Weather prediction using Machine Learning algorithms

Nuzhat Rahman, 21301538, Mostakim Mahmud Mugdhha, 24141249.
Department of Computer SCience and Engineering, BRAC University.

***Abstract*—Weather forecasting has always been a dominant and well-demanded topic. Our day-to-day life is quite dependant on the weather. Although there are many ways to track the daily state of weather but predicting it accurately is quite difficult. Accurate daily predictions can ease one's everyday life and help people prepare better for the day. In this project, we have used machine learning to predict the weather based on the previous day. The processes includes some basic algorithms with the help of Python libraries.**

***Index Terms*—keywords, temperature, SVR, Gradient Boosting, Random Forest, etc.**

I. Introduction

WEATHER forecasting has always been the subject of interest for people of statistics and data science. Wind pattern, temperature, humidity, etc. are some of the many features based on which the weather can be predicted and analyzed. Because of the availability of large-scale datasets, researchers can meet their thirst of quest and end up with some satisfying and useful results.

Data scientists and people related to statistics have tried to utilize several techniques and methods to analyze the datasets and come with a more satisfactory result which will help them to predict the weather or the air pattern. They have collected data and the research on air pattern or weather prediction was never ending. So, there are vast amount of dataset to work in this sector. All that leaves are the models or algorithms which can be picked or trained to achieve the better prediction value. Although weather changes very rapidly, the basic prediction predictions can be done using the most recent state of the atmosphere. These predictions can heavily affect many fields, including farming and even the textile and fashion industry. Sales of many products may also be significantly effected by the climate.

II. RELATED RESEARCHES

For the project purpose, we have gone through some of the fine works by the researchers on the weather. Inspection on some research papers were conducted and we found out some excellent work on the matter. Mihir Bhawsar, Vandan Tewari, Preeti Khare [1] tried to create a survey on weather forecasting system using Machine Learning and Deep Learning algorithms. They collected data from the meteorologist's center and eventually removed the missing and damaged data. The data file was saved as the CSV format. For proper analyzation , they used Machine Learning algorithms such as Backpropagation and KNN. They also went for some clustering algorithms such as K-mean and K-media. Structural structures such as trees representing decision sets were also been used. These decisions generate data classification rules. Additionally, some deep learning methods were used to predict the weather i.e. Convolution Network (CN), Conditional Restricted Boltzmann Machines (CRBM), Recurrent Neural Network. As limitations we can focus on the fact that the dataset use had incomplete data. Also, as limitations we can focus on the fact that the dataset use had incomplete data.

Bogdan Bochenek and Zbigniew Ustrnul [2] wokred on title paper ”Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives”. The dataset used for the study consists of 500 scientific articles published since 2018, obtained from Google Scholar, with half related to numerical weather prediction and machine learning, and the other half related to climate and machine learning. The models discussed in the paper include Artificial Neural Networks (ANN), Deep Learning (DL), Random Forest (RF), XGBoost (XGB), K-means Clustering, and Principal Component Analysis (PCA) as methods commonly used by atmospheric scientists for data analysis. The study focused on post-processing NWP data for renewable energy projections by analyzing research publications pertaining to machine learning techniques and NWP models. The most often used terms were ”Ensemble Forecasting” and ”Wind Forecasting,” with ”Data Assimilation” and ”Extreme Events” coming in first. Well-known articles on climate change and weather forecasting were also examined in the study. With nearly 140 counts, ”Climate Change” was the most frequently occurring phrase in the second group. ”Climate Impact” and ”Global Climate Models” were the most frequently used terms, but ”Coupled Models,” ”Convection,” and ”Calibration” were less frequently used. For limitations, earch engine used may favor publications with specific phrases in titles or abstracts, potentially missing relevant papers. Also, since the articles were collected manually, their overall number was only 500, far less than the requirement of several thousand, as suggested in other text mining publications related to searching for patterns in scientific articles.

Wei Fang, Qiongying Xue, Liang Shen and Victor S. Sheng published papers on Extreme Weather Prediction using Deep Learning methods. For datasets, they used simulation predictions, database framework, deep learning pattern recognition techniques using Capsule Neural Networks (CapsNets). For better results they used Convolutional Neural Network and Capsule Neural Network models. The outcome is effective using deep learning technology. As a result, weather forecasts in the future may be more accurate. Collaboration between deep learning and meteorology can be extended.

Dae-Jun Kim, Jin-Hee Kim , Eun-Jeong Yun, Dae Gyoon Kang and Eunhye Ban discussed a study funded by the Rural Development Administration in the Republic of Korea about weather risk prediction techniques in agriculture. The study aims to provide early warning services for weather risk management in the agricultural sector. Data from the study can be accessed through the "early warning service for weather risk management in the agricultural sector" website. The technology of electronic climate mapping using the FS-GeST geospatial correction system is covered. In addition, the study makes use of weather radar data and spatial statistical methods based on the mountain precipitation model (PRISM) to enhance the resolution of weather data, which has been downscaled from 5 km to a more accurate 270 m grid. At the farm or orchard level, these techniques aid in the forecasting of regional weather patterns and their effects on crops. Regarding the study, the authors declare that they have no conflicts of interest.

III. Methodology

A machine learning-based approach was used to predict the target. Among the many available models, Random Forest, Support Vector Regression and Gradient Boosting were initially used to train and test the dataset. These models were specifically chosen as all three models can predict non-linear relationships and are quite robust to outliers.

Weather data from multiple regions and countries, spanning over for nearly a year was collected from the Global Weather Repository which was submitted to the free dataset platform. This dataset initially included over 40 features which was then checked and cut down to 15 features and 1 target column. This was done during the pre-processing steps which included normalization and handling of missing values.Features such as temperature, humidity, wind speed, and atmospheric pressure were selected based on domain knowledge and feature importance analysis.

The dataset was divided into training (80 percent) and test (20 percent) sets using a random split with Python libraries such as scikit-learn and pandas.Each machine learning model was initially trained using the training dataset with default hyperparameters and further tuned by checking their cluster diagram. For Random Forest, we experimented with varying tree depths and a number of estimators.

Model performance was checked by calculating the value for mean-squared error. Random forest was later chosen as the final model as it showed the best results in both cluster diagrams and model performance. Random Forest showed the best performance due to its ability to handle non-linear relationships and interactions between features. However, it should be mentioned that Gradient boosting, while slightly less accurate, showed promising results for capturing complex patterns. The cluster diagrams provided below show the performance of models used.

Random Forest: While using this method, we came up with the error value near 151 but the graph showed us the fittest diagram. On the other hand, The minimum error can be detected as near 77.87 as we add some more parameters such as, n-estimators= 3000, random-state=150, max-depth= 500, bootstrap= True, min-samples-split = 2, max-features =30

IV. Results

As Random Forest is the only model that gave us the proper illustration, it can be declared that this is the optimal model which can predict the forecast best.

figure. For Random Forest, "Figure 1".

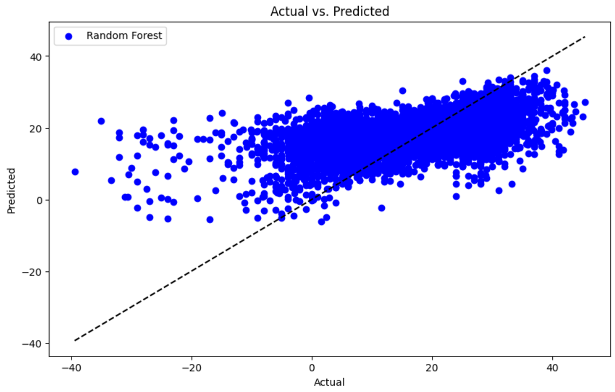

Fig. 1. The Graph Plot of Random Forest Classifier

Another better-plotted Graph can be given below. Here the predicted error is 151.36757978193418. But the plotting is the best fitted according to the graph below.

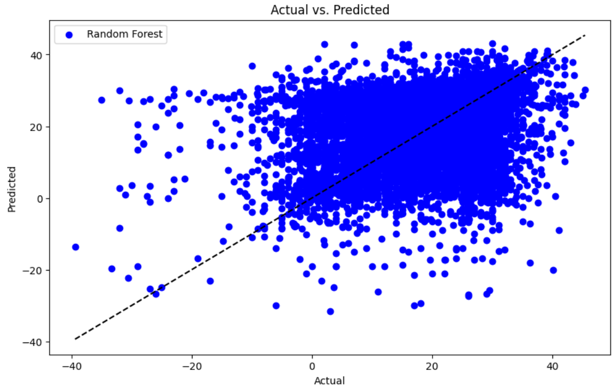

Fig. 2. The Graph Plot of Random Forest Classifier

V. Conclusion

Using machine learning for weather forecasting has made us discover quite a few key findings and showed us the importance of machine learning in this field of science.

Our results demonstrate the effectiveness of machine learning models in predicting temperature with high accuracy. Among the models used, Random Forest emerged as the most promising, exhibiting better performance in capturing non-linear relationships and interactions between weather variables. The predictive capabilities of these models have significant implications for various sectors, including agriculture, transportation and urban planning.

While the study did produce results, some limitations must be addressed. The availability and quality of data pose challenges for model training and generalization. Moreover, drastic changes in the weather conditions from region to region also make it difficult for the models to find a pattern.

Looking ahead, there are several avenues for future research that merit exploration. Incorporating additional meteorological features could enhance the predictive power of the models. Moreover, investigating ensemble techniques that combine multiple machine-learning algorithms may offer further improvements in forecast accuracy and robustness.

In conclusion, our study underscores the importance and potential of using machine learning in weather forecasting. By investing in better data collection methods and refining models to be used in the datasets, we can continue to improve our understanding of weather dynamics and produce more accurate results. This research attempts to contribute to the ongoing efforts to harness technology for mitigating the impacts of weather-related events and enhancing resilience in the face of changing weather conditions.

VI. References

Acknowledgment

The authors would like to thank...

[1] Mihir Bhawsar, Vandan Tewari, Preeti Khar ”A Survey of Weather Forecasting based on Machine Learning andDeep Learning Techniques”

[2]Bogdan Bochenek and Zbigniew Ustrnu, ”Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives”

Example of scientific journal paper:

[3] Wei Fang, Qiongying Xue, Liang Shen and Victor S, “Survey on the Application of Deep Learning in Extreme Weather Prediction”

Example of conference paper proceedings:

[4] Dae-Jun Kim, Jin-Hee Kim , Eun-Jeong Yun, Dae Gyoon Kang and Eunhye Ban , ”Farmstead-Specific Weather Risk Prediction Technique Based on High-Resolution Weather Grid Distribution”