A Survey on Weather Prediction

ABSTRACT

Weather prediction aims to help people and minimize weather-related damage and improve overall well-being. It helps ensure public health and safety and supports economic growth. Weather prediction plays a crucial role in various sectors. It includes agriculture, transportation, energy management, and biological field. Nowadays methods of weather forecasting have limitations in accurately capturing the tough dynamics of the atmosphere.

In recent years, the collaboration of machine learning techniques has shown promising results in improving the accuracy and solidity of weather predictions. This survey report provides an overview of the application of machine learning in weather prediction, including the techniques used, data sources put to use, Advanced mathematics applied, and real-world applications. Additionally, the report discusses the challenges faced in leveraging machine learning for weather forecasting and explores emerging trends and future directions in the field. Through case studies and a literature review, this report highlights the significant potential of machine learning in revolutionizing weather prediction, ultimately contributing to more precise and actionable forecasts for societal benefit.

Earlier the models were made up using a high number of nodes and they used to consume a very high amount of energy to predict weather conditions but our model predict it in very short periods using modern advanced mathematics and calculation. We collect data from the neighboring regions and train our model on that data keeping in mind that RMSE(Root Mean Square Error) is minimized.

In this paper, we will predict the temperature by feeding weather features along with the temperature that comes under **Supervised Learning** as the feature we want to predict is present in the dataset.

In Temperature prediction –**Linear Regression Algorithm**, **Random Forest Algorithm, Polynomial Regression Algorithm and Support Vector Machine Algorithm (SVM)** are used.

The results of these models are compared based on the Root Mean Squared Error (RMSE) between the actual and predicted values. Graphs are plotted for each model to provide a clearer depiction of accuracy.

Python, NumPy, Anaconda Navigator, Jupiter Notebook, Matplotlib, Scikit-learn, and Panda will be used in this project. After the research, we have found that the **Random Forest Regression** performs best in weather prediction with an accuracy of 86%.

Keywords: *Machine Learning,Regression, Forecasting techniques, challenges, applications, case studies, Real-world impact, Support Vector Machine, Random Forest, Supervised Learning.*

1. INTRODUCTION

Predicting the weather conditions is challenging. It has been a critical aspect of human life that influences various sectors such as transportation, agriculture, and natural disaster management. Being able to predict weather conditions of the coming days will be helpful for us to protect lives and our economic damage, reliable crop bowing, and events related to weather.

In recent years, Machine Learning models have trained through very complex mathematical equations, which results in weather prediction with good accuracy. Machine learning is a subset of Artificial intelligence. It involves the algorithm and statistics to draw the graph and analyze it.

ML reads the historical weather data and patterns by itself. This survey report will explore the application of machine learning techniques in predicting weather conditions. It will cover all the algorithms used in weather forecasting, advantages, and disadvantages. In this model, we will train the historical weather data of a region and then test it. We will find accuracy by predicting the data through each model set and the model set providing the maximum accurate data will be taken into account.

2. LITERATURE SURVEY

2.1 PREVIOUS WORK

Traditionally, weather predictions are performed with the help of large complex models of physics, which utilize different atmospheric conditions over a long period. These conditions are often unstable because of perturbations of the weather system, causing the models to provide inaccurate forecasts. In recent years, weather prediction has seen a variety of approaches. However, many of these fail to capture the complex relationships between various factors that affect weather.

2.2 METHODOLOGY

The following steps were performed to implement the Machine Learning model:

1. Setup

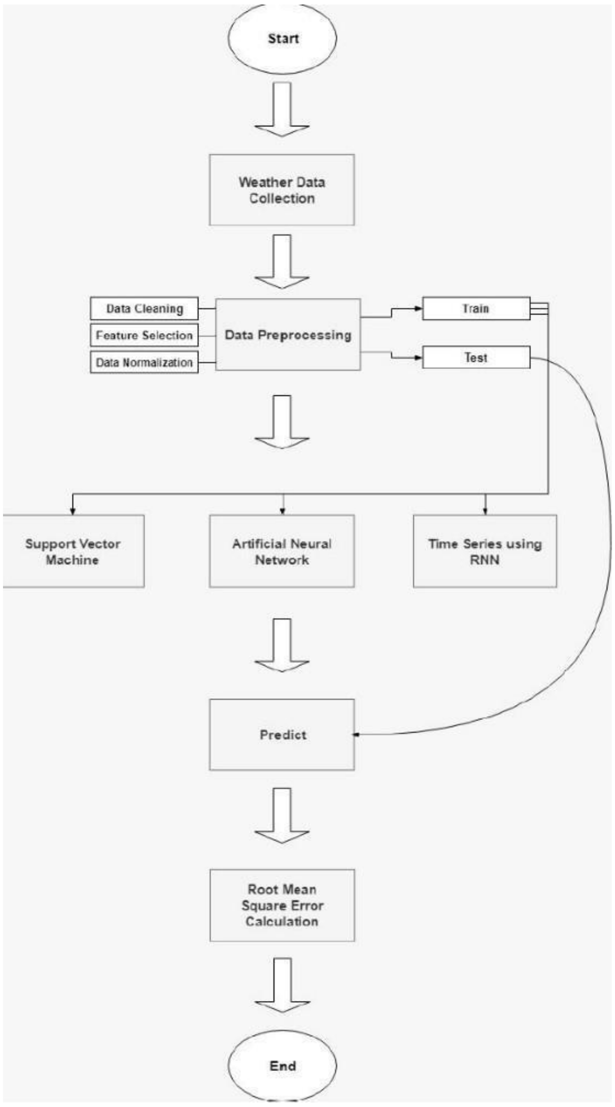
2. Data Collection

3. Data preprocessing

4. Splitting dataset into Train and Test Data

5. Model Deployment.

6. Comparison of Model.



2.2.1 SETUP

1. <u>SOFTWARE USED</u>

- **Anaconda Navigator:** Anaconda Navigator is a desktop graphical user interface (GUI) included in the Anaconda distribution, which allows users to easily manage conda packages, environments, and launch applications without the need to use command line commands.
- **Jupyter Notebook:** It is an open-source web-based interactive computing notebook environment. It is widely used in data science, machine learning, and

scientific computing for interactive data analysis, and visualization.

2. LANGUAGE USED

- **Python 3.12** : Python is an interpreted, high-level, general-purpose programming language known for its clear and readable syntax. It offers a vast library ecosystem, making it highly suitable for data mining and predictive analytics.

3. LIBRARIES USED

- **Numpy**: For numerical computations and handling arrays.
- **Pandas**: For data manipulation and analysis.
- **Matplotlib**: For data visualization.
- **Seaborn:** For statistical data visualization.
- **Scikit-learn:** For implementing machine learning algorithms and models.

2.2.2 DATA COLLECTION

To implement the Machine Learning model, we collected weather dataset of Kolkata city from **Kaggle.com.** The dataset consists of several attributes such as precipitation, wind speed, maximum and minimum temperature, humidity, wind direction, etc. The dataset has been downloaded in a CSV file. The weather data was collected in Kolkata from 2017 to 2022.

2.2.3 DATA PREPROCESSING

This involves transforming raw, unorganized, and unstructured data into a structured, usable format. The steps are as follows:

1. Data Cleaning

Data cleaning is crucial as datasets can be messy, with empty values, missing or additional values, invalid fields, etc. Before applying any model the is to be cleaned.

In this figure, some of the features like Wind Chill, Wind Speed, Heat Index, and Info have more than half of their values as NaN. Therefore, we dropped these columns.For the Sea Level Pressure feature, there was only one NaN value, so we dropped that particular row.After these adjustments, we dropped some non-significant features. Since none of the selected features have string values after feature selection, no encoding was required.

2. Feature Selection

The data we have collected has many unwanted attributes which will not be needed in our project. Hence, we will select the significant features for temperature and rainfall prediction .

3. Data Normalization

The goal of normalization is to change the values of numeric columns in the dataset to a common scale without affecting differences in the ranges of values or losing any information. Normalization helps in faster training of models. We normalized the dataset, except for the temperature, to a 0-1 range.

2.2.4 TRAINING AND TESTING OF DATA

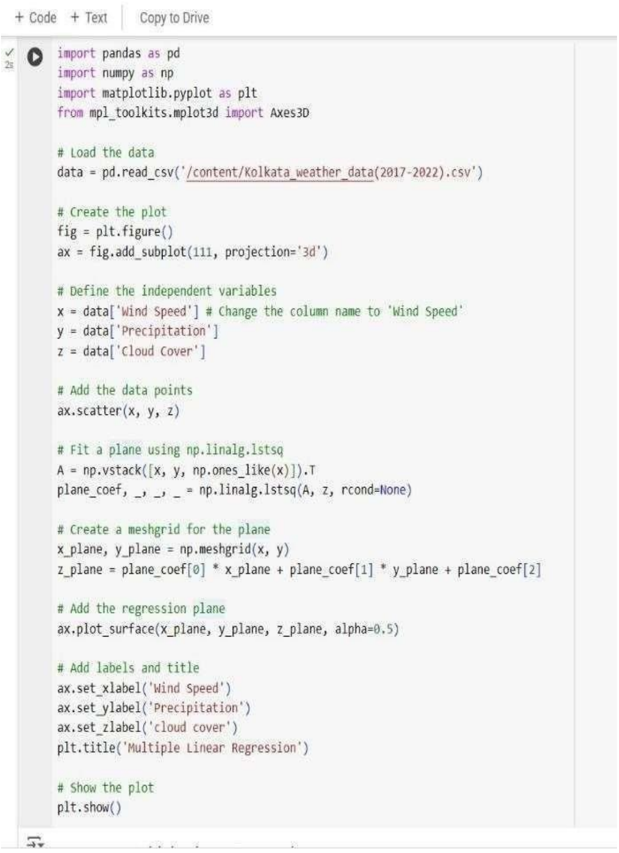
Once everything is done, we will split the data in train and test set.

● Training dataset (usually 80%) is used to train the machine learning model and to get the best-fit shape.

● Testing dataset (usually 20%) is used to test the model formed.

This process of feeding data helps in knowing the accuracy of the model. The data cleaning and all these processes are carried out before separating data for training and testing

Code to Test and Train Data



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

# Load the data
data = pd.read_csv('/content/Kolkata_weather_data(2017-2022).csv')

# Create the plot
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')

# Define the independent variables
x = data['Wind Speed'] # Change the column name to 'Wind Speed'
y = data['Precipitation']
z = data['Cloud Cover']

# Add the data points
ax.scatter(x, y, z)

# Fit a plane using np.linalg.lstsq
A = np.vstack([x, y, np.ones_like(x)]).T
plane_coef, _, _, _ = np.linalg.lstsq(A, z, rcond=None)

# Create a meshgrid for the plane
x_plane, y_plane = np.meshgrid(x, y)
z_plane = plane_coef[0] * x_plane + plane_coef[1] * y_plane + plane_coef[2]

# Add the regression plane
ax.plot_surface(x_plane, y_plane, z_plane, alpha=0.5)

# Add labels and title
ax.set_xlabel('Wind Speed')
ax.set_ylabel('Precipitation')
ax.set_zlabel('cloud cover')
plt.title('Multiple Linear Regression')

# Show the plot
plt.show()
```

2.2.5 ALGORITHMS USED

1. Linear Regression Algorithm:

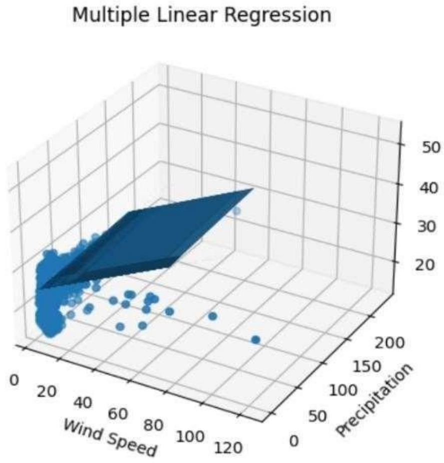
Linear Regression can be simple or multiple, Linear Algorithm Regression is a method of modeling a target value based on independent prediction. This method is mostly used for forecasting and finding out cause and effect relationship between variables.

The first method we used for temperature prediction is Multiple Linear Regression, a Supervised learning algorithm that derives a relationship between a dependent variable and one or more independent variables. The variable to be predicted is called the **dependent variable**, while the variables used for prediction are called **independent variables**. In our case, we aim to predict the apparent temperature based on some features like precipitation, wind speed, relative humidity visibility etc. Linear regression fits a straight line that minimizes the discrepancies between predicted and actual output values using a “least squares” method to discover the best-fit line.

y = mx+ C

Where ‘y’ is a dependent variable and ‘x’ is an independent variable , ‘m’ is the slope and ‘C’ is intercept

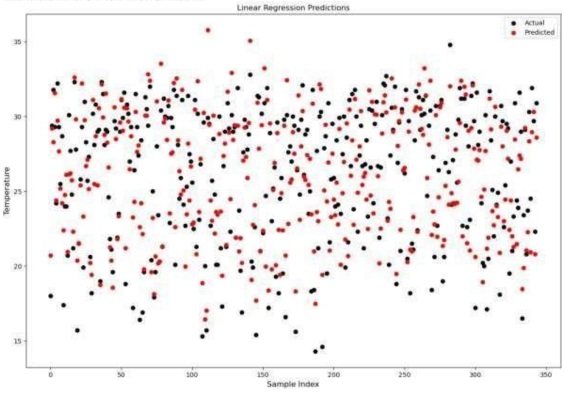
Diagram graph of Multiple linear regression



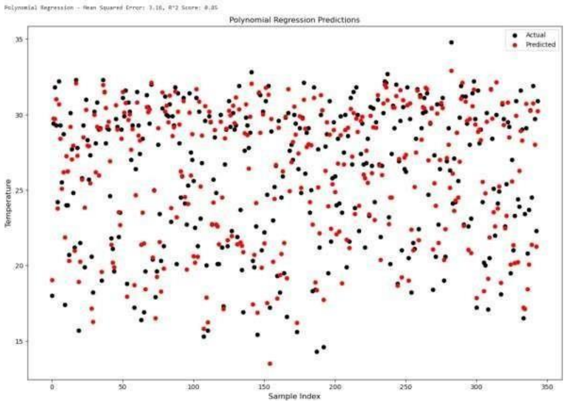
Accuracy: 3.79%

1. Linear regression Algorithm

Linear Regression - Mean Squared Error: 4.62, R^2 Score: 0.78

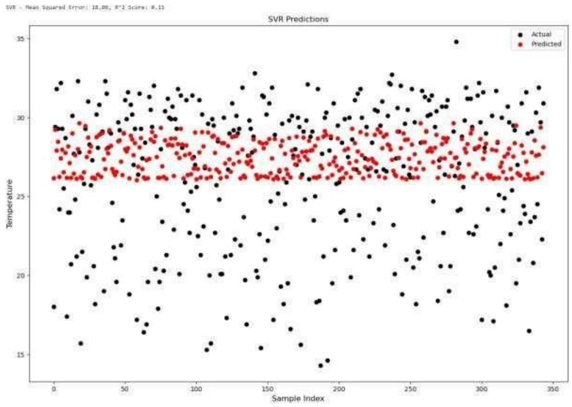


2. Polynomial Regression Algorithm



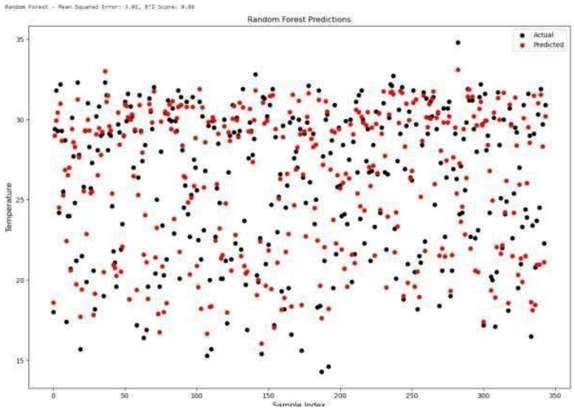
Polynomial Regression - Mean Squared Error: 3.16, R^2 Score: 0.85

3. Support Vector Regression Algorithm



SVR - Mean Squared Error: 18.06, R^2 Score: 0.15

4.Random Forest Algorithm.



Random Forest - Mean Squared Error: 3.01, R^2 Score: 0.86

3. OPTIMIZATION TECHNIQUES

In recent years, machine learning (ML) techniques have been increasingly adopted to improve the accuracy, efficiency, and scope of weather predictions. ML models offer advantages by learning patterns from historical data and can be used either independently or in conjunction with traditional physical models to enhance forecasting capabilities.

Data Preprocessing and Feature Engineering

Before applying any machine learning we have to do preprocessing of data which in can be done in following ways.

The first step in applying machine learning to weather prediction involves preprocessing large amounts of meteorological data and engineering features that can be used by ML algorithms.

1. Data Cleaning: Involves handling missing data, outliers within weather dataset.

2. Feature Selection: Choosing the most relevant meteorological parameters (temperature, humidity, wind speed, pressure and precipitation.) that affects the temperature.

3. Dimensionality Reduction: Techniques such as Principal Component Analysis (PCA) are used to reduce the dimensionality of data while retaining most of the information, which helps in managing computational costs and improving model performance.

4.Time-Series Data Handling: Since weather data is sequential, handling time-series data appropriately using techniques such as lag features, rolling averages, or differencing is crucial.

Supervised Learning Techniques

Supervised learning models are trained on historical weather data to predict temperature of the given dataset. We have used the following models.

1.Linear Regression: Despite being a simple model, linear regression can be useful for predicting certain weather parameters like temperature. However, but it don’t yield correct answers for non linear relationship

2. **Decision Trees and Random Forests:** These are non-linear models that can capture complex interactions between weather variables. Random Forests, in particular, improve prediction accuracy by averaging the results of multiple decision tree.

3. **Support Vector Machines (SVM):** SVMs are used for classification tasks in weather prediction, such as predicting whether it will rain or not. They work well for small to medium-sized datasets.

4. **Neural Networks:** These models, particularly feedforward networks, have been applied to tasks like rainfall prediction and temperature forecasting. They can capture complex non-linear relationships but require careful tuning and large datasets.

Ensemble Methods

Ensemble methods combine multiple machine learning models to create a more robust and accurate prediction system. These techniques reduce the risk of overfitting and improve generalization.

1. **Bagging:** Techniques like Bootstrap Aggregating (Bagging) are used to create multiple instances of a model using different subsets of the data and then aggregate their predictions. Random Forest is a popular example of bagging.

2. **Boosting:** Boosting algorithms like Gradient Boosting Machines (GBM) and XGBoost sequentially build models by focusing on the errors of the previous models. These are particularly effective for weather prediction tasks that require high precision.

3. **Stacking:** This technique involves training different types of models on the same dataset and then using another model to combine their predictions. This can be especially powerful when different models capture different aspects of the weather data.

Hybrid Models

- **Quantile Regression**: This technique provides a range of possible outcomes instead of a single prediction, offering a more comprehensive view of the uncertainty.

- **Bayesian Approaches:** These methods provide a probabilistic framework for predictions, allowing for the incorporation of prior knowledge and the updating of forecasts as new data becomes available.

Despite the advancements, there are still challenges in optimizing weather predictions using machine learning:

1. **Data Quality and Availability**: High-quality, high-resolution weather data is essential for training accurate models. Data sparsity in certain regions or for certain weather phenomena can limit the effectiveness of ML models

2. **Integration with Existing Systems**: Integrating ML models with existing weather forecasting systems and ensuring their scalability and reliability remains a challenge.

In summary, machine learning techniques are transforming weather prediction by enabling more accurate, efficient, and high-resolution forecasts. As these techniques continue to evolve, they will play an increasingly important role in mitigating the impacts of extreme weather events and aiding in climate resilience efforts globally.

4.DISCUSSION

the weather forecasting analysis done in our report compares various seven day period forecasting models .It mainly focusses on the main differences between a weather forecasting services and the model that has been created .It also check the performance difference between linear regression and functional regression
models. This result shows that the professional weather forecasting services are more better in finding the output than the created model throughout the week, with a larger discrepancy in the early days this can mainly occur because of high accuracy in the short term predictions than long term, whereas machine learning algorithms also deals with initial(starting) conditions better and might excel over longer periods. The linear regression model showed up better results than the functional regression model. This can be proved through the fact that the forecasts were based on the weather data for the last two days, which are too short to capture significant results in the weather trends.
After analysing the results we can say that the linear regression would be a better model than
functional regression and it was also mentioned by Occam's razor that the simpler lin- ear regression model is better than the funtional regression .If we increase the number of days in the forecast from two days to four to five days then, funtional regression should have outperformed linear regression and then perhaps there would be evident trends in the weather that functional regression could capture.

5. CONCLUSION

In this paper, we have introduced a technology that applies machine learning techniques to provide accurate weather forecasts. These machine-learning models are simpler and less resource-intensive than traditional physical models, and they can be run on any computer or mobile device. Our evaluation results indicate that these models are accurate enough to compete with traditional models. We also incorporate historical data from surrounding areas to predict weather patterns for a specific location. This approach is more effective than focusing solely on the area in question. Weather forecasts can be used to save lives, such as by making arrangements for local residents in areas where famine is predicted. AI may not be able to prevent disasters, but new scientific tools like machine learning, image recognition, and computational modeling can help us prepare for them. In the future, we plan to use low-cost IoT devices, such as temperature and humidity sensors, to collect weather data from various parts of a city. The use of multiple sensors would increase the number of local features in the training dataset, which would further enhance the performance of our prediction models.

6. REFERNCES

1. https://ieeexplore.ieee.org/document/9862337

2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3836085

3. https://www.researchgate.net/publication/362517661_WEATHER_PREDICTION_USING_MACHINE_LEARNING

4. https://janaksenevirathne.medium.com/building-a-weather-prediction-model-with-machine-learning-a-step-by-step-guide-9eaf768171be

5. https://www.kaggle.com/

6. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

7. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

8. https://www.researchgate.net/publication/372365275_Polynomial_Regression_Method_and_Support_Vector_Machine_Method_for_Predicting_Disease_Covid-19_in_Indonesia

9. https://towardsdatascience.com/machine-learning-basics-support-vector-regression-660306ac5226

10. https://www.mdpi.com/2079-9292/12/8/1789

11. https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558

12. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

13. "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

14. "Machine Learning Yearning" by Andrew Ng

15. "Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili

16. "Applied Predictive Modeling" by Max Kuhn and Kjell Johnson

17. "Introduction to Machine Learning with Python: A Guide for Data Scientists" by Andreas C. Müller and Sarah Guido

18. "Data Science for Business" by Foster Provost and Tom Fawcett

19. "Weather Prediction by Numerical Process" by Lewis Fry Richardson

20. "Numerical Weather and Climate Prediction" by Thomas Tomkins Warner

21. "Practical Time Series Analysis: Prediction with Statistics and Machine Learning" by Aileen Nielsen

22. Eberhart, R., & Kennedy, J. (1995). "A new optimizer using particle swarm theory." In Proceedings of the Sixth International Symposium on Micro Machine and Human Science (pp. 39-43). IEEE. https://ieeexplore.ieee.org/document/488968

23. Zhang, Y., & Ding, S. (2011). "A hybrid particle swarm optimization approach for the gene regulatory network reconstruction." Neurocomputing, 74(17), 3129-3136. https://www.sciencedirect.com/science/article/abs/pii/S0925231211004426

24. Kennedy, J., & Eberhart, R. C. (2001). "Swarm Intelligence." The Morgan Kaufmann Series in Artificial Intelligence. https://www.elsevier.com/books/swarm-intelligence/kennedy/978-1-55860-595-4

25. Yang, X. S. (2010). "A new metaheuristic bat-inspired algorithm." In Nature Inspired Cooperative Strategies for Optimization (NICSO 2010) (pp. 65-74). Springer. https://link.springer.com/chapter/10.1007/978-3-642-12538-6_6

26. Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). "Grey wolf optimizer." Advances in Engineering Software, 69, 46-61. https://www.sciencedirect.com/science/article/abs/pii/S0965997813001853

27. Karaboga, D. (2005). "An idea based on honey bee swarm for numerical optimization." Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department. https://www.researchgate.net/publication/228721295_An_Idea_Based_on_Honey_Bee_Swarm_for_Numerical_Optimization

28. Lee, C. P., & Lin, Y. H. (2011). "Gene regulatory network inference using a hybrid GA/PSO approach with an S-system model." In 2011 IEEE International Conference on Systems, Man, and Cybernetics (pp. 947-952). IEEE.

https://ieeexplore.ieee.org/document/6083731

29. Liu, Y., Gao, L., & Zhang, L. (2017). "Improved PSO algorithm for solving high-dimensional optimization problems." Soft Computing, 21(3), 677-685.

https://link.springer.com/article/10.1007/s00500-015-1830-4

30. Khan, I., Khan, F. A., & Jan, M. A. (2014). "Bat algorithm for autonomous discovery of trust communities in mobile ad-hoc networks." Journal of Networks, 9(8), 2015.http://www.academypublisher.com/ojs/index.php/jnw/article/view/jnw09082010

31. Babayigit, B. (2015). "Modified artificial bee colony algorithm for large-scale optimization problems." Computers & Industrial Engineering, 87, 112-122.https://www.sciencedirect.com/science/article/abs/pii/S0360835215002001

32. https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/

33. https://www.r-bloggers.com/2017/06/weather-forecast-with-regression-models-part-4/amp/

34. 34. https://www.researchgate.net/figure/Regression-of-Weather-Variables-against-Temperature-variability-Observations-versus_fig1_312043501

35. Linear Regression Model to Predict Maximum Temperaturehttps://www.linkedin.com/pulse/linear-regression-model-predict-maximum-temperature-ayush-ruel-das?utm_source=share&utm_medium=member_android&utm_campaign=share_via