Determining Baltimore Area Weather Patterns

Group 2
Ethan Lee, Matthew Jensen, Ingrid Nkenlifack, Michael Kelican, Chirag Shah
IS428 Fall 2020
December 8, 2020

Abstract

Weather can have a significant impact on people, especially in the case of severe and hazardous weather conditions. Predicting weather has been a hot topic for a long time, and with the recent advancements in data mining, there is a chance to make even more accurate predictions on future weather. An increase in accuracy can lead to people knowing what is to come regarding future weather and prepare accordingly. By being able to accurately predict severe weather events, it could help people prepare for hazardous weather and natural disasters, and even save lives. For example, weather prediction in the agricultural industry has helped farmers save produce from harmful conditions. If people know there is going to be a hurricane, snowstorm, or other major weather event ahead of time, they can avoid the confusion and hysteria that comes with it.

This project aims to assess and analyze existing weather data in order to accurately predict severe weather events for the Baltimore area. Using data collected by the National Oceanic and Atmospheric Administration's National Centers for Environmental Information (NCEI), various data mining techniques such as regression models and classification algorithms were applied in an attempt to discover patterns in extreme weather cases. When focusing on temperature and wind as precursors to extreme weather, limitations within the dataset prevented several algorithms from producing the desired results.

Background

Although the idea of "weather" has been around since 3000 BC, the first official weather forecast was published in *The Times* on August 1st, 1861. Early weather forecasts consisted of basic observations written down and then used for predictions. The modern-day meteorologist then started utilizing weather balloons and as technology kept improving, things such as radars started being implemented. In addition to all these technological advancements, soon enough data mining was applied to weather forecasting. Neural networks, regression, and decision tree algorithms became the main data mining methods/tools used by meteorologists and scientists. Data mining application to weather forecasting at the basic level occurs by taking into account past weather data and climate patterns with attributes such as temperature, wind speed, visibility, air pressure, etc. We will be using these basic principles to predict Baltimore weather. As data mining and machine learning is improved upon, we believe more complex attributes will be taken into consideration and weather forecasting will continue to become more accurate.

Data

Our original topic of analyzing weather was rather broad and provided us with various aspects to focus on. Our first task was to decide on a geographic location to focus on, as weather conditions can vary dramatically based on the location. There were several options we considered, such as several major cities in the United States including Washington DC, Manhattan, and Los Angeles. However, given Baltimore's relevance to our lives as members of the UMBC community and its variance in weather throughout the year, it was the ideal candidate for performing our task.

Because of the nature of our project, it was important for us to source our dataset from an accurate and reliable source in order for the result of our data mining tasks to have any practical application. Thus, the dataset was selected from the National Oceanic and Atmospheric Administration's National Centers for Environmental Information (NCEI) database. NOAA is a government agency that was founded in 1970, after the merging of the U.S. Coast and Geodetic Survey, Weather Bureau, and the U.S. Commission of Fish and Fisheries. The agency conducts research on various aspects of Earth's atmosphere and oceans and has 9 key focus areas, climate being one of them. Due to their reliability as a government agency and expansive history of collecting data on weather and climate, we determined that one of their datasets would be the best fit for our task. NOAA indexes their datasets by each station that collects the data. There were 2 options that were within the vicinity of the city of Baltimore: Clyburn, MD and the Baltimore Washington International (BWI) Airport. Given BWI is closer and a more notable location between the two, it was chosen as the station to analyze.

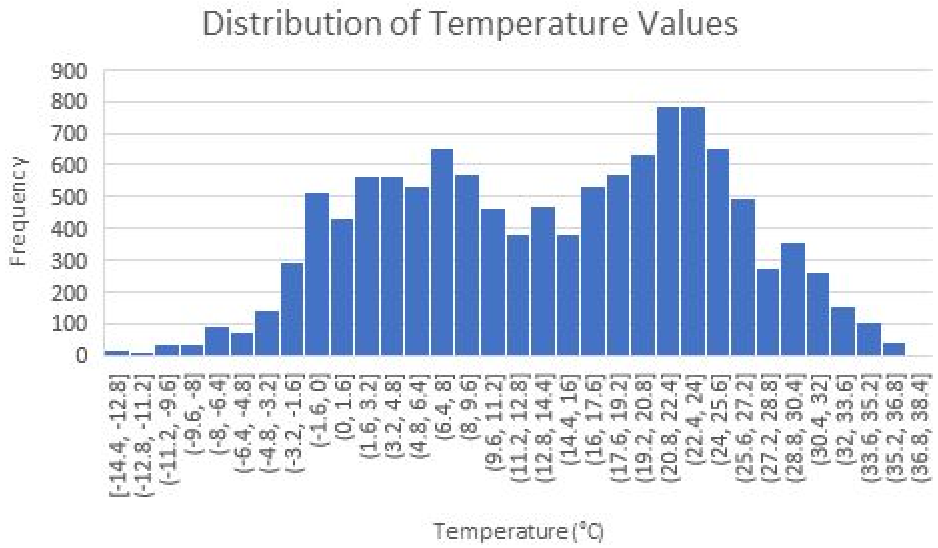
The original dataset had 16 main attributes which includes: station ID, date, source, latitude, longitude, elevation, station name, report type, call sign, quality control, wind observation, sky condition, visibility observation, air temperature observation, dew points, and sea level pressure. There were also many attributes that contained notes for various attributes, however they are not consistently used for all instances. The wind observation, sky condition, visibility observation, sky condition, air temperature observation, dew point, and sea level pressure all had multiple values combined that specified various quality codes and variability codes along with the initial measurement. This would later need to be cleaned in order to have any meaningful data mining tasks completed. The date range for this dataset spanned from January 1, 2019 to January 1, 2020 and had instances for at least every hour of each day, sometimes even more frequently than that. This resulted in a massive number of instances for this dataset. Overall, there were a total of 12,272 instances, 11,893 having complete data and 379 having missing data.

Because of common traits between hazardous weather occurrences in the Baltimore area, we highlighted certain instances of extreme weather to focus on. Extreme temperatures, hot or cold, and high winds were our focus out of the various forms of extreme weather. The dataset chosen contains the attributes that are closely related to the occurrences of our highlighted forms of extreme weather. High winds are often a precursor to severe storms and even hurricanes and tornadoes, and temperature is the main factor to extreme heat and cold. We believe that the data will be able to give us insights into patterns that indicate whether these weather scenarios will occur.

Pre-processing

At first glance, it was clear that our dataset would not be fit to run any tasks without any preprocessing being done first. First, there were several cleaning tasks that needed to be done. As mentioned above, there were several additional attributes that had notes for some of the main 16 attributes. Because they were not consistent and would not add much value to our analysis, we decided to remove them from our dataset. In addition to this, we removed any attribute that was repetitive or redundant. These include station ID, source, latitude, longitude, elevation, station name, report type, and call sign. Most of these attributes were the same for all instances throughout the dataset.

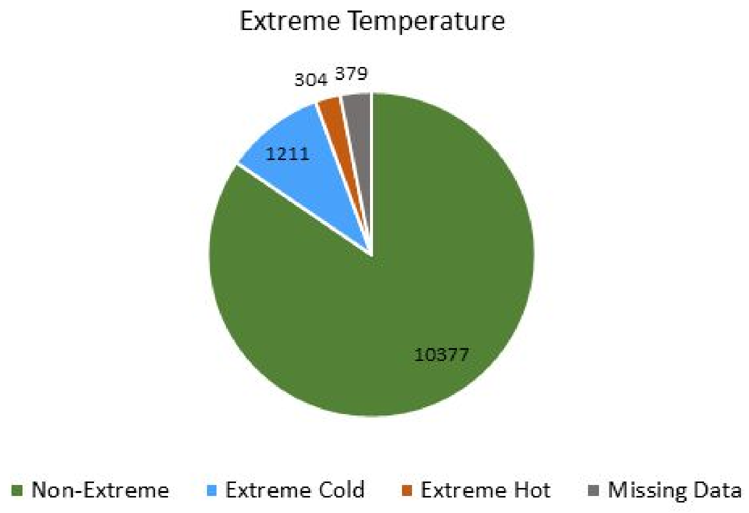
Because of our decision to focus on the 3 forms of extreme weather, our focus shifted to two attributes, wind observation and air temperature observation. The wind observation attribute contains the direction angle, direction quality code, type code, speed rate, and speed quality code, while the air temperature observation includes the air temperature and air temperature quality code. In order for Weka to properly analyze the data, we altered the attribute to only include the most important value for each attribute. Thus, the values in wind speed observation were simplified to the speed rate and the values in air temperature observation were simplified to the air temperature, both being numeric values. Wind speed rate was measured in meters per second while air temperature was measured in Celsius. *Figure 1* displays the distribution of air temperature values after simplification.



(Figure 1)

We have also created 2 additional attributes that classify the wind speed as extreme or non-extreme and the temperature as extreme hot, extreme cold, and non-extreme. Extreme winds were classified as anything greater or equal to 17.8 meters per second, extreme hot as 32.2 degrees Celsius and over, and extreme cold as 0 degrees Celsius and under. These thresholds were decided based on existing definitions created by the NWS for Baltimore/Washington. Extreme winds would equate to a high wind warning, extreme hot as a heat advisory, and extreme cold as a freeze warning. We believed that these set of definitions would be best as it would be most applicable in a practical setting.

After creating the subclasses, we noted that there were a very limited number of instances that were classified as extreme weather. For temperature, about 10.182% of the usable data or 1211 instances were classified as extreme cold and a mere 2.556% were classified as extreme hot (*Figure 2*). Most importantly, there were no instances of extreme wind in our dataset. Unfortunately this would impact the capabilities of our testing. However after some consideration, we ultimately decided to maintain the definitions that classified extreme weather. This was due to several reasons. The first being it would bias our results, since there would be more instances of extreme weather than in reality. Second, the results would hold no value in practical application because the definition would be different from the NWS, who actually issue the warnings.



(Figure 2)

After performing several tasks and realizing some limitations with our data, we prepared to shift focus to produce some form of results with our dataset. We attempted to predict the upcoming temperature values given the data for a certain period of time. To prepare for this, we created a test set by extracting the date and temperature values for the first 5 days of January. All preprocessing procedures were done using excel.

Methods

Linear Regression

Initially we performed a linear regression on the added attributes of extreme wind and extreme temperature before compiling the final dataset. It was during that process when we discovered there were no instances of extreme wind, so as part of preparing for the final analyses the extreme wind attribute was abandoned in favor of general wind data. To prep the final dataset, we traded out the extreme temperature attribute for a more general “extreme weather” one containing a binary yes/no, then moving the results of that shift down by a number of rows roughly equivalent to seven days.

Classification

From there, we used a shotgun approach to classification, with an emphasis on Bayesian models (so new data could be predicted with probabilities rather than binary choice) and deep learning via WEKA's multilayer perceptron generator (as the binary nature of the final class was conducive to this type of analysis). Despite that emphasis, however, we checked every other model that we had available to us that could be constructed in a reasonable amount of time, mainly to see if any surprising patterns arose from running the less well related algorithms on the data.

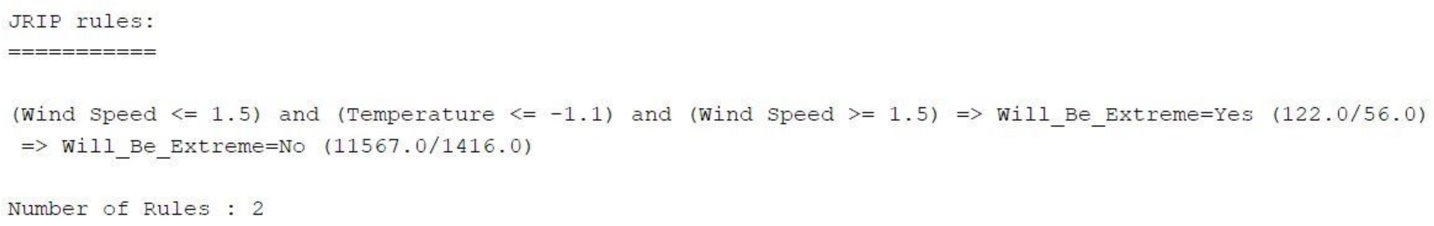
Time Series Analysis

Finally, we used time series analysis to forecast data given a certain excerpt of data. We extracted the temperature and timestamp values for January 1, 2019 to January 5, 2019, which was 173 instances of our dataset. Then, using the time series forecasting tool in WEKA, we attempted to forecast the temperature values for the following day, January 6, 2019, using linear regression.

Results and Findings

The linear regression for temperature gave us a correlation coefficient of 0.9975, mean absolute error of 0.0302, root mean squared error of 0.7407, relative absolute error of 0.3333%, and root relative squared error of 7.0761%. Linear regression for wind speed gave us a correlation coefficient of 0.9461, mean absolute error of 0.0932, root mean squared error of 0.7575, relative absolute error of 5.2582%, and root relative squared error of 32.6091%.

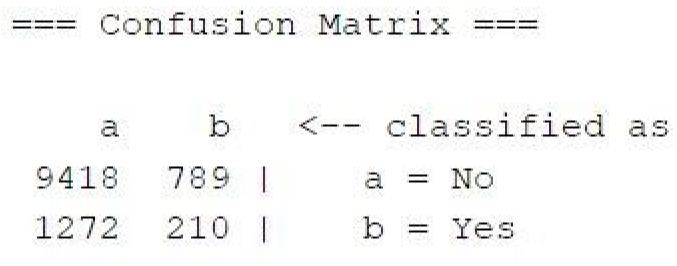
The most reliable classification model we used was JRIP's rule generator with an accuracy of 87.3813%. The nature of rules-based classification and the unpredictability of weather in general, however, made this more of an interesting aside than something genuinely useful. On the subject of interesting asides, 0R managed to attain an accuracy of 87.3214% by simply predicting the majority class. This indicated to us that we needed to change our metric of success, as a weather model that simply states extreme weather never happens is an exceedingly unhelpful one. While flat accuracy is nice, we decided it would be more important to select a less accurate model that would be more likely to identify severe weather to more closely align with the goals of our project.



```
JRIP rules:
===========

 (Wind Speed <= 1.5) and (Temperature <= -1.1) and (Wind Speed >= 1.5) => Will_Be_Extreme=Yes (122.0/56.0)
  => Will_Be_Extreme=No (11567.0/1416.0)

Number of Rules : 2
```

(Figure 3) The rules generated by JRIP

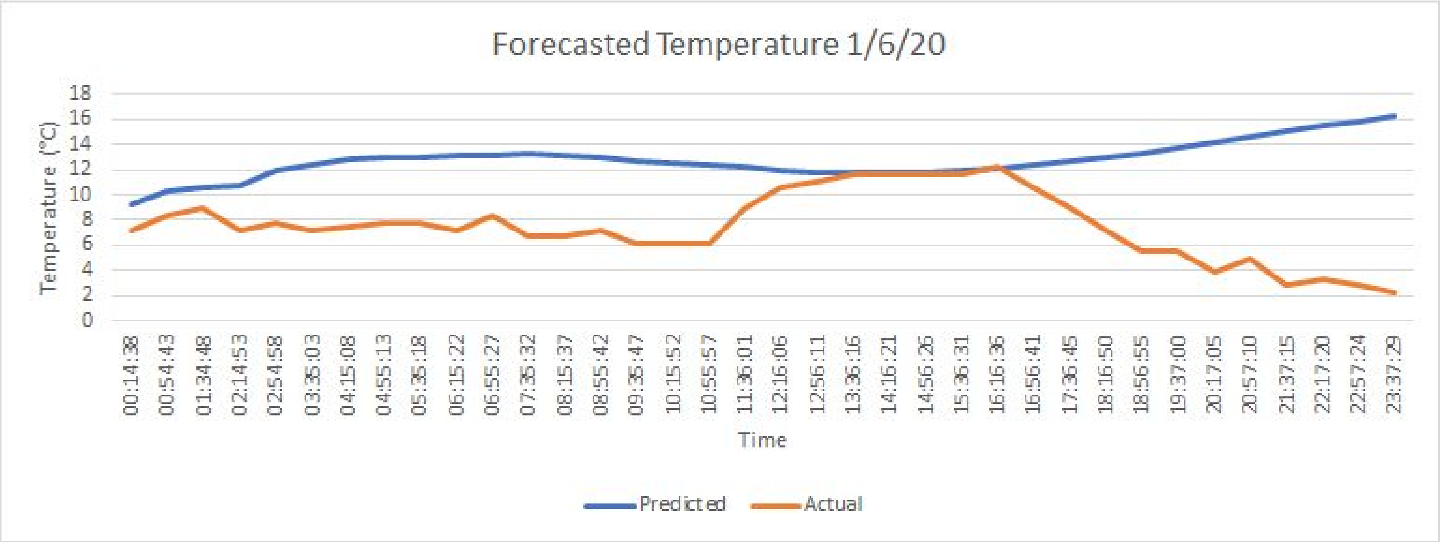
From that point forward, we started measuring the success of our models based on a ratio of false negatives (that being, the number of instances of severe weather that the model did not predict) against accurate positives. With that in mind, our most successful model was the Bayes network classifier, which correctly predicted 14% of all tested cases as being precursors to severe weather. The neural network, on the other hand, also chose to predict the majority class, so it was useless by our updated metric.



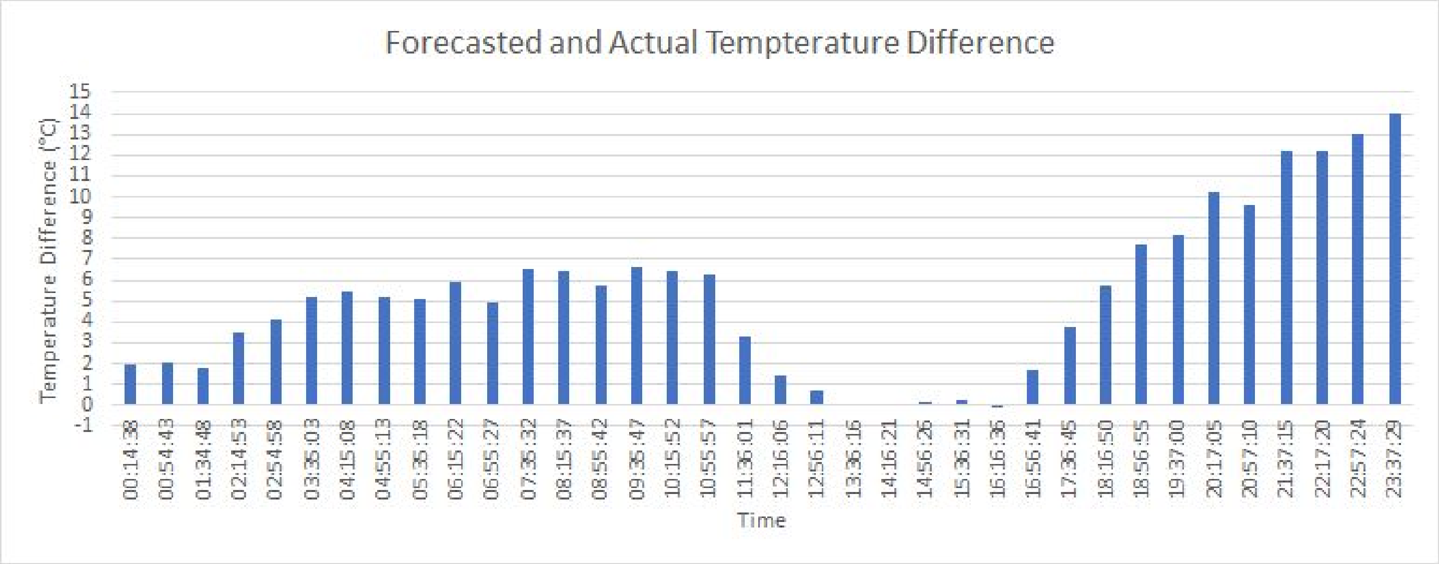
```
=== Confusion Matrix ===

    a    b   <-- classified as
 9418  789 |    a = No
 1272  210 |    b = Yes
```

(Figure 4) The confusion matrix of the Bayesnet classifier, as presented by WEKA.
The bottom two values are what have been chosen to be relevant to our project.

After running the time series analysis on 5 days' worth of data, WEKA was able to forecast the temperatures for the following day. *Figure 4* visualizes the forecasted temperature compared to the actual temperature from the dataset. It is clear that the actual temperature has a distinct difference in between night and day times. Comparatively the curve of the predicted temperatures is much more subtle, and roughly inverse in slope to the curve of the actual temperatures. The differences range from as little as 0.0751°C to as great as 14.0302℃, the closest being in the afternoon and greatest differences at night. The full range of differences in temperature can be seen in *Figure 6*.



(Figure 5)



(Figure 6)

Discussion

In this paper, we've introduced several deep learning techniques and methods we've used to explore and attempt to predict the weather patterns and storm tendencies in Baltimore, Maryland. Our findings show that the 0R and Naive Bayes Multinomial Text techniques gave us the highest accuracy percentages for prediction, with an accuracy of 87.3214%. While accuracy is a key indicator of the model's effectiveness, it should not be taken at face value and it was determined that the limitations of the data are likely the main cause for this result.

Implications of the Data

As shown in the results above, there seemed to be a strong correlation between the wind and temperature in determining weather patterns throughout distinctive time intervals. Our results do not seem to support nor refute existing studies, but are likely in the right direction of successful studies that show the correlation of weather patterns. For example in the research paper, Temperature Forecasting Based on Neural Network Approach, researchers Mohsen Hayati & Zahra Mohebi utilize Artificial Neural Networks (ANN) for the one-day ahead prediction of temperature, while one of our methods used time series analysis. Despite using different and likely more thoroughly complex deep learning methods, the researchers' mean squared error (MSE) results were similar to ours (between 0 and 2 MSE) which attests to the fact that despite limitations in our data, we can safely say that our results are a good starting point to expand on other successful applications of weather forecasting.

Limitations in Findings

Factors that may have limited the results of the experiment were the limitations of the dataset itself and our means of classifying the data.

The limitations of our dataset was the biggest factor. Despite trying to narrow down the attributes of the data to temperature and wind, the most focal aspects in determining weather patterns, the process of running the data in WEKA was very limiting. We determined that it would not be beneficial to lower the threshold of the data as it would not serve much purpose in practical application, and this resulted in our dataset having limited instances of extreme weather to base off.

Another limitation was in how we needed to classify the data. To achieve an 87.3214% accuracy using 0R and naive Bayes, we classified most attributes as non-extreme. While we are not sure whether the graph of our results -- i.e the graph of the forecasted predicted temperature and actual temperature -- mirror the graphs or results of our predecessors in the weather pattern analysis field, what we do know for certain is that former machine learning projects have proven that diverse and reliable data is directly correlated to the number of iterations, or epochs, that your dataset runs through and the ability to process multi-dimensional data in a very large amount in order to uncover their underlying patterns. So it is likely that given the extended time, data mining resources, and flexibility to classify beyond “extreme” and “non-extreme”, our results would be more accurate in time.

Recommendations

Nevertheless, the results drawn are still important. While it did not allow us to truly make a prediction about Baltimore weather, it has helped us to understand what factors contribute to making a true prediction. Further studies should take into account the benefit of certain techniques over others, such as clustering and naive Bayes Theorem. Former proposals from other researchers seem to support using the Naive Bayes model on large data sets such as ours since it is easy to build and particularly useful for real-world application, as well as the K-medoids algorithm, which is an algorithm related to the K-means clustering algorithm that “minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster.”

Conclusion

Weather prediction has a widespread impact on all of society and affects everyone from all aspects of life. Having accurate predictions has benefits that can range from optimizing construction schedules against rain, to giving people ample time to prepare or evacuate ahead of a hurricane. The ability to potentially save lives reinforces the importance of having the most accurate predictions possible.

Overall, the results were inadequate for producing any strong correlations between the data and cases of extreme weather, while also failing to predict future temperatures after shifting focus. Despite this, our work provides a starting platform for further research and testing. While the dataset we used was very limited in extreme weather cases, there could be further tests done by incorporating other data of extreme weather to find correlations between them. Also, the work done on time series forecasting was done at the latter stages of our research. Therefore, there was limited time to explore all the available options for forecasting data.

There is much more potential for further work to be done with forecasting by refining the testing parameters to create a more accurate prediction.

References

Encyclopedia Britannica Inc. (n.d). Numerical Weather Prediction (NWP) Models. Britannica.

https://www.britannica.com/science/weather-forecasting/Numerical-weather-prediction-NWP-models

National Centers for Environmental Information. (n.d.). *Federal climate complex*. National Centers for Environmental Information. https://www.ncei.noaa.gov/data/global-hourly/doc/isd-format-document.pdf

National Centers for Environmental Information. (n.d.). *Index of /data/global-hourly/access/2019*. National Centers for Environmental Information. https://www.ncei.noaa.gov/data/global-hourly/access/2019/

National Oceanic and Atmospheric Administration. (n.d.). *About our agency*. National Oceanic and Atmospheric Administration. https://www.noaa.gov/about-our-agency

National Weather Service. (n.d.). *Watch/warning/advisory definitions*. National Weather Service. https://www.weather.gov/lwx/WarningsDefined

Biradar, P., Ansari, S., Paradkar, Y., & Lohiya, S. (2017). Weather Prediction Using Data Mining. International Journal of Engineering Development and Research, 5(2), 4.