A Machine Learning Model For Predicting Outdoor Air Temperature

Phase One

CSCE 3602 Fundamentals of Machine Learning

Amr Abdelbaky
Computer Science
The American University In Cairo
Cairo, Egypt
amrkhaled122@aucegypt.edu

Ali Eissa
Computer Science
The American University In Cairo
Cairo, Egypt
alieissa@aucegypt.edu

Introduction

Weather forecasting plays a role in aspects of contemporary society impacting sectors like energy generation, farming strategies and readiness for potential disasters. Given that air temperature has an impact on well being, ecosystems and industrial operations it stands out as a key factor in weather prediction. The primary objective of this initiative is to enhance the accuracy of air temperature forecasts by leveraging a weather dataset encompassing meteorological parameters.

Motivation

The reason for forecasting air temperature stems from the necessity of dependable weather predictions. Precise temperature forecasts impact safety, energy usage and the smooth functioning of industries. Grasping how air temperature relates to weather factors enhances climate modeling and aids in making decisions in areas, like farming, transportation and city planning.

Problem Specification

The primary goal of this study is to develop a predictive model for air temperature based on a diverse set of meteorological parameters. The dataset at our disposal includes time-related features like Datetime and day length, solar radiation measurements such as GHI Pyranometer, Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI), as well as additional parameters like humidity, wind speed, wind direction, barometric pressure, and sensor cleaning information. The complexity of the issue stems from the interactions and intricate relationships, among these factors. For example factors like radiation, wind speed and humidity can collectively impact air temperature. Their effects can differ based on varying conditions and geographical locations. The key lies in comprehending these relationships while creating a model that performs effectively when applied to new data .To tackle this challenge we require modeling machine learning algorithms and advanced data analysis techniques. Accurately predicting air temperature can enhance our understanding of climate patterns, boost the precision of weather predictions and offer insights for numerous applications reliant on temperature forecasts. The goal of this project is to support continued efforts to enhance weather prediction models and promote developments in the field of meteorology.

Literature Review

This literature review delves into the diverse methodologies employed in outdoor air temperature prediction models. We explore a spectrum of approaches, from simpler statistical models to complex deep learning architectures, highlighting their strengths, weaknesses, and contributions to the field. Particular attention will be paid to studies that showcase unique methodologies, address specific challenges, or demonstrate significant performance improvements. Through this review, we aim to provide a comprehensive understanding of the current state of the art in this domain and identify potential directions for implementation methodologies.

Traditional Approaches

Traditional methods for weather prediction-related problems were solved using numerical methods. The study [4] discusses Numerical Weather Prediction (NWP) which is a numerical model that solves the equations describing the atmospheric processes and how the atmosphere changes with time. NWP models divide the atmosphere into 3D cubes and solve weather parameter equations for each

atmospheric variable at each grid point as shown in Fig. 1. The higher the resolution of the model, the more accurate the results, but also the more computational power is needed.

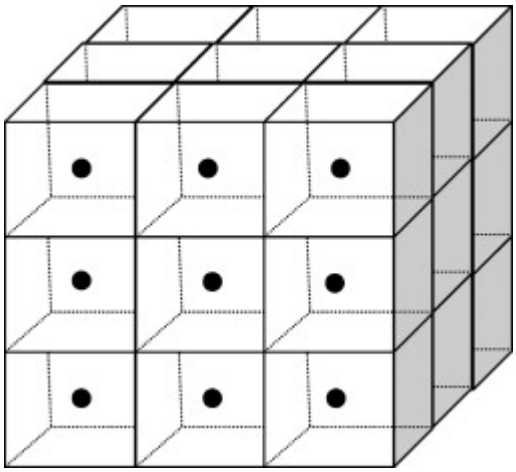

Fig. 1. 3D cubes of the atmosphere used by NWP models. [4]

The journey begins with crafting the initial state of the atmosphere, akin to assembling the initial configuration of a giant puzzle. This crucial stage, known as data assimilation, merges diverse real-world observations (temperature, pressure, wind, satellite data) with the previous model's output. Specialized algorithms meticulously analyze this amalgam to construct the initial conditions for each grid point within the NWP model. This initial state serves as the launchpad for the subsequent calculations. Such models rely on a set of fundamental equations derived from physics and fluid dynamics to govern the atmospheric dance.

These governing equations, while fundamental, are analytically intractable for large-scale atmospheric systems. To overcome this hurdle, NWP models employ a technique called discretization. The atmosphere is divided into a grid of 3D points, transforming it into a vast, numerical canvas. At each grid point, the complex equations are approximated using numerical methods, effectively translating them into solvable algebraic equations. This approach simplifies the problem but introduces inherent errors due to the approximations made.

With the equations discretized, the model embarks on its computational journey. Iterative calculations are performed at each grid point, predicting how atmospheric variables (temperature, pressure, wind) evolve over small time steps (typically seconds or minutes). The calculated changes at each point are then used to update the state of the atmosphere for the next time step. This process iterates until the desired forecast time is reached, essentially marching through time and unraveling the future state of the atmosphere. Fig. 2. Shows the complete architecture for NWP.

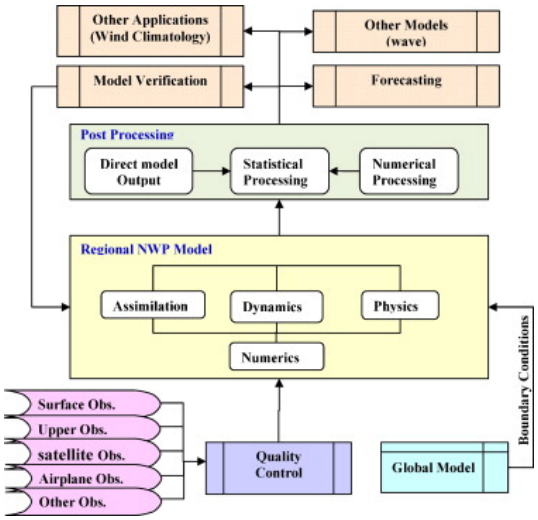

Fig. 2. NWP model Processes data flow. [1]

One of the advantages of NWP models is that they are freely available for download from the Internet. Another advantage is that they can provide gridded data for the whole model domain, so data is not limited to the weather station location. NWP models can also provide wind values at higher altitudes than just 10 meters. Unlike real world measurement where we cannot go back in time, NWP models provide flexibility to go back and forward in time. NWP models can be rerun for any historical event or period of time, and they can also provide one year of data within a shorter period of time than it would take to collect one year of real world measurements.

Machine Learning Approaches

Modernly, with the introduction of machine learning, Air temperature and weather prediction becomes an ideal application for such models as they provide classification and function approximation solutions with more reasonable computation times, with much less data analytical work, relying only on prepared real-world data. [1] investigates the use of machine learning (ML) models for short-term weather forecasting, focusing on the benefits of incorporating data from surrounding areas.

The aim of the paper was to develop a weather prediction technique that utilizes readily available historical data and

simple ML models, offering an alternative to complex, resource-intensive traditional methods. They emphasize the potential for such models to run on less powerful machines, making them more accessible and practical.

Compared to existing works that primarily leverage data from a single location, this study highlights the advantage of incorporating weather information from neighboring regions to enhance prediction accuracy. [1] used weather data from ten cities and trained several machine learning models including Ridge Regression, Support Vector Regressor, Multi-Layer Perceptron Regressor, Random Forest Regressor and Extra-Tree Regressor to predict the temperature of the next day for the city of Nashville, Tennessee.

They evaluated the model's performance by comparing its Root Mean Squared Error (RMSE) on test data, both when trained with data from all ten locations and when limited to Nashville alone. The study demonstrates that the RFR model trained with data from multiple locations achieves a significantly lower RMSE (35% less) compared to the model using only local data. Authors found that it is more effective to use historical data from surrounding areas to predict weather of a particular area, because some seasonal change or bad weather situation of a neighboring region might affect a particular region. This finding underlines the value of incorporating broader regional information for improved short-term weather prediction accuracy. The presented method offers a potentially more accessible and accurate alternative to traditional techniques, paving the way for further advancements in short-term weather prediction.

Another paper proposes a Multiple Linear Regression (MLR) based model, named MLRM, to predict the average temperature of a specific location for upcoming days [6]. The proposed MLRM model leverages multiple linear regression to predict temperature based on weather data from the past three days, an architecture of the model is illustrated in Fig. 4. The meteorological data is collected using Weather Underground's API, including features like mean temperature, mean humidity, precipitation, etc., for the past three days. The model employs two steps for feature selection: Correlation Analysis, which ensures a linear relationship by removing features with a Pearson correlation coefficient less than 0.6 with the target variable (mean temperature), and Backward Elimination, which iteratively removes insignificant features based on their p-values, using a significance level of 5%.The remaining data features are used to train the MLR model by splitting the data into (80% training and 20% testing), and aiming to minimize the mean squared error between predicted and actual values.

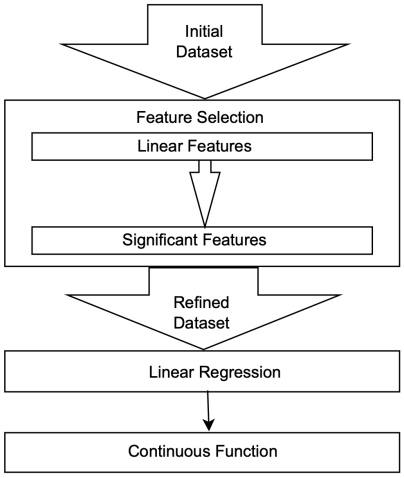

Fig. 4 Block diagram of proposed model [6]

Performance

The MLRM model achieves an absolute mean error (AME) of 2.8 degrees Celsius on the test set, indicating an accurate prediction of the average temperature with an error margin of +/- 2.8 degrees. The authors visualize the results using a scatter plot, where the closer the data points are to the y = x line, the higher the accuracy. The model however, poses limitations in its long term accuracy as it doesn't consider dynamic factors like weather patterns or climate change.

Accurately predicting outdoor air temperature plays a crucial role in diverse fields, impacting sectors ranging from agriculture and energy management to public safety and disaster preparedness. While traditional numerical weather prediction models offer valuable insights, their complexity and computational demands can pose limitations. Machine learning (ML) has emerged as a promising alternative, offering efficient and potentially more accurate temperature prediction models.

What Is Intended To Be Done

The intended model is going to utilize supervised learning models and data preprocessing techniques to predict one or more temperature samples in a day based on atmospheric data such as:humidity, time of day, windspeed, air pressure, and wind direction numeric data from 5 previous consecutive days. A variety of applicable supervised learning models will be explored to apply this problem

Prior to model development, we will conduct normalization on the numerical features and address any missing or outlier values through careful data preprocessing. We will conduct a thorough analysis of the dataset, examining statistical distributions, and evaluating feature correlations with the target variable.

Following the data preparation phase, we will proceed to a pilot study where multiple supervised machine learning models will be applied to the cleaned and preprocessed dataset. This comparative analysis will involve experimental setup, parameter selection, and performance evaluation. The top-performing models will undergo a detailed comparative analysis to determine the most suitable approach.

Moving into the model design phase, insights from the pilot study will guide the selection of the most fitting algorithm. We will design the model with considerations for data nature, size, and initial performance. Additionally, a utility application will be designed to enable predictions based on user-friendly inputs.

In the implementation phase, the chosen model will be developed from scratch, trained on the dataset, and deployed in the utility application. The entire process will be documented, including implementation choices, preprocessing components, and potential retraining methods, aligning with the specified deliverables.

Our main source of inspiration for the architecture of the model will be [6] for the similarity of the nature of the problem and the aim to solve it with a machine learning model. Based on the datasets found, it is possible that we combine some data from different regions to better enhance the generalization of the model.

Datasets

The quality of the training dataset for a machine learning model is crucial to the performance and generalization of the model model performance on real world samples. One dataset that caught our attention was “Pakistan Solar Radiation Dataset" , it contains **479,618** samples that were collected starting from October 10th 2014 with a 10-minute window between each sample till May 1st 2017. The dataset contains about **16** meteorological features that are as follows :

Date time , GHI Pyranometer , Direct normal irradiance (DNI), Diffuse horizontal irradiance (DHI), Temperature, Humidity, Wind speed , Speed of gust, Wind direction standard deviation , Direction, Pressure, Sensor cleaning, Station comments , Datetime, Day length (sunset - sunrise) , GHI Rotating Shadowband Irradiometer (RSI). The size of this dataset is about **44.6 MB** and it is available at kaggle [3].

This is a visualization for the first sample of that dataset :

```
Features for the first sample with units:
time = 10/25/2014 13:10 Date time
ghi_pyr = 628.3 GHI Pyranometer
dni = 652.3 Direct normal irradiance (DNI)
dhi = 191.3 Diffuse horizontal irradiance (DHI)
air_temperature = 25.5 Temperature (°C)
relative_humidity = 46.9 Humidity (%)
wind_speed = 3.9 Wind speed (m/s)
wind_speed_of_gust = 6.4 Speed of gust (m/s)
wind_from_direction_st_dev = 12.7 Wind direction standard deviation
wind_from_direction = 266.0 Wind direction (degrees)
barometric_pressure = 948.3 Pressure (mb)
sensor_cleaning = 0.0 Sensor cleaning
comments = nan Station comments
actual_date = 10/25/2014 Datetime
day_lenght = 662 Day length (s)
ghi_rsi = nan GHI Rotating Shadowband Irradiometer (RSI)
```

A Second Dataset we also came up upon was “Indian Weather Repository (Daily Updating)” , where it currently contains **92,623** sample that are still being collected from August 29th 2023 till now with **42** meteorological features that are as follows: Country ,Location Name, Region, Latitude, Longitude, Timezone, Last Updated Epoch(Unix timestamp of the last data update) ,Last Updated(Local time of the last data update * Indian time) , Temperature in Celsius, Temperature Fahrenheit, Condition Text(Weather condition description), Wind speed in Mph, Wind speed in Kph, Wind Direction in degrees,, Wind Direction(as in a 16-point compass), Pressure Measured in MB(millibars) , Pressure measured in In (inches),Precipitation amount in millimeters, Precipitation amount in inches, Humidity, Cloud(Cloud cover as a percentage), Feels-like temperature (in Celsius), Feels-like temperature(in Fahrenheit), Visibility (in kilometers), Visibility (in miles) ,UV Index, Wind gust in Mph, Wind gust in Kph,Air quality measurement(WRT Carbon Monoxide), Air quality measurement(WRT Ozone), Air quality measurement(WRT Nitrogen Dioxide), Air quality measurement(WRT Sulphur Dioxide), Air quality measurement(using PM2.5),Air quality measurement(using PM10), Air quality measurement(using US EPA Index), Air quality measurement (using GB DEFRA Index), Local time of sunrise ,Local time of sunset ,Local time of moonrise ,Local time of moonset ,Current moon phase ,Moon illumination percentage. The size of this dataset is about **23.97 MB** and it is available at kaggle [2].

This is a visualization for the first sample of that dataset :

```
Features for the first sample of Indian Weather Repository with units:
country = India Country
location_name = Ashoknagar Location Name (City)
region = Madhya Pradesh Administrative Region
latitude = 24.57 Latitude
longitude = 77.72 Longitude
timezone = Asia/Kolkata Timezone
last_updated_epoch = 1693286100.0 Last Data Update (Epoch)
last_updated = 2023-08-29 10:45 Local Time of Last Data Update
temperature_celsius = 27.5 Temperature (°C)
temperature_fahrenheit = 81.5 Temperature (°F)
condition_text = Partly cloudy Weather Condition Description
wind_mph = 12.8 Wind Speed (mph)
wind_kph = 20.5 Wind Speed (kph)
wind_degree = 281.0 Wind Direction (Degrees)
wind_direction = WNW Wind Direction (16-point compass)
pressure_mb = 1008.0 Pressure (mb)
pressure_in = 29.77 Pressure (inches)
precip_mm = 0.0 Precipitation Amount (mm)
precip_in = 0.0 Precipitation Amount (inches)
humidity = 67.0 Humidity (%)
cloud = 26.0 Cloud Cover (%)
feels_like_celsius = 29.7 Feels-like Temperature (°C)
feels_like_fahrenheit = 85.5 Feels-like Temperature (°F)
visibility_km = 10.0 Visibility (km)
visibility_miles = 6.0 Visibility (miles)
uv_index = 7.0 UV Index
gust_mph = 14.8 Wind Gust (mph)
gust_kph = 23.8 Wind Gust (kph)
air_quality_Carbon_Monoxide = 243.7 Air Quality: Carbon Monoxide
air_quality_Ozone = 45.8 Air Quality: Ozone
air_quality_Nitrogen_dioxide = 1.7 Air Quality: Nitrogen Dioxide
air_quality_Sulphur_dioxide = 3.1 Air Quality: Sulphur Dioxide
air_quality_PM2.5 = 12.6 Air Quality: PM2.5
air_quality_PM10 = 18.5 Air Quality: PM10
air_quality_us-epa-index = 1.0 Air Quality: US EPA Index
air_quality_gb-defra-index = 2.0 Air Quality: GB DEFRA Index
sunrise = 05:59 AM Local Time of Sunrise
sunset = 06:41 PM Local Time of Sunset
moonrise = 05:42 PM Local Time of Moonrise
moonset = 03:38 AM Local Time of Moonset
moon_phase = Waxing Gibbous Current Moon Phase
moon_illumination = 93.0 Moon Illumination (%)
```

A third dataset that we found also was “Climate Weather Surface of Brazil - Hourly” where it contains about **11.6 Million** Sample that were collected in Brazil starting from May 7th 2000 and stopping at April 30th 2021 with an hour window between each sample , this dataset has about **27** meteorological features that are as follows :

Date (YYYY-MM-DD), Time (HH:00), Amount of precipitation in millimeters (last hour), Atmospheric pressure at station level (mb), Maximum air pressure for the last hour (mb), Minimum air pressure for the last hour (mb), Solar radiation (KJ/m2), Air temperature (instant) (°c), Dew point temperature (instant) (°c), Maximum temperature for the last hour (°c), Minimum temperature for the last hour (°c), Maximum dew point temperature for the last hour (°c), Minimum dew point temperature for the last hour (°c), Maximum relative humid temperature for the last hour (%), Minimum relative humid temperature for the last hour (%), Relative humid (% instant), Wind direction (radius degrees (0-360)), Wind gust in meters per second, Wind speed in meters per second ,Brazilian geopolitical regions, State (Province) ,Station Name (usually city location or nickname), Station code (INMET number), Latitude, Longitude, Elevation. The size of this dataset is about 1.9 GB and it is available at kaggle [5].

This is a visualization for one of samples from that dataset :

```
Features for the first sample of Central West dataset:
index                                                  138998
Data                                                2017-12-20
Hora                                                     14:00
PRECIPITAÇÃO TOTAL, HORÁRIO (mm)                           0.0
PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)    899.6
PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)          900.0
PRESSÃO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)         899.6
RADIACAO GLOBAL (Kj/m²)                                 3391.0
TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)               26.5
TEMPERATURA DO PONTO DE ORVALHO (°C)                       17.7
TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)                 26.5
TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)                 24.4
TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)           18.3
TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)           16.5
UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)                   65.0
UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)                   57.0
UMIDADE RELATIVA DO AR, HORARIA (%)                        59.0
VENTO, DIREÇÃO HORARIA (gr) (° (gr))                       39.0
VENTO, RAJADA MAXIMA (m/s)                                  9.6
VENTO, VELOCIDADE HORARIA (m/s)                             3.9
region                                                       CO
state                                                        DF
station                                  PARANOA (COOPA-DF)
station_code                                               A047
latitude                                            -16.011111
longitude                                           -47.5575
height                                                  1043.0
```

In conclusion, After going through the above-mentioned three datasets , we came to consider the first dataset due to many different factors. One of the factors is due to its sample Size and file size , as it contains **479,618** samples and its also **44.6 MB** in file size compared to the Second and the Third datasets with **92,623 Samples** and a size of **23.97 MB** and **11.6 Million Samples** and a size of **1.9 GB** respectively. This size would fit our time limitation for this project . Furthermore , the First dataset 10 Minutes window between each sample allows for a better training than the third dataset with a one hour window between samples to support a wide range on inputs from the user (i.e 14:10 and 14:30 has samples in the first dataset m whereas its only 14 and 15 in the third dataset), In addition to that , the dataset being about the solar radiation plays a crucial role in predicting the temperature with its 16 given features which are not too much to train on that the model may overfit or too little to train on that it may underfit.

On the other hand , it's important that one of the limitations that must be addressed is the location of the dataset , where the results that will be obtained should be limited to countries that are in the region of Pakistan (The temperate climate zone) and this model cannot be generalized to all the other regions around the world , which is acceptable in Air Temperature prediction models .

References

[1] A H M Jakaria, Md Mosharaf Hossain, and Mohammad Ashiqur Rahman. 2020. Smart Weather Forecasting Using Machine Learning:A Case Study in Tennessee. Retrieved February 18, 2024 from https://arxiv.org/pdf/2008.10789.pdf

[2] Eligriyewithana,N.(2023, August). Indian Weather Repository (Daily Updating), Version 170. Retrieved February 18, 2024 from

https://www.kaggle.com/datasets/nelgiriyewithan/indian-weather-repository-daily-snapshot/data

[3] Farooq, U. (2014, October). Pakistan Solar Radiation Dataset, Version 1. Retrieved February 18, 2024 from https://www.kaggle.com/datasets/muhammadusmanfarooq/pakistan-solar-radiation-dataset/data

[4] Sultan Al-Yahyai, Yassine Charabi, and Adel Gastli. 2010. Review of the use of Numerical Weather Prediction (NWP) Models for wind energy assessment. Renewable and Sustainable Energy Reviews 14, 9 (2010), 3192–3198. Retrieved February 18, 2024 from https://doi.org/https://doi.org/10.1016/j.rser.2010.07.001

[5] INMET (National Meteorological Institute - Brazil). (2000, January). Climate Weather Surface of Brazil - Hourly, Version 9. Retrieved February 18, 2024 from https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region

[6] Ishu Gupta, Harsh Mittal, Deepak Rikhari, and Ashutosh Kumar Singh. 2022. MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of A Day. Retrieved February 18, 2024 from https://arxiv.org/abs/2203.05835