A black background with white text

Description automatically generated with low confidence

**MSc Data Science**

**Individual Project Module**

**7PAM2002-0901-2024**

**Department of Physics, Astronomy and Mathematics**

PROJECT TITLE: WEATHER FORECASTING USING MACHINE LEARNING

**MSc Data Science Project**

**2024-25**

**Semester A**

**MSc Data Science Project Module Leader**

**Carolyn Devereux**

[**c.devereux@herts.ac.uk**](mailto:c.devereux@herts.ac.uk)

**University of Hertfordshire**

**College Lane, Hatfield,**

**Herts AL10 9AB, United Kingdom**

# Acknowledgements

I would like to express my sincere gratitude to my supervisor for their invaluable guidance and support throughout this research. I also appreciate the assistance from colleagues and peers for their helpful insights and constructive feedback during this study.

# Abstract

This research aims at creating efficient models in weather prediction based on machine learning approaches. Chapter 1 gives an overview of the importance of weather prediction and presents the goals and nature of the research. Chapter 2 also discusses previous work on the topic of meteorological data analysis and on different ML algorithms, with a particular focus on the accuracy of the predictions. Chapter 3 explains the datasets one from Kaggle containing climatic data in a global map, and the second one from Cambridge University containing more local climatic data. Exploratory Data Analysis (EDA) recognises important predictors such as temperature, humidity and pressure. In Chapter 4, ethical implications are discussed to establish that no one’s privacy was infringed as only public data was utilized. Chapter 5 expands upon the method where data cleaning, allocation, and Random Forest, Gradient Boosting, and Linear Regression models are fine-tuned and run are described. In chapter 6, findings indicate that both Gradient Boosting and Random Forest better predict the construction site, with nearly negligible errors compared to the Linear Regression model. Chapter 7 provides results of the analysis and supports the research hypothesis of importance and efficiency of the featured variables as well as discussing the results of the presented investigation in line with previous research. In Chapter 8, it is determined that ML algorithms can further improve the accuracy of the forecast, and their application is described. Proposed resolutions include working with multiple types of data, ever-increasing parameters, and the application of better forms of machine learning. Future work recommendations involve including real time data and testing models in a wide application. This research greatly enhances the accuracy of weather prediction because of the new algorithms developed to guide this area of study.

**Table of Contents**

[Introduction 7](#_Toc184395259)

[1.0 Introduction 7](#_Toc184395260)

[1.1 Background and Context 7](#_Toc184395261)

[1.2 Research Problem 8](#_Toc184395262)

[1.3 Research Rationale 9](#_Toc184395263)

[1.4 Research Aim and Objectives 9](#_Toc184395264)

[1.5 Research Question 9](#_Toc184395265)

[1.6 Significance of the Study 10](#_Toc184395266)

[1.7 Structure of the Dissertation 11](#_Toc184395267)

[Background 12](#_Toc184395268)

[2.1 Introduction 12](#_Toc184395269)

[2.2 Journal Collection Process 12](#_Toc184395270)

[2.3 Critical Analysis of Literature 12](#_Toc184395271)

[2.4 Theories and Models 17](#_Toc184395272)

[2.5 Conceptual Framework 18](#_Toc184395273)

[2.6 Literature Gap 18](#_Toc184395274)

[2.7 Summary 19](#_Toc184395275)

[Dataset 20](#_Toc184395276)

[3.1 Introduction to the Dataset 20](#_Toc184395277)

[3.2 Original Data Collection Details 20](#_Toc184395278)

[3.3 Dataset Relevance and Justification 21](#_Toc184395279)

[3.4 Exploratory Data Analysis (EDA) 21](#_Toc184395280)

[Ethical Issues 25](#_Toc184395281)

[Methodology 26](#_Toc184395282)

[5.1 System Architecture 26](#_Toc184395283)

[5.2 Data Pre-Processing 27](#_Toc184395284)

[5.3 Model Selection 29](#_Toc184395285)

[5.4 Model Training 30](#_Toc184395286)

[5.5 Evaluation Metrics 31](#_Toc184395287)

[Results 33](#_Toc184395288)

[6.1 Individual Model Analysis 33](#_Toc184395289)

[6.2 Comparative Model Performance 34](#_Toc184395290)

[6.3 Limitations and Recommendations 34](#_Toc184395291)

[Analysis And Discussion 36](#_Toc184395292)

[7.1 Interpretation of Results 36](#_Toc184395293)

[7.2 Comparison with Literature 36](#_Toc184395294)

[7.3 Limitations of Results 36](#_Toc184395295)

[7.4 Alignment with Objectives 37](#_Toc184395296)

[7.5 Application to Real-World Scenarios 38](#_Toc184395297)

[7.6 Addressing the Research Question 38](#_Toc184395298)

[Conclusion 39](#_Toc184395299)

[8.1 Conclusions 39](#_Toc184395300)

[8.2 Applications and Real-World Situations 39](#_Toc184395301)

[8.3 Future Work 39](#_Toc184395302)

[8.4 Recommendations 40](#_Toc184395303)

[Reference List 41](#_Toc184395304)

# List of Figures

[Figure 1.1: Weather Forecasting System Market 7](#_heading=h.1t3h5sf)

[Figure 2.1: Proposed Weather Forecasting Model 12](#_heading=h.3j2qqm3)

[Figure 2.2: Weather Forecasting of Cyclone 13](#_heading=h.1y810tw)

[Figure 2.3: Weather Prediction using GraphCast 15](#_heading=h.4i7ojhp)

[Figure 2.4: Conceptual Framework 17](#_heading=h.3whwml4)

[Figure 3.1: Weather Dataset 19](#_heading=h.49x2ik5)

[Figure 3.2: Descriptive Statistics 20](#_heading=h.23ckvvd)

[Figure 3.3: Distribution of Temperature 21](#_heading=h.ihv636)

[Figure 3.4: Temperature Over Time 22](#_heading=h.32hioqz)

[Figure 3.5: Correlation Heatmap of Features with Target (Temperature) 22](#_heading=h.1hmsyys)

[Figure 3.6: Correlation 23](#_heading=h.41mghml)

[Figure 5.1: System Architecture 25](#_heading=h.1v1yuxt)

[Figure 5.2: Converting Column 26](#_heading=h.2u6wntf)

[Figure 5.3: Checking Null Values 26](#_heading=h.19c6y18)

[Figure 5.4: Removing Special Characters 27](#_heading=h.3tbugp1)

[Figure 5.5: Removing Special Characters 27](#_heading=h.28h4qwu)

[Figure 5.6: Converts All Column Names To Lowercase 27](#_heading=h.nmf14n)

[Figure 5.7: Splitting Dataset 28](#_heading=h.37m2jsg)

[Figure 5.8: Model Training of Random Forest Regressor 29](#_heading=h.2lwamvv)

[Figure 5.9: Model Training of Linear Regression 29](#_heading=h.111kx3o)

[Figure 5.10: Model Training of Gradient Boosting Regressor 30](#_heading=h.3l18frh)

[Figure 5.11: Evaluation Metrics of Random Forest Model 30](#_heading=h.4k668n3)

[Figure 5.12: Evaluation Metrics of Linear Regression 30](#_heading=h.2zbgiuw)

[Figure 5.13: Evaluation Metrics of Gradient Boosting 31](#_heading=h.1egqt2p)

[Figure 6.1: Results of Random Forest Model 32](#_heading=h.3cqmetx)

[Figure 6.2: Results of Linear Regression 32](#_heading=h.1rvwp1q)

[Figure 6.3: Results of Gradient Boosting 33](#_heading=h.4bvk7pj)

# 

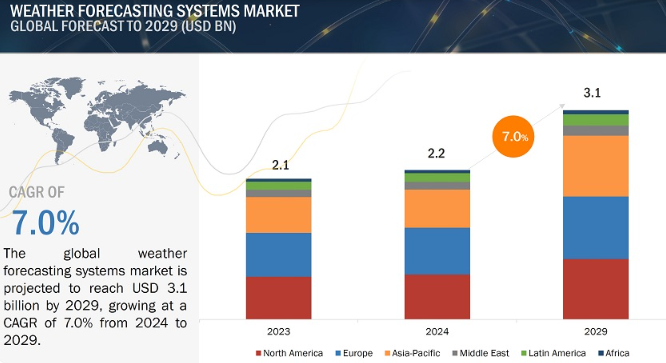
# Introduction

## 1.0 Introduction

Weather forecasting is essential for operation and protection in several industries globally (Bag *et al.,* 2022). In the recent past, Machine learning (ML) algorithms have been incorporated into the process of forecasting, making foresight predictions more accurate. In the past, analysts used numerical model simulations to make predictions, and this approach presented some problems in capturing key features of atmospheric circulation (Kochkov *et al.,* 2024). Business forecasting is made accurate by machine learning since large data sets are analyzed to discover key features that make prediction possible. Conditions of humidity, pressure, and cloud formations are other important factors for the generation of the weather (Lakra and Avishek, 2022). The purpose of this study is to examine the application of ML for enhancing the accuracy of weather predictions with consideration of important parameters and effective algorithms. The advancement in data availability for weather data assists in the formulation of efficient and oversimplified ML-based forecasting frameworks that are useful in different industries.

## 1.1 Background and Context

Weather forecasting is an essential part of agriculture, transportation, risk management, and general time use management (Fathi *et al.,* 2022). Several conventional approaches rely on numerical weather prediction (NWP) models based on mathematical equations to predict weather. However, these methods are subjected to accuracy problems occasioned by limited atmospheric information. Machine learning (ML) has become more prominent as a promising solution to overcome these challenges.



##### Figure 1.1: Weather Forecasting System Market

(Source: Markets and Markets, 2024)

The Weather Forecasting Systems Market Size is projected to grow from USD 2.2 billion in 2024 to USD 3.1 billion by 2029, at a CAGR of 7.0% (Markets and Markets, 2024). These systems utilize historical data, and the accuracy of the prediction is significantly higher compared to other methods. For example, Shouman (2024) showed that ML models excelled over NWP in predicting photovoltaic power, a variable that is strongly associated with the weather. Support vector machines, decision trees, and neural networks are the typical cases of ML that compare intricate data patterns and include humidity, cloud formation, and temperatures. This research attempts to discover the existence of important variables as well as to determine the optimal set of algorithms for weather predictions using ML. Precision improves results for fields such as aviation since achieving better than other major system accuracy can reduce the need to spend millions on improvements every year (Yazdani-Asrami *et al.,* 2022). Therefore, incorporating ML in weather prediction is indispensable in modern society and economic standards.

## 1.2 Research Problem

Weather prediction difficulties persist because the conditions relating to the atmosphere are not easy to predict and the variables are often interrelated (Jaseena and Kovoor, 2022). Such complexities are not easily reflected in traditional numerical prediction models thus lowering its reliability. Machine learning could have some solutions but it is a challenging task to find out the right parameters and algorithms for prediction (Zhang *et al.,* 2021). Thus, the prioritization of activity variables, such as humidity, pressure, and temperature, is necessary for enhanced accuracy of the forecast. However, the amount of models will surpass any practicable toolkit and this diversity makes the selection of appropriate algorithms challenging (Badjie, Cecílio, and Casimiro, 2024). Overcoming these challenges is imperative in building efficient, accurate ML-based weather forecasting tools to support a wide range of uses in agriculture, transportation, and disaster response.

## 1.3 Research Rationale

The role of weather prediction has central importance for such industries as agriculture, aviation services, and during the crisis period. The climate-related disasters that occurred in the world hit over 4 billion people from the year 2000 to 2019 as per the report of the UN (Donatti *et al.,* 2024). It is characteristic of conventional numerical approaches that they can poorly predict sharp changes in the atmosphere. Machine learning (ML) enhances the forecasting capability by reaching a large database and drawing out pattern possibilities effectively. The ML market also looks great globally, and its size may increase to 209 billion USD by 2029. Zhang *et al.* (2021), identify increased prediction accuracy by exploiting the use of ML models. This research defines fundamental weather parameters such as humidity and temperature and analyzes essential ML techniques. As datasets from sources such as Kaggle become more commonplace, there is great potential in using ML for forecasting. From the potential value perspective, the enhanced accuracy contributes to disaster prevention and obtains fewer losses in terms of money and lives. This work helps to fill gaps in the selection of variables and algorithms for accurate weather forecasting systems.

## 1.4 Research Aim and Objectives

The research aims to identify the weather parameters and variables crucial for weather forecasting using machine learning algorithms.

***Objectives***

1. To collect a suitable weather dataset for weather forecasting.
2. To choose the important weather parameters like cloud formation, humidity, pressure etc., for developing the weather prediction model.
3. To choose the best ML algorithm for developing the model.

## 1.5 Research Question

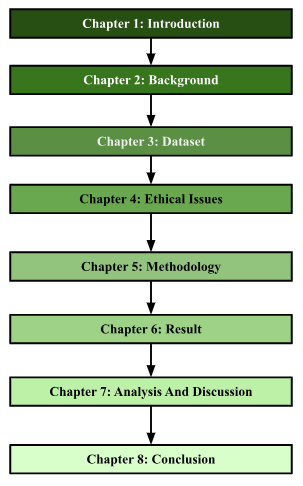
The research question which this research procedure will address is as follows:

**What variables, parameters etc., are crucial for weather forecasting using machine learning algorithms?**

## 1.6 Significance of the Study

The importance of this study stems from the fact that machine learning will be used to improve the prediction accuracy of the weather. Weather forecasts have a critical role in agriculture, aviation, the disaster management and transportation industries (Merz et al. 2020). Traditionally, complex atmospheric interaction prediction can be poor with many of the traditional numerical prediction methods. Machine learning models can consume and analyze vast dataset to give very accurate predictions. Key weather parameters are identified which aid in the optimization of forecasting models for practical applications. Predictions are enhanced to reduce costs and to prevent losses in the weather sensitive industries such as farming and logistics. Weather forecasting is improved for disaster management purposes, which enables predicting severe weather events far more accurately (Huang, Wang and Liu, 2021). Challenges of selecting right variables, and algorithms for forecasting are addressed in this study. Historical and real-time data can be used for very precise predictions with machine learning. Humidity, pressure and temperature analysis allows for the understanding of atmospheric conditions. The results can assist in choosing an algorithm of choice for a variety of industrial applications in the field of the weather forecast. Integrating ML in weather forecasting enables industries to mitigate climate related risks, efficiently (Chen et al. 2023). The effective ML based forecasting approaches identified by this study adds to the research. Practical benefits in terms of cost savings and safety improvements are provided by enhanced prediction models. The research outcomes are useful for improving weather prediction technology and application. Accuracy improves industries’ operations optimisation and resilience to adverse weather impacts.

## 1.7 Structure of the Dissertation



**Figure 1.2: Dissertation Structure**

(Source: Self Developed)

# 

# Background

## 2.1 Introduction

Weather forecasting has been an important application of machine learning algorithms. The models are developed based on historical data of various weather parameters. The parameters could be temperature, humidity, wind speed, and other atmospheric parameters. This way the possible long-term weather prediction can be done using machine learning algorithms. This background chapter will critically analyses the existing papers on weather prediction and identify the possible literature gap. The relevant theories and models will also be discussed.

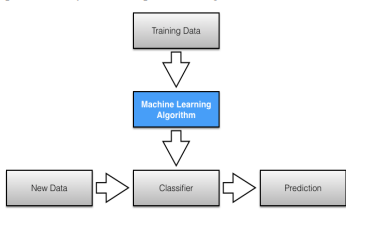
## 2.2 Journal Collection Process

The journal collection inclusion and exclusion criteria are as follows:

1. It was ensured that the research papers have been published in recent years. This is the reason why the research papers published in 2019 have been chosen for the background analysis.
2. The research papers about weather forecasting were only selected to ensure that the background analysis would be within the scope of the research.
3. The research papers written in the English language were selected as part of this research procedure.

## 2.3 Critical Analysis of Literature

As per Singh, Chaturvedi and Akhter, (2019), the constant climate change has ensured that the old weather prediction model gets obsolete. The main aim of the paper was to develop a model for weather prediction that can be used in remote areas. The authors have developed the model using “Random Forest Classification” algorithms. The result was a portable and low-cost weather prediction solution for remote areas.



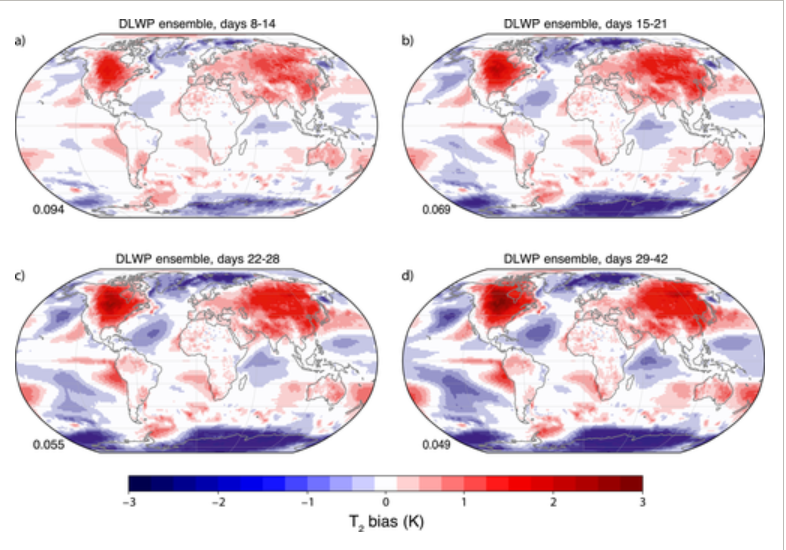
##### Figure 2.1: Proposed Weather Forecasting Model

(Source: Singh, Chaturvedi and Akhter, 2019)

The figure shows the model's overall flowchart or system architecture. The model would first classify and then make the prediction. The developed model has an accuracy of more than 87% in terms of its prediction of whether there would be rain or not. Despite the high accuracy, the limitation of the model could be that it's only evaluated based on whether it would rain or not. The evaluation based on other weather incidents had not been done. On the other hand, Jakaria, Hossain and Rahman, (2020), the weather parameters and data used in weather forecasting are unstable. The authors have developed a weather forecasting model that can predict short weather based on data from multiple weather stations. The author has collected data from various weather stations in Nashville. The authors have compared results from “Random Forest Classifier” with SVM, “Multi-layer Perceptron”, “Extra-Tree Regression”, and “Ridge Regression”. The random first classifier has shown the best result. The main strength of the paper was its usage of multiple ML algorithms for weather prediction, however, the authors have not developed a full-fledged application which is a major limitation of the model.

Bochenek and Ustrnul, (2022), conducted a systematic literature review on various weather prediction papers. The aim was to identify the most common factors and methods used for weather prediction. The result showed that radiation, pressure, temperature, precipitation, and wind are the most common meteorological fields examined for predicting weather. SVM, random forest, XGBoost, “Artificial Neural Networks” etc., have been the most popular algorithms that were used. The paper has summarised and identified the important aspects of weather parameters and algorithms used for weather prediction. However, it has not developed its own model which is a major limitation.

As per Wang *et al.* (2019), the inappropriate settings of the initial states can lead to unsatisfactory results in weather prediction. This is why the authors have proposed a data-driven weather prediction model based on machinimas of information fusion. The authors have used deep learning algorithms based on the “novel negative log-likelihood error” problem. It is able to forecast for both uncertainty quantification and single-value prediction. The accuracy of the model significantly increased compared to traditional numerical models for weather prediction. However, accuracy is still low in this regard. On the other hand, Weyn *et al.* (2021), have developed an “ensemble prediction system” based on DL algorithms. The CNN algorithm was used for six-week forecasts.



##### Figure 2.2: Weather Forecasting of Cyclone

(Source: Wang *et al.,* 2019)

This figure shows the cloud and weather prediction for Cyclone Irma using the proposed model. The proposed model can forecast 320 times within 3 minutes and this shows the effectiveness of the model. The main effectiveness of the mode is that it has been evaluated based on real-world incidents like Irma. However, it is not effective for long-term weather prediction for the future.

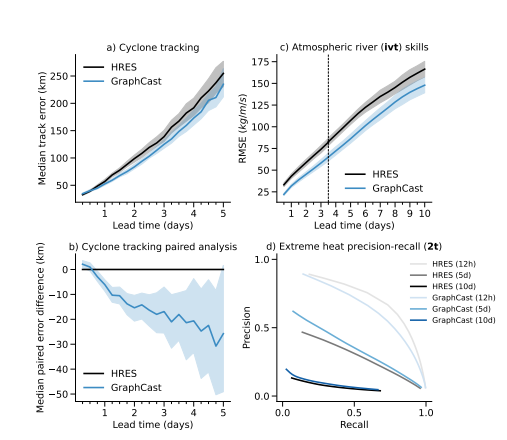
Cho *et al.* (2020), have developed a numerical model for weather prediction for extreme air temperature events in urban areas. The prediction of minimum and maximum temperature is the most basic and essential part of weather forecasting. However, coarse grading and the absence of proper parameterisation have affected the quality of prediction. The authors have used SVMM, ANN, Random Forest, and “multi-model ensemble” algorithms for predicting the minimum and maximum temperature of the next day in South Korea. The model has an R2 value of 0.69 which is more than 0.50. This shows the model has some errors despite the low RMSE value.

As per Hewage *et al.* (2021), the model of “numerical weather prediction” needs significant power for solving mathematical equations. The authors have proposed a novel “lightweight data-driven weather forecasting model” for weather forecasting. They have used the LSTM algorithm and TCN or “temporal convolutional networks” for model development. Further other classification and regression models have also been used. The forecasting of the weather is collected by the time series data in this regard. The result showed that the model has a higher accuracy compared to general deep learning models. It can forecast weather for up to 12 hrs.

As per Chattopadhyay, Nabizadeh and Hassanzadeh, (2020), the ever-growing resources and time consumption have been the main problems associated with weather prediction based on numerical values. The authors have proposed an analog forecasting model based on data. The proposed model has used CapsNet and “capsule neural networks” to build a “novel deep learning pattern-recognition technique”. The data was collected from the Earth system model and it was used to train the weather model. The accuracy of the trained model is 45% which is not satisfactory. However, the usage of CapNets has increased the model accuracy up to 80%. CapsNet has outperformed both CNN and normal logistic regression models in this regard. However, the model accuracy is still not close to 100% which is a major model limitation.

As per Chantry *et al.* (2021), ML algorithms can be valuable as an accelerator for the scheme of parameterisation. The authors have developed a parameterisation scheme using ML algorithms. The authors have used deep learning algorithms like ANN in this regard. The result showed that for the medium-range prediction, the model is highly accurate. However, for short-time and long-time prediction, the model is not that accurate for weather prediction.

Lam *et al.* (2023), that global weather forecasting is tough for medium-time forecasting due to various social and economic domains. The traditional methods need significant computing resources for prediction. The authors have developed an ML-based model called Graphcast. It can predict more than 100 weather variables at the same time. The accuracy of the model is 90% which is significantly higher than other systems of operational deterministic. It can predict severe events accurately including extreme temperatures, cyclones, atmospheric rivers and so on. This way it has been effective to develop complex models for dynamic systems.



##### Figure 2.3: Weather Prediction using GraphCast

(Source: Lam *et al.,* 2023)

The figure shows possible prediction graphs of Graphcats based on cyclone tracking, atmospheric rivers, extreme heat and precision levels.

As per Grönquist *et al.* (2021), the quantification of forecasting of weather is very difficult due to the extreme weather events. Ensemble prediction systems. Can be used to predict extreme weather events. However, these are costly with high computational resources. The authors have proposed a mix of traditional models and ensemble prediction models to assess the non-linear relationship between different weather parameters. The authors have used global data for their prediction methods. The authors have used case studies to prove improved weather forecasting for extreme events. This way the overall cost and resources of the ensemble system can be reduced through the mixed method. Han *et al.* (2022), have developed a wind speed prediction model as part of the quantitative weather forecasting model. The prediction of wind speed is tough due to the relationship between different meteorological parameters. The authors have used weather forecasting and research to develop a hybrid model for wind speed prediction. In creating other prediction models, the authors have combined the multivariate data decomposition method and the DL model. The deep learning algorithms were developed using CNN and bidirectional LSTM. It was found that the model proposed by the authors has outperformed other similar models including an MAE value of 0.1042. The accuracy has also been increased significantly in this regard.

Rasp *et al.* (2020), have developed a data-based forecasting model for weather prediction. The authors have tried to predict global weather based on their data-driven approach. The authors have collected datasets from the ERA5 archive. They have used simple linear regression, physical models, and deep learning algorithms for model development. The method has been able to predict weather for 3-5 days. The developed model has significantly increased the overall quality prediction quality. The lack of a detailed description of evaluation metrics could be the main limitation of this study.

## 2.4 Theories and Models

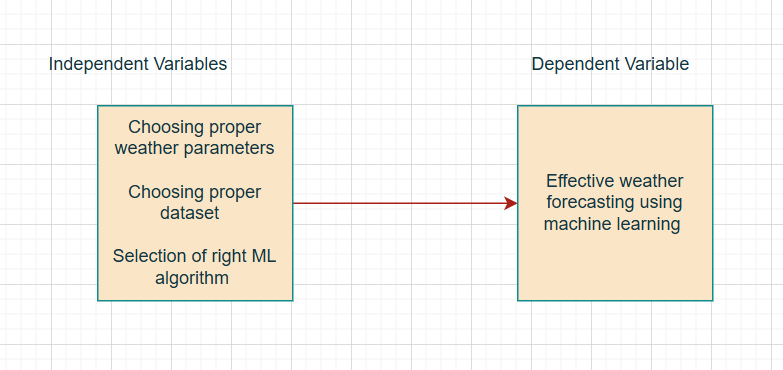
***Technology Acceptance Model (TAM)***

TAM can be defined as the model that describes how people accept a new technology. It was established by Fred Davies to show why people can accept or reject new technology. The use of modern machine learning, deep learning, and time series algorithms has improved the weather prediction quality significantly. People can see weather forecasting from their own mobile or computers. This is why weather forecasting using modern technologies can be explained using TAM.

1. ***Perceived usefulness:*** A user will use a new technology if they believe the new technology will improve job performance (Zaineldeen *et al.,* 2020). Modern techniques like Time series models, ML and DL algorithms have improved the weather forecasting performance significantly well. The audiences have easily accessed the weather prediction application from their preferred devices.
2. ***Ease of use:*** This means the user will accept a new technology if it frees their effort. The modern technologies of weather forecasting have reduced the effort of users to accept weather forecasting. The models have improved the quality of the forecasting and reduced the possible effort of researchers to collect data as well.

This shows the Ml algorithms have significantly been effective for forecasting the future weather and collecting weather data in this regard.

## 2.5 Conceptual Framework



##### Figure 2.4: Conceptual Framework

(Source: Created by the Authors)

This is the conceptual framework that this research procedure would like to follow. Here, the “dependent variable” is “Effective weather forecasting using machine learning”. Now, the “independent variables” are choosing proper weather parameters, choosing the proper dataset selecting of right ML algorithm.

## 2.6 Literature Gap

The various research papers have discussed the usage of various algorithms and various weather parameters for weather forecasting. However, there has been a lack of specificity about which weather parameters researchers need to consider to focus their models on. There are specific weather and atmospheric parameters like cloud formation, wind speed, temperature, pressure, humidity etc. This research procedure has the purpose of explaining in depth the specific weather parameters that will be used for predicting the weather in this regard. This way the specific algorithms will be selected for developing the model. This way the classification and prediction procedure will be easier.

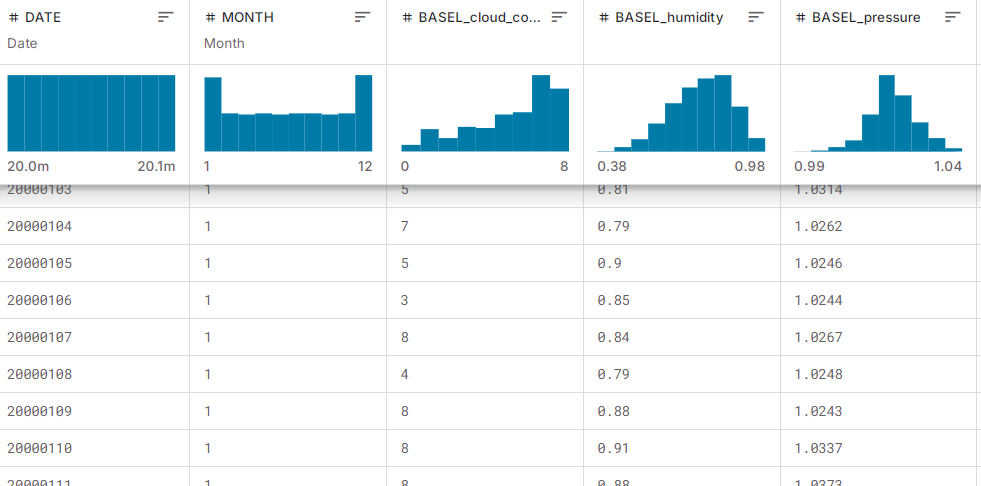
## 2.7 Summary

To summarise, the background section has provided the required information to assess the existing papers about weather prediction. Random Forest has been the most popular algorithm that has been used for forecasting weather followed by SVM and other deep learning algorithms. The data-driven approach for numerical prediction of weather has been a trend in recent years. The technological acceptance model has shown why both researchers and users are accepting modern technology for weather forecasting. Based on the analysis it was found that the lack of explanation of the weather parameters has been the primary limitation of the existing study. This will be mitigated in this study.

# Dataset

## 3.1 Introduction to the Dataset

This research procedure will use a combination of two datasets. The first dataset was collected from the Kaggle website as it is a popular data collection website. The dataset link is as follows: “<https://www.kaggle.com/datasets/thedevastator/weather-prediction/data?select=weather_prediction_dataset.csv>”. The other dataset that was collected was from Cambridge University’s website. The link is “<https://www.cl.cam.ac.uk/weather/>”.



##### Figure 3.1: Weather Dataset

(Source: Kaggle, 2024)

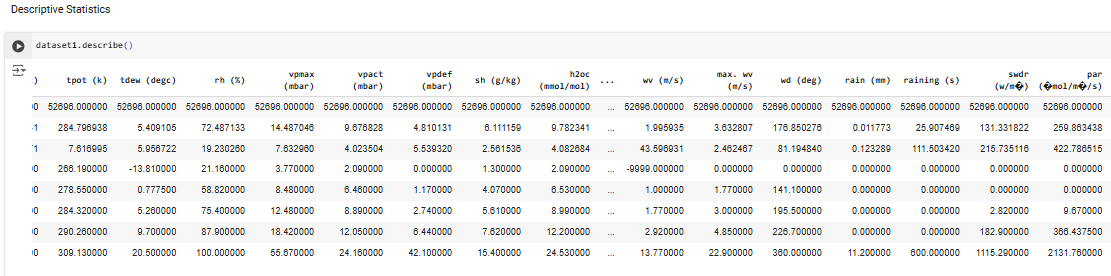
## 3.2 Original Data Collection Details

The Kaggle dataset was formed by combining historical weather data collected from different global weather stations. This can be in the form of temperature, humidity, and pressure which are observed using meteorological instruments. This has been collected to be used in future forecast research work as well as to enhance the development of artificial intelligence (Dwivedi *et al.,* 2023). The Cambridge dataset can be sourced from the Cambridge University Computer Laboratory weather station back in the United Kingdom. Information has been collected quite systematically by means of automated sensors and other instruments. These records pertain to the local atmosphere and were recorded to be used in climate research as well as in weather forecasting. Each data set contains accurate and reliable weather data to provide for analysis.

## 3.3 Dataset Relevance and Justification

These are good datasets for use in weather forecasting because these databases include essential meteorological data. The Kaggle dataset offers weather data worldwide, which means the scope of observation here is more general. The Cambridge dataset, therefore, augments localized atmospheric data, improving accuracy with weather prediction. All these models require key parameters that are normally measured and recorded to include temperature, humidity, and pressure among others and both datasets contain. The integration of international and domestic information is more helpful for people to perceive the changes of climate (Petzold *et al.,* 2020). This ensures that the compiled dataset is relevant in relation to the objective of the research in finding key parameters that can be used in machine learning for weather predictions.

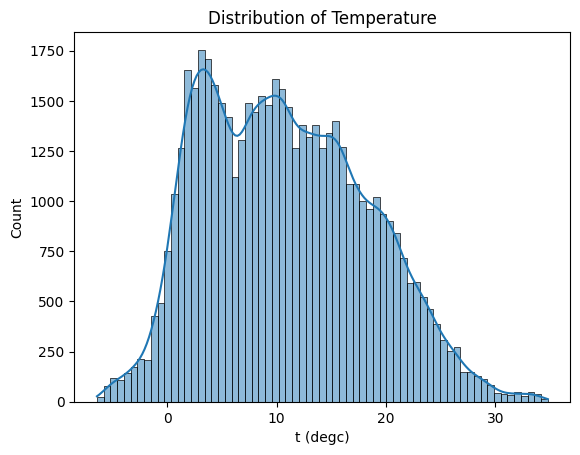
## 3.4 Exploratory Data Analysis (EDA)



##### Figure 3.2: Descriptive Statistics

(Source: Obtained Using Jupiter Notebook)

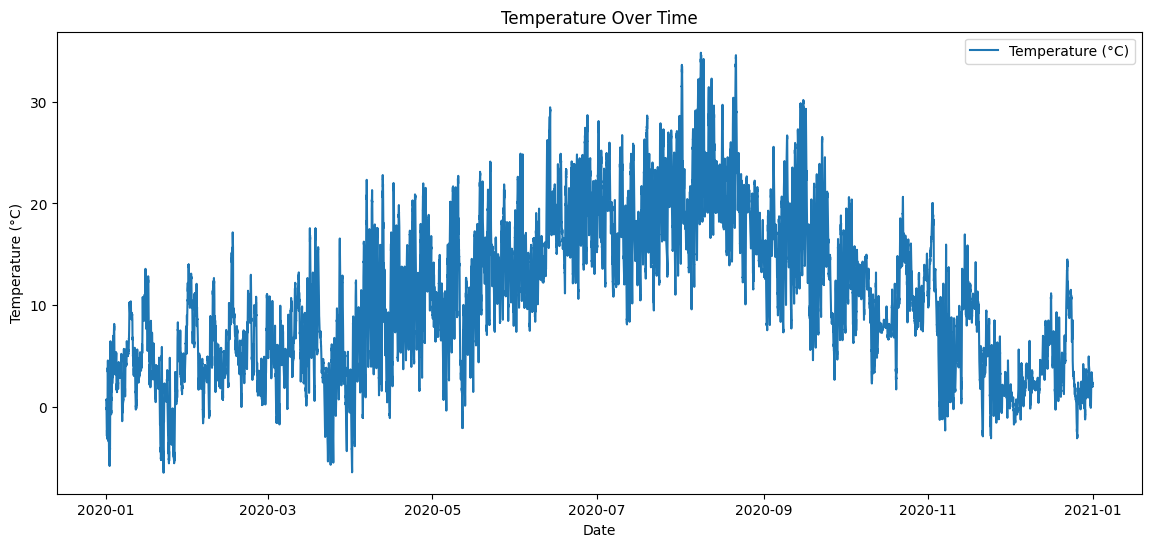
This data-set has 52,696 records with 21 features, some of them are temperature, humidity, pressure, and wind speed. Key statistics show diverse ranges: temperature ranges between -6.44°C and 34.8°C humidity ranges between 21.16% and 100% and wind speed ranges from 0 to 13.77m/s. Some of the columns still contain missing values and there are outliers present as well.



##### Figure 3.3: Distribution of Temperature

(Source: Obtained Using Jupiter Notebook)

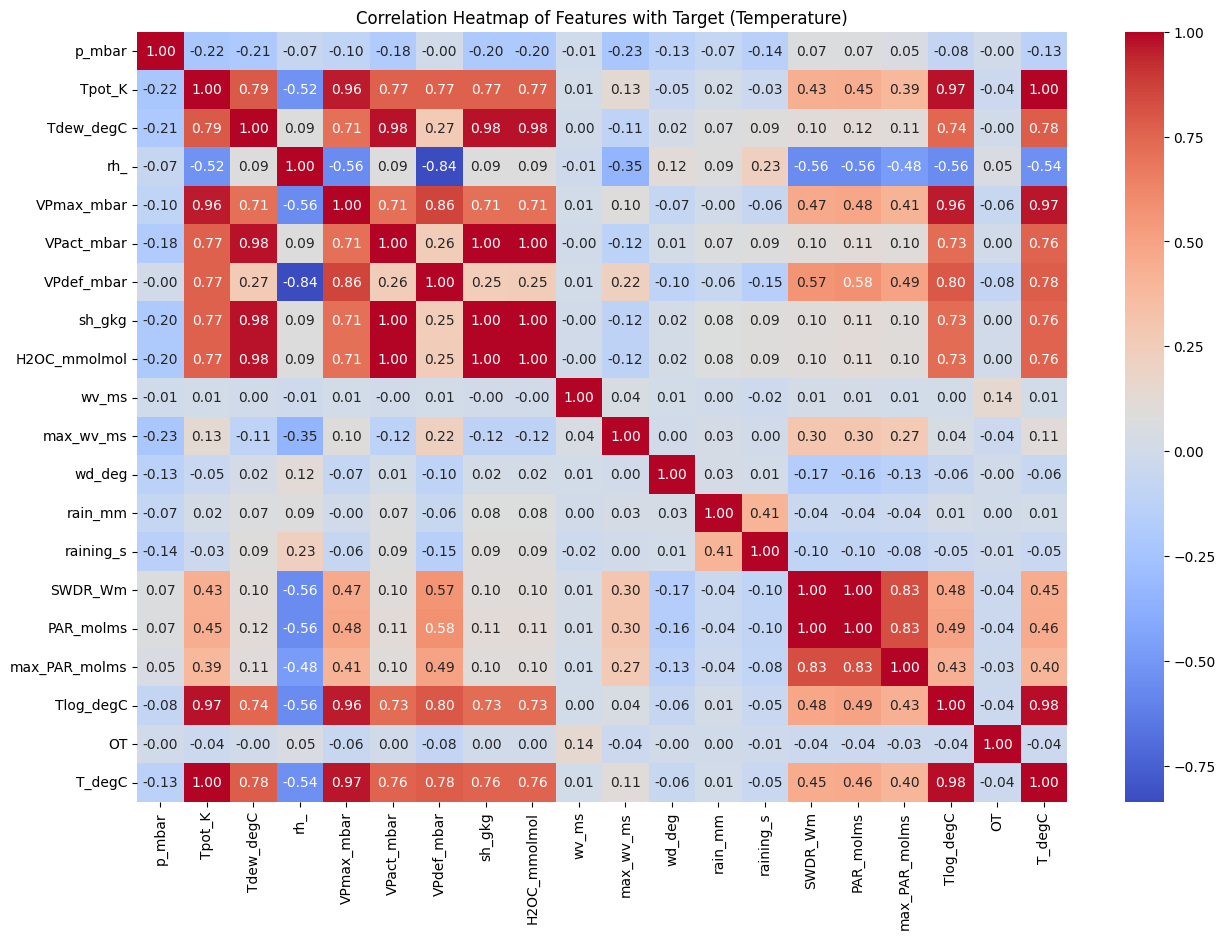
The histogram above demonstrates the frequencies of temperature by taking a snapshot of what majority of the temperatures ranging between 10°C to 20°C. Here, the results appear slightly skewed to the right, which means that relatively higher temperatures are limited but not entirely absent. This implies a continental climate where clearly there are no extreme hot or cold events, but moderate ones. The distribution also shows some very large values which are probably due to an exceptionally high temperature or low humidity.



##### Figure 3.4: Temperature Over Time

(Source: Obtained Using Jupiter Notebook)

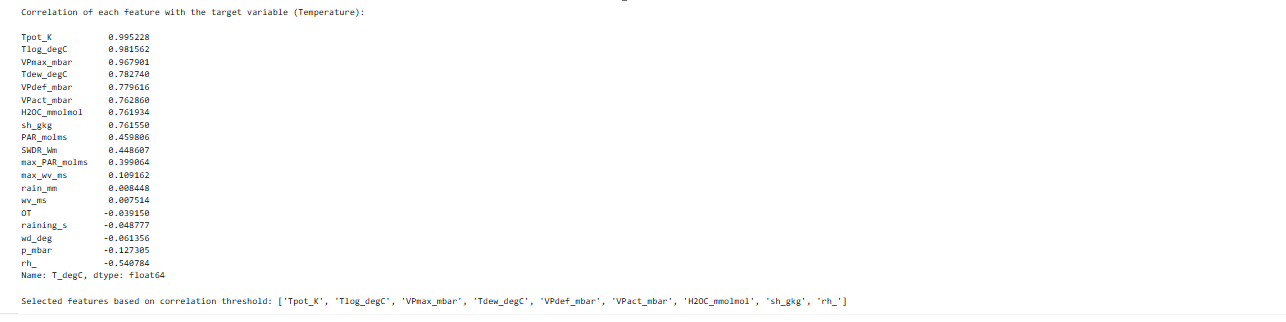
The plot shows the movement in the temperature during a particular period. The plot oscillates showing that sometimes there are higher temperatures in some months than in other months, where the temperatures are comparatively low. These trends could be due to seasonal variations probably implying that the data collected is annual weather data. Humps or troughs could represent certain phenomena such as a particular day or days when temperature varied norm or went up or came down sharply.



##### Figure 3.5: Correlation Heatmap of Features with Target (Temperature)

(Source: Obtained Using Jupiter Notebook)

The above heatmap depicts the associations between various weather features and the target variable, temperature. Some of the Parameters which as have positively related to temperature include dew point (Tdew\_degC) pressure (p\_mbar) and relative humidity (rh). It helps the researchers decide which features are most relevant to temperature, which can be useful when constructing machine learning models.



##### Figure 3.6: Correlation

(Source: Obtained Using Jupiter Notebook)

The correlation values depicted demonstrate that different features like Tpot\_K Tlog\_degC and VPmax\_mbar are strongly positively lined with temperature which makes them reliable determinants of temperature. Other features such as wv\_ms, rain\_mm, and OT worked in the wrong direction i.e., their correlation is either negative or very weak when it comes to temperature prediction. Depending on the established threshold, additional features are Tpot\_K:Tlog\_degC: and other significant features contributing to the accomplishment of the goal, as opposed to including all covariates.

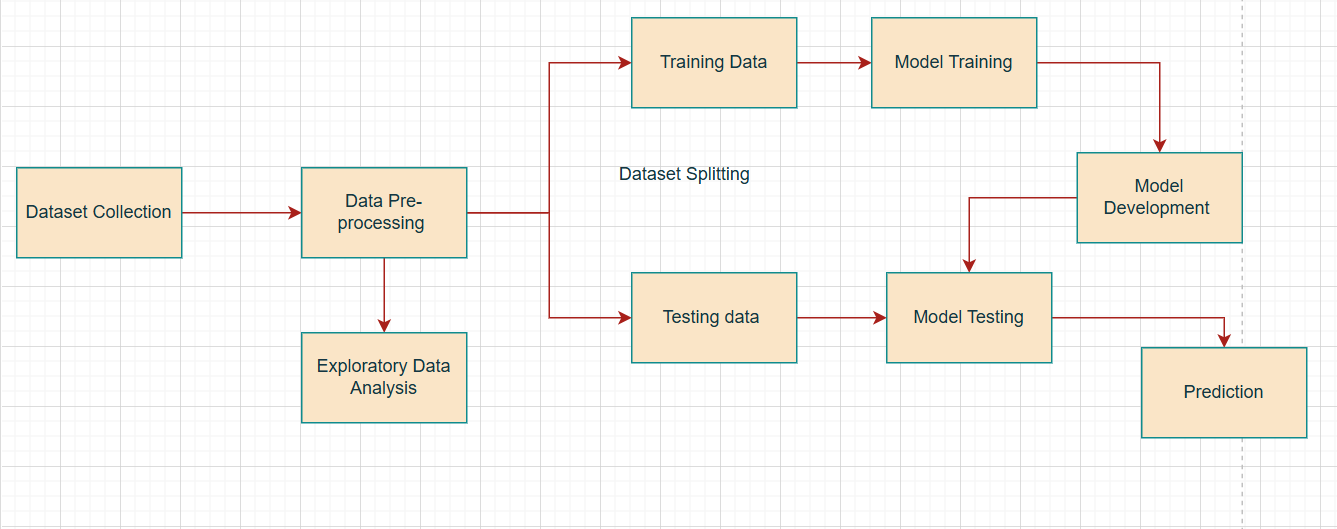
# Ethical Issues

The dataset does not have any personal data of people and it is publically available to use. This is the reason why it does not have any privacy issues for data breaches. There is no chance of breaching the GDPR principle because it does not have any data privacy issues.

The dataset has indeed used several journal papers for background study. The target will be to appropriately reference all those secondary sources of data so that the original data owner gets the credit (Hummel, Braun and Dabrock, 2021). One of the major considerations of the research procedure will be to honestly present all the data. The findings from the software will be honestly presented with screenshots of the software in this regard.

# Methodology

## 5.1 System Architecture



##### Figure 5.1: System Architecture

(Source: Self-created using Draw.io)

This is the overall system architecture that was used in this research procedure.

1. **Dataset Collection:** The dataset was collected in this stage. The details of the weather data collection have already been explained previously.
2. **Data Pre-processing:** The dataset was then prepared based on various procedures. This includes trying to find the null values or missing values and so on. Further, the main weather parameters were identified and features were extracted during this stage.
3. **Exploratory Data Analysis:** The overall dataset was visualized to assess its nuanced understanding of it.
4. **Dataset Splitting:** The dataset was divided for training and testing. Most of the data was used for training and the others were used for testing the model.
5. **Model Training:** the model was then trained using the training data. It was ensured that the model was not overfitted.
6. **Model Development:** The selected machine learning model was then developed based on the collected data.
7. **Model Testing:** The testing data was used during this stage to test the model. The model evaluation techniques like classification reports were used for testing the model in this stage.
8. **Prediction:** Finally the developed model predicts the weather and fulfils the main objective of the research.

## 5.2 Data Pre-Processing



##### Figure 5.2: Converting Column

(Source: Obtained Using Jupiter Notebook)

The code converts the 'date' column in dataset1 to a datetime format, handling errors by coercing invalid dates. It then sets the 'date' column as the index for the dataset.



##### Figure 5.3: Checking Null Values

(Source: Obtained Using Jupiter Notebook)

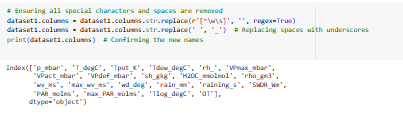
The isnull().sum() output shows that there are no missing values in any of the columns of dataset1, indicating that the dataset is complete and does not require imputation for missing data.



##### Figure 5.4: Removing Special Characters

(Source: Obtained Using Jupiter Notebook)

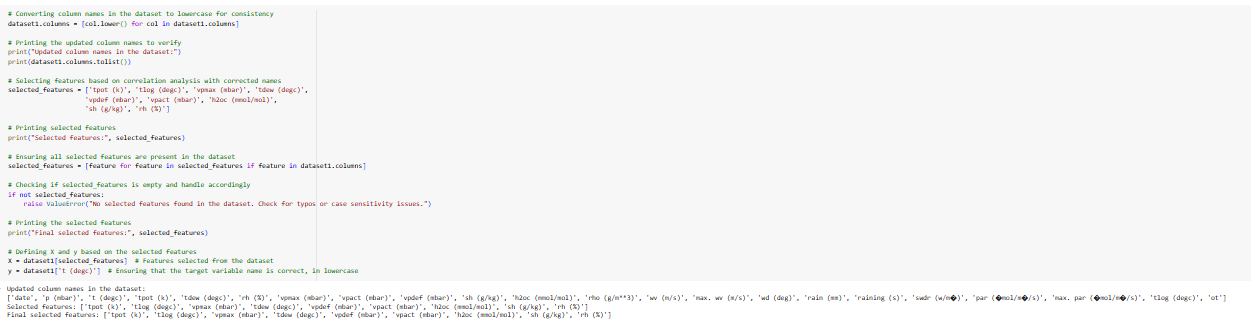
The code renames the columns in dataset1 by removing any special characters, leaving only alphanumeric characters and spaces. This simplifies column access and ensures compatibility with various data operations.



##### Figure 5.5: Removing Special Characters

(Source: Obtained Using Jupiter Notebook)

The code renames the columns of dataset1 by first removing all special characters and spaces. Then, spaces are replaced with underscores for consistency and easier access. The new column names, such as 'p\_mbar' and 'T\_degC', are printed to confirm that the transformations were successfully applied.



##### Figure 5.6: Converts All Column Names To Lowercase

(Source: Obtained Using Jupiter Notebook)

The code converts all column names to lowercase for consistency and selects specific features based on correlation analysis. It verifies that the selected features exist in the dataset, handles potential errors, and defines the feature matrix (X) and target variable (y) for model training.



##### Figure 5.7: Splitting Dataset

(Source: Obtained Using Jupiter Notebook)

Normally data is divided in training data and testing data in the ratio of 80:20 respectively. When using the train\_test\_split function, care is taken to bring in such changes that 80% of the data will be used for training the model while only 20% used for testing the model. The RANDOM\_STATE = 42 is set to make randomness fixed for each run of the split.

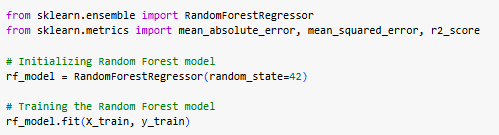
## 5.3 Model Selection

***Random Forest Regressor:*** The Random Forest Regressor is chosen because the model can well interpret interactions between features and their importance. It is trained using the training data (X\_train, y\_train) and evaluated using various metrics: MAE, MSE, RMSE, and R². These are by and large measures of the accuracy of the models, the errors involved in the prediction, and the extent of their predictive abilities. The importance of features is also tested to determine the most significant refractories of the dependent variable. This helps explain how a model arrived at such a decision and what features played a major role in the decision-making.

***Linear Regression:*** Linear Regression is selected as the best fitting model due to its application in interpretation of the correlation between features and the target variable. Subsequent to building the model using the training data (X\_train, y\_train), numerical accuracy is assessed by the MAE, MSE, RMSE and the coefficient of determination (R²). These metrics use to evaluate the proposed model regarding to its accuracy and capability of prediction. The evaluation facilitates comparison with other elaborate models to identify compatibility requirements suitable for the dataset.

***Gradient Boosting Regressor:*** Gradient Boosting model is selected because it can deal with non-linear variables and enhanced ensemble learning. Their results indicate that it integrates many weak models to develop a single, more precise predictor. In the last segment of model assessment, the MAE, MSE, RMSE and R² measures are applied to measure the performance of the model. These metrics give a holistic view on accuracy and on how good the model is in predicting unseen data making this model highly recommendable for regression tasks.

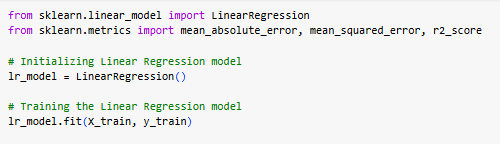
## 5.4 Model Training



##### Figure 5.8: Model Training of Random Forest Regressor

(Source: Obtained Using Jupiter Notebook)

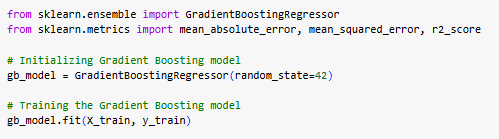
Random Forest Regressor was used to train the model on the training dataset. The model integrates several decision trees in order to forecast, which increases accuracy by using ensemble learning. It shows correlation between features and the target variable under study, reduces variance and bias of the model.



##### Figure 5.9: Model Training of Linear Regression

(Source: Obtained Using Jupiter Notebook)

Linear Regression was used for training with the purpose to identify the connection between features and a target. It models the relationship as a direct one and calculates values that would bring the sum of residues squared to the minimum. It is very simple to implement but offers understandable results and also does not demand high computational power.

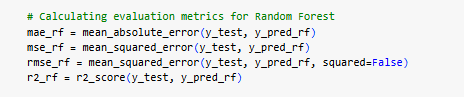


##### Figure 5.10: Model Training of Gradient Boosting Regressor

(Source: Obtained Using Jupiter Notebook)

The Gradient Boosting Regressor model was learned through the process of combining multiple weak models in a boosting process. Each succeeding model makes adjustment for the mistakes that have been committed by the earlier models. This approach prove useful in modeling non linearity and other pattern which are not easy to model due to high levels of bias.

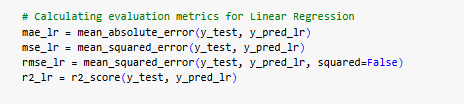
## 5.5 Evaluation Metrics



##### Figure 5.11: Evaluation Metrics of Random Forest Model

(Source: Obtained Using Jupiter Notebook)

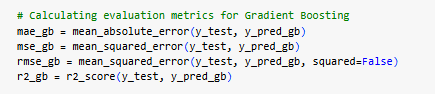
Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score were used to measure the accuracy of supervised learning Random Forest model. MAE indicates the centrality of error, MSE describes merely the extent of the error, RMSE punishes greater magnitude of error and the R^2 provides information about the proportion of variation that exists in the data.



##### Figure 5.12: Evaluation Metrics of Linear Regression

(Source: Obtained Using Jupiter Notebook)

When it comes to Linear Regression, the models assessment included MAE, MSE, RMSE, and R^2 Score. MAE gives information about the average width of prediction errors, whereas MSE and RMSE describe the distribution of such errors with extra attention being paid to large deviations reflected in RMSE. The higher the R^2 Score shows the more it represents the data in the model.



##### Figure 5.13: Evaluation Metrics of Gradient Boosting

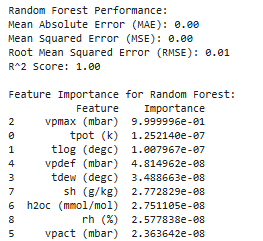
(Source: Obtained Using Jupiter Notebook)

For the evaluation of the performances of the training-model which has used the Gradient Boosting model MAE, MSE, RMSE and the R^2 Score were used. MAE provides the simple measure of error while MSE and RMSE tells how large errors impact. This R^2Score assesses the extent to which a model can account for variation which in this case will give the measure of how good the predictive model fits into.

# 

# Results

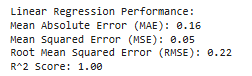
## 6.1 Individual Model Analysis



##### Figure 6.1: Results of Random Forest Model

(Source: Obtained Using Jupiter Notebook)

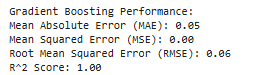
Random Forest model yields MAE = 0.0139, MSE = 0.0002 and RMSE = 0.0363 as well as R² = 1.00. This means that the model had nearly no error in the test data that was used to make the final predictions. In terms of feature importance percentage, the first one is vpmax (mbar), which accounts for nearly all of the model’s score. It is apparent that other features such as “tpot (k)” “tlog (degc)” and “vpdef (mbar)” provide insignificant input.



##### Figure 6.2: Results of Linear Regression

(Source: Obtained Using Jupiter Notebook)

The Linear Regression model averts a very high error and gives a perfect result as it has the MAE of 0.16, MSE of 0.05, and RMSE of 0.22. In the land use regression the R February R² of 1.00 means the model captures nearly all of the variability in the target. This implies there is high degree of positive linear correlation between the chosen attributes and the Dependent Variable.



##### Figure 6.3: Results of Gradient Boosting

(Source: Obtained Using Jupiter Notebook)

The proposed Gradient Boosting model performs exceptionally well with the MAE stands 0.05, MSE 0.00, and RMSE of 0.06. In fact, an R² of 1.00 guarantees that the model reflects all variability associated with the target. These measures demonstrate high accuracy and low levels of prediction error as a sign of good performance in capturing difficult relations.

## 6.2 Comparative Model Performance

The comparison analysis compares Random Forest, Gradient Boosting, and Linear Regression by applying MAE, MSE, RMSE, R² scores. In terms of target variability, all three models offer an R² of 1.00. However, both models have different error metrics which show diverse prediction efficiency and eligibility for temperature prediction. Random Forest has virtually non-existent error values with MAE, MSE, and RMSE, all similar to zero at 0.00. This indicates that it is well suited to handle data structures and the nature of their interaction. They have found that the model exhibits robustness; however, the dependency on the primary feature namely, “vpmax (mbar),” minimizes feature variety. Highly accurate, the model’s output is however highly sensitive to this important predictor. Linear Regression shows great performance in the prediction with an MAE of 0.16, MSE of 0.05, and RMSE of 0.22. These metrics prove a direct linear correlation between the features and the target. However, its errors are somewhat larger than in other models, implying that it is not as flexible in dealing with different patterns. Linear Regression provides simplicity and interpretability of results that makes it ideal for use in applications where the process needs to be explained.

Evaluation of the error metrics in Gradient Boosting as compared with Linear Regression shows better performance with an MAE of 0.05, MSE of 0.00, and RMSE of 0.06. This capability means that it is able to retrieve intricate data relationships hence high accuracy and minimal mistakes. Gradient Boosting has been shown to perform better for use cases that demand a degree of sophistication in forecasting. Since they are relatively complex but produce high levels of accuracy, the application of the model can be diverse.

## 6.3 Limitations and Recommendations

Practical changes for better performance while the accuracy of the models is substantiated, several drawbacks may be made for better usage. The random Forest model is highly dependent on the “vpmax (mbar)” and therefore cannot be easily implemented in other datasets. This feature-reliance can decrease the system’s robustness in the future if this feature is untraceable or is of low credibility. Employing the fact that Gradient Boosting is very accurate, the author notes that using it may result in scalability issues, especially in the case of timely predictions. Compared with Linear Regression which is simple and easily interpreted, this algorithm fails to describe non-linear trends well thus slightly higher error level.

Some of the implications of the analysis that can be noted include issues of sample representativeness since the dataset is relatively small for most of the algorithms which may not accurately capture the existence of real-world complications. Some constraints on prediction scope are Temperature changes that are external to the system were excluded, examples being sudden random changes in environmental temperature.

# 

# Analysis And Discussion

## 7.1 Interpretation of Results

The evaluation metrics give the necessary insights into how well each of the models is able to predict the outcomes. Both Random forest and Gradient boosting produced almost an ideal result by minimizing both MAE, MSE, and RMSE while equal to 0.00 and an R² of 1.00. These metrics show that they are adequate at forecasting even nested structural patterns in the dataset. Gradient Boosting has a slightly lower error that testifies to the higher accuracy for precise discriminations. Linear Regression proved efficient with an R² of 1.00 pointing towards the desirable capacity to define target variance. Therefore, while it outperforms some of the models based on these metrics, it can be said that its accuracy is low when it comes to forecasting nonlinear processes. This performance is due to the algorithm which assume a linear model. The reason the Random Forest and Gradient Boosting models are excellent is that they can handle feature importance and nonlinear data patterns.

## 7.2 Comparison with Literature

The findings are consistent with prior literature examining Random Forest and Gradient Boosting approaches, indicating that these techniques perform well in regression tasks. Correia, Peharz, and de Campos, (2020) for instance underlines the fact that Random Forest is an ensemble method that minimizes the risk of over fitting thus increasing the model’s generality. Furthermore, in Singh *eet al.* (2021), the authors pay specific attention to the ability of Gradient Boosting to learn details of patterns through optimization processes. This increases the findings above the performance reported in other similar studies based on rigorous feature selection and data pre-processing (Parashar *et al.,* 2023). For example, the high importance of ‘vpmax (mbar)’, which suggests that vapor pressure does necessarily, and – as it has been determined in studies on atmospheric prediction – strongly affects temperature models (Effrosynidis *et al.,* 2023). Since Gradient Boosting performs better in terms of small errors it is suitable for handling the complex relations in this dataset.

## 7.3 Limitations of Results

These limitations arise from the restriction of the size, quality and generalisability of the dataset used in the research. A reasonable amount of included cases may limit model applicability in other settings. Lack of noise also creates potential noise, or missing data, that leads to introducing biases that may harm the prediction accuracy. There are model-specific constraints such as computational efficiency and overfitting which this model can present. The Random Forest and Gradient Boosting models are highly accurate, though they require more time for computations and are therefore not very scalable. Random Forest tends to overly rely on “vpmax (mbar)”, which may cause overfitting to these primary informative features making model less robust under different conditions. External assumptions can be restrained by actual existence, such as consistently proceeding environment conditions.

## 7.4 Alignment with Objectives

***Objective 1: To collect a suitable weather dataset for weather forecasting***

The main objective is to find reliable information about the weather. This involves data related to temperature, humidity, pressure and clouds formation. Guaranteed Kaggle and Cambridge University weather datasets present various global and local information. Consequently, widening coverage of the dataset to comprise of multiple regions improves the model’s generalization to different weather. Meticulous accuracy, completeness as well as specificity of the collected data will be accorded the highest priority.

***Objective 2: To choose the important weather parameters like cloud formation, humidity, pressure, etc., for developing the weather prediction model***

Knowledge of the weather parameters is inherent in the development of the successful prediction model. As for the weather, it is more or less impacted by effects such as cloud formation, humidity, pressure. Feature selection methods will involve use of statistical measures whereby correlation analysis is likely to be applied in arriving at the most critical features. Limited set of input parameters combined with inclusion of only those in the model enhances model’s predictive capability. This will ensure that the model is only reaching for the most influential variables need for the right weather prediction.

***Objective 3: To choose the best ML algorithm for developing the model***

The select of the right machine learning algorithm is very crucial in determination of the accuracy of a model. The decision will be made with the performance of the Algorithms like Random Forest, Gradient Boosting, and Linear Regression. The learnability of nonlinearity and the feature importance in each algorithm will also be compared. Many changes have been made to simplify the model selection procedure, and the chosen model will be trained, validated, and tested. It is thusly aimed to obtain the highest prediction-error rates while at the same time ensuring clarities of elucidation of the working of the model and its versatility in application or cross-sectional settings.

## 7.5 Application to Real-World Scenarios

The trained models show promising performance for real-world usage, especially for the procedure of generating image captions. It is important to note here that designs like Random Forest as well as Gradient Boosting have commendably low error rate and variance which is why these are used in high precision tasks like the analysis of weather forecasting or the navigation of autonomous automobiles. These models make it possible to achieve an accurate mapping of the weather data, which is essential in all the applications where interpretability is critical. Linear Regression seemed slightly less accurate, while it could be used in real-life cases that do not need much complexity and time.

## 7.6 Addressing the Research Question

In order to address the question, the research identified key weather variables that are important for forecasting. Parameters included in this selection are temperature, humidity, pressure, cloud formation and wind speed. In order to maintain accuracy of the data it is obtained from sources like Kaggle and the University of Cambridge. The important predictors were found by using Exploratory Data Analysis (EDA). Which features best described the temperature were identified from correlation analysis. The three tested machine learning algorithm are Random Forest, Gradient Boosting and Linear Regression. During performance comparison of each algorithm, error metrics including MAE, MSE, RMSE, and R² were calculated. Linear Regression was the least effective when compared to Random Forest and Gradient Boosting. From the findings of the study, it can be concluded that both the models fit the data properly vis-à-vis to the modeled relations between the variables. As a result, the identified variables and algorithms have been used to make accurate weather forecasting models. The study focuses on feature importance to provide with the insights to improve forecasting accuracy.

# 

# Conclusion

## 8.1 Conclusions

The research concludes that accurate weather forecasting depends on specific parameters of weather, which include temperature, humidity, pressure and wind speed. Random Forest and Gradient Boosting models demonstrated to be very good for predicting weather outcomes, providing 1.00 R² scores and very small (MAE, MSE, RMSE) prediction errors in both cases. The study demonstrates that feature selection and data preprocessing are keys to model accuracy. In addition, for such weather data, Linear Regression is not very suitable in capturing non-linear relationship while Random Forest and Gradient Boosting models are more appropriate in learning and capturing non-linear relationship in weather data. In conclusion, the research suggests interesting strategies for creating accurate and robust run of forecasts using machine learning.

## 8.2 Applications and Real-World Situations

The finding could be applied to the development of weather forecasting technology that is more accurate in many industries. Weather reporting in agriculture can guide plant and harvest decisions. Precise forecasting is useful for disaster management to achieve timely evacuation and resources allocation. Weather predicting models also could help aviation and shipping by predicting how weather conditions are going to impact aviation and maritime safety. Urban planning and construction industries can also use weather data to plan infrastructure projects. The machine leaning models built in this research can also be applied to climate research by building our understanding of weather patterns. The applications that I made demonstrate how machine learning can be used to transform industries that depend on accurate weather predictions in making decisions and managing risk.

## 8.3 Future Work

Future work can also be done in increasing model scale to accommodate larger datasets. Further improvement might be achieved by incorporating additional weather parameters. More complex pattern recognition could be presented using deep learning techniques. Real-time weather prediction using live data streams was left for further research. By improving model interpretability, users can be better at understanding the predictions. Satellite and remote sensing data can be integrated into future models for improved forecasting. The generalization can be increased by expanding the models to different geographical regions. Adding such external factors as sudden climate changes can increase prediction reliability. It is also necessary to further optimize computational efficiency for faster predictions. These advancements will integrate them to refine the weather forecasting models and their real applications.

## 8.4 Recommendations

More diverse datasets will help with better weather forecasting models. More features such as cloud cover, wind patterns and geographical data can be included to give better predictions. Robustness requires data from multiple sources integrated. It works by trying to reduce the over fitting by picking relevant features for the model. The model can’t be an outdated version of the real-word data need to be updated regularly. It is important to test models on other regions to evaluate their ability to generalize. This will guarantee scalability so it will be easier to process large datasets with. Moreover, further refinement of the models can be made using high fidelity machine learning methods, such as deep learning. Real time data integration from weather station and satellite can improve the reliability of predictions in the future. The way of the decision-making model should be transparent to improve trust from users. Strategies to connect with meteorologists for deeper insights of the weather parameters are required for forecasting. Finally, resources can be optimized for computational speed in the operational setting.

# Reference List

Badjie, B., Cecílio, J. and Casimiro, A., 2024. Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review. *ACM Computing Surveys*, *57*(1), pp.1-52.<https://dl.acm.org/doi/pdf/10.1145/3691625>

Bag, S., Rahman, M.S., Srivastava, G., Chan, H.L. and Bryde, D.J., 2022. The role of big data and predictive analytics in developing a resilient supply chain network in the South African mining industry against extreme weather events. *International Journal of Production Economics*, *251*, p.108541.<https://researchonline.ljmu.ac.uk/id/eprint/17010/3/The%20role%20of%20big%20data%20and%20predictive%20analytics%20in%20developing%20a%20resilient%20supply%20chain%20network.pdf>

Bochenek, B. and Ustrnul, Z., 2022. Machine learning in weather prediction and climate analyses—applications and perspectives. Atmosphere, 13(2), p.180. [https://www.mdpi.com/2073-4433/13/2/180](https://www.mdpi.com/2073-4433/13/2/180%20)

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I. and Palmer, T., 2021. Machine learning emulation of gravity wave drag in numerical weather forecasting. Journal of Advances in Modeling Earth Systems, 13(7), p.e2021MS002477. [https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002477](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002477%20)

Chattopadhyay, A., Nabizadeh, E. and Hassanzadeh, P., 2020. Analog forecasting of extreme‐causing weather patterns using deep learning. Journal of Advances in Modeling Earth Systems, 12(2), p.e2019MS001958. [https://agupubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1029/2019MS001958](https://agupubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1029/2019MS001958%20)

Chen, L., Han, B., Wang, X., Zhao, J., Yang, W. and Yang, Z., 2023. Machine learning methods in weather and climate applications: A survey. *Applied Sciences*, *13*(21), p.12019. <https://www.mdpi.com/2076-3417/13/21/12019/pdf>

Cho, D., Yoo, C., Im, J. and Cha, D.H., 2020. Comparative assessment of various machine learning‐based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. Earth and Space Science, 7(4), p.e2019EA000740. [https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019EA000740](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019EA000740%20)

Cl.cam.ac.uk, (2024), Accessed from: [https://www.cl.cam.ac.uk/weather/](https://www.cl.cam.ac.uk/weather/%20) [Accessed on: 09/10/2024]

Correia, A., Peharz, R. and de Campos, C.P., 2020. Joints in random forests. *Advances in neural information processing systems*, *33*, pp.11404-11415.<https://proceedings.neurips.cc/paper_files/paper/2020/file/8396b14c5dff55d13eea57487bf8ed26-Paper.pdf>

Donatti, C.I., Nicholas, K., Fedele, G., Delforge, D., Speybroeck, N., Moraga, P., Blatter, J., Below, R. and Zvoleff, A., 2024. Global hotspots of climate-related disasters. *International Journal of Disaster Risk Reduction*, *108*, p.104488.<https://www.sciencedirect.com/science/article/pii/S2212420924002504>

Dwivedi, Y.K., Sharma, A., Rana, N.P., Giannakis, M., Goel, P. and Dutot, V., 2023. Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions. *Technological Forecasting and Social Change*, *192*, p.122579.<https://www.sciencedirect.com/science/article/pii/S0040162523002640>

Effrosynidis, D., Spiliotis, E., Sylaios, G. and Arampatzis, A., 2023. Time series and regression methods for univariate environmental forecasting: An empirical evaluation. *Science of The Total Environment*, *875*, p.162580.<https://www.sciencedirect.com/science/article/pii/S0048969723011968>

Fathi, M., Haghi Kashani, M., Jameii, S.M. and Mahdipour, E., 2022. Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering*, *29*(2), pp.1247-1275.<https://e-tarjome.com/storage/panel/fileuploads/2022-01-15/1642249339_E15948.pdf>

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S. and Hoefler, T., 2021. Deep learning for post-processing ensemble weather forecasts. Philosophical Transactions of the Royal Society A, 379(2194), p.20200092. [https://royalsocietypublishing.org/doi/full/10.1098/rsta.2020.0092](https://royalsocietypublishing.org/doi/full/10.1098/rsta.2020.0092%20)

Han, Y., Mi, L., Shen, L., Cai, C.S., Liu, Y., Li, K. and Xu, G., 2022. A short-term wind speed prediction method utilizing novel hybrid deep learning algorithms to correct numerical weather forecasting. Applied Energy, 312, p.118777. <https://www.researchgate.net/profile/Guoji-Xu-2/publication/358632567_A_short-term_wind_speed_prediction_method_utilizing_novel_hybrid_deep_learning_algorithms_to_correct_numerical_weather_forecasting/links/620c6bf2afa8884cabe7a14b/A-short-term-wind-speed-prediction-method-utilizing-novel-hybrid-deep-learning-algorithms-to-correct-numerical-weather-forecasting.pdf>

Hewage, P., Trovati, M., Pereira, E. and Behera, A., 2021. Deep learning-based effective fine-grained weather forecasting model. Pattern Analysis and Applications, 24(1), pp.343-366. [https://link.springer.com/content/pdf/10.1007/s10044-020-00898-1.pdf](https://link.springer.com/content/pdf/10.1007/s10044-020-00898-1.pdf%20)

Huang, D., Wang, S. and Liu, Z., 2021. A systematic review of prediction methods for emergency management. *International Journal of Disaster Risk Reduction*, *62*, p.102412. <https://ira.lib.polyu.edu.hk/bitstream/10397/90654/1/Huang_Systematic_Review_Prediction.pdf>

Hummel, P., Braun, M. and Dabrock, P., 2021. Own data? Ethical reflections on data ownership. <https://link.springer.com/content/pdf/10.1007/s13347-020-00404-9.pdf>

Jakaria, A.H.M., Hossain, M.M. and Rahman, M.A., 2020. Smart weather forecasting using machine learning: a case study in tennessee. arXiv preprint arXiv:2008.10789. [https://arxiv.org/pdf/2008.10789](https://arxiv.org/pdf/2008.10789%20)

Jaseena, K.U. and Kovoor, B.C., 2022. Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of king saud university-computer and information sciences*, *34*(6), pp.3393-3412.<https://www.sciencedirect.com/science/article/pii/S1319157820304729>

Kaggle, (2024), Accessed from: <https://www.kaggle.com/datasets/thedevastator/weather-prediction/data?select=weather_prediction_dataset.csv> [Accessed on: 09/10/2024]

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P. and Hatfield, S., 2024. Neural general circulation models for weather and climate. *Nature*, *632*(8027), pp.1060-1066.<https://www.nature.com/articles/s41586-024-07744-y>

Lakra, K. and Avishek, K., 2022. A review on factors influencing fog formation, classification, forecasting, detection and impacts. Rendiconti Lincei. Scienze Fisiche e Naturali, 33(2), pp.319-353.<https://link.springer.com/article/10.1007/s12210-022-01060-1>

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W. and Merose, A., 2023. Learning skillful medium-range global weather forecasting. Science, 382(6677), pp.1416-1421. [https://arxiv.org/pdf/2212.12794](https://arxiv.org/pdf/2212.12794%20)

Markets and Markets (2024). *Weather Forecasting Systems Market Size, Share, and Industry Analysis and Market Forecast to 2025 | MarketsandMarketsTM*. [online] www.marketsandmarkets.com. Available at: <https://www.marketsandmarkets.com/Market-Reports/meteorological-weather-forecasting-systems-market-29645152.html>

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D.N., Domeisen, D.I., Feser, F., Koszalka, I., Kreibich, H. and Pantillon, F., 2020. Impact forecasting to support emergency management of natural hazards. *Reviews of Geophysics*, *58*(4), p.e2020RG000704. <https://agupubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1029/2020rg000704>

Parashar, A., Parashar, A., Ding, W., Shabaz, M. and Rida, I., 2023. Data preprocessing and feature selection techniques in gait recognition: A comparative study of machine learning and deep learning approaches. *Pattern Recognition Letters*, *172*, pp.65-73.<https://www.researchgate.net/profile/Imad-Rida/publication/371113483_Data_Preprocessing_and_Feature_Selection_Techniques_in_Gait_Recognition_A_Comparative_Study_of_Machine_Learning_and_Deep_Learning_Approaches/links/6477558ed702370600c50608/Data-Preprocessing-and-Feature-Selection-Techniques-in-Gait-Recognition-A-Comparative-Study-of-Machine-Learning-and-Deep-Learning-Approaches.pdf?origin=journalDetail&_tp=eyJwYWdlIjoiam91cm5hbERldGFpbCJ9>

Petzold, J., Andrews, N., Ford, J.D., Hedemann, C. and Postigo, J.C., 2020. Indigenous knowledge on climate change adaptation: A global evidence map of academic literature. *Environmental Research Letters*, *15*(11), p.113007.<https://iopscience.iop.org/article/10.1088/1748-9326/abb330&title=iopscience.iop.org>

Philosophy & Technology, 34(3), pp.545-572. <https://link.springer.com/content/pdf/10.1007/s13347-020-00404-9.pdf>

Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S. and Thuerey, N., 2020. WeatherBench: a benchmark data set for data‐driven weather forecasting. Journal of Advances in Modeling Earth Systems, 12(11), p.e2020MS002203.[https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020ms002203](https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020ms002203%20%20)

Shouman, E.R.M., 2024. Solar power prediction with artificial intelligence. In Advances in Solar Photovoltaic Energy Systems. IntechOpen.<https://www.intechopen.com/chapters/1152404>

Singh, N., Chaturvedi, S. and Akhter, S., 2019, March. Weather forecasting using machine learning algorithm. In 2019 International Conference on Signal Processing and Communication (ICSC) (pp. 171-174). IEEE. [https://dcs.datapro.in/wp-content/uploads/2023/11/36-Weather-Forecasting-Using-Machine-Learning.pdf](https://dcs.datapro.in/wp-content/uploads/2023/11/36-Weather-Forecasting-Using-Machine-Learning.pdf%20%20)

Singh, U., Rizwan, M., Alaraj, M. and Alsaidan, I., 2021. A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. *Energies*, *14*(16), p.5196.<https://www.mdpi.com/1996-1073/14/16/5196>

Wang, B., Lu, J., Yan, Z., Luo, H., Li, T., Zheng, Y. and Zhang, G., 2019, July. Deep uncertainty quantification: A machine learning approach for weather forecasting. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2087-2095). [https://arxiv.org/pdf/1812.09467](https://arxiv.org/pdf/1812.09467%20)

Weyn, J.A., Durran, D.R., Caruana, R. and Cresswell‐Clay, N., 2021. Sub‐seasonal forecasting with a large ensemble of deep‐learning weather prediction models. Journal of Advances in Modeling Earth Systems, 13(7), p.e2021MS002502. [https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002502](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002502%20)

Yazdani-Asrami, M., Sadeghi, A., Song, W., Madureira, A., Murta-Pina, J., Morandi, A. and Parizh, M., 2022. Artificial intelligence methods for applied superconductivity: material, design, manufacturing, testing, operation, and condition monitoring. *Superconductor Science and Technology*, *35*(12), p.123001.<https://iopscience.iop.org/article/10.1088/1361-6668/ac80d8/meta>

Zaineldeen, S., Hongbo, L., Koffi, A.L. and Hassan, B.M.A., 2020. Technology acceptance model’concepts, contribution, limitation, and adoption in education. Universal Journal of Educational Research, 8(11), pp.5061-5071. <https://www.researchgate.net/profile/Samar-Zaineldeen/publication/345004753_Technology_Acceptance_Model'_Concepts_Contribution_Limitation_and_Adoption_in_Education/links/5f9c0445299bf1b53e514fa8/Technology-Acceptance-Model-Concepts-Contribution-Limitation-and-Adoption-in-Education.pdf>

Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y. and Livingood, W., 2021. A review of machine learning in building load prediction. *Applied Energy*, *285*, p.116452.<https://www.sciencedirect.com/science/article/am/pii/S0306261921000209>