IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WEATHER FORECASTING USING MACHINE LEARNING

Aniket Gupta, Aryan Gonsalves, Auxilia Arockiasamay, Imran Ali Mirza
Student, Student, Student, Professor
Dept of Computer Engineering ,
Don Bosco Institute Of Technology, Mumbai, India

Abstract: A lot of businesses, especially those in the agriculture sector, depend a lot on certain weather patterns to run their operations. But the effects of climate change make earlier climate models outdated, so weather forecasts must constantly be improved. The ramifications of imprecise forecasts surpass their impact on enterprises; they also have an effect on individuals' livelihoods and the country's economy. By improving the forecasting of the weather, this project seeks to address these problems, with an emphasis on delivering trustworthy forecasts for remote locations. The strategy makes use of machine learning and data analysis methods, like using random forest classification to forecast weather.

Index Terms **- Machine Learning, Random Forest, Data Analysis, Climate Change, and Weather Forecasting**

I. INTRODUCTION

Weather forecasting is a significant machine learning application that makes predictions about the upcoming weather based on historical weather data. Accurate weather pattern forecasting for a specific location is now feasible thanks to advancements in machine learning algorithms and the growing amount of data available. Machine learning-based weather forecasting applications can deliver precise and timely weather conditions information, which is crucial for a range of activities like transportation, tourism, and agriculture.

The science and technology of meteorology is used to forecast local and future weather patterns. Informal weather forecasting dates back thousands of years, while successful weather prediction dates back at least to the nineteenth century. The practice of anticipating how the weather will change over the next several hours by obtaining reliable information about the current weather and applying a working knowledge of atmospheric dynamics is known as weather forecasting.

Weather forecasts have a wide range of end users. Because they are intended to save lives and property, weather warnings are significant forecasts. Weather and temperatures forecasts are crucial for agriculture, and consequently, for commodities dealers in stock exchanges. Utility providers frequently utilize temperature forecasts to project demand for the upcoming days. People regularly utilize weather forecasts to plan their outfits for the upcoming day. Forecasts can be used to schedule activities around these events, prepare for them, and withstand them, as torrential rains have recently severely limited outdoor activities in Uyo, Nigeria, for example.

II. DATA COLLECTION AND DATA PREPROCESSING

A. Data Collection

We focused on OpenWeatherMap as our main data source for our weather forecasting project and used its API to gather a large dataset over a period of at least ten years. This dataset includes important meteorological variables like temperature, humidity, wind speed, and type of precipitation, among others. We carried out thorough checks to ensure consistency over time and cross-referenced the data against other reliable sources in order to ensure its reliability. Maintaining the quality of the data required prompt resolution of any discrepancies. The task of managing a large dataset was accomplished with well-considered methods. We put strategies into place to retrieve data in digestible chunks and built a strong storage system that could handle the volume. Because of the size of the dataset, regular backups were put in place as a crucial precaution against data loss.

During the whole data collection process, documentation was essential. Extensive logs were kept, documenting minute details of the trip, such as sources, methods of retrieval, and difficulties encountered. The aforementioned documentation is an invaluable asset that will guarantee transparency and traceability in the later phases of our weather forecasting endeavor. The resultant dataset, which was carefully curated and meticulously examined, provides the foundation for sensible and accurate weather forecasts.

B. Data Quality Check and Preprocessing

The first and most important step in our project was to confirm the accuracy of our weather data. We carefully went over the dataset, searching for any odd values, outliers, or missing data that would affect our projections. Our dataset was complete and prepared for analysis by using an interpolation technique to fill in the gaps. To ensure that our data remains accurate, we took great care to identify and eliminate outliers—unusual data points that have the potential to cause problems. In order to ensure a consistent scale for simpler comparisons, we also normalized our numbers to make sure they all worked well together.

We also applied a special technique to categorical data, such as various weather conditions: we encoded them so that our forecasting model could interpret the words as numbers. Our weather forecasting efforts are built on a solid and dependable foundation thanks to this extensive preprocessing and quality check of the data.

We also took into account time and space factors as we dug deeper into the preprocessing stage. To provide our model with a better understanding of temporal patterns, we extracted important time-related information, such as the day of the week and time of day. We also took geography into account to understand regional weather nuances. This thorough approach sets the stage for accurate and insightful predictions in the later stages of our research by preparing our data to capture the complexities of weather patterns.

C. Feature Engineering and Splitting

We wanted to give our forecasting model a deeper comprehension of the data during the Feature Engineering and Reduction phase of our weather forecasting project. Using the existing dataset, feature engineers created new elements like lag features to capture historical patterns and crucial temporal features like the day of the week. These improvements gave our model useful context, enabling it to produce more accurate predictions. Conversely, feature reduction aimed to streamline the dataset while preserving important details. Principal Component Analysis (PCA) techniques were used to reduce redundant data and enhance computational efficiency by concentrating on the most important features. This simplified method guarantees that our forecasting model stays flexible, effective, and primed to provide precise and accurate weather forecasts.

III. DATA ANALYSIS

For our weather forecasting project, we examined the performance of two models—Logistic Regression and the Random Forest algorithm—during the data analysis phase. Both models were assessed for accuracy, with the Logistic Regression approach scoring 0.5 and the Random Forest technique scoring 0.8. Given the significant disparity in accuracy ratings, the strategic choice was made to use the Random Forest algorithm for the remaining stages of the project. The higher performance of the algorithm drove this decision, providing more accurate forecasts and increasing the overall dependability of our weather forecasting efforts.

The choice to prioritize the Random Forest algorithm reflects the importance of precision in producing reliable weather forecasts. Selecting a model with a higher accuracy score puts us in a better position to manage the complexity of meteorological data and better capture complicated patterns. This calculated action demonstrates our dedication to using cutting-edge modeling methods to produce more accurate and successful weather forecasting results.

IV. SYSTEM ARCHITECTURE

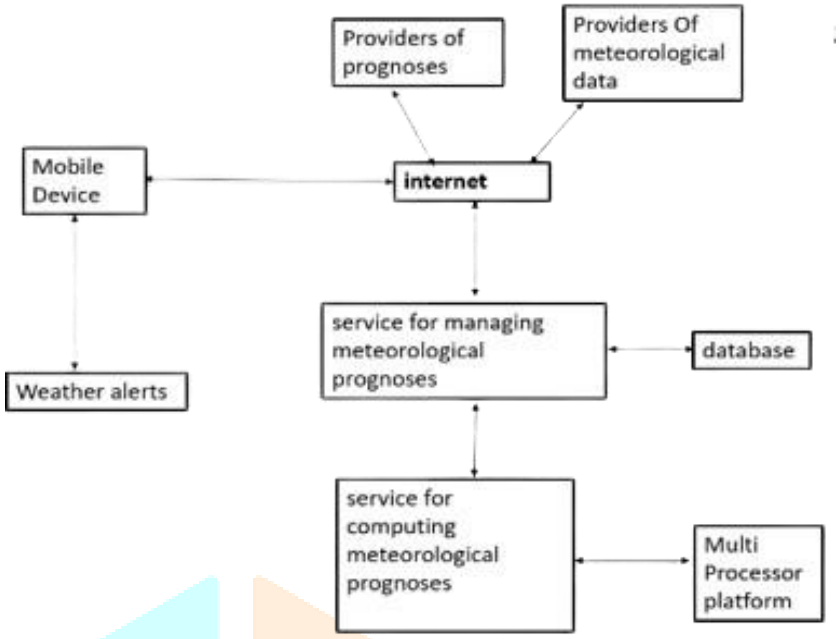

Fig. 1. Following diagram represents the working of our prediction model

V. METHODOLOGY

Predicting future atmospheric conditions using data from the past and present is a complicated process known as weather forecasting. Weather forecasting techniques vary, each with advantages and disadvantages of their own. We will provide a summary of the various approaches to weather predicting in this article, such as nowcasting, climatology forecasting, sky condition forecasting, persistence forecasting, and using forecasting models.

Since predicting future atmospheric conditions using past and present data is a complicated technique, and each weather forecasting technique has its own benefits and drawbacks, there are numerous variations that are used. An overview of various weather forecasting techniques is given in this article, including the usage of barometers, existing forecasts, forecasting models, sustainability forecasting, climate forecasting, sky conditions, and sky conditions.

A. Resilience Forecasting

The simplest forecasting technique, known as a resilience forecast, makes the assumption that the existing weather patterns will persist into the future. In order to forecast steady state weather, like summer in the tropics, it depends on the conditions of the moment. The presence of stationary weather regimes is crucial for this forecasting technique to function. For both long-term and short-term forecasting, this can be helpful.

Local forecasters use persistence forecasts as a way to predict whether the current weather will persist in their area and if it will return to its current state in the future. Changes in the production of new weather phenomena or in the strength of existing weather phenomena are not taken into consideration by persistence forecasting. For instance, a persistence forecast would indicate rain for tonight if it is raining currently. Persistence projections fall down within twelve hours to a day at most because to these constraints and the extreme speed at which weather patterns change in most geographical areas. The easiest method for creating a forecast is the persistence method.

B. Climatology Forecasting

While persistence forecasting is most accurate during short periods of time (before change factors work), the best approach for determining the weather over a longer period of time at a specific location is to use the average value of past measurements taken there at that time of day and year.

The premise behind climatology forecasting is the observation that local weather conditions on a given day tend to remain relatively constant from year to year. By averaging the climate statistics gathered over time, climatology forecasting uses weather data that has been collected from a specific location over a number of years to determine the current weather at that location. For instance, we would average all of the meteorological

data that has been recorded in Mumbai on January 1st over the years in order to forecast the weather there on that day.

Due to this, one can usually estimate the quality of the weather for a given day or month by looking at the long-term average weather for that day or month. ”Cold in December and warm in July (popular July school holidays)” is the most accurate climate prediction for Nigeria. These forecasts can be made without the need for meteorological expertise. Climate statistics are still used as a ”reality check” in today’s numerical forecasting techniques. They help prevent climatologically extreme computer models and decrease the effect of outliers on the model.

C. Viewing the Sky

One of the most essential meteorological criteria available for mountain weather forecasting is the utilization of sky conditions in conjunction with barometric pressure trends. When the clouds thicken or break up into higher clouds, it signifies that rain is on the way. Clear conditions preceding rainy weather where wind or clouds prevent fog from forming are referred to as morning fog. A thunderstorm line approaching may signify the onset of a cold front. A clear sky signals that good weather is on the way. The employment of sky coverings for weather forecasting has resulted in a wide range of weather knowledge over the years.

D. Use of the barometer

Since the late 19th century, forecasting has made use of barometric pressure and barometric trends, or variations in barometric pressure over time. The higher the difference in pressure, greater the weather change, especially if it exceeds 2. Weather fluctuations are more likely at 54 mmHg. The faster the pressure drops, the nearer you are to a low-pressure system, increasing the likelihood of rain. Clear sky and other better weather tend to coincide with sharp rises in air pressure.

VI. FORECASTING PROBLEMS

Accurate prediction is a fundamental goal of all scientific endeavors. When performing an experiment in the lab, a physicist or chemist seeks to identify basic principles that can be applied to predict the results of further investigations. An experiment is based on these principles.

The majority of scientific laws are actually just extremely precise forecasts of how particular types of tests would turn out. However, because to the difficulty of weather prediction, few scientists deal with more difficult or complex prediction problems than meteorologists. First of all, because the meteorological laboratory is global in scope, measuring the state of the atmosphere at any one time is quite difficult. Furthermore, the land and ocean surfaces of the globe are unevenly mixed, and each has a unique solar reaction. Furthermore, the earth’s energy balance is influenced by a multitude of gaseous, liquid, and solid components that make up the atmosphere itself. One such element that alternates states frequently is water.

Moreover, the magnitude of the atmosphere’s circulations varies, ranging from minuscule vortices that last a few seconds to extraordinarily large ones that could endure for weeks or months. The challenge of forecasting, according to Ayado and Burt (2001), is attempting to monitor, evaluate, and anticipate the numerous interactions among the sun’s energy supply, the earth’s physical features, and the characteristics and movements of the atmosphere. This is the reason why today’s weather forecasts are still inaccurate. Forecasts still prove to be inaccurate, as noted by Ackerman and Knox (2003), who list the following constraints that are directly related to the numerical forecast models used today:

I. Imperfect data: Today’s numerical models continue to rely heavily on radiosonde observations. However, the number of radiosonde sites in countries has decreased in recent decades. The industrialized world now spends more money on deploying weather satellites than it does on boring weather balloons. On average, satellite data is global in nature, but data digestion researchers are still striving to figure out how to effectively digest these data by the model. Important meteorological objects, particularly above the ocean, are still undetected. The model’s results are only as good as the data under beginning conditions.

II. Incorrect ”Vision” and ”Fiction”: The forecasts of today also incorporate the unavoidable compromise between forecast length and horizontal resolution. This is due to the fact that better resolution implies more points need to be calculated. It takes a lot of computer time to do this. Millions or maybe billions of calculations would be needed to make predictions about the far future.

When the tremendous resolution is paired with the distant prediction of this problem, even the fastest modern supercomputers will be overwhelmed. One week will not suffice to receive a forecast. Future enhancements to the calculation will help to accelerate the process. However, some models continue to be unable to capture or ”see” small-scale phenomena such as clouds, showers, and snowflakes. The computer code provides crude approximations of the invisible to compensate for these blurry ”visions” of the pattern. They

are known as parameterizations (). Despite the fact that a great deal of science has been put in them, these approximations do not accurately capture the phenomenon's complicated reality, because the simplest things are frequently the most hardest to comprehend. So, it is not an insult to meteorologists' competence to say that parametrization is a "fiction" of a real phenomenon.

III. Chaos: It's astonishing to learn that no better predicting result would be obtained even if a supercomputer capable of performing quadrillions of computations per second were to be created. There are limitations to brute force numerical weather forecasting with incredibly fine precision. These restrictions are caused by an odd characteristic of intricate, dynamic systems like the atmosphere. It is referred to as "Sensitive dependence on initial conditions" and is a feature of chaos theory. Instead than implying that everything is in disarray, chaos in the atmosphere just indicates that the initial conditions, whether they are slightly different in real life or in a computer model, may read quite differently.

The chaos indicates that there is a decreasing daily resemblance between the model's prediction and reality since we are never certain of the exact atmospheric circumstances at any particular time under regular circumstances, like a tropical summer. The forecasting technique heavily relies on the presence of weather regimes that are stationary. For both short- and long-term forecasting, it can be helpful. Local meteorologists utilize sustainability predictions to predict things like thunderstorms approaching a region.

Persistence forecasts do not account for variations in the intensity or kind of weather phenomena that may occur, i.e. they do not predict the emergence of new sorts of weather conditions. Because of these constraints, as well as the sheer speed with which weather conditions change in most geographical regions, persistent forecasting fails to perform effectively after twelve hours, or at most a day, as previously stated.

VII. IMPLEMENTATION

In this project, the code was written in Python, leveraging various libraries such as NumPy, Pandas, Seaborn, and Matplotlib. The initial step involved data extraction, followed by a thorough dataset training process aimed at rectifying errors and handling null values. This process essentially constituted a comprehensive data cleaning operation, ensuring the dataset's quality and reliability. Subsequently, Seaborn was employed to visually represent the data through diverse plots, including bar plots, line plots, and pair plots, providing a meaningful exploration of the dataset's characteristics as shown in Figures 2 ,3 and 4.

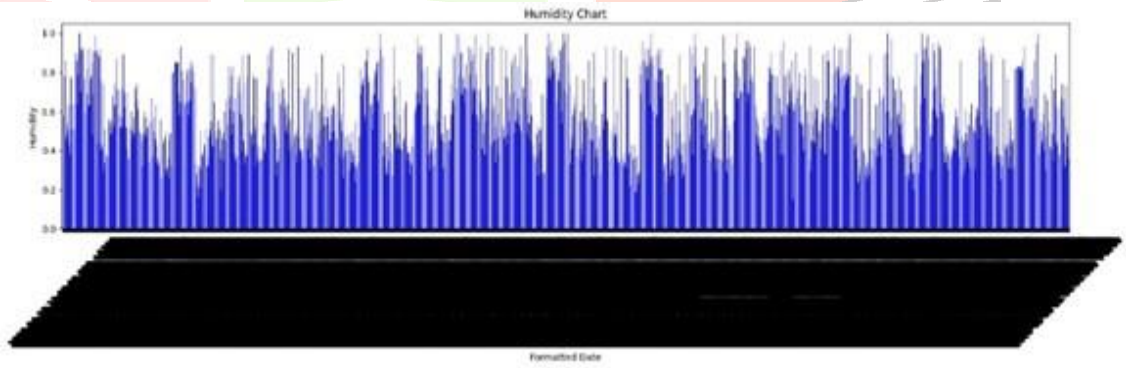

Fig. 2. Graphical representation of humidity trends over observation period

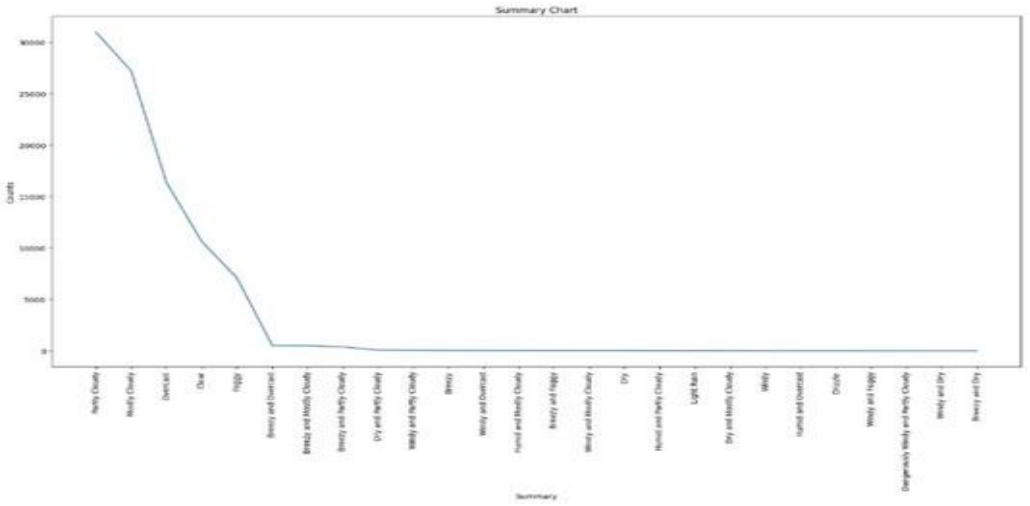

Fig. 4. Summary representing the count of different weather conditions that occurred during the observation period

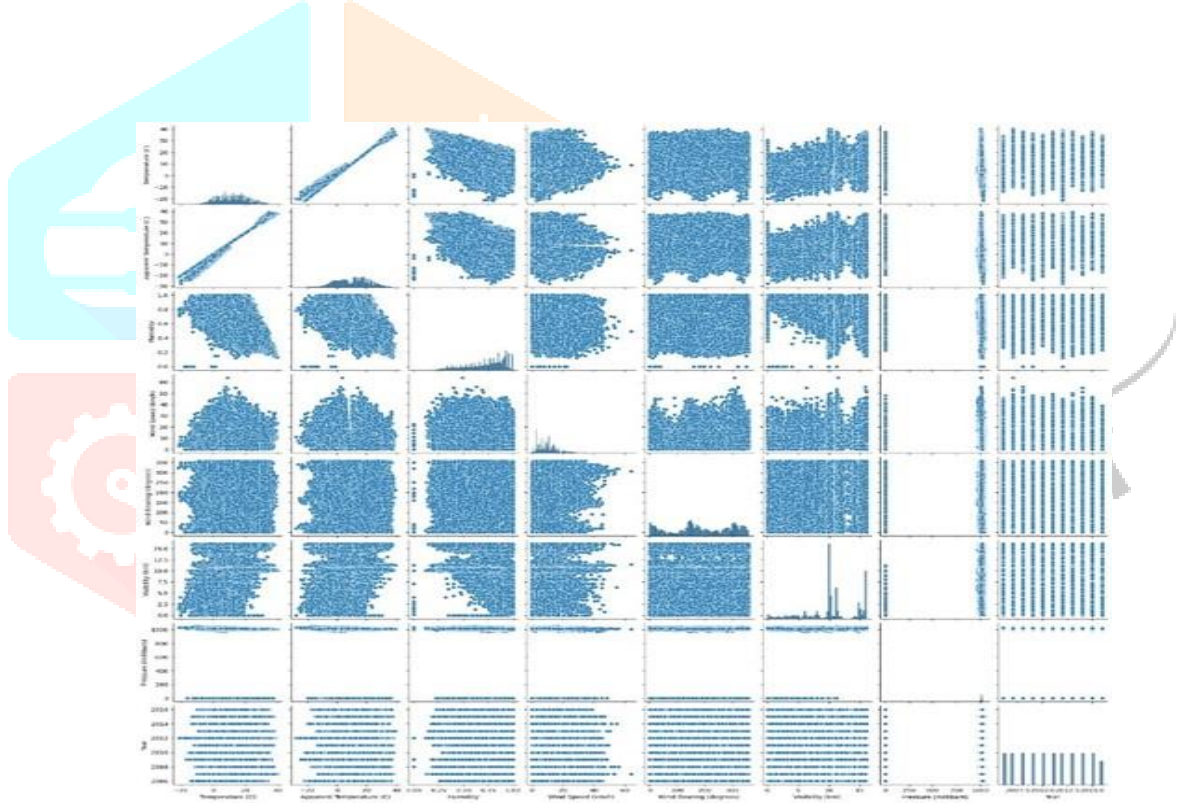

Fig. 5. Graph representing the relationship between all meteorological variables used by our prediction model

Moving forward, the trained dataset underwent testing, and the evaluation was conducted using two distinct machine learning models: logistic regression and a random forest classifier. The comparison of the accuracy scores revealed that the Random Forest Classifier outperformed the logistic regression model. Consequently, the Random Forest Classifier was selected for the final evaluation of the data.

Utilizing the chosen machine learning models, the dataset was evaluated for predictive purposes. The predicted outputs were then stored and organized in a CSV file as shown in Figure 5, providing a tangible and accessible record of the model's predictions. This comprehensive approach—from data cleaning and visualization to model selection and final evaluation—demonstrates a systematic and rigorous methodology

employed in the data analysis and machine learning processes, ultimately yielding valuable insights and predictive outcomes.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Formatted	Summary	Precip Typ	Temperatu	Apparent T	Humidity	Wind Spee	Wind Bear	Visibility (k	Cloud Cov	Pressure (r	Daily Summary			
2	2006-04-0	Partly Clou	rain	9.472222	7.388889	0.89	14.1197	251	15.8263	0	1015.13	Partly cloudy throughout the day.			
3	2006-04-0	Partly Clou	rain	9.355556	7.227778	0.86	14.2646	259	15.8263	0	1015.63	Partly cloudy throughout the day.			
4	2006-04-0	Mostly Clo	rain	9.377778	9.377778	0.89	3.9284	204	14.9569	0	1015.94	Partly cloudy throughout the day.			
5	2006-04-0	Partly Clou	rain	8.288889	5.944444	0.83	14.1036	269	15.8263	0	1016.41	Partly cloudy throughout the day.			
6	2006-04-0	Mostly Clo	rain	8.755556	6.977778	0.83	11.0446	259	15.8263	0	1016.51	Partly cloudy throughout the day.			
7	2006-04-0	Partly Clou	rain	9.222222	7.111111	0.85	13.9587	258	14.9569	0	1016.66	Partly cloudy throughout the day.			
8	2006-04-0	Partly Clou	rain	7.733333	5.522222	0.95	12.3648	259	9.982	0	1016.72	Partly cloudy throughout the day.			
9	2006-04-0	Partly Clou	rain	8.772222	6.527778	0.89	14.1519	260	9.982	0	1016.84	Partly cloudy throughout the day.			
10	2006-04-0	Partly Clou	rain	10.82222	10.82222	0.82	11.3183	259	9.982	0	1017.37	Partly cloudy throughout the day.			
11	2006-04-0	Partly Clou	rain	13.77222	13.77222	0.72	12.5258	279	9.982	0	1017.22	Partly cloudy throughout the day.			
12	2006-04-0	Partly Clou	rain	16.01667	16.01667	0.67	17.5651	290	11.2056	0	1017.42	Partly cloudy throughout the day.			
13	2006-04-0	Partly Clou	rain	17.14444	17.14444	0.54	19.7869	316	11.4471	0	1017.74	Partly cloudy throughout the day.			
14	2006-04-0	Partly Clou	rain	17.8	17.8	0.55	21.9443	281	11.27	0	1017.59	Partly cloudy throughout the day.			
15	2006-04-0	Partly Clou	rain	17.33333	17.33333	0.51	20.6885	289	11.27	0	1017.48	Partly cloudy throughout the day.			
16	2006-04-0	Partly Clou	rain	18.87778	18.87778	0.47	15.3755	262	11.4471	0	1017.17	Partly cloudy throughout the day.			
17	2006-04-0	Partly Clou	rain	18.91111	18.91111	0.46	10.4006	288	11.27	0	1016.47	Partly cloudy throughout the day.			
18	2006-04-0	Partly Clou	rain	15.38889	15.38889	0.6	14.4095	251	11.27	0	1016.15	Partly cloudy throughout the day.			
19	2006-04-0	Mostly Clo	rain	15.55	15.55	0.63	11.1573	230	11.4471	0	1016.17	Partly cloudy throughout the day.			
20	2006-04-0	Mostly Clo	rain	14.25556	14.25556	0.69	8.5169	163	11.2056	0	1015.82	Partly cloudy throughout the day.			
21	2006-04-0	Mostly Clo	rain	13.14444	13.14444	0.7	7.6314	139	11.2056	0	1015.83	Partly cloudy throughout the day.			
22	2006-04-0	Mostly Clo	rain	11.55	11.55	0.77	7.3899	147	11.0285	0	1015.85	Partly cloudy throughout the day.			
23	2006-04-0	Mostly Clo	rain	11.18333	11.18333	0.76	4.9266	160	9.982	0	1015.77	Partly cloudy throughout the day.			
24	2006-04-0	Partly Clou	rain	10.11667	10.11667	0.79	6.6493	163	15.8263	0	1015.4	Partly cloudy throughout the day.			
25	2006-04-0	Mostly Clo	rain	10.2	10.2	0.77	3.9284	152	14.9569	0	1015.51	Partly cloudy throughout the day.			
26	2006-04-1	Partly Clou	rain	10.42222	10.42222	0.62	16.9855	150	15.8263	0	1014.4	Mostly cloudy throughout the day.			
27	2006-04-1	Partly Clou	rain	9.911111	7.566667	0.66	17.2109	149	15.8263	0	1014.2	Mostly cloudy throughout the day.			
28	2006-04-1	Mostly Clo	rain	11.18333	11.18333	0.8	10.8192	163	14.9569	0	1008.71	Mostly cloudy throughout the day.			
29	2006-04-1	Partly Clou	rain	7.155556	5.044444	0.79	11.0768	180	15.8263	0	1014.47	Mostly cloudy throughout the day.			

Fig. 5. Training Dataset

VIII. CONCLUSION

For people, businesses, and organizations to make well informed decisions based on precise and trustworthy weather predictions, machine learning-based weather forecasting can be a useful tool. Large volumes of historical weather data can be analyzed by machine learning algorithms to find trends and forecast future weather conditions with accuracy. This ability to prepare ahead, reduce risk, and increase cost-effectiveness benefits both individuals and organizations. This system can also be used to deliver timely weather alerts, which can assist people and organizations in taking the necessary precautions to guarantee safety and reduce potential harm.

The credibility and dependability of the system can be further increased by accreditation from independent organizations, government agencies, academic institutions, or industry certifications. In general, a variety of consumers and stakeholders stand to gain significantly from the creation and usage of machine learning-based weather forecasting systems. Applications for machine learning-based weather forecasting have enormous potential to enhance human capacity for weather event prediction and preparation. But rather than taking the place of human expertise, they ought to be utilized as an additional tool. It is critical that we keep funding both human expertise and machine learning technologies to increase our comprehension of weather patterns and strengthen our defenses against the effects of catastrophic weather.

IX. ACKNOWLEDGMENT

It is our pleasure to express our deep sense of gratitude and thanks to our project guide Prof. Imran Ali Mirza, HOD Dr. Shaikh Phiroj for providing the guideline with continuous advice and feedback throughout the duration of this project and the entire third year department of computer engineering for contributing their valuable time and effort also, providing us with their valued advice and guidance for this project.

We also thank Dr. Prasanna Nambiar (principal of Don Bosco Institute of Technology, Mumbai) for providing us the opportunity to embark on this project. We would also like to place on record our sincere thanks to our seniors and all people who directly or indirectly helped us in this project work.

REFERENCES

[1] N. Hasan, M. T. Uddin, and N. K. Chowdhury, "Automated weather event analysis with machine learning," in Proc. IEEE 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2016, pp.

[2] L. L. Lai, H. Braun, Q. P. Zhang, Q. Wu, Y. N. Ma, W. C. Sun, and L. Yang, "Intelligent weather forecast," in Proc. IEEE 2004 International Conference on Machine Learning and Cybernetics, 2004, pp. 4216-4221.

[3] A. G. Salman, B. Kanigoro, and Y. Heryadi, "Weather forecasting using deep learning techniques," in Proc. IEEE 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2015, pp. 281-285.

[4] D. Ahijevych, J. O. Pinto, J. K. Williams and M. Steiner, "Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique", Weather and Forecasting, vol. 31, no. 2, pp. 581-599, 2016

[5] H. Murphy, "Probabilistic weather forecasting", Probability statistics and decision making in the atmospheric sciences, pp. 337-377, 2019

[6] 17) E. B. Abrahamsen, O. M. Brastein and B. Lie, "Machine learning in python for weather forecast based on freely available weather data", 2018.

[7] G. Salman, B. Kanigoro and Y. Heryadi, "Weather forecasting using deep learning techniques", Proc. IEEE 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 281-285, 2015.

[8] 20) S. Mathur and A. Paras, "Simple weather forecasting model using mathematical regression", Indian Res J Exten Educ: Special, vol. 1, 2012.

[9] M. Kannan, S. Prabhakaran and P. Ramachandran, "Rainfall forecasting using data mining technique", International Journal of Engineering and Technology, vol. 2, no. 6, pp. 397-401, 2010.

[10] Prathyusha, Zakiya, Savya, Tejaswi, N. Alex and S. C C, "A Method for Weather Forecasting Using Machine Learning," 2021 5th Conference on Information and Communication Technology (CICT), Kurnool, India, 2021, pp. 1-6, doi: 10.1109/CICT53865.2020.967240