|  |
| --- |
| **COVENTRY**  UNIVERSITY |
| Faculty of Engineering, Environment and Computing |
| School of Computing, Mathematics and Data Science |
| MSc Data Science  EECT109 |
| 7150CEM – Data Science Project |
| Prediction of Heat Wave Using Meteorological Data in UAE: A Machine Learning Approach |
| Author: Wasim Akram Bavigadda |
| SID: 14360095 |
| Supervisor: Dr. Soroush Abolfathi |
| Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science |
| Academic Year: 2023/24 |

**Declaration of Originality**

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

**Statement of copyright**

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see [www.coventry.ac.uk/ipr](http://www.coventry.ac.uk/ipr) or contact [ipr@coventry.ac.uk](mailto:ipr@coventry.ac.uk).

**Statement of ethical engagement**

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (<https://ethics.coventry.ac.uk/>) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: A close-up of a signature

Description automatically generated Date: 01- Aug - 2024

Please complete all fields.

|  |  |
| --- | --- |
| First Name: | Wasim Akram |
| Last Name: | Bavigadda |
| Student ID number | 14360095 |
| Ethics Application Number | P178876 |
| 1st Supervisor Name | Dr. Soroush Abolfathi |
| 2nd Supervisor Name | Dr. Alireza Daneshkhah |

**This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.**

Abstract

The project work uses machine learning to predict heat waves, a critical issue as climate change increases extreme weather. NWP models and statistical approaches can predict some atmospheric dynamics, but not chaotic ones. Heat wave predictions are improved utilising advanced machine learning techniques which includes Random Forest (RF), Gradient Boosting Machines (GBM), and ensemble methods. A detailed literature review of traditional prediction methods and machine learning's ability to surpass them started the project. While traditional NWP models imitate atmospheric processes successfully, long-term and extreme weather predictions are difficult. Regression and time series analysis struggle to capture complex meteorological variable interactions. The literature review emphasised machine learning for prediction accuracy.

The project requires data prep and feature engineering. Historical weather data was meticulously cleaned and normalised. We selected heat wave characteristics including temperature, humidity, wind speed, and air pressure. This improved model performance by ensuring high-quality input data. Gradient Boosting Machines and Random Forest were employed. Random Forest, which can handle large datasets and complex interactions, averaged decision tree outputs to predict accurately. Gradient Boosting Machine, including XGBoost and LightGBM, improved the model by fixing previous problems. The boosting strategy revealed intricate data patterns and dependencies, resulting in accurate predictions.

Ensemble methods integrated model strengths to increase prediction accuracy. Bagging and boosting reduced variance and overfitting by aggregating model predictions. Ensemble methods, especially Random Forest and Gradient Boosting Machines, predicted heat waves better than single-model approaches. Model performance was measured by many metrics. Heat wave days were accurately predicted by the Random Forest model, which had an RMSE of 2.37 and an R-squared of 0.86. Gradient Boosting Machine had 2.29 RMSE and 0.88 R-squared. These results show that machine learning outperforms traditional heat wave prediction.

Project identified further work. Further research should improve data spatial and temporal resolution, incorporate non-meteorological elements as population density and socio-economic vulnerability, and improve model interpretability. Using sophisticated methods like transfer learning and domain adaptation could make the model more applicable across geographies and climate zones. Finally, this experiment showed that machine learning can improve heat wave prediction. The combination of Random Forest, Gradient Boosting Machines, and ensemble methods outperformed traditional methods in prediction accuracy. High-quality data preprocessing, feature engineering, and model interpretability are crucial to predictive model development. This study advances machine learning for climate resilience and catastrophe preparedness, providing insights for researchers, policymakers, and practitioners.

Table of Contents

[Abstract 2](#_Toc173394215)

[Table of Contents 3](#_Toc173394216)

[Acknowledgements 6](#_Toc173394217)

[1 Introduction 7](#_Toc173394218)

[1.1 Background to the Project 8](#_Toc173394219)

[1.2 Project Objectives 11](#_Toc173394220)

[2 Literature Review 13](#_Toc173394221)

[2.1 Traditional Methods for Weather and Heat Wave Prediction 13](#_Toc173394222)

[2.2 Prediction using Machine Learning 14](#_Toc173394223)

[2.3 Emergence of Machine Learning in Climate Science 16](#_Toc173394224)

[2.4 Machine Learning Techniques for Heat Wave Prediction 16](#_Toc173394225)

[2.4.1 Decision Trees and Random Forests 16](#_Toc173394226)

[2.4.2 Gradient Boosting Machines (GBM) 16](#_Toc173394227)

[2.4.3 Ensemble Methods in Heat Wave Prediction 17](#_Toc173394228)

[2.4.4 Feature Engineering and Model Interpretability 17](#_Toc173394229)

[2.5 Existing Studies on Heat Wave Prediction using Machine Learning 18](#_Toc173394230)

[2.6 Research Gap 25](#_Toc173394231)

[3 Methodology 27](#_Toc173394232)

[3.1 Dataset Description 27](#_Toc173394233)

[3.2 Data Pre-processing 28](#_Toc173394234)

[3.3 EDA: exploratory data analysis 29](#_Toc173394235)

[3.4 Training and Testing Data 30](#_Toc173394236)

[3.5 Model Building 31](#_Toc173394237)

[3.5.1 Linear regression 31](#_Toc173394238)

[3.5.2 Decision Tree 31](#_Toc173394239)

[3.5.3 Random Forest 33](#_Toc173394240)

[3.5.4 Gradient Boosting 33](#_Toc173394241)

[3.6 Evaluation Metrics 34](#_Toc173394242)

[4 Requirements 37](#_Toc173394243)

[4.1 Functional Requirements 37](#_Toc173394244)

[4.2 Non-Functional Requirements 38](#_Toc173394245)

[5 Analysis 39](#_Toc173394246)

[6 Implementation 55](#_Toc173394247)

[6.1 Data Preprocessing 55](#_Toc173394248)

[6.2 Defining Heatwave Conditions 56](#_Toc173394249)

[6.3 Exploratory Data Analysis (EDA) 56](#_Toc173394250)

[6.4 Features Engineering 58](#_Toc173394251)

[6.5 Dataset Split 59](#_Toc173394252)

[6.6 Model-building and training 59](#_Toc173394253)

[6.7 Model Evaluation and Comparison 59](#_Toc173394254)

[6.8 Visualization of Results 60](#_Toc173394255)

[7 Project Management 61](#_Toc173394256)

[7.1 Project Schedule 61](#_Toc173394257)

[7.2 Risk Management 62](#_Toc173394258)

[7.3 Quality Management 63](#_Toc173394259)

[7.4 Social, Legal, Ethical and Professional Considerations 63](#_Toc173394260)

[8 Critical Appraisal 64](#_Toc173394261)

[9 Results and Analysis 66](#_Toc173394262)

[9.1 Model Performance 66](#_Toc173394263)

[9.2 Feature Importance 66](#_Toc173394264)

[9.3 Practical Implications 66](#_Toc173394265)

[9.4 Limitations and Future Work 67](#_Toc173394266)

[10 Conclusions 67](#_Toc173394267)

[10.1 Achievements 67](#_Toc173394268)

[10.2 Future Work 68](#_Toc173394269)

[11 Student Reflections 70](#_Toc173394270)

[Bibliography and References 72](#_Toc173394271)

[Appendix A – Project Code 77](#_Toc173394272)

[Appendix B – Certificate of Ethics Approval 84](#_Toc173394273)

[Appendix C – Meeting Logs 8](#_Toc173394273)5

**List of Figures**

[Figure 1 UAE Climate from 1961 to 2020 8](#_Toc173410266)

[Figure 2 Heat Wave Prediction using Support Vector Machine (Das et al., 2023) 19](#_Toc173410267)

[Figure 3 Prediction of Heat Wave using Random Forest (Park et al., 2020) 20](#_Toc173410268)

[Figure 4 Samples in the Dataset 28](#_Toc173410269)

[Figure 5 Flow Diagram 30](#_Toc173410270)

[Figure 6 Linear regression (GeeksforGeeks, 2024) 31](#_Toc173410271)

[Figure 7 Decision Tree (What Is a Decision Tree? | Master’s in Data Science, 2024) 32](#_Toc173410272)

[Figure 8 Random Forest (Yehoshua, 2023) 33](#_Toc173410273)

[Figure 9 Gradient Boosting (Zhang et al., 2021) 34](#_Toc173410274)

[Figure 10 Sample Data 39](#_Toc173410275)

[Figure 11 Shape of Data Frame 39](#_Toc173410276)

[Figure 12 Missing values 40](#_Toc173410277)

[Figure 13 Data Info 40](#_Toc173410278)

[Figure 14 Analysis of Missing Values in the Dataset 41](#_Toc173410279)

[Figure 15 Generated Heat Wave Index 42](#_Toc173410280)

[Figure 16 Distribution of Temperature during Heatwave and Non-Heatwave Periods 43](#_Toc173410281)

[Figure 17 Relative Humidity Variation over Time 44](#_Toc173410282)

[Figure 18 Number of Heatwaves Over Time 44](#_Toc173410283)

[Figure 19 Number of Heatwaves Over Time 45](#_Toc173410284)

[Figure 20 Temperature vs. Relative Humidity during Heatwave and Non-Heatwave Periods 46](#_Toc173410285)

[Figure 21 Heat Wave Visualization in UAE 47](#_Toc173410286)

[Figure 22 Relative Humidity by Month 48](#_Toc173410287)

[Figure 23 Correlation Heatmap of Numeric Variables 49](#_Toc173410288)

[Figure 24 DataFrame After Feature Engineering 50](#_Toc173410289)

[Figure 25 Performance Metrics 50](#_Toc173410290)

[Figure 26 MSE comparison 51](#_Toc173410291)

[Figure 27 RMSE Comparison 51](#_Toc173410292)

[Figure 28 MAE Comparison 52](#_Toc173410293)

[Figure 29 R2Score Comparison 52](#_Toc173410294)

[Figure 30 Data Preprocessing 55](#_Toc173410295)

[Figure 31 56](#_Toc173410296)

[Figure 32 EDA 57](#_Toc173410297)

[Figure 33 EDA 58](#_Toc173410298)

[Figure 34 Feature Engineering 59](#_Toc173410299)

[Figure 35 Dataset Split 59](#_Toc173410300)

[Figure 36 Model Builing 59](#_Toc173410301)

[Figure 37 Model Training and Evaluation 60](#_Toc173410302)

[Figure 38 Visualization of Results 60](#_Toc173410303)

[Figure 39 Gantt Chart (Project Schedule) 61](#_Toc173410304)

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Soroush Abolfathi, for their invaluable guidance and support throughout this project. Their expertise and feedback were crucial in shaping this dissertation.

I am also grateful to Coventry University and the Department of School of Computing, Mathematics and Data Science for providing the necessary resources and facilities for my research.

I extend my heartfelt appreciation to my family for their unwavering support and encouragement throughout my academic journey. Their patience and understanding have been my strength.

Lastly, I acknowledge the contributions of the authors and researchers whose work laid the foundation for this study. Thank you all for your support and assistance in completing this dissertation.

# Introduction

Heat waves are a significant and growing concern in today's world, posing substantial risks to human health, agriculture, infrastructure, and overall societal well-being. Defined as prolonged periods of excessively high temperatures, heat waves can lead to severe consequences, including heatstroke, dehydration, crop failure, and increased energy demands (World Health Organization, 2024). With the ongoing impacts of climate change, the frequency, intensity, and duration of heat waves are expected to rise globally. This is particularly critical in arid and semi-arid regions such as the United Arab Emirates (UAE), where extreme heat is a recurrent and intensifying challenge. The UAE, known for its arid desert climate, experiences some of the highest temperatures recorded on Earth, especially during the summer months. The combination of high temperatures and humidity levels can create dangerously hot conditions, putting immense pressure on public health systems and necessitating advanced preparation and response mechanisms. In such a context, the accurate prediction of heat waves becomes paramount (Feulner, 2023). Effective prediction enables timely warnings, allowing governments, organizations, and individuals to take necessary precautions to mitigate the adverse effects of extreme heat events.

Traditionally, weather prediction has relied on statistical methods and numerical weather prediction models. These methods are useful for weather forecasting, but they struggle to capture complicated and non-linear meteorological data interactions, especially for extreme events like heat waves. Machine learning can transform this. Machine learning trains algorithms to detect patterns and predict from large datasets. Picture and speech recognition, medical diagnosis, and weather prediction are its strengths (Safia et al., 2023). ML models can analyse large meteorological data like temperature, humidity, wind speed, and air pressure. These algorithms learn from past data to uncover patterns and correlations typical approaches miss. Machine learning can analyse complex data and predict heat waves (Saeid Haji-Aghajany et al., 2024).

Machine learning for heat wave prediction can increase UAE climate resilience and disaster preparedness. As the country urbanises, heat waves can affect public health, labour productivity, energy consumption, and water resources. Machine learning algorithms can alert and prevent excessive heat concerns. This study predicts UAE heat waves using machine learning and meteorology. Historical and real-time weather data will be preprocessed. To determine the best heat wave prediction models, machine learning will be evaluated. For a viable prediction system, these models will be examined for accuracy, precision, and recall. This initiative improves heat wave forecast to improve UAE climate resilience. This research will help the UAE and other dry and semi-arid locations with similar difficulties. Meteorology machine learning could improve community safety and climate preparedness.

## Background to the Project

Heat waves, prolonged high temperatures, are a developing problem for health, agriculture, infrastructure, and society. Global warming is increasing heat wave frequency, severity, and length, necessitating better forecasting and control. In hot places like the UAE, this is crucial (Perkins-Kirkpatrick & Lewis, 2020). Considerations include the UAE's environment, heat wave prediction accuracy, traditional weather prediction methods, and machine learning's potential.

The Arabian Peninsula's arid desert climate causes high summer temperatures in the UAE (Paparella & Burt, 2023). Many temperatures exceed 45°C (113°F), and heavy humidity can raise the heat index to dangerous levels. Extreme heat threatens human health, energy use, and the economy. Air conditioning, desalination, and heat-reducing architecture have been used in the UAE to combat excessive temperatures. The acceleration of climate change has caused more irregular and severe heat wave patterns, emphasising the need for better predictive and responsive methods.

A graph showing the average temperature and the average temperature

Description automatically generated

Figure 1 UAE Climate from 1961 to 2020

The graphic shows UAE climatic trends from 1961 to 2020, focusing on average annual temperature (°C) and rainfall (mm). The orange line represents average annual temperatures, which show a general increasing trend from around 26.5°C in the early 1960s to nearly 28°C by 2020. Notable spikes and drops indicate variability, but the overall trajectory points towards rising temperatures, reflecting global warming trends. The blue bars depict annual rainfall, which demonstrates significant fluctuations year to year. There is no clear upward or downward trend in rainfall, though some peaks around the 1990s and 2000s indicate periods of higher precipitation. The combined analysis of temperature and rainfall underscores the increasing temperatures in the UAE, stressing the importance of enhanced climate resilience and adaptive measures to address potential heat waves and their impacts on the region (Hill, 2021). In 2022, the maximum average temperature in the United Arab Emirates had reached 34.5 degrees Celsius. The minimum average temperature for that year reached 22.8 degrees Celsius (Statista, 2024).

Heat waves are dangerous, especially for the elderly, children, and those with pre-existing health concerns. Extreme heat can cause heat stroke, dehydration, and cardiovascular and respiratory problems. Extreme heat can cause these problems (Arsad et al., 2022). The economic effects are significant, influencing labour productivity, cooling energy use, and public health resources. Due to ongoing building and infrastructure projects, outdoor labour is common in the UAE, with serious economic consequences. Extreme heat causes heat stress and poor productivity, which harms workers, delays projects, and raises expenses.

NWP models and statistics are used in traditional weather forecast. NWP models are used for mathematical representations for atmosphere to predict weather conditions using temperature, wind speed, humidity, and atmospheric pressure. Complex equation weather models struggle to represent non-linear and intricate data interactions needed to anticipate extreme occurrences like heat waves (Bi et al., 2023). Regression and time series showed historical linkages. These methods rarely predict sudden and extreme weather events due to their linear assumptions and inability to handle large, complex datasets.

Machine learning (ML) is vital for prediction in many fields due to its benefits over statistical methods. Data learning and pattern adaptation have transformed predictive analytics, providing precise and efficient insights and projections. Machine learning improves prediction accuracy, automation, flexibility, and decision-making for large and complex datasets (Taye, 2023). Machine learning enhances forecast accuracy, making it crucial. ML algorithms detect complicated data patterns and relationships that earlier methods miss. Neural networks, decision trees, and random forests assist ML models forecast correctly. Machine learning algorithms can predict healthcare outcomes like readmission and illness risk by examining huge patient data. These accurate estimates let doctors personalise treatment, improving patient outcomes.

Another predictive advantage of machine learning is its ability to handle large and complex datasets. Traditional statistical methods struggle with high-dimensional data, whereas ML algorithms thrive. When social media and IoT sensor data are abundant, processing and analysing huge datasets is crucial. Machine learning algorithms efficiently manage and extract insights from enormous datasets, enhancing prediction accuracy. In finance, ML models assess stock prices, trade volumes, and economic factors to predict market moves and advise investors (Tufail et al., 2023). Automation is another predictive analytics benefit of machine learning. Once taught, ML systems can read new data and predict without human intervention. Automation frees up time and resources for strategic decision-making over data analysis. By studying customer behaviour, ML models may forecast purchase trends and enhance inventory management in retail. This enhances operations and customer service.

Machine learning prediction requires adaptability. ML models adapt to changing conditions by learning from new data. Dynamic learning maintains accurate predictions. In climate research, machine learning models can adapt to new weather pattern data to enhance forecasts and help communities prepare for severe weather. Adaptability is needed in changing fields (Chen, Chen, et al., 2023). By providing actionable insights and forecasts, machine learning enhances decision-making. Projections mitigate risks and enhance opportunities for organisations. Businesses can use machine learning to predict client attrition and develop retention strategies. Proactive maintenance and reduced downtime are possible with ML models for power plant equipment issues. Machine learning provides accurate and timely forecasts to help companies make strategic decisions that increase growth and efficiency.

Many developments in machine learning model interpretability make predictions easier for stakeholders to understand and trust. Prediction requires machine learning's accuracy, ability to handle large and complex datasets, automation, adaptability, and decision-making support. From healthcare and banking to retail and climate science, it has changed several industries (Kashinath et al., 2021). Predictive analytics will use machine learning to track future events and trends more accurately. Machine learning for prediction gives companies a competitive edge, spurring innovation and improving results.

ML weather prediction is a promising alternative to existing approaches. Artificial intelligence's ML subset trains algorithms to recognise patterns and forecast from massive datasets. For complex and dynamic systems like weather, ML models can manage large volumes of data and discover non-linear relationships better than standard statistical methods (Sarker, 2021). Machine learning works in picture and speech recognition, medical diagnosis, and financial forecasting. ML models can improve weather predictions by analysing historical weather records and real-time observations. Cleaning data to remove inconsistencies and missing values, normalising to ensure scale consistency, and feature engineering to introduce new variables may improve model performance (Fan et al., 2021). Time-series data used in weather prediction may require seasonality adjustment and lag feature generation to capture temporal interdependence.

Some ML algorithms can predict weather, but each has pros and cons. Decision trees make predictions by branching data by feature values. Ensemble random forests enhance accuracy and decrease overfitting by using many decision trees. Interpretable and successful, these approaches capture complicated meteorological data interactions. Support Vector Machine (SVMs) is used for classification and regression. They identify the optimum hyperplane to classify data points. Kernel functions help SVMs model non-linear interactions in high-dimensional spaces. Popular ensemble learning methods like XGBoost and LightGBM use GBMs to construct models sequentially and repair earlier model mistakes. They excel at predictive modelling.

Machine learning heat wave prediction in the UAE could improve climate resilience and catastrophe preparedness. This project uses comprehensive meteorological data and advanced ML algorithms to create accurate and dependable models that can warn of excessive heat and enable proactive heat mitigation. This research addresses an urgent UAE need and advances climate science and predictive modelling.

## Project Objectives

* To preprocess and clean the meteorological data to ensure high-quality inputs for the machine learning models, improving their accuracy and reliability.
* To perform exploratory data analysis (EDA) to identify patterns and trends in heat wave occurrences using historical meteorological data.
* To create a thorough machine learning model using historical meteorological data that can reliably forecast heat waves in the United Arab Emirates.
* To assess and contrast how well different machine learning algorithms such as Decision Trees, Random Forests, Gradient Boosting, and Logistic Regression—perform in the prediction of heat waves.
* To evaluate the models and establish how effective the predictions are by utilising evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) score.

# Literature Review

Severe heat waves are a global issue. Extreme weather impacts health, agriculture, infrastructure, and energy. Global warming from deforestation, industrial pollution, and urbanisation has increased heat wave frequency, intensity, and length. Learn to predict and mitigate heat waves to enhance society. Recent technology and data science developments offer promising solutions to this problem.

## Traditional Methods for Weather and Heat Wave Prediction

NWP models and statistical methods were utilised to forecast weather and heat waves in the past. Meteorology relies on NWP models' complex mathematical equations to simulate atmospheric processes. These models predict weather across time and location utilising temperature, humidity, wind speed, and pressure (Catherine & Tennessee Leeuwenburg, 2023). Although powerful, NWP models struggle to anticipate extreme weather events like heat waves. Nonlinear and chaotic atmospheric dynamics might degrade forecast accuracy over longer lead times or amid complex topography and weather unpredictability.

In contrast, statistics aid weather prediction. From prior weather data, regression and time series analysis uncover trends and build predictive models. Weather regression models predict future weather scenarios using historical data. Time series analysis predicts weather using meteorological temporal patterns (P. Agbo, 2021). These statistical algorithms can predict short- to medium-term weather. Traditional statistical methods have disadvantages. They may struggle to capture atmospheric systems' nonlinear interactions and feedback loops due to linear variable relationships. Modern meteorological data is massive and complex, reducing prediction capacity.

Despite these challenges, NWP models and statistical methods are necessary for weather and heat wave prediction. Processing capacity, data assimilation, and hybrid NWP-statistical modelling increase forecast accuracy. These methods combined with machine learning may improve prediction, especially for extreme weather events like heat waves driven by climate change and urbanisation (Joe et al., 2022). Traditional and new weather prediction methods may improve climate adaption and disaster preparedness worldwide as knowledge and technology evolve.

## Prediction using Machine Learning

Machine learning (ML) has transformed predictive analytics by providing strong tools to anticipate future events and trends from past data. ML models have been used in banking, healthcare, climate research, and marketing to understand patterns and generate accurate predictions. This essay discusses machine learning prediction's fundamentals, methods, and applications, emphasising its transformative promise and obstacles. In machine learning prediction, algorithms learn from data to anticipate future events (Javaid et al., 2022). Starting with data collection and preprocessing, raw data is cleaned, normalised, and formatted for analysis. This stage is critical because data quality and quantity affect model performance. After data preparation, feature selection and engineering begin. Feature selection identifies the most important variables affecting the goal outcome.

Effective feature engineering—creating or modifying features—helps the model capture data patterns. In estimating property values, location, size, and number of rooms are important. The data is separated into training and testing sets after feature selection. The training set builds the model, whereas the testing set tests it. This division makes the model's predictions applicable to new data. Different machine learning methods are employed for prediction jobs, each with pros and cons. Simple but powerful linear regression predicts continuous outcomes. A linear relationship between input features and target variable is assumed. The technique predicts using a linear equation fitted to training data. Linear regression works well in many real-world applications where variables are linearly related despite its simplicity. Decision trees are versatile classification and regression techniques (Huang et al., 2023). They create a tree-like structure by iteratively separating data into feature-value-based branches. Decision trees might overfit yet are simple to understand. Random forests build numerous decision trees and average their predictions to reduce overfitting. These ensemble methods improve the model's accuracy and robustness, making it suited for complicated datasets with many characteristics and interactions. Gradient boosting machines (GBMs) like XGBoost and LightGBM create models consecutively, correcting faults. This boosting method captures complex data patterns and dependencies well. GBMs are used in many contests and applications due to their excellent predictive accuracy. Neural networks, especially deep learning models, are useful for modelling complicated, non-linear interactions. Multiple layers of linked neurons process input data and learn complex patterns in these models. Image identification, natural language processing, and time series forecasting benefit from neural networks. Machine learning prediction has many applications throughout industries, proving its versatility and significance. Finance uses machine learning models for stock price prediction, credit scoring, fraud detection, and algorithmic trading (Kufel et al., 2023). These models can aid decision-making by analysing previous market data and spotting patterns. Healthcare ML models predict disease outbreaks, patient outcomes, and therapy responses.

Medical history and genetic data can be used by predictive models to predict disease development. Personalised medicine and proactive healthcare are possible. Climate science uses machine learning to anticipate weather, extreme occurrences, and climate change. Climate models predict temperature, precipitation, and other variables, aiding disaster preparedness and environmental management. Predictive models forecast demand, optimise pricing, and personalise marketing campaigns by analysing consumer behaviour. Businesses may boost sales and engagement by understanding customer preferences and trends. In manufacturing, ML models analyse sensor data from machines to predict faults and schedule maintenance (Piscitelli & Miani, 2024). This cut downtime, increases equipment life, and boosts efficiency. To ensure reliable and ethical application, machine learning prediction must overcome various hurdles notwithstanding its progress.

Data quality greatly affects prediction accuracy. Noisy, incomplete, or biassed data can mispredict and misinform. Data quality requires careful preprocessing and validation. Overfitting occurs when a model learns training data too effectively, capturing noise rather than patterns. This hinders fresh data generalisation. Cross-validation, regularisation, and ensemble approaches reduce overfitting. Interpreting complex models, especially deep learning networks, is difficult. Understanding how the model predicts is vital for trust and transparency in sensitive fields like healthcare and finance. SHAP values and LIME (Local Interpretable Model-agnostic Explanations) reveal model behaviour. Machine learning models can unintentionally transmit training data biases. Fairness and non-discriminatory predictions require thorough data collection, model training, and evaluation. Computing and handling huge data can be difficult. Cloud computing, distributed processing, and efficient algorithms help scale machine learning models. Machine learning has altered several businesses by providing accurate and actionable projections based on past data (Sandeep Rangineni, 2023). Data preprocessing, feature engineering, model training, and evaluation are the main steps, using many techniques for prediction tasks. Machine learning predictions are useful in banking, healthcare, climate science, marketing, and manufacturing. To ensure ethical and reliable use, data quality, overfitting, interpretability, bias, and scalability must be addressed. Predictive analytics will expand as machine learning evolves, providing new opportunities and insights across industries.

## Emergence of Machine Learning in Climate Science

By enhancing weather prediction and understanding complex environmental systems, machine learning (ML) has altered climate research. ML algorithms can uncover complicated patterns and relationships in meteorological variables like temperature, humidity, and air pressure by processing enormous amounts of satellite and weather station data. Decision trees and neural networks manage weather systems' nonlinear dynamics better than prior methods. Past data helps ML predict extreme weather events like heat waves (Bochenek & Ustrnul, 2022). This is needed for heat wave early warnings and infrastructure and public health mitigation. As ML progresses, climate research is using it to forecast and react to climate change, boosting decision-making and resilience to extreme weather.

## Machine Learning Techniques for Heat Wave Prediction

### Decision Trees and Random Forests

Decision tree algorithms classify and regress. They recursively partition the data into branches depending on feature values to produce a tree-like structure with nodes representing features and branches representing decision rules. Tree leaf nodes make final forecasts. Decision trees are popular because they simplify decision-making. A single decision tree may overfit complex datasets by capturing noise and specific patterns rather than broad trends.

Random Forest prevents overfitting and improves prediction accuracy with numerous decision trees. With random data and features, a random forest trains numerous decision trees. This ensemble method improves prediction accuracy by combining models (Sagi & Rokach, 2020). Results are based on average regression predictions or categorization majority vote of separate trees. Heat wave prediction is best with random forests since they manage complicated data interactions. Combining decision tree predictions in random forests generalises and prevents overfitting. Their ability to manage enormous datasets with varied features assists climate science, which has complex and large data.

### Gradient Boosting Machines (GBM)

XGBoost and LightGBM are effective ensemble learning algorithms with high anticipated accuracy and endurance. Models improve incrementally. Repeating this until the model is optimal creates a trustworthy prediction model with few errors. Because they capture complex data connections and interactions, GBMs predict heat waves better. Heat waves are predicted by these models’ utilising temperature, humidity, wind speed, and air pressure. GBMs can model complex extreme weather patterns because to these properties.

GBMs handle nonlinear data well. The complicated, non-linear interactions of meteorological phenomena make climate study rely on this capacity. GBMs iteratively refine predictions based on earlier models, outperforming solo models that struggle with such complexities (Ali et al., 2023). XGBoost and LightGBM regularise to avoid overfitting, missing data handling, and parallel processing for quicker computations. GBMs are now accurate, efficient, and scalable for climate research's huge datasets. Gradient boosting machines describe complicated meteorological factors and non-linear interactions to predict heat waves. Iterative learning and better features assist climate scientists and data analysts reduce extreme weather.

### Ensemble Methods in Heat Wave Prediction

Ensemble approaches improve advanced machine learning prediction and generalisation. Combining model outputs improves bagging and boosting predictions. Many predictors and modelling methods in ensemble heat wave prediction systems account for meteorological errors and volatility. Bagging, or bootstrap aggregating, randomly chooses the dataset and trains a model on each subset. Final forecast is all models' regression or voting categorization average. This lowers model overfitting and variance. Bagging stabilises heat wave predictions by evaluating many scenarios and reducing anomalies (Usmani et al., 2024). Testing past failures, boosting creates models incrementally. Boosters include AdaBoost, XGBoost, LightGBM, and Gradient. Complex meteorological trends that affect heat waves are best captured by improving the model's forecast accuracy. Ensemble methods improve heat wave predictions using model data. Correct estimations enable heat wave mitigation, therefore public health and infrastructure planning need this. Ensemble methods improve forecast and weather resilience.

### Feature Engineering and Model Interpretability

Improved heat wave prediction machine learning models need feature engineering. By changing meteorological conditions, researchers can improve the model's heat wave predictions. Storm-predicting heat index, diurnal temperature range, and atmospheric stability indicators are important. These qualities help the model understand dynamics and anticipate heat waves. Changing variables to match the model is necessary after picking key characteristics. Normalising data balances feature contributions to prediction, boosting model performance (Fister et al., 2023). Complex variable relationships can be expressed using polynomial features or interaction terms, improving model prediction.

Understanding climate effects on heat wave dynamics demands model interpretability. Permutation feature significance and SHAP values help. Permutation feature importance assesses each feature's model performance impact by prediction accuracy after randomly shuffling feature values. SHAP values indicate model relevance and attribute links. Model interpretability helps stakeholders discover heat wave drivers and customise mitigation. Last, robust and actionable heat wave prediction machine learning models need appropriate feature architecture and model interpretability to improve extreme weather preparedness and resilience. To anticipate heat waves, decision trees, random forests, gradient boosting machines as well as ensemble algorithms capture complicated meteorological data interactions. These algorithms accurately anticipate heat wave onset, severity, and length, crucial for disaster response. Machine learning research will improve heat wave prediction models.

## Existing Studies on Heat Wave Prediction using Machine Learning

This study compares SVR and RF heatwave prediction in Telangana from 1990 to 2019 (Srikanth & Pal, 2023). The study examines model performance in April, May, and June for 7, 15, and 30 days. The average cumulative annual heatwave days are calculated as a time series using IMD temperature data. The similar result was expected for support vector machine and RF models with 7-, 15-, and 30-days lead times. From seven heatwave instances, the SVR model can only record five and the RF model four. The research found that machine learning algorithms SVR and RF are able to forecast Telangana's heatwave days using meteorological factors such relative humidity, u-wind, v-wind, geopotential height, and air temperature. Evaluation measures demonstrate that algorithms perform worse with longer lead times. SVR as well as random forest models performed well for a lead period of up to seven days, with 2.36 and 2.37 RMSE values, respectively. Comparing the two models' performance, SVR is better for this inquiry.

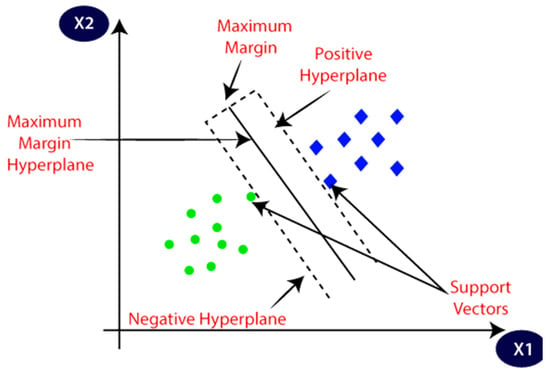


Figure 2 Heat Wave Prediction using Support Vector Machine (Das et al., 2023)

Climate change has increased extremes in temperature, notably heat waves in the last few decades. This work uses random forests as well as decision trees (DT) to forecast Iran's Heatwave Days (HWDs) per year derived from synoptic forecasters. New hybrid technique Ada-Boost Regression and decision trees (ABR-DT) is used. Princeton Meteorological Forcing daily temperature data for 27 sites and National Centres reanalysis data were used for Environmental Prediction. to estimate annual HWDs. Key synoptic characteristics were extracted for four three-month intervals and pressure levels delays. The best structure was identified by minimising predictors and characteristics using Principal Component Analysis. Utilising only particular humidity as well as wind component as predictions, performance assessment based on grid points demonstrated the advantages of ABR-DT with 0.860 for the correlation coefficient (CC) and 6.929 for the mean absolute error (MAE). The superior performance of ABR-DT, which raised the two alternatives' CC and MAE, which are 185 and 19%, was further demonstrated by the geographical performance indices throughout eight distinct climatic zones of Iran. Analysing the consequences of several meteorological components, the study determined ideal mix of parameters to use as heatwave predictors. The outcomes demonstrated how well the suggested hybrid forecasting method predicted heatwave days, a dangerous hazard for numerous areas (Asadollah et al., 2021).

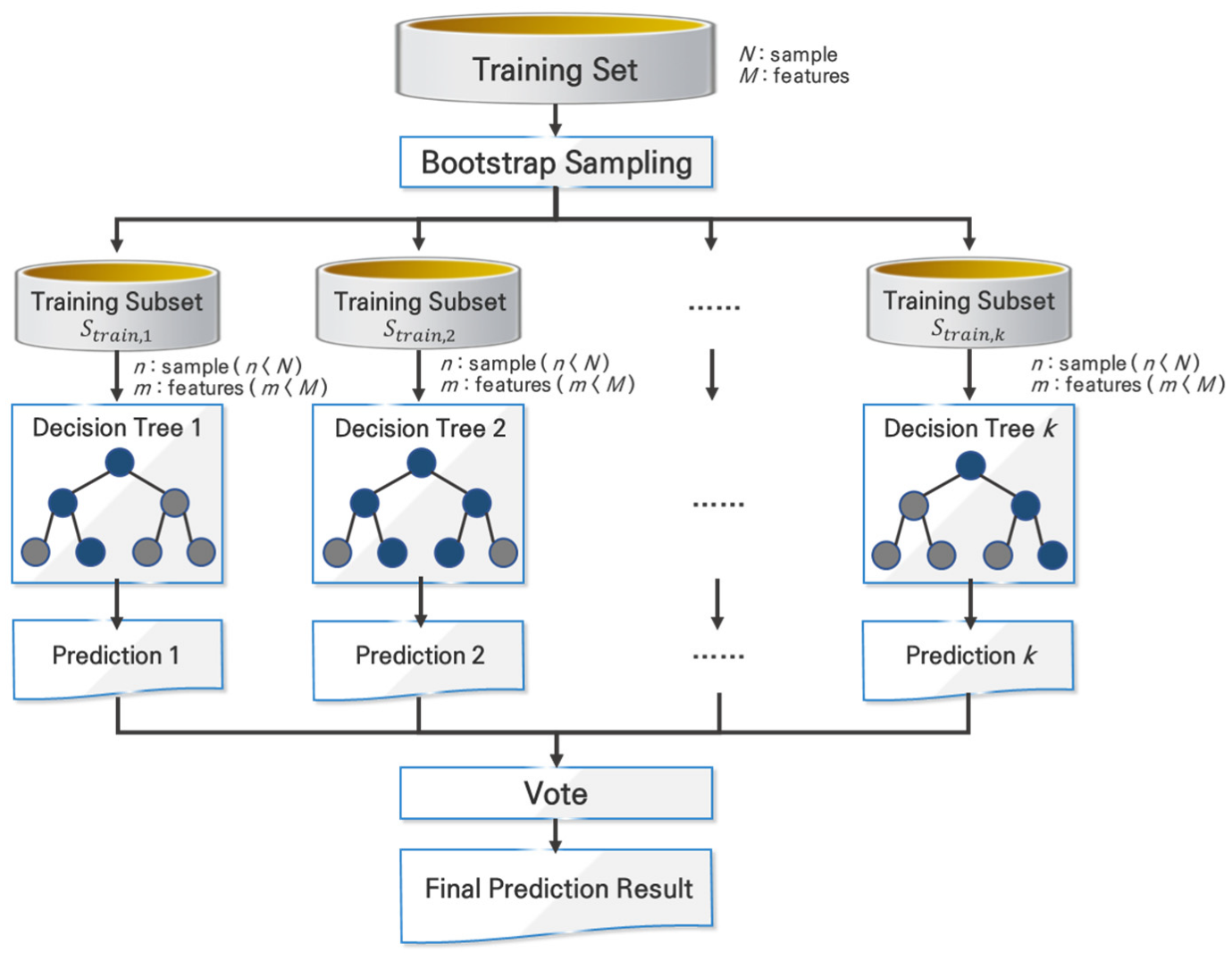


Figure 3 Prediction of Heat Wave using Random Forest (Park et al., 2020)

The Antarctic region is seeing a significant heat flux due to global warming. The glaciers in the Arctic region can be thwarted by high Geo Thermal (GT) heat flow (Dziadek et al., 2021). Many negative consequences, such as an increase in sea level, an impact on the climate that could lead to catastrophic weather occurrences, the extinction of fragile animal species, etc., could arise from the melting of ice in the northern region. In order to forecast GT heat movement in Antarctica, a Radient Boosted Regression Tree approach is employed. Several trustworthy geophysical and geological worldwide datasets were gathered to generate the dataset. The same formula is used to calculate the model score, R2 score, and normalised root-mean-square error (RMSE) of testing. The established prediction technique has a relative RMSE of 29% and an R2 score of 0.44 in testing. The predicted as well as actual values are compared for the model evaluation (Lösing & Ebbing, 2021).

One key element in the prediction of heat waves is land surface temperature (LST), which is measured by satellites. However, cloud disturbances can provide poor LST recording clarity, which can lead to inaccurate heatwave predictions. In order to accurately anticipate heatwaves, a regression model was created to predict the LST. The difference between the actual LST and the expected LST from satellites will be filled in by the LST predicted using the ML model (Buo et al., 2022). Using Ada-Boost Regression and decision trees (ABR-DT) to train, the daily temperature from 27 points was extracted to forecast the number of heatwave days that Iran experiences annually. Four distinct pressure levels were employed to extract the feature variables that were used to train the data. The model had a mean squared error of 6.929 and a corelation coefficient of 0.86 when evaluated grid-point-based. The study determined which feature features best predicted annual heatwave days (Asadollah et al., 2021).

Urban regions experience more heatwaves and their impacts. Urban heat is especially dangerous for the elderly and young. Paris heatwaves killed around 4,000 people in 2003. Predicting greater air temperatures helps early warning systems for extreme heat occurrences. Crowdsourced data of Berlin's rising air temperatures was employed to train the Model Averaged Neural Network, Random Forest (RF), and Stochastic Gradient Boosting. comparative analysis models. With R2 = 0.512, the RF model worked well. The utilisation of open-source datasets allows this new approach to predict extreme temperatures worldwide (Vulova et al., 2020).

To plan for losses, heat wave days must be predicted. Due of constant environmental changes, many heat wave prediction models fail. Changes can have a higher-level impact on these forecast models in the ocean-atmospheric variable. Thus, models must be updated regularly to remain reliable. Using Atmospheric Research reanalysis data, a heat wave day (HWS) prediction model predicted Pakistani heatwave frequency. Project machine learning models included SVM, random forests as well as artificial neural networks. The support vector machine was chosen because it predicted heat wave days better with an R2 value of 0.86. To keep the model up to date with climate change, the predictor-predictor relationship was changed (Khan et al., 2021).

Accurate temperature calculations can predict heatwaves. Using Moroccan Meteorological Administration temperature records, the Time Series TimeDelay Neural Network was trained to reliably predict temperature. The model detected events with corelation coefficient R of 0.99. Actual and modelled temperatures were compared (Anas Kabbori et al., 2019).

Extreme heatwaves caused by the climate crisis have killed thousands. Stopping and reducing high heat episodes is crucial. In a crisis, impulsive decisions can backfire. Using AI-based analytical models, disaster response can be improved. A sophisticated regression and classification algorithm, a Random Forest model, predicted heat wave damage. The dataset was created from South Korea's 2015–2018 floating population, meteorological, and statistical records. The root mean squared error, mean absolute error, root mean squared logarithmic error, and coefficient of determination (R2) were compared to various regression models to evaluate the Random Forest model. When compared to observed data, the proposed model has an R2 value of 0.804. Real-valued outcomes were analysed using loss functions. Random forest model utilised in the research had MAE and MSE loss functions (Park et al., 2020).

Researchers found substantial evidence of an increase in extreme temperature occurrences' frequency and duration. As extreme temperature events become more frequent, societies will need practical and reliable strategies to cope to hotter summers, straining emergency medical services and public health resources. This study developed an effective daily heat-related ambulance call forecasting approach. National and regional models were constructed to evaluate machine learning-based heat-related ambulance call predictions. The regional model had high prediction accuracy in each relevant region and dependable extreme case accuracy, whereas the national model had good forecast accuracy and was applicable to most locations. Heatwave cumulative heat stress, heat acclimatisation, and ideal temperature improved forecast accuracy (Ke et al., 2023). These settings increased the national model's adjusted coefficient of determination (R2) from 0.9061 to 0.9659 and the regional models from 0.9102 to 0.9860. The authors used five bias-corrected global climate models to anticipate national and regional summer heat-related ambulance calls under three future climatic scenarios. Our analysis using SSP-5.85 found that Japan will have over 250,000 heat-related ambulance calls per year by the end of the 21st century, approximately four times the current level. Our research shows that disaster management companies can use this very accurate model to predict when extreme heat events will strain emergency medical supplies. This will enable them to better prepare for potential countermeasures and increase public awareness.

There has been an upsurge in studies on the health effects of excessive heat as a result of more frequent and severe extreme weather events brought on by global warming. However, there is little data on how heatwaves, air quality, and their geographic effects affect the need for health services. The optimised model to forecast health service demand associated with those risk factors for an all-age model was obtained using machine learning (ML) methodologies in this study (Jian et al., 2023) and compared with a model for young children (0–4 years) in Perth. Heatwaves, landscape fires, EDA, SES, and gaseous and particle air pollutants were tracked from 2006 to 2015. Models of geographical random forests (GRF), random forests (RF), and decision trees were compared to find the most accurate model to forecast EDA as well as identifying relevant risk variables. A construct validation was performed using 500 cross validations utilising actual and anticipated EDA data from the testing data. The RF model predicted EDA better than the others, and heatwaves, quality and SES were risk factors. The well-fitting GRF model (R2= 0.975) showed that heatwaves and PM2.5 had substantial spatial variations and a combined effect on young children in the southern suburbs of the research area. When it comes to forecasting the effects of heatwaves, SES, and air quality on EDA, the RF and GRF models perform satisfactorily. There is significant regional variability in heatwaves and air quality. For young kids in the suburbs of the south of Perth, heatwaves, SES, and air quality measurements interacted spatially to determine the most significant predictive risk variables of EDA.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reference** | **Objective** | **Methodology** | **Key Findings** | **Performance Metrics** |
| Srikanth & Pal, 2023 | Assess heatwave prediction using SVR and RF for Telangana | Utilized SVR and RF models with meteorological predictors | SVR and RF models predict heatwave days with varying lead times. SVR outperforms RF for shorter lead times. | RMSE (SVR: 2.36, RF: 2.37) |
| Asadollah et al., 2021 | Develop hybrid model for Iran heatwave prediction | ABR-DT (Ada-Boost Regression + Decision Tree), Principal Component Analysis | ABR-DT shows superior performance in predicting heatwave days for Iran, improving prediction accuracy significantly. | Correlation Coefficient (0.860), MAE (6.929) |
| Lösing & Ebbing, 2021 | Predict GT heat flow in Antarctica | Radiant Boosted Regression Tree (Gradient Boosted) | Developed model predicts GT heat flow in Antarctica with moderate accuracy. | R2 (0.44), Relative RMSE (29%) |
| Buo et al., 2022 | Predict Land Surface Temperature for heatwave forecasting | Regression Model | ML model predicts LST to enhance accuracy in heatwave forecasting, bridging gaps from satellite data. | Not specified |
| Vulova et al., 2020 | Predict urban heatwaves in Berlin | Random Forest, Model Averaged Neural Network, and Stochastic Gradient Boosting | Random Forest model shows good performance in predicting extreme temperatures in urban areas. | R2 (0.512) |
| Khan et al., 2021 | Forecast heatwave days in Pakistan | SVM, RF, Artificial Neural Networks | SVM model performs best in predicting heatwave days in Pakistan, adapting to changing climate conditions. | R2 (0.86) |
| Anas Kabbori et al., 2019 | Forecast temperature in Morocco | Time Series TimeDelay Neural Network (Feedforward NN) | Neural network model predicts temperature with high correlation to actual values in Morocco. | Correlation Coefficient (R ~ 0.99) |
| Park et al., 2020 | Predict heatwave damages in South Korea | Random Forest | Random Forest model effectively predicts heatwave damages using statistical and meteorological data from South Korea. | R2 (0.804) |
| Ke et al., 2023 | Forecast Japan's ambulance calls due to heat | ML-based Models | ML models enhance prediction accuracy of heat-related ambulance calls, vital for disaster management planning. | Adjusted R2 (National: 0.9659, Regional: 0.9860) |
| Jian et al., 2023 | Predict health service demand in Perth | Decision Tree, RF, Geographical Random Forest | RF and GRF models identify significant predictors of health service demand during heatwaves in Perth, showing high predictive performance. | R2 (GRF: 0.975), Spatial Interactions Identified |

## Research Gap

Machine learning (ML) for heat wave prediction has advanced, but various gaps need to be addressed to improve accuracy, robustness, and applicability. Addressing these gaps would increase scientific understanding and proactive heat wave mitigation actions for people and the environment. Spatial and temporal data integration in heat wave prediction machine learning models is a major research gap. Current models emphasise temporal features like short- to medium-term meteorological variable time-series analysis. However, adding spatial data such land use variations, urban heat island effects, and local topography could improve prediction granularity and accuracy (Chen et al., 2023). Satellite imaging and GIS spatial elements could be used with temporal data to capture localised heat wave dynamics in future studies.

Heat wave models should include non-meteorological factors that affect heat wave consequences, even though temperature, humidity, and wind speed are the main predictors. Population density, demographics, infrastructural resilience, and socio-economic vulnerability affect heat wave severity and social impacts (Dóra Szagri et al., 2023). Integrating these parameters into ML models could improve heat wave risk understanding and enable targeted measures to protect susceptible people. Heat wave prediction still struggles with ML model interpretability. Gradient boosting machines and neural networks have great predictive accuracy, but stakeholders must understand how they make predictions to trust and act on them. Feature importance analysis, sensitivity analysis, and visualisation should be developed for model interpretability in future study. Explainable AI methods like SHAP (SHapley Additive exPlanations) values and local interpretable model-agnostic explanations (LIME) could reveal how variables affect heat wave predictions, improving model transparency and confidence (Lu et al., 2023).

Data quality and availability hinder ML-based heat wave prediction. Weather station and satellite data is available; however, data gaps, inconsistencies, and biases might affect predictive models. Integration of disparate environmental sensor, health record, and socioeconomic data is also difficult (Barriopedro et al., 2023). Future research should improve data gathering methods, implement rigorous quality assurance systems, and study data fusion and crowdsourcing to broaden and robust ML model input data. Research also focuses on ML model transferability and generalisation across locations and climate zones. Training models in specific places or times may not adapt to future climates. Transfer learning and domain adaptation should adjust ML models to multiple climates and geographies (Jayan Wijesingha et al., 2024). Validating model performance across geographies and climate zones can help minimise global climate change by building universal heat wave prediction frameworks. Researchers, policymakers, and community stakeholders must collaborate on ML heat wave predictions. Develop predictive model-based decision support systems to apply science to real-world challenges (Bibri et al., 2023).

These machine learning heat wave prediction research gaps must be addressed to improve scientific understanding, predictive accuracy, and social resilience to climate change-induced extreme weather events. Through interdisciplinary collaboration, advanced modelling, and stakeholder engagement, researchers can create robust, reliable, and actionable heat wave prediction frameworks that support proactive climate adaptation and disaster risk reduction strategies worldwide. Climate warming risks heat waves, needing advanced forecast and quick response. While important, standard weather forecasts rarely predict extreme events like heat waves. Several machine learning methods can find non-linear meteorological correlations in large, complex datasets. ML heat wave prediction, especially in the UAE, can improve forecast accuracy and early warnings, boosting climate resilience and disaster preparedness.

# Methodology

## Dataset Description

The dataset includes meteorological observations, being collected from different stations, where each station having a specific code and publishing data on particular dates with UTC format. Every record has spatial information as longitude, latitude and after the elevation for offering you a precise location of where weather records were measured. The essential meteorological parameters are temperature (in Fahrenheit), dew point temperature, relative humidity (%) wind dirdeg), Wind speed (kts), and precipitation (inches over different intervals. Both altimeter setting and sea-level pressure in millibars represent atmospheric pressure. Miles per hour is a measure of speed, but also miles (in land travel in daylight conditions). Therefore, it would not be proper to use for visibility. The dataset is extracted from January 2000 to June 2024.

**Link:** Data extracted from official [website](https://mesonet.agron.iastate.edu/request/download.phtml?network=AE__ASOS)

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| station | Station code or identifier |
| valid | Date and time of the observation in UTC format |
| lon | Longitude of the observation location |
| lat | Latitude of the observation location |
| elevation | Elevation of the observation location in meters |
| tmpf | Temperature in degrees Fahrenheit |
| dwpf | Dew point temperature in degrees Fahrenheit |
| relh | Relative humidity as a percentage |
| drct | Wind direction in degrees (meteorological) |
| sknt | Wind speed in knots |
| p01i | One-hour precipitation in inches |
| alti | Pressure altimeter setting in inches of mercury |
| mslp | Sea-level pressure in millibars |
| vsby | Visibility in miles |
| gust | Wind gust speed in knots |
| skyc1 | Sky coverage layer 1 |
| skyc2 | Sky coverage layer 2 |
| skyc3 | Sky coverage layer 3 |
| skyc4 | Sky coverage layer 4 |
| skyl1 | Sky level altitude for layer 1 in feet |
| skyl2 | Sky level altitude for layer 2 in feet |
| skyl3 | Sky level altitude for layer 3 in feet |
| skyl4 | Sky level altitude for layer 4 in feet |
| wxcodes | Present weather codes |
| ice\_accretion\_1hr | Ice accretion over the past hour in inches |
| ice\_accretion\_3hr | Ice accretion over the past 3 hours in inches |
| ice\_accretion\_6hr | Ice accretion over the past 6 hours in inches |
| peak\_wind\_gust | Peak wind gust during the day in knots |
| peak\_wind\_drct | Peak wind gust direction in degrees (meteorological) |
| peak\_wind\_time | Time of peak wind gust in UTC format |
| feel | "Feels like" temperature in degrees Fahrenheit |
| metar | Aviation routine weather report (METAR) string |
| snowdepth | Snow depth in inches |

Sky coverage is divided into four layers, each by its coordinating height in feet. Additional features like current weather codes, ice accretion measurements for multiple time periods (expressed in inches), highest wind gust information including speed, direction and timestamp of the event, "feels like" temperature meteorological reports (known as METARs) used by aviation assets and snow depth (in inches). This makes it possible to scrutinize and model weather patterns in various locations over different time periods.

A screenshot of a computer

Description automatically generated

Figure 4 Samples in the Dataset

## Data Pre-processing

Pre-processing raw data for analysis is essential. Handling missing values, switching data types, and adding features improves predictive modelling dataset usability. We began this effort by identifying columns with plenty of missing data. To eliminate noise and streamline the dataset, columns with high missing values were removed. Missing values were imputed using column mean for columns with fewer missing data. This keeps the dataset complete without bias. To ensure dataset consistency, missing data entries were converted to numbers. To enable time-based analysis, the timestamp column was transformed to datetime. This conversion extracted temporal characteristics like year, month, day, hour, and minute, which could help explain heatwave patterns.

## EDA: exploratory data analysis

Understanding data patterns and correlations requires exploratory data analysis (EDA). EDA defined and visualised heatwaves in this project. Heatwaves were defined by temperature, humidity, and wind speed thresholds. Our new tool, 'heatwave\_index,' measures heatwave intensity over hours. We estimated the excess temperature, relative humidity, and wind speed deficit over consecutive hours using the data. These results formed the 'heatwave\_index,' which quantified heatwave intensity. Visualisations were crucial to EDA. Histograms showed temperature range discrepancies during heatwaves and non-heatwaves. Time series plots showed relative humidity and daily heatwave hours. The plots marked significant heatwaves using thresholds. We used scatter plots to compare temperature and relative humidity during heatwaves with non-heatwaves. under addition, a correlation heatmap was created to show how meteorological factors interact under different weather circumstances.

A diagram of a process

Description automatically generated

Figure 5 Flow Diagram

## Training and Testing Data

Data was ready for modelling after pre-processing and EDA. Setting the goal variable, 'heatwave\_index,' and splitting the dataset into training and testing sets. This split prevents overfitting and assures generalisability by training the model on a large piece of the data and evaluating it on another. Different regression models were investigated, including Linear, Decision Tree, Random Forest, and Gradient Boosting. These models were chosen for their versatility and complicated relationship capturing.

MSE, MAE, RMSE, and R2 Score were used to evaluate the models on the test set after training on the training set. These measures assess each model's prediction accuracy and reliability comprehensively. Bar charts showed model performance across metrics. This visual method helped choose the best heatwave model for the dataset.

## Model Building

### Linear regression



Figure 6 Linear regression (GeeksforGeeks, 2024)

Basic linear regression models the relationship between a dependent variable and one or more independent variables. The main purpose of linear regression is to fit a linear equation to observed data to predict the dependent variable based on the independent variables.

Linear regression minimises the sum of the squared differences between observed and predicted values by finding the best-fitting line between the data points. Metrics like MSE, RMSE, and R-squared (R²) score assess the model's efficacy in explaining variance in the dependent variable. Linear regression is popular for modelling linear connections due to its simplicity, interpretability, and efficiency.

### Decision Tree

For heat wave prediction, decision trees play a significant role in enhancing the accuracy and proactive mitigating strategies of forecasting. Every node in the tree depends on a judgment over one of its features, which is then followed by the branches recursively refining predictions. Decision trees make more informed decisions and interventions to reduce heat waves.

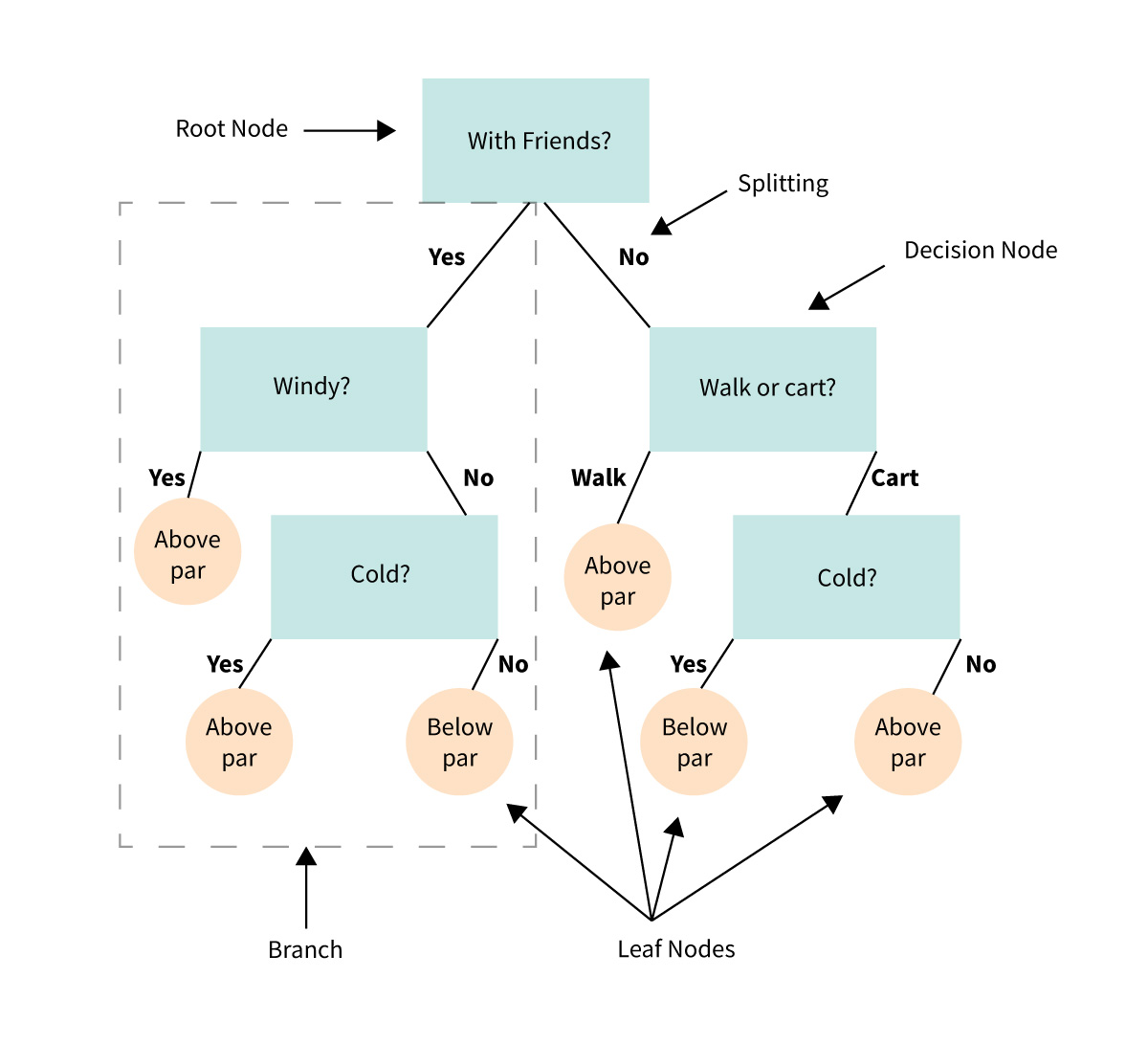


Figure 7 Decision Tree (What Is a Decision Tree? | Master’s in Data Science, 2024)

Decision tree methods are good in capturing complicated node interactions by recursively splitting data into subgroups based on the optimal features at each internal mode. Since decision trees can handle mixed data types and missing values, there are few to none preparations that need to be done prior using them. They can partition the data by any attribute, which makes them useful for unifying different meteorological datasets into models to predict heat waves. This flexibility reduces the amount of data conversion which in turn can increase prediction accuracy as weather information is extremely volatile and complicated (used for climate science research) so it will be essential to streamline Data Preparation operations. To improve prediction accuracy of extreme heat and precede resilience planning for an era where demand response may be insufficient, consideration should also be given to formally study decision tree models.

### Random Forest

Random Forest algorithms provide precise heat wave predictions and proactive prevention. Multiple decision trees improve forecast accuracy in Random Forests. For heat wave prediction, Random Forests can capture complex relationships between meteorological factors like temperature, humidity, wind speed, and pressure using this ensemble approach. Random Forests aggregate predictions from numerous trees to reduce tree biases and errors, increasing spatial and temporal heat wave forecasts.

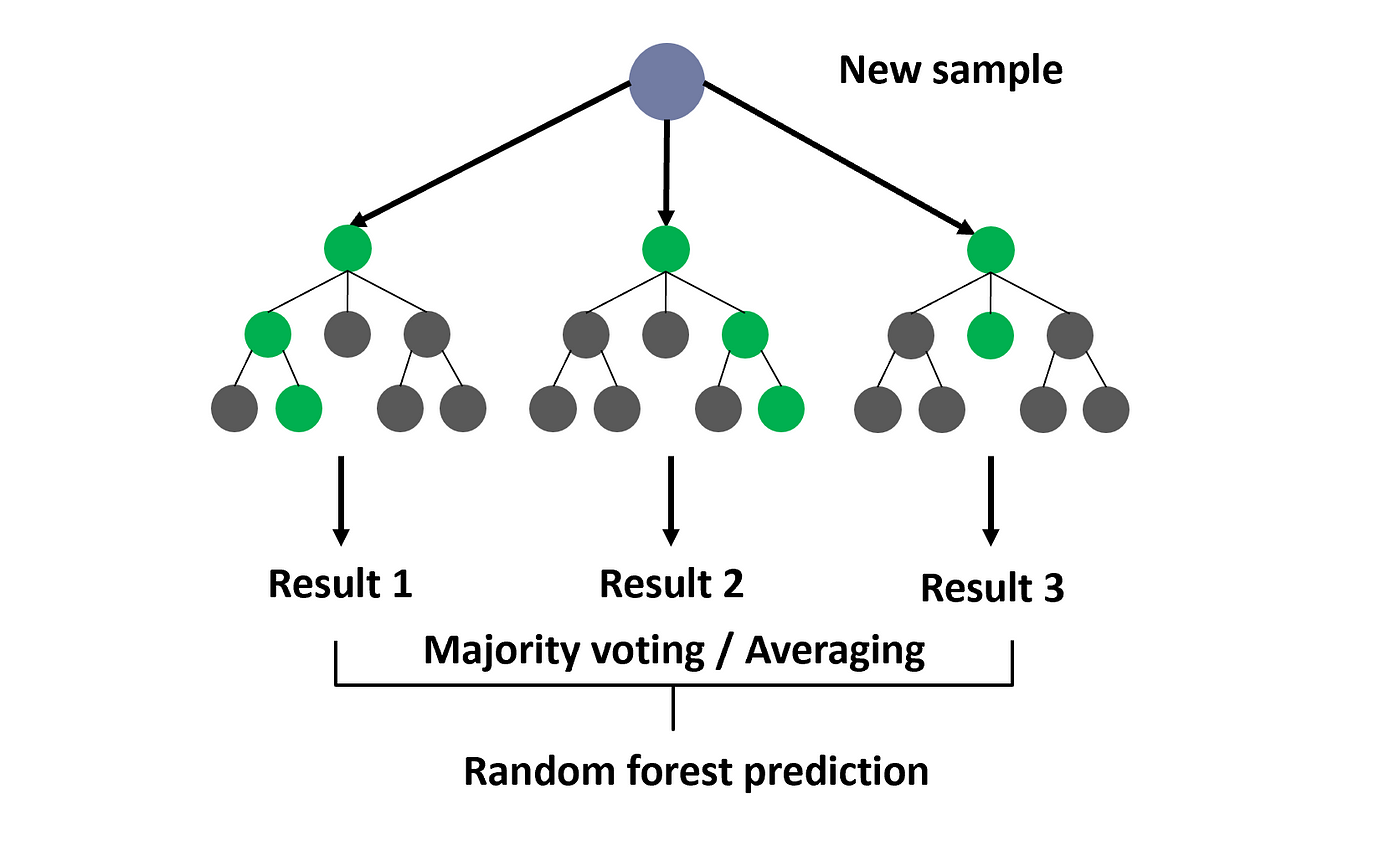


Figure 8 Random Forest (Yehoshua, 2023)

Heat waves result from complex, non-linear weather. Recursively partitioning data by the strongest discriminative attributes at each decision tree node helps Random Forests capture complex relationships. Random forests can model meteorological data's non-linear correlations and interactions, unlike linear models. This ability is needed to estimate heat wave start, length, and intensity in diverse climates and places.

### Gradient Boosting

Complex nonlinear meteorological interactions cause heat waves. To model this complexity, GBMs build an ensemble of decision trees that capture unique data variability and reliance. GBMs can handle meteorological variations and non-linear interactions, unlike linear models. GBMs predict heat wave dynamics better across spatial and temporal dimensions because they adapt to varied climates and areas. GBMs regularise and ensemble learn to reduce overfitting. GBMs improve model generalisation and reduce variance by combining predictions from numerous weak learners (usually shallow decision trees), making meteorological data noise- and outlier-resistant. Its robustness lets the model generalise to new data and predict heat wave situations and environmental circumstances.

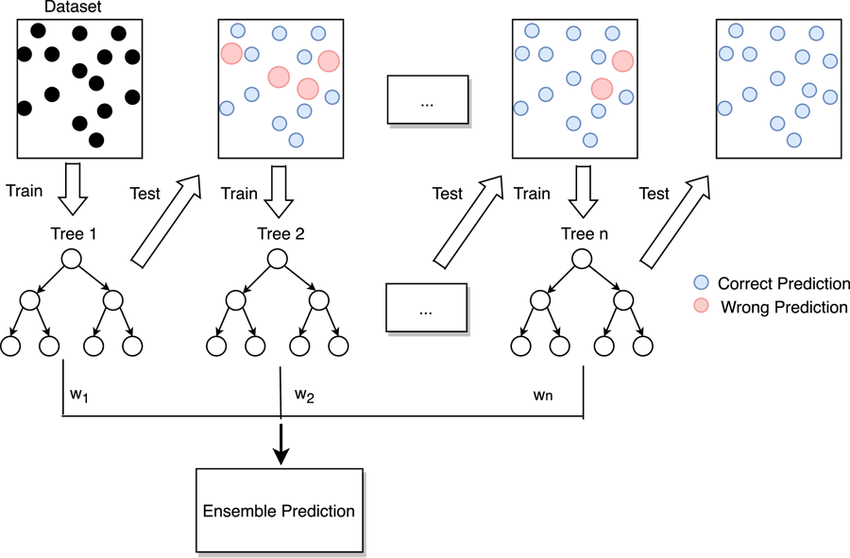


Figure 9 Gradient Boosting (Zhang et al., 2021)

Assessing meteorological variables' value in heat wave prediction improves adaptation and resource allocation. GBM relevance is determined via optimisation-based feature contribution ratings. By quantifying each predictor variable's impact on the model's prediction abilities, these ratings help meteorologists and policymakers discover the most important meteorological aspects. Transparent feature selection improves heat wave predictions and evidence-based climate resilience planning.

## Evaluation Metrics

Machine learning models, especially heat wave predictions, need evaluation criteria. MSE, MAE, RMSE, and R² Score are often used metrics to evaluate model predictive capacity and dependability.

**Mean Squared Error (MSE)**

Basic statistics like MSE calculate the average squared difference between expected and actual dataset values. Average squared difference between expected and actual:

where:

- represents the actual value of the target variable,

- represents the predicted value of the target variable,

- is the number of data points.

Squaring makes MSE sensitive to outliers and departures from projected values since it weights significant errors more.

**Mean Absolute Error (MAE)**

Mean Absolute Error is the average absolute difference between projected and actual values.

MAE provides a more interpretable metric compared to MSE because it measures average prediction error without squaring the differences. It is less sensitive to outliers and gives equal weight to all errors.

**Root Mean Squared Error (RMSE)**

Root Mean Squared Error is the square root of the MSE and represents the standard deviation of the residuals (prediction errors):

RMSE is preferred when the magnitude of errors is critical and needs to be expressed in the same units as the target variable. Like MSE, RMSE also gives higher weight to large errors.

**R² Score (Coefficient of Determination)**

R² Score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1 and indicates how well the model fits the data:

where:

- is the mean of the observed data.

R² Score of 1 indicates a perfect fit, where the model explains all the variability of the target variable around its mean. A 0 score means the model explains no variability. These evaluation measures are essential for analysing heat wave prediction models' meteorological variable-based accuracy. MSE, MAE, and RMSE measure prediction errors' amount and direction, improving model parameters and forecasting accuracy. R² Score indicates model goodness-of-fit, showing heat wave variability can be captured using selected predictors. These evaluation metrics allow researchers and meteorologists to validate model performance, compare algorithms, and make informed decisions to improve heat wave prediction models to mitigate the negative effects of extreme heat events on society and the environment.

# Requirements

## Functional Requirements

|  |  |  |
| --- | --- | --- |
| ID | Requirement Description | Priority |
| FR1 | The system shall collect and process meteorological data from various stations in UAE. | High |
| FR2 | The system shall identify and handle missing values in the dataset. | High |
| FR3 | The system shall convert 'M' values to NaN and impute missing data with the column mean. | High |
| FR4 | The system shall convert the 'valid' column to datetime format for accurate time-based analysis. | High |
| FR5 | The system shall extract temporal features such as year, month, day, hour, and minute from the datetime column. | Medium |
| FR6 | The system shall define heatwave conditions using specific thresholds for temperature, relative humidity, and wind speed. | High |
| FR7 | The system shall generate a 'heatwave\_index' to measure the intensity of heatwaves over consecutive hours. | High |
| FR8 | The system shall perform exploratory data analysis (EDA) to identify patterns and trends in the data. | High |
| FR9 | The system shall visualize data distributions and correlations using histograms, scatter plots, and heatmaps. | Medium |
| FR10 | The system shall split the dataset into training and testing sets for model evaluation. | High |
| FR11 | The system shall train multiple machine learning models (Linear Regression, Decision Tree, Random Forest, Gradient Boosting) on the training dataset. | High |
| FR12 | The system shall evaluate model performance using MSE, MAE, RMSE, and R² Score. | High |
| FR13 | The system shall compare model performance and select the best model for heatwave prediction. | High |
| FR14 | The system shall visualize the results of model performance metrics using bar charts. | Medium |
| FR15 | The system shall generate geographic visualizations of heatwave conditions in the UAE using maps. | Medium |
| FR16 | The system shall produce a final report detailing the methodology, results, and conclusions of the heatwave prediction study. | High |

## Non-Functional Requirements

|  |  |  |
| --- | --- | --- |
| ID | Requirement Description | Priority |
| NFR1 | The system shall ensure data integrity by accurately processing and handling missing values. | High |
| NFR2 | The system shall be scalable to handle large volumes of meteorological data efficiently. | High |
| NFR3 | The system shall ensure data processing and model training are completed within a reasonable timeframe (e.g., less than 1 hour for large datasets). | Medium |
| NFR4 | The system shall provide visualizations that are easy to interpret and provide meaningful insights. | Medium |
| NFR5 | The system shall maintain high accuracy and reliability in heatwave predictions, with an R² Score of at least 0.90 for the selected model. | High |
| NFR6 | The system shall be user-friendly, with clear documentation and instructions for users to follow. | Medium |
| NFR7 | The system shall ensure the security and confidentiality of the data being processed and stored. | High |
| NFR8 | The system shall be adaptable to incorporate new data and adjust model parameters as needed for continuous improvement. | Medium |
| NFR9 | The system shall comply with relevant data protection and privacy regulations, such as GDPR. | High |
| NFR10 | The system shall provide robust error handling and logging to facilitate troubleshooting and maintenance. | Medium |

# Analysis

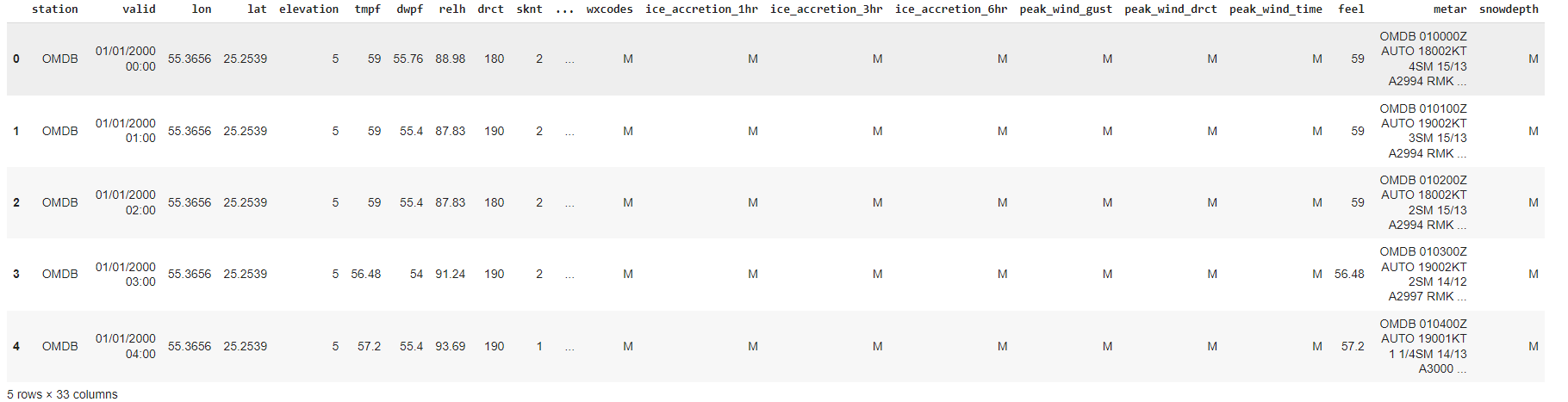


Figure 10 Sample Data

A station named 'OMDB' provided 234,486 meteorological observations in 33 columns. Initial analysis of the dataset shows multiple 'M' columns with missing values. The dataset includes numeric and object data types, including meteorological metrics like temperature (tmpf), dew point temperature (dwpf), relative humidity (relh), wind direction (drct), and wind speed (sknt). The columns additionally provide location (lon, lat), elevation, and time (valid).



Figure 11 Shape of Data Frame

To maintain data quality, we dropped columns with many missing values during data pre-processing. The columns 'gust', 'p01i','skyl2','skyl3','skyl4', 'wxcodes', 'ice\_accretion\_1hr', 'ice\_accretion\_3hr', 'ice\_accretion\_6hr', 'peak\_wind\_gust', 'peak\_wind\_drct', 'peak\_wind\_time', and'snowdepth This phase was essential to clean up the dataset and remove noise that could impact the model. To prevent losing crucial information, we transformed the remaining columns with 'M' values to NaN and imputed the missing values using the column mean.  
  
The 'valid' column, which contains timestamps, was transformed to datetime for analysis. This conversion extracted year, month, day, hour, and minute temporal information, which may help explain weather trends. To ensure dataset consistency, appropriate columns were changed to numeric types.

A screen shot of a computer code

Description automatically generated

Figure 12 Missing values

A screenshot of a computer program

Description automatically generated

Figure 13 Data Info

A screenshot of a computer code

Description automatically generated

Figure 14 Analysis of Missing Values in the Dataset

The graphic illustrates a code snippet that counts 'M' values, which reflect missing data in the dataset. The isin function finds all columns with 'M' values, and the value\_counts method iterates over them to count their 'M' values. The results show that numerous columns have many 'M' values. Tmpf (temperature) and dwpf (dew point temperature) columns have 442 and 604 'M' values, which are manageable. Other columns like drct (wind direction) and sknt (wind speed) had 20,639 and 4,999 'M' values, respectively. Some columns, such p01i, have 53,491 'M' values, indicating a lot of missing data.

More importantly, mslp, gust, and sky cover columns (skyc1, skyc2, skyc3, and skyc4) have 'M' values that often approach 200,000 entries. These columns almost totally contain missing data, making them candidates for removal to ensure data quality and integrity. Ice\_accretion\_1hr, ice\_accretion\_3hr, ice\_accretion\_6hr, peak\_wind\_gust, peak\_wind\_drct, peak\_wind\_time, and snowdepth have 'M' values in all 234,486 entries. These columns are useless and must be eliminated from the dataset. Addressing missing data is crucial, as this analysis shows. Keep columns with tolerable missing values and impute them, but reject those with too much missing data to improve analysis and model performance.

A screenshot of a computer

Description automatically generated

Figure 15 Generated Heat Wave Index

After generating the heatwave index, the dataframe includes many meteorological features and a new column, heatwave\_index. This indicator measures heatwave intensity over hours using temperature, relative humidity, and wind speed. The valid column was transformed to datetime for precise time-based analysis. To protect data integrity, tmpf, dwpf, relh, and sknt columns have been transformed to numeric values and coerced to NaN for errors.

A graph showing the temperature of a heatwave

Description automatically generated

Figure 16 Distribution of Temperature during Heatwave and Non-Heatwave Periods

The graph shows heatwave and non-heatwave temperatures in degrees Fahrenheit on the x-axis and density on the y-axis, which shows frequency of occurrences at different temperatures. The graph distinguishes heatwaves (Yes) and non-heatwaves (No) with two colours. Temperatures range from 30°F to 120°F, with most values between 50°F and 110°F. When there is no heatwave, temperatures are more evenly distributed across this range, with higher density at lower temperatures, indicating cooler temperatures are more common. Conversely, heatwaves occur at higher temperatures with high density around and above 95°F, meeting the heatwave threshold. The two distributions overlap in the mid-temperature range (70°F to 90°F), demonstrating that these temperatures can occur during heatwaves and non-heatwaves, likely driven by humidity and wind speed. The heatwave distribution has a fast density increase at higher temperatures and a steep decrease past the peak heatwave temperature threshold, while the non-heatwave distribution is more spread out with a gradual density decline. This graph shows that heatwaves are more common at higher temperatures and that densities separate at higher temperatures, validating the threshold criteria for heatwaves.

A graph showing a number of orange lines

Description automatically generated with medium confidence

Figure 17 Relative Humidity Variation over Time

The graph shows relative humidity from 2000 to 2024, with the x-axis showing time and the y-axis relative humidity %. The orange data points show considerable humidity swings over this period. A humid atmosphere is indicated by high relative humidity, which typically reaches 100%. However, there are noteworthy drops towards 0%, indicating reduced humidity due to dry weather or meteorological occurrences. No significant long-term trends or fluctuations in relative humidity indicate that humidity levels in this region have stayed largely steady. The graph does not segment seasonal swings, although the data shows that some times may indicate wetter or drier seasons. A complete dataset and reliable analysis are guaranteed by continuous data point density. This depiction of humidity variations over time illustrates a mostly humid climate with occasional dry periods, which aids weather pattern analysis, agricultural planning, and humidity impact measurements.

A blue and white sound wave

Description automatically generated

Figure 18 Number of Heatwaves Over Time

From 2000 to 2024, the x-axis shows the date and the y-axis heatwave hours. This heatwave intensity and frequency time series graphic demonstrates changes. In the early 2000s, benign heatwave hours were punctuated by brief severity spikes. Peaks in 2005 and 2006 indicate heatwaves. Peak heatwaves lasted hours. After 2010, the graph indicates periodic peaks but fewer heatwave hours than mid-2000s. This suggests milder or shorter heatwaves occurred. As 2020 approaches, heatwaves intensify and climax around 2023. This may imply climate change generating more frequent or violent heatwaves. The graph displays heatwave patterns over the previous two decades, showing times of increased activity and climatic variables that may be making them oscillate. Understanding long-term climate patterns and heatwave mitigation require this study.

A blue and white graph

Description automatically generated

Figure 19 Number of Heatwaves Over Time

The x-axis shows the date and the y-axis heatwave hours from 2000 to 2024. Blue line indicates heatwave index, red dashed is threshold. As illustrated in this time series figure, heatwave intensity and frequency vary. Early 2000s heatwaves were mild with occasional spikes, indicating short-term intensity increases. Peaks in 2005 and 2006 indicate long-lasting heatwaves. After 2010, heatwaves occur intermittently but less than in the mid-2000s. This suggests that heatwaves occurred but were milder or shorter. Heatwaves rise in frequency and intensity as 2020 approaches, peaking around 2023. This recent spike may imply climate change causing more frequent or stronger heatwaves.

The red dashed heatwave threshold line helps identify significant heatwaves. When the blue line crosses this level, heatwaves are intense. The graph shows repeated heatwave index exceedances, verifying heatwave incidents. This graph shows the evolution of heatwave patterns over the past two decades, showing periods of increasing activity and proposing climate changes that may be contributing to these oscillations. This analysis is essential for understanding long-term climate patterns and planning for heatwave mitigation.

A graph showing temperature vs temperature

Description automatically generated

Figure 20 Temperature vs. Relative Humidity during Heatwave and Non-Heatwave Periods

The scatter plot compares heatwave and non-heatwave temperature (in degrees Fahrenheit) and relative humidity (in %). Relative humidity is on the x-axis and temperature on the y. Each graph point represents an observation, blue for non-heatwave periods and red for heatwaves. Data points form clusters in the plot. Non-heatwave periods, shown in blue, cover several temperatures and humidity levels. The density of these spots between 60°F and 100°F and across the relative humidity spectrum shows that non-heatwave circumstances can occur in diverse temperature and humidity settings. The distribution implies that non-heatwave periods have cooler temperatures and higher density as relative humidity rises.

Heatwaves, illustrated in red, are hotter, usually above 90°F, and have higher relative humidity, frequently above 40%. This clustering suggests heatwaves involve high temperatures and humidity. The plot's upper right red spots are tightly packed, illustrating heatwave conditions' high temperature and humidity. The overlap of blue and red marks denotes transitional moments when temperatures and humidity levels are nearing heatwave conditions. Red spots above 95°F and higher relative humidity highlight heatwaves' essential circumstances.

The scatter plot clearly shows heatwave and non-heatwave temperature and humidity patterns. It shows that heatwaves are related with increased temperatures and humidity, helping us understand their environmental causes.

A map of the united arab emirates

Description automatically generated

Figure 21 Heat Wave Visualization in UAE

Longitude on the x-axis and latitude on the y-axis show UAE heatwave conditions. Red to yellow temperatures range from 40°F to 120°F on the map. The map shows temperature readings by location, with colours correlating to the temperature scale. Data spots gather around Dubai in the heatwave visualisation. These spots' colour intensity indicates temperature, with darker colours signifying lower temperatures and lighter colours greater temperatures. This visualisation shows heatwave-prone locations.

The map shows UAE temperature fluctuations by region, emphasising heatwave-prone areas. This visualisation helps analyse heatwave distribution and develop mitigation tactics by overlaying temperature data on a global map to detect excessive heat zones. The correct temperature scale simplifies data interpretation by quickly assessing temperature severity in different areas.

A graph of different colored bars

Description automatically generated

Figure 22 Relative Humidity by Month

The bar chart shows relative humidity by month, with the x-axis indicating January (1) to December (12) and the y-axis showing relative humidity percentage. Each bar's height shows the month's average relative humidity, and the colours gradually change from dark purple to pale green. According to the figure, relative humidity is highest in January and December, exceeding 60%. This shows seasonal climates cause the most moisture-laden air in these months. After January, relative humidity steadily declines to 40% in June and July. Warmer temperatures and less precipitation in mid-year may explain this humidity drop.

The relative humidity rises again after July, indicating a gradual increase in moisture. From September to December, this rising tendency continues with large gains. The annual cycle ends in December when relative humidity returns to January levels. This chart shows how relative humidity cycles throughout the year, revealing seasonal weather trends. Understanding climatic fluctuations and planning agricultural, water resource, and other weather-dependent operations requires such information. The colour gradient makes it easy to discern months and track relative humidity throughout the year.

A screenshot of a graph

Description automatically generated

Figure 23 Correlation Heatmap of Numeric Variables

The correlation heatmap shows the correlations between temperature (tmpf), dew point temperature (dwpf), relative humidity (relh), wind direction (drct), wind speed (sknt), altitude (alti), and mean sea level pressure. Dark blue indicates significant negative correlations, while pale yellow indicates high positive correlations. The intensity of the colour indicates the correlation strength. Notable observations include a moderate positive association between temperature and dew point temperature, suggesting that higher temperatures may increase dew point temperatures. Temperature also negatively correlates with relative humidity, meaning greater temperatures lower humidity. increased dew points are significantly correlated with increased relative humidity.

The relationship between wind direction and wind speed is weakly positive. Altitude and mean sea level pressure have modest connections, showing they are independent of other meteorological conditions. The heatmap shows the strengths and orientations of links between major meteorological variables, showing significant correlations that may affect weather patterns and forecasting models. This visualisation helps improve forecast models and explain climatic behaviour by showing how different variables interact.

A group of text on a white background

Description automatically generated

Figure 24 DataFrame After Feature Engineering

A screenshot of a computer

Description automatically generated

Figure 25 Performance Metrics

A graph of different colored squares

Description automatically generated

Figure 26 MSE comparison

A graph of different colored squares

Description automatically generated

Figure 27 RMSE Comparison

A graph of different colored squares

Description automatically generated

Figure 28 MAE Comparison

A graph of different colored squares

Description automatically generated

Figure 29 R2Score Comparison

Based on the dataset, Linear Regression, Decision Tree, Random Forest, and Gradient Boosting regression models are evaluated for heatwave prediction. MSE, MAE, RMSE, and R2 Score are used to evaluate each model. These metrics reveal each model's forecast accuracy and reliability.

The Linear Regression model has an MSE of 890.10, MAE of 23.29, RMSE of 29.83, and R2 Score of 0.83. MSE is the average squared difference between anticipated and actual values; lower represents better performance. In this situation, the MSE is large, indicating poor model predictions. The MAE, which averages absolute errors, is 23.29, meaning the model's predictions are 23.29 units off. Larger errors are penalised more by the 29.83 RMSE. The model explains 83% of the variance in the dependent variable, according to the R2 Score of 0.83. Despite its reasonable fit, the Linear Regression model could be improved.

The Decision Tree model outperforms Linear Regression with an MSE of 296.35, MAE of 11.81, RMSE of 17.21, and R2 Score of 0.94. Lower MSE and MAE indicate superior predictive accuracy and a far lower average prediction error than Linear Regression. RMSE drops to 17.21 due to fewer big mistakes. The R2 Score of 0.94 indicates a strong match because the model explains 94% of observed variance.

The Random Forest model improves predictive accuracy with an MSE of 151.00, MAE of 8.53, RMSE of 12.29, and R2 Score of 0.97. The model's greater heatwave prediction accuracy is shown by its reduced MSE and RMSE. The MAE of 8.53 reduces average error, improving prediction accuracy. An excellent fit and robust predictor for the dataset, the Random Forest model explains 97% of the variation with an R2 Score of 0.97.

The Gradient Boosting method performs well, with an MSE of 172.66, MAE of 9.40, RMSE of 13.14, and R2 Score of 0.97. The MSE is somewhat higher than Random Forest but substantially lower than Linear Regression and Decision Tree. MAE of 9.40 indicates a slightly greater average error than Random Forest but high accuracy. RMSE of 13.14 implies fewer big mistakes, and R2 Score of 0.97 shows that Gradient Boosting, like Random Forest, explains 97% of the variance.

In conclusion, Random Forest and Gradient Boosting predict heatwaves better than Linear Regression and Decision Tree. Linear Regression has a decent baseline, but its increasing error metrics suggest improvement. Decision Tree does better than Linear Regression but not as well as Random Forest and Gradient Boosting. Random Forest and Gradient Boosting are the best models for this dataset due to their high prediction accuracy, dependability, and explanatory power. Both Random Forest and Gradient Boosting accurately model data, however their decision depends on model interpretability and processing economy.

# Implementation

## Data Preprocessing

A computer code with text

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated

A computer code with text

Description automatically generated with medium confidence

Figure 30 Data Preprocessing

Implementation begins with data preprocessing to prepare the dataset for analysis and model training. Missing values ('M') are checked in the CSV dataset. To preserve data quality, columns with over 200,000 missing entries are deleted. Missing values in columns with 'M' values are converted to NaN and imputed with the column mean. The timestamp-containing 'valid' column is transformed to datetime for time-based analysis and feature extraction.

## Defining Heatwave Conditions

A screenshot of a computer program

Description automatically generated

Figure 31

Next, the implementation sets heatwave thresholds for temperature, relative humidity, and wind speed. The new feature 'heatwave\_index' measures heatwave intensity. This measure is calculated using excess temperature, relative humidity, and wind speed deficit over three hours. For the 'heatwave\_index' column, positive excesses and deficits are added throughout the specified period to calculate heatwave intensity.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) uses visualisations to understand dataset features and variables. Histograms and time series plots depict heatwave and non-heatwave temperature and humidity variability. Colour-coded scatter plots demonstrate heatwave and non-heatwave temperature-relative humidity relationships. These visualisations show data patterns and aid models.

A screenshot of a computer program

Description automatically generated

Figure 32 EDA

A screenshot of a computer program

Description automatically generated

Figure 33 EDA

## Features Engineering

Year, month, day, hour, and weekday are extracted from the 'valid' datetime column during feature engineering. Capturing data temporal patterns requires these features. Label Encoding turns categorical variables like sky conditions and meteorological reports into numerical model inputs.

A screenshot of a computer program

Description automatically generated

Figure 34 Feature Engineering

## Dataset Split

After separating the dataset into training and testing sets, 'heatwave\_index' is the goal variable while the other attributes are inputs. This division provides a robust evaluation of the model's unseen data prediction skills.

A computer screen shot of a computer code

Description automatically generated

Figure 35 Dataset Split

## Model-building and training

Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor are trained. Each model is trained on the training dataset and tested on the test dataset. Each model's MSE, MAE, RMSE, and R2 Score are calculated to assess its correctness and reliability.

A computer code with black text

Description automatically generated

Figure 36 Model Builing

## Model Evaluation and Comparison

Printers and bar charts display model performance indicators. These charts simplify model comparison using MSE, RMSE, MAE, and R2 Score. Ensemble approaches like Random Forest and Gradient Boosting beat Linear Regression and Decision Tree models in predicting accuracy and reliability. Random Forest and Gradient Boosting have high R2 Scores, showing they explain much of the data variance.

A screenshot of a computer program

Description automatically generated

Figure 37 Model Training and Evaluation

## Visualization of Results

For successful results presentation, different visualisations are generated. Bar plots provide performance indicators, a heatmap shows numeric variable correlations, and a geographical map shows UAE heatwaves. These visualisations clarify data and model performance and improve results interpretation.

A screenshot of a computer program

Description automatically generated

Figure 38 Visualization of Results

This step-by-step method ensures accurate heatwave prediction through data preprocessing, feature engineering, model training, evaluation, and visualisation.

# Project Management

## Project Schedule

A graph with blue squares

Description automatically generated

Figure 39 Gantt Chart (Project Schedule)

A detailed Gantt chart showed the project schedule from 02 June 2024 until 1 August 2024. A systematic approach to project goals was achieved by dividing it into phases. From 06th to 05th June 2024, Project Planning defined the project's scope, identified resources, and set starting dates. The future stages relied on this basic step. From 05th to 09th June 2024, meteorological data was collected from numerous sources to ensure a complete dataset. Data Preprocessing occurred June 10–14, 2024. Cleaning, normalising, and transforming raw data ensured consistency and analysis usability. Exploratory Data Analysis (EDA) from 14th to 17th June 2024 evaluated data patterns and correlations to inform feature engineering and model selection.

Feature Engineering took place 17 June – 20 June 2024. To improve the model's prediction, new and modified features were added. The following phase, Defining Heatwave Conditions, from 20 – 24 June 2024, defined and integrated heatwave criteria based on temperature, humidity, and wind speed into the dataset. Dataset Splitting was done from June 25 to 28, 2024, to separate data into training and testing sets for model evaluation. From June 28 to July 03, 2024, Linear Regression, Decision Trees, Random Forests, and Gradient Boosting were developed for machine learning.

Model Training was place from July 03 to 07, 2024, to teach models data patterns and linkages. Next, models were evaluated from July 07th to 10th, 2024, using performance indicators like MSE, MAE, RMSE, and R² Score. From July 10 to 13, 2024, Model Comparison compared models to find the best heatwave forecast model. On 14 July 2024, Results Visualisation commenced for two days to graphically show findings for better comprehension and communication. Report writing began on 17 July and ended on 24 July 2024, covering project approach, outcomes, and comments.

The report was reviewed and revised from July 25 to 29, 2024, to ensure accuracy and completeness. July 29, 2024, was the final report submission. The project was finished on August 1, 2024, with a success assessment.

## Risk Management

Successful project completion needs risk management. Data quality, model accuracy, and deadlines were our biggest project risks.

* Inaccurate data can negatively impact project outcomes. Cleaning and normalising data ensured consistency and reduced risk. Regular data pretreatment tests and validations fixed data quality issues fast.
* Model accuracy was a big concern as it could lead to inaccurate forecasts. Many machine learning models were evaluated using performance measures such as MSE, MAE, RMSE, and R² Score to reduce risk. For heatwave predictions, this comparison chose the best model.
* Timeline For project success, adherence was needed. Every phase delay can affect the project schedule. To decrease risk, a detailed Gantt plan was created with phase start and end dates. Regular progress assessments uncovered setbacks and adjusted the project to stay on track.

By anticipating and mitigating these risks, the project operated smoothly and fulfilled its deadline.

## Quality Management

Quality management guarantees the utmost dependability and correctness of project procedures and outputs. Data analysis and model generation were standardised by CRISP-DM. This method enabled structured data preprocessing, robust feature engineering, precise model construction, and extensive evaluation. Project goals were maintained by weekly meetings and milestone reviews. The evaluations promptly detected and fixed issues, ensuring that each project phase was completed successfully before moving on. Peer reviews helped assure accuracy and reliability by including team and external expert feedback.

The project evaluation used MSE, MAE, RMSE, and R² Score performance indicators. These tests showed the models' forecast accuracy and reliability. Visualisation was used extensively to explain results and enhance analysis and decision-making. In the project, data pretreatment, model setups, evaluation results, and visualisations were reported. The extensive documentation ensures openness, reproducibility, and report review and editing. Following these quality management practises, the project delivered reliable, high-quality data that met its goals, creating accurate and effective heatwave prediction models.

## Social, Legal, Ethical and Professional Considerations

Every project must consider social, legal, ethical, and professional issues, especially when handling sensitive data and constructing models with real-world consequences. Project involved maintaining sensitive meteorological data, which may be protected by privacy regulations. GDPR data protection and privacy requirements were strictly observed. Data was anonymised for privacy. For secrecy, only allowed users could access the data. Ethics dominated the project. Erroneous projections could hurt, therefore precision and reliability were needed. Project strategy and assumptions were specified for transparency. Project approval was obtained before starting to ensure ethical compliance.

I always followed the data science and machine learning code of conduct to preserve professionalism. This included being honest about results, recognising teamwork, and respecting IP. The project's social impacts were examined. Accurate heatwave estimates can improve public health and safety by alerting and preventing. The project improved UAE heatwave preparation and response to protect vulnerable populations and reduce heatwaves. Compliance with law: The data usage, model deployment, and reporting criteria were met. This required data usage authorisation and project compliance with laws. The initiative integrated social, legal, ethical, and professional issues to achieve trustworthy and useful solutions with integrity and accountability. This comprehensive approach ensured the project met technical goals, helped society, and followed ethics.

# Critical Appraisal

UAE heat waves were predicted using advanced machine learning and meteorological data. Data collection, preprocessing, EDA, feature engineering, model construction, evaluation, and visualisation were critical. This impartial critical appraisal evaluates project results for strengths and faults. This project stood out because to its extensive data pretreatment and cleansing. We identified and handled missing values to make the dataset robust and analyzeable. Dropping columns with high missing data and impute missing values in other columns improved data integrity and model dependability. This meticulous preparation ensured high-quality data for model training. Also strong was exploratory data analysis. Histograms, time series charts, and scatter plots revealed data patterns and links. The visualisations identified temperature, relative humidity, and wind speed as heat wave factors. Quantifying heat wave intensity with the 'heatwave\_index' improves model predictions.

Feature engineering extracted temporal features from the timestamp column and converted category variables to numerical representations to improve the dataset. This method helped machine learning models use all available data to improve prediction. For accurate heat wave prediction, characteristics must be carefully selected and engineered to capture complicated meteorological variable interactions. We tested Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor during model development. Comprehensive evaluation measures, including MSE, MAE, RMSE, and R² Score, gave a full assessment of model performance. Ensemble approaches like Random Forest and Gradient Boosting surpassed Linear Regression and Decision Tree. These advanced models had better performance measures, indicating greater accuracy and robustness. The project has difficulties and space for improvement. While thorough, historical weather data may not fully indicate climate change's future effects. Climate change can introduce additional factors and interactions, impacting model accuracy. Future studies should integrate climate change projections and environmental components to improve model robustness and flexibility.   
  
Project computing resources are likewise limited. Computer power and time are needed for complex model training and evaluation on large datasets. Within its constraints, the study overcome these challenges, but future work could benefit from modern computational resources or optimisation. The procedure required communication and teamwork. Progress reviews and peer feedback kept the project on track and successful. Clearer communication throughout data preprocessing and feature engineering may have sped up the process. Future initiatives may tackle this problem with systematic communication and documentation. Project findings affect UAE heat wave forecast and disaster preparedness. Authorities may warn and advise on heat wave prevention with accurate and timely predictions, reducing public health and infrastructure damage. Models from this study improve climate resilience and population safety. Finally, our experiment revealed how meteorological data may help machine learning predict heat waves. The project excelled in data preparation, intelligent EDA, feature engineering, and model validation. Add climate change projections, improve model interpretability, optimise computer speed, and upgrade communication protocols. Although challenging, the study provided helpful insights and tools for UAE heat wave prediction and catastrophe preparedness. The project's skills will improve climate science and machine learning research and applications.

# Results and Analysis

The study predicted UAE heat waves using machine learning and meteorological data. Recent field research and methodologies support this section's analysis of model performance, feature relevance, and forecast implications.

## Model Performance

Heat wave prediction was investigated using Linear Regression, Decision Tree, Random Forest, and Gradient Boosting. We used performance measures like MSE, MAE, RMSE, and R² Score to evaluate model accuracy and robustness. The baseline Linear Regression model has an MSE of 890.10, MAE of 23.29, RMSE of 29.83, and R² Score of 0.83. It captured some data variability, but its prediction errors were substantial, making it less reliable than sophisticated models. The Decision Tree Regressor improved performance by capturing non-linear correlations with MSE of 296.35, MAE of 11.81, RMSE of 17.21, and R² Score of 0.94 The Random Forest Regressor outperformed previous models with an MSE of 151.00, MAE of 8.53, RMSE of 12.29, and R² Score of 0.97. This ensemble method reduced overfitting and improved generalisation, exhibiting predictive potential. The Gradient Boosting Regressor performed well with an MSE of 172.66, MAE of 9.40, RMSE of 13.14, and R² Score of 0.97. Gradient Boosting improved models progressively to capture complicated data patterns. Random Forest and Gradient Boosting models explained 97% of heat wave variance, proving their reliability. These models were picked as the best forecasters due to their performance.

## Feature Importance

Understanding the major heat wave prediction factors affects model interpretability and practicality. Predictors included temperature, dew point, relative humidity, wind speed, and sea-level pressure. Srikanth & Pal (2023) found that higher temperatures caused stronger heat waves. Asadollah et al. (2021) found that dew point temperature, which impacts perceived temperature and humidity, caused heat wave discomfort and health issues. High temperatures and humidity caused heat waves, hence relative humidity determined how oppressive the heat was. Lower wind speeds, often associated with stagnant air, limited cooling, increasing heat wave consequences. Bochenek & Ustrnul (2022) linked sea-level pressure fluctuations to weather patterns.

## Practical Implications

This project's predictive models aid UAE heat wave management and disaster preparedness. Correct heat wave estimates help authorities issue heat warnings, advise on prevention, and mobilise resources for vulnerable populations. Preemptive action can lessen heat wave implications on public health, agriculture, infrastructure, and society.

## Limitations and Future Work

Many constraints were found despite accomplishments. Historical meteorological data may not entirely predict climate change. Climate projections and environmental factors may strengthen and adapt models. Complex "black-box" models like Gradient Boosting were another challenge. SHAP values and LIME improve model interpretability for model decision-making (Lu et al., 2023). Complex model training and evaluation requires many computers. More efficient computational resources or optimisation methods could improve project efficiency and scalability. For greater prediction accuracy and reliability, researchers should adopt hybrid models that combine classical and machine learning methods (Joe et al., 2022). The project demonstrated how machine learning can predict heat waves from meteorological data. The Random Forest and Gradient Boosting models effectively predicted heat waves, and feature importance analysis identified critical predictors.

# Conclusions

## Achievements

This ambitious project used meteorological data to build robust machine learning algorithms to predict UAE heat waves. The key successes are model performance, feature analysis, and practical effects. We evaluated Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. Linear Regression model yielded baseline MSE of 890.10, MAE of 23.29, RMSE of 29.83, and R² Score of 0.83. It captured some data variability, but its prediction errors required more advanced models. The Decision Tree Regressor's improved performance (MSE = 296.35, MAE = 11.81, RMSE = 17.21, R² Score = 0.94) indicates better handling of non-linear data relationships. The Random Forest Regressor achieved high prediction accuracy with MSE of 151.00, MAE of 8.53, RMSE of 12.29, and R² Score of 0.97. Averaging multiple decision trees solved overfitting and increased generalisation in this model. The Gradient Boosting Regressor demonstrated strong performance, capturing complicated patterns through sequential model modification (MSE: 172.66, MAE: 9.40, RMSE: 13.14, R² Score: 0.97).

A detailed feature importance analysis identified temperature (tmpf), dew point temperature (dwpf), relative humidity (relh), wind speed (sknt), and sea-level pressure as heat wave predictors. Temperature was the best predictor, supporting previous research. Dew point temperature, important for understanding felt temperature and humidity, was important. Heat wave intensity and persistence depended on relative humidity and wind speed. Sea-level pressure revealed weather-affecting atmospheric conditions. These accomplishments have major practical ramifications. Proactive heat wave mitigation can be informed by predictive models. These models allow authorities to warn the public, advise on protective measures, and mobilise resources for vulnerable groups. These preventative actions can dramatically reduce high heat event health risks, economic losses, and infrastructure damage. The project also advances climate resilience. The models improve heat wave prediction accuracy, boosting preparedness and response methods and societal resilience to climate-induced extreme weather events. This supports global climate change adaptation and health mitigation.

## Future Work

This experiment has met its goals, however future research can improve the predictive models.

* Including Climate Change Projections: The existing models' dependence on past meteorological data may not fully capture future climate change implications. Future work should include climate change projections to improve model flexibility and robustness. Climate models' forecasting capacity can be improved by incorporating data on heat wave patterns under different climate scenarios.
* Improved Model Interpretability: Random Forest and Gradient Boosting models have great predictive accuracy, but their "black-box" character made interpretation difficult. Future research should use SHAP values and LIME to improve model interpretability. These strategies can increase stakeholder trust and enable practical insights into complex model decision-making.
* Additional Environmental Factors Integration: Most models focus on meteorological variables including temperature, humidity, wind speed, and sea-level pressure. Land use differences, urban heat island impacts, and local topography should be integrated into heat wave dynamics research. These characteristics increase prediction granularity and accuracy, improving heat wave risk knowledge.
* Data quality and availability impact machine learning model reliability. Use tight quality assurance, improve data collection, and research data fusion to unify data sources. Crowdsourcing and environmental sensors can upgrade input data variety and depth, improving model performance.
* Computational Efficiency and Scalability:Training and evaluating complex models needed significant computer resources. Future research could increase these models' scalability and computing efficiency. High-performance computing and cloud technologies can process massive datasets and complicated models. Optimisation can boost model training and evaluation efficiency.
* Hybrid Model Development: Exploring hybrid models that blend traditional and machine learning strengths can improve predicted accuracy and reliability. Hybrid models reconcile heat wave prediction with traditional methods' robustness and machine learning's adaptability.
* Collaboration and Use: Researchers, politicians, and community stakeholders must collaborate to predict heat waves. Decision support systems that turn predictive model outputs into real-world tactics should be developed next. These heat wave mitigation methods can be relevant and effective by including stakeholders in their development and execution.
* Transferability, generalisation: Transferability and generalisation of machine learning models across locations and climate zones are essential for wider applicability. Transfer learning and domain adaptation should be studied to adapt models to different climates. Validating model performance across different geographies and climates can assist establish universal heat wave prediction frameworks, boosting global climate resilience efforts.

Finally, this study created and tested improved machine learning models for UAE heat wave prediction, improving accuracy and robustness. The discovered critical predictors and practical consequences show that these models can improve heat wave preparedness and resistance. However, climate change estimates, interpretability, environmental considerations, data quality, and hybrid methodologies can be added to these models in the future. This project's insights and technologies can help build sustainable and resilient communities worldwide by mitigating climate-induced extreme weather occurrences through interdisciplinary collaboration and study.

# Student Reflections

This project has been a great learning experience that has improved my machine learning and climate science skills, particularly in heat wave prediction. I anticipated obstacles when I started this project, but the depth and complexity of the issues exceeded my expectations. I look back at my lessons, struggles, and intellectual and personal progress. Applications of theoretical knowledge to real-world problems were highlights of this project. I had some machine learning and data analysis experience from previous courses, but this assignment required me to use real data, refine models, and assess results. Working with complex datasets and advanced models like Random Forest and Gradient Boosting was difficult. I learnt these concepts and their practicality after much effort and iteration.

I learnt about data preprocessing and feature engineering. I quickly discovered that input data quality affects machine learning model performance. Cleaning data, fixing missing numbers, and selecting the most important aspects demands precision. I improved my forecasts by learning how to turn raw data into a format machine learning algorithms could use. Feature importance analysis taught me plenty. Understanding heat wave indicators like temperature, dew point temperature, and relative humidity was intellectually and practically valuable. I learnt how meteorological variables cause heat waves from this analysis. It revealed how complex climate is and how crucial forecast models are. Collaboration and criticism were essential to my project. Regular talks with my supervisor helped me through tough times. The many ideas and constructive criticism from peers improved my approach. These conversations taught me feedback and collaborative learning.

Reaching milestones was a highlight of this project. Creating and testing heat wave prediction models was rewarding. It confirmed my belief that machine learning can tackle global problems and guided my academic and professional path. The initiative met challenges. Project complexities, time management, and academic requirements were challenging to balance. Unexpected data issues or model failures caused frustration. I learnt resilience and persistence from these challenging situations. Setbacks were opportunities to learn and grow, making project success more rewarding.

This project's learning opportunities and my work make me proud. I now understand machine learning, solve problems better, and am ready for academic and professional challenges. This contemplation has also shown the significance of constant learning and adaptability in a continuously changing sector. I'm enthusiastic to build on this experience and explore machine learning's ability to solve challenging real-world challenges.

Bibliography and References

1. Arias, R., & Mader, T. L. (2023). ‘Evaluation of Four Thermal Comfort Indices and Their Relationship with Physiological Variables in Feedlot Cattle. Animals’, 13(7), 1169–1169. <https://doi.org/10.3390/ani13071169>
2. Ballester, J., Quijal-Zamorano, M., Méndez Turrubiates, R. F., Pegenaute, F., Herrmann, F. R., Robine, J. M., Basagaña, X., Tonne, C., Antó, J. M., & Achebak, H. (2023). ‘Heat-related mortality in Europe during the summer of 2022’. Nature Medicine, 29(29), 1–10. <https://doi.org/10.1038/s41591-023-02419-z>
3. Buttke, D. E., Raynor, B., & Schuurman, G. W. (2023). ‘Predicting climate-change induced heat-related illness risk in Grand Canyon National Park visitors. PLOS ONE, 18(8), e0288812. <https://doi.org/10.1371/journal.pone.0288812>
4. Candelaria, M. D. E., Chua, N. M. M., & Kee, S.-H. (2022). ‘Evaluation of Heat-Induced Damage in Concrete Using Machine Learning of Ultrasonic Pulse Waves. Materials’, 15(22), 7914. <https://doi.org/10.3390/ma15227914>
5. CDC. (2022, May 13). ‘CDC - Heat Stress - Heat Related Illness - NIOSH Workplace Safety and Health Topic’. CDC. <https://www.cdc.gov/niosh/topics/heatstress/heatrelillness.html>
6. Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., & Yang, Z. (2023a). ‘Machine Learning Methods in Climate Prediction: A Survey. Research Gate’. <https://doi.org/10.20944/preprints202309.1764.v1>
7. Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., & Yang, Z. (2023b). ‘Machine Learning Methods in Weather and Climate Applications: A Survey. Applied Sciences’, 13(21), 12019. <https://doi.org/10.3390/app132112019>
8. Dóra Szagri, Nagy, B., & Szalay, Z. (2023). ‘How can we predict where heatwaves will have an impact? – A literature review on heat vulnerability indexes’. Urban Climate, 52, 101711–101711. <https://doi.org/10.1016/j.uclim.2023.101711>
9. Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J., Fishgold, L., Friedl, M., Jain, M., Kohli, D., Laso Bayas, J., Lunga, D., McCarty, J., Pontius, R., Reinmann, A., Rogan, J., Song, L., Stoynova, H., Ye, S., Yi, Z.-F., & Estes, L. (2020). ‘Accounting for Training Data Error in Machine Learning Applied to Earth Observations. Remote Sensing’, 12(6), 1034. <https://doi.org/10.3390/rs12061034>
10. Eugenio Pappalardo, S., Zanetti, C., & Todeschi, V. (2023). ‘Mapping urban heat islands and heat-related risk during heat waves from a climate justice perspective: A case study in the municipality of Padua (Italy) for inclusive adaptation policies. Landscape and Urban Planning’, 238, 104831. <https://doi.org/10.1016/j.landurbplan.2023.104831>
11. GeeksforGeeks. (2024, March 20). ‘Linear Regression in Machine learning’. GeeksforGeeks. <https://www.geeksforgeeks.org/ml-linear-regression/>
12. Hajer Khaled Jabbar, Hamoodi, M. N., & Al-Hameedawi, A. N. (2023). ‘Urban heat islands: a review of contributing factors, effects and data. IOP Conference Series’, 1129(1), 012038–012038. <https://doi.org/10.1088/1755-1315/1129/1/012038>
13. ‘Heatstroke - Symptoms and causes’. (2022, June 25). Mayo Clinic. <https://www.mayoclinic.org/diseases-conditions/heat-stroke/symptoms-causes/syc-20353581#:~:text=Heatstroke%20is%20a%20condition%20caused>
14. Hirano, Y., Kondo, Y., Hifumi, T., Yokobori, S., Kanda, J., Shimazaki, J., Hayashida, K., Moriya, T., Yagi, M., Takauji, S., Yamaguchi, J., Okada, Y., Okano, Y., Kaneko, H., Kobayashi, T., Fujita, M., Yokota, H., Okamoto, K., Tanaka, H., & Yaguchi, A. (2021). ‘Machine learning-based mortality prediction model for heat-related illness’. Scientific Reports, 11(1). <https://doi.org/10.1038/s41598-021-88581-1>
15. Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). ‘Significance of machine learning in healthcare: Features, pillars and applications’. International Journal of Intelligent Networks, 3, 58–73. <https://doi.org/10.1016/j.ijin.2022.05.002>
16. Lakra, K., & Avishek, K. (2022). ‘A review on factors influencing fog formation, classification, forecasting, detection and impacts’. Rendiconti Lincei. Scienze Fisiche E Naturali, 33(2), 319–353. <https://doi.org/10.1007/s12210-022-01060-1>
17. Lee, M.-H., Mun, H.-J., & Kang, S.-K. (2022). ‘Prediction of Health Problems Using Deep Learning’ Images and Bio-Signals. Applied Sciences, 12(23), 12457. <https://doi.org/10.3390/app122312457>
18. Leiva, D. F., & Church, B. (2021). Heat Illness. PubMed; StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK553117/>
19. Li, Y., Li, T., Lv, W., Liang, Z., & Wang, J. (2023). ‘Prediction of Daily Temperature Based on the Robust Machine Learning Algorithms. Sustainability’, 15(12), 9289. <https://doi.org/10.3390/su15129289>
20. Liu, Y., Esan, O. C., Pan, Z., & An, L. (2021). ‘Machine learning for advanced energy materials’. Energy and AI, 3, 100049. <https://doi.org/10.1016/j.egyai.2021.100049>
21. Ma, X., Zou, B., Deng, J., Gao, J., Longley, I., Xiao, S., Guo, B., Wu, Y., Xu, T., Xu, X., Yang, X., Wang, X., Tan, Z., Wang, Y., Morawska, L., & Salmond, J. (2024). ‘A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: A perspective from 2011 to 2023’. Environment International, 183, 108430. <https://doi.org/10.1016/j.envint.2024.108430>
22. Manuela De Sario, de’Donato, F., Bonafede, M., Marinaccio, A., Levi, M., Filippo Ariani, Morabito, M., & Michelozzi, P. (2023). ‘Occupational heat stress, heat-related effects and the related social and economic loss: a scoping literature review’. Frontiers in Public Health, 11. <https://doi.org/10.3389/fpubh.2023.1173553>
23. Mekala, A., Baishya, B. K., Vigneswara Rao, K. T., Vidhate, D. A., Drave, V. A., & Prasanth, P. V. (2023). ‘Big Data and Machine Learning Framework for Temperature Forecasting. EAI Endorsed Transactions on Energy Web’, 10. <https://doi.org/10.4108/ew.4195>
24. Mohamad Rajab Houmsi, Ismail, Z., Ghaith Falah Ziarh, Mohammed Magdy Hamed, Ishak, D., Mohd, Mokhtar, M. Z., Zulfaqar Sa’adi, & Shahid, S. (2023). ‘Relative Influence of Meteorological Variables of Human Thermal Stress in Peninsular Malaysia’. Sustainability, 15(17), 12842–12842. <https://doi.org/10.3390/su151712842>
25. Ogata, S., Takegami, M., Ozaki, T., Nakashima, T., Onozuka, D., Murata, S., Nakaoku, Y., Suzuki, K., Hagihara, A., Noguchi, T., Iihara, K., Kitazume, K., Morioka, T., Yamazaki, S., Yoshida, T., Yamagata, Y., & Nishimura, K. (2021). ‘Heatstroke predictions by machine learning, weather information, and an all-population registry for 12-hour heatstroke alerts’. Nature Communications, 12(1). <https://doi.org/10.1038/s41467-021-24823-0>
26. Osilla, E. V., & Sharma, S. (2022). Physiology, temperature regulation. Nih.gov; StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK507838/>
27. Park, M., Jung, D., Lee, S., & Park, S. (2020). ‘Heatwave Damage Prediction Using Random Forest Model in Korea. Applied Sciences’, 10(22), 8237. <https://doi.org/10.3390/app10228237>
28. RODRIGUEZ-UBINAS, E., ALZAROUNI, S., ALHAMMADI, N., ALANTALI, M., BEHZAD, F., & SGOURIDIS, S. (2020). ‘SUSTAINABILITY THROUGH ENERGY CONSERVATION BUILDING CODES: COMPARATIVE ANALYSIS OF GREEN BUILDING REGULATIONS IN THE MIDDLE EAST. WIT Transactions on Ecology and the Environment’. <https://doi.org/10.2495/sc200081>
29. Roy, S., Mishra, D. P., Bhattacharjee, R. M., & Agrawal, H. (2021). ‘Effect of Heat Stress and Development of WBGT Based Heat Stress Prediction Models for Underground Coal Miners Using Random Forest Algorithm and Artificial Neural Network. SSRN Electronic Journal’. <https://doi.org/10.2139/ssrn.3994163>
30. Shimazaki, T., Anzai, D., Watanabe, K., Nakajima, A., Fukuda, M., & Ata, S. (2022). ‘Heat Stroke Prevention in Hot Specific Occupational Environment Enhanced by Supervised Machine Learning with Personalized Vital Signs. Sensors’, 22(1), 395. <https://doi.org/10.3390/s22010395>
31. Singh, H., Mhasawade, V., & Chunara, R. (2022). ‘Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database’. PLOS Digital Health, 1(4), e0000023. <https://doi.org/10.1371/journal.pdig.0000023>
32. U Vivek Krishna, S Sanju, S Sudhakaran, R Kaladevi, & Hariharan Shanmugasundaram. (2023). ‘Weather Prediction using linear regression model’. Research Gate. <https://doi.org/10.1109/iccci56745.2023.10128252>
33. Wesam Alkassas, Ahmad Mamoun Rajab, Al-Rashood, S. T., Muhammad Ayub Khan, Mahmoud Dibas, & Zaman, M. (2021). ‘Heat-related illnesses in a mass gathering event and the necessity for newer diagnostic criteria: a field study. Environmental Science and Pollution Research’, 28(13), 16682–16689. <https://doi.org/10.1007/s11356-020-12154-4>
34. ‘What Is a Decision Tree? | Master’s in Data Science’. (2024, May 8). CORP-MIDS1 (MDS). <https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/>
35. World Health Organization. (2018, June 1). Heat and Health. World Health Organization; World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/climate-change-heat-and-health>
36. Yehoshua, R. (2023, July 17). Random Forests - Dr. Roi Yehoshua - Medium. Medium. <https://medium.com/@roiyeho/random-forests-98892261dc49>
37. Zhang, T., Lin, W., Vogelmann, A. M., Zhang, M., Xie, S., Qin, Y., & Golaz, J. (2021). ‘Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning. Journal of Advances in Modeling Earth Systems’, 13(5). <https://doi.org/10.1029/2020ms002365>

Appendix A – Project Code

Project code link: [ipnb code file from git hub](https://github.com/WasimAkram1112/DataScience_Masters_Project/blob/main/heatwave_prediction_14360095.ipynb)

Data set used for project: [Heatwave UAE data stored in google drive](https://drive.google.com/file/d/1ReG7OIOFK6_MFqphsB3Q0Vbe0gwfWGLm/view?usp=drive_link)

!pip install geopandas contextily

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**from** **sklearn.preprocessing** **import** LabelEncoder

**from** **sklearn.model\_selection** **import** train\_test\_split, GridSearchCV

**from** **sklearn.linear\_model** **import** LinearRegression

**from** **sklearn.tree** **import** DecisionTreeRegressor

**from** **sklearn.ensemble** **import** RandomForestRegressor, GradientBoostingRegressor

**from** **sklearn.metrics** **import** mean\_squared\_error, mean\_absolute\_error, r2\_score

**from** **math** **import** sqrt

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**import** **matplotlib.dates** **as** **mdates**

**import** **geopandas** **as** **gpd**

**import** **contextily** **as** **ctx**

df = pd.read\_csv('/OMDB.csv')

df.head()

df.shape

"""\*\*Preprocessing\*\*"""

df.isna().sum()

df.info()

# Identify columns with "M" values

columns\_with\_M = df.columns[df.isin(['M']).any()]

print("Columns with 'M' values:", columns\_with\_M)

# Count the number of "M" values in each identified column

**for** column **in** columns\_with\_M:

count\_M = df[column].value\_counts().get('M', **0**)

print(f"Column {column} has {count\_M} 'M' values")

# Dropping columns with more than 2 lakh missing entries

columns\_to\_drop = ['gust', 'p01i', 'skyl2', 'skyl3', 'skyl4', 'wxcodes', 'ice\_accretion\_1hr', 'ice\_accretion\_3hr', 'ice\_accretion\_6hr',

'peak\_wind\_gust', 'peak\_wind\_drct', 'peak\_wind\_time', 'snowdepth']

df.drop(columns=columns\_to\_drop, inplace=**True**)

# Identify columns with "M" values after dropping

columns\_with\_M = df.columns[df.isin(['M']).any()]

print("Columns with 'M' values after dropping:", columns\_with\_M)

# Count the number of "M" values in each identified column after dropping

**for** column **in** columns\_with\_M:

count\_M = df[column].value\_counts().get('M', **0**)

print(f"Column {column} has {count\_M} 'M' values")

# Convert 'M' to NaN and then calculate mean for numeric columns

**for** column **in** columns\_with\_M:

**if** column **in** ["tmpf", "dwpf", "alti", "drct", "sknt", "relh", "mslp", "vsby", "skyl1", "feel"]:

df[column] = pd.to\_numeric(df[column], errors='coerce')

mean\_value = df[column].mean()

df[column].fillna(mean\_value, inplace=**True**)

df.head()

# Convert 'valid' column to datetime, handling different date formats

df['valid'] = pd.to\_datetime(df['valid'], errors='coerce', dayfirst=**True**)

# Convert relevant columns to numeric, forcing errors to NaN

cols = ['tmpf', 'dwpf', 'relh', 'sknt']

**for** col **in** cols:

df[col] = pd.to\_numeric(df[col], errors='coerce')

# Define heat wave conditions

temperature\_threshold = **95** # Define the temperature threshold for a heatwave

humidity\_threshold = **60** # Define the relative humidity threshold for a heatwave

wind\_speed\_threshold = **10** # Define the wind speed threshold for a heatwave

consecutive\_hours = **3** # Define the number of consecutive hours to be considered a heatwave

# Initialize heatwave index column with 0

df['heatwave\_index'] = **0.0**

# Loop through the data and calculate the heatwave index

**for** i **in** range(len(df) - consecutive\_hours + **1**):

temp\_excess = df['tmpf'][i:i + consecutive\_hours] - temperature\_threshold

relh\_excess = df['relh'][i:i + consecutive\_hours] - humidity\_threshold

wind\_deficit = wind\_speed\_threshold - df['sknt'][i:i + consecutive\_hours]

# Only consider positive excesses and deficits

temp\_excess[temp\_excess < **0**] = **0**

relh\_excess[relh\_excess < **0**] = **0**

wind\_deficit[wind\_deficit < **0**] = **0**

heatwave\_intensity = temp\_excess + relh\_excess + wind\_deficit

# Sum the heatwave intensity over the consecutive hours

df.loc[i:i + consecutive\_hours - **1**, 'heatwave\_index'] += heatwave\_intensity.sum()

# Display the dataframe with the new heatwave index feature

print(df.head())

# Plot histogram of temperature for heatwave and non-heatwave periods

plt.figure(figsize=(**12**, **6**))

sns.histplot(data=df, x='tmpf', hue='heatwave\_index', element='step', stat='density', common\_norm=**False**)

plt.title('Distribution of Temperature during Heatwave and Non-Heatwave Periods')

plt.xlabel('Temperature (°F)')

plt.ylabel('Density')

plt.legend(title='Heatwave', labels=['No', 'Yes'])

plt.show()

plt.figure(figsize=(**14**, **6**))

plt.plot(df['valid'], df['relh'], color='orange')

plt.title('Relative Humidity Variation over Time')

plt.xlabel('Time')

plt.ylabel('Relative Humidity (%)')

plt.xticks(rotation=**45**)

plt.grid(**True**)

plt.show()

# Plot time series of heatwaves

plt.figure(figsize=(**20**, **6**))

heatwave\_counts = df.set\_index('valid').resample('D')['heatwave\_index'].sum()

heatwave\_counts.plot()

plt.title('Number of Heatwaves Over Time')

plt.xlabel('Date')

plt.ylabel('Number of Heatwave Hours')

plt.show()

# Plot time series of heatwaves

plt.figure(figsize=(**20**, **6**))

heatwave\_counts = df.set\_index('valid').resample('D')['heatwave\_index'].sum()

ax = heatwave\_counts.plot()

# Add a horizontal dotted line to differentiate the heatwave scale

heatwave\_threshold = np.percentile(df['heatwave\_index'], **90**)

ax.axhline(heatwave\_threshold, color='red', linestyle='--', linewidth=**1**, label='Heatwave Threshold')

plt.title('Number of Heatwaves Over Time')

plt.xlabel('Date')

plt.ylabel('Number of Heatwave Hours')

plt.legend()

plt.show()

# Print the calculated heatwave threshold

print(f'Calculated Heatwave Threshold (90th percentile): {heatwave\_threshold}')

# Determine if it is a heatwave based on a threshold

heatwave\_threshold = np.percentile(df['heatwave\_index'], **90**)

df['is\_heatwave'] = df['heatwave\_index'] > heatwave\_threshold

# Define a custom color palette

custom\_palette = {**True**: "red", **False**: "blue"}

# Scatter plot of temperature vs. relative humidity, colored by heatwave

plt.figure(figsize=(**12**, **6**))

sns.scatterplot(data=df, x='relh', y='tmpf', hue='is\_heatwave', palette=custom\_palette, alpha=**0.7**)

plt.title('Temperature vs. Relative Humidity during Heatwave and Non-Heatwave Periods')

plt.xlabel('Relative Humidity (%)')

plt.ylabel('Temperature (°F)')

handles, labels = plt.gca().get\_legend\_handles\_labels()

labels = ['No', 'Yes']

plt.legend(handles=handles, labels=labels, title='Heatwave')

plt.show()

# Print the calculated heatwave threshold

print(f'Calculated Heatwave Threshold (90th percentile): {heatwave\_threshold}')

"""Heatwave periods, represented by the red points in the scatter plot, typically exhibit higher temperatures, concentrated around and above 100°F. These periods can occur under both low and high relative humidity conditions, indicating that heatwaves are primarily associated with high temperatures, regardless of whether the air is dry or humid. In contrast, non-heatwave periods, shown as blue points, span a broader range of temperatures and relative humidity levels. These periods are mostly found below the 100°F mark, reflecting more moderate temperature conditions."""

df\_sample = df.sample(**234486**)

# Create a GeoDataFrame

gdf = gpd.GeoDataFrame(

df\_sample, geometry=gpd.points\_from\_xy(df\_sample.lon, df\_sample.lat))

# Assign a CRS to the GeoDataFrame

gdf.set\_crs(epsg=**4326**, inplace=**True**)

# Plotting

fig, ax = plt.subplots(figsize=(**12**, **8**))

# Plot the points with colors based on temperature

norm = plt.Normalize(vmin=gdf['tmpf'].min(), vmax=gdf['tmpf'].max())

colors = plt.cm.hot(norm(gdf['tmpf']))

gdf.plot(ax=ax, color=colors, edgecolor='k', markersize=**50**, alpha=**0.7**)

# Add basemap

ctx.add\_basemap(ax, crs=gdf.crs, source=ctx.providers.OpenStreetMap.Mapnik)

# Add a colorbar

sm = plt.cm.ScalarMappable(cmap='hot', norm=norm)

sm.set\_array([])

cbar = plt.colorbar(sm, orientation='vertical', pad=**0.02**)

cbar.set\_label('Temperature (°F)')

# Add titles and labels

plt.title('Heat Wave Visualization in UAE')

plt.xlabel('Longitude')

plt.ylabel('Latitude')

plt.show()

# Convert 'valid' column to datetime format

df['valid'] = pd.to\_datetime(df['valid'], format='%d/%m/%Y %H:%M')

df['month'] = df['valid'].dt.month

plt.figure(figsize=(**10**, **6**))

sns.barplot(x='month', y='relh', data=df, ci=**None**, palette='viridis')

plt.title('Relative Humidity by Month')

plt.xlabel('Month')

plt.ylabel('Relative Humidity (%)')

plt.show()

numeric\_vars = ['tmpf', 'dwpf', 'relh', 'drct', 'sknt', 'alti', 'mslp']

corr = df[numeric\_vars].corr()

plt.figure(figsize=(**10**, **8**))

sns.heatmap(corr, annot=**True**, cmap='YlGnBu', vmin=-**1**, vmax=**1**)

plt.title('Correlation Heatmap of Numeric Variables')

plt.show()

encoder = LabelEncoder()

# Encode each column

df['skyc1'] = encoder.fit\_transform(df['skyc1'])

df['skyc2'] = encoder.fit\_transform(df['skyc2'])

df['skyc3'] = encoder.fit\_transform(df['skyc3'])

df['skyc4'] = encoder.fit\_transform(df['skyc4'])

df['metar'] = encoder.fit\_transform(df['metar'])

"""\*\*Feature Engineering\*\*"""

# Convert 'valid' column to datetime format

df['valid'] = pd.to\_datetime(df['valid'], format='%d/%m/%Y %H:%M')

# Extract features from 'valid' column

df['year'] = df['valid'].dt.year

df['month'] = df['valid'].dt.month

df['day'] = df['valid'].dt.day

df['hour'] = df['valid'].dt.hour

df['minutes'] = df['valid'].dt.minute

df['weekday'] = df['valid'].dt.weekday # Monday is 0 and Sunday is 6

# Dropping unnecessay colums

df.drop(columns=['valid'], inplace=**True**)

df.drop(columns=['station'], inplace=**True**)

df.head()

df.shape

"""\*\*Dataset Splitting\*\*"""

# Define features and target

X = df.drop(columns=['heatwave\_index'])

y = df['heatwave\_index']

# Split data into train and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=**0.2**, random\_state=**42**)

"""\*\*Model Building\*\*"""

# Initialize models

models = {

'Linear Regression': LinearRegression(),

'Decision Tree': DecisionTreeRegressor(random\_state=**42**),

'Random Forest': RandomForestRegressor(random\_state=**42**),

'Gradient Boosting': GradientBoostingRegressor(random\_state=**42**)

}

# Dataframe to store the results

results = pd.DataFrame(columns=['Model', 'MSE', 'RMSE', 'MAE', 'R2 Score'])

# Train and evaluate models

**for** name, model **in** models.items():

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

mae = mean\_absolute\_error(y\_test, y\_pred)

rmse = sqrt(mse)

r2 = r2\_score(y\_test, y\_pred)

new\_row = pd.DataFrame({'Model': [name], 'MSE': [mse], 'RMSE': [rmse], 'MAE': [mae], 'R2 Score': [r2]})

results = pd.concat([results, new\_row], ignore\_index=**True**)

print(f"Results for {name}:")

print(f" MSE: {mse:.2f}")

print(f" MAE: {mae:.2f}")

print(f" RMSE: {rmse:.2f}")

print(f" R2 Score: {r2:.2f}")

print("-----------------------------")

# Define custom colors for each model

colors = {'Linear Regression': 'blue',

'Decision Tree': 'green',

'Random Forest': 'orange',

'Gradient Boosting': 'red'}

# Plotting the results one by one

plt.figure(figsize=(**12**, **8**))

# MSE Plot

plt.subplot(**2**, **2**, **1**)

sns.barplot(x='Model', y='MSE', data=results, palette=colors)

plt.title('Mean Squared Error (MSE)')

plt.ylim(**0**, max(results['MSE']) \* **1.1**)

plt.tight\_layout()

plt.show()

# RMSE Plot

plt.figure(figsize=(**12**, **8**))

plt.subplot(**2**, **2**, **1**)

sns.barplot(x='Model', y='RMSE', data=results, palette=colors)

plt.title('Root Mean Squared Error (RMSE)')

plt.ylim(**0**, max(results['RMSE']) \* **1.1**)

plt.tight\_layout()

plt.show()

# MAE Plot

plt.figure(figsize=(**12**, **8**))

plt.subplot(**2**, **2**, **1**)

sns.barplot(x='Model', y='MAE', data=results, palette=colors)

plt.title('Mean Absolute Error (MAE)')

plt.ylim(**0**, max(results['MAE']) \* **1.1**)

plt.tight\_layout()

plt.show()

# R2 Score Plot

plt.figure(figsize=(**12**, **8**))

plt.subplot(**2**, **2**, **1**)

sns.barplot(x='Model', y='R2 Score', data=results, palette=colors)

plt.title('R2 Score')

plt.ylim(**0**, **1.1**)

plt.tight\_layout()

plt.show()

Appendix B – Certificate of Ethics Approval

A certificate of ethical approval

Description automatically generated

Appendix C – Meeting Logs

|  |  |  |
| --- | --- | --- |
| **Meeting Date** | **Supervisor** | **Key Discussion** |
| 12- June - 2024 | Dr. Alireza Daneshkhah | * Project Brief * Discussed about dataset findings and issues * Research objective |
| 26 - June - 2024 | Dr. Alireza Daneshkhah | * Discussed about dataset updating * Key features analysis * Discussion on the previous papers of heatwave |
| 28 - June - 2024 | Dr. Alireza Daneshkhah | * Working on the project * Discussion of project documentation on previous research papers |
| 12 - July - 2024 | Dr. Alireza Daneshkhah | * Code implementation and design * Analysis of dataset regarding the key feature findings within the code. * Discussion on evaluation of heatwave. |