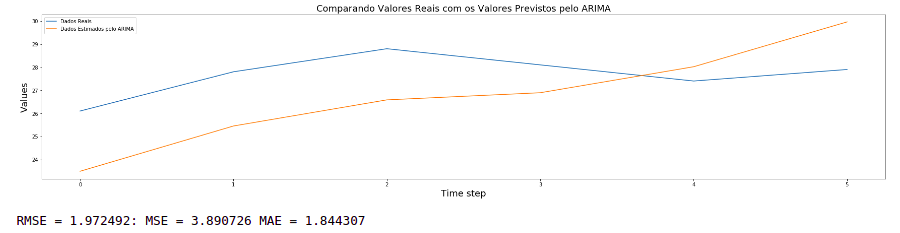
Fonte: Autor

Figura : Valores de Erro Obtidos pelo LSTM



Fonte: Autor

Figura : Resultados Obtidos Pelo ARIMA

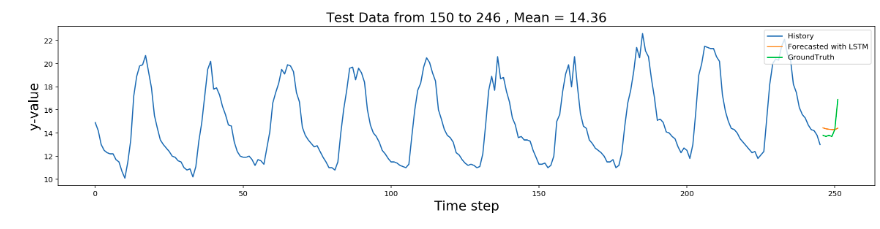


Fonte: Autor

Observou-se que o ARIMA também não acompanhou com precisão as mudanças de taxa de variação dos dados reais, tendo não obtido resultados consideravelmente precisos em regiões próximas à pontos de inflexão.

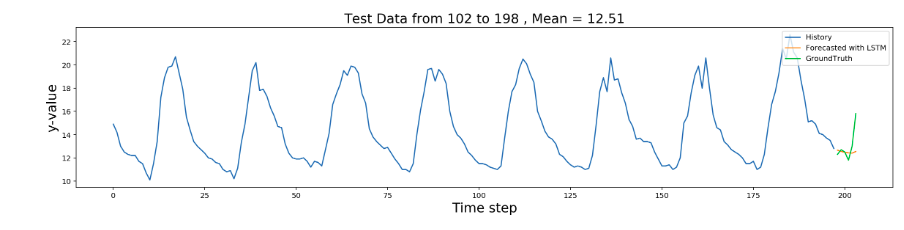
Mesmo tendo-se obtido menores erros com o Modelo LSTM, cabe salientar que este algoritmo não lidou com grande precisão com regiões próximas à pontos de inflexão, sendo que a taxa de variação dos dados previstos não conseguiu acompanhar com exatidão a taxa de variação dos dados reais, como segue nos exemplos abaixo.

Figura : Análise dos Resultados Obtidos pelo LSTM



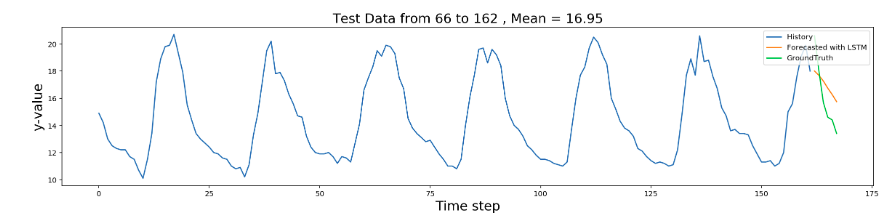
Fonte: Autor

Figura : Análise Parcial dos Dados Obtidos pelo LTSM



Fonte: Autor

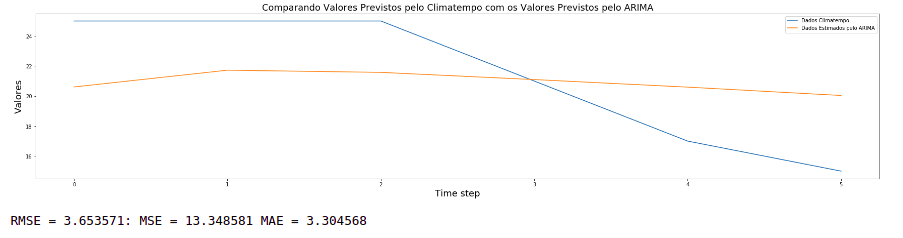
Figura : : Análise Parcial dos Dados Obtidos pelo LTSM



Fonte: Autor

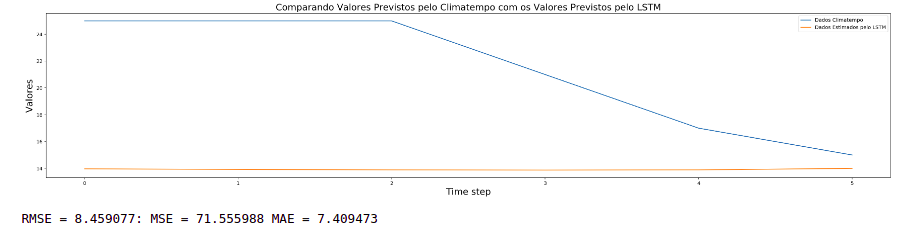
Para fins de visualização e comparação, comparamos os resultados obtidos pelo ARIMA e o LSTM com dados previstos pelo ClimaTempo, detalhados previamente no presente trabalho no ítem referente à coleta de dados.

Figura : Gráfico Comparando dados Previstos pelo Climatempo com Dados Previstos pelo ARIMA



Fonte: Autor

Figura : Gráfico Comparando dados Previstos pelo Climatempo com Dados Previstos pelo LSTM



Fonte: Autor

Observou-se um elevado valor de RMSE para o ARIMA e o LSTM, tendo o ARIMA apresentado melhor resultado. Entretanto, não se pode afirmar os dados obtidos pela API do Clima Tempo foram obtidos com os mesmos equipamentos de medição que os dados fornecidos pelo INMET, além do fato que não foi informado pela API Clima Tempo qual é a região da cidade onde foram obtidos os dados das medições, tendo essa comparação apenas um valor de visualização das respostas dos modelos propostos com dados obtidos através de métodos numéricos tradicionais.

Possivelmente, desenvolvendo um modelo com mais variáveis independentes relacionadas à temperatura do ar pode-se obter um modelo LSTM com menores erros, porém isso aumentaria o custo computacional e de armazenagem dos dados.

# 7. Links

Link para o vídeo:

https://www.youtube.com/watch?v=veCo2a3IxmA

Link para o repositório:

https://github.com/MarceloPereiraCunha/TCC\_DATASCIENCE\_MARCELOCUNHA

# REFERÊNCIAS

Yonue, Rita Yuri. **Meteorologia.** Cidade: São Paulo, Editora USP/UNIVESP, 2012.

Preduna, T.G. **Neuro Model for Weather Forecast.** Bucarest: IEEE, 2020.

Booz, Jarret. **A Deep Learning-Based Weather Forecast System**

**for Data Volume and Recency Analysis.** Frotsburg: ICNC, 2019.

Mahmood, Hisham. **Short-Term Photovoltaic Power Forecasting Using**

**an LSTM Neural Network and Synthetic Weather Forecast.** Lakeland: IEEE Access, 2020.

Haupt, Sue Ellen. **Machine Learning for Applied Weather Prediction.** Cidade: IEEE, 2018.

Salman, Afam Galih. **Weather Forecasting using Deep Learning Techniques.** Jakarta: IEEE, 2018.