Developing An Ensemble Modal to Predict the Effect of Heat Waves

Nitish Maity, Gaurav Kumar Yadav

Department of Computer Science and Engineering,
Institute of Technical Education and Research,
Siksha 'O' Anusandhan University, Bhubaneswar, Odisha

ORCID ID: 0000-0001-7022-290X

Abstract:

Heat waves pose significant risks to human health and the environment. With the increasing frequency and intensity of heat waves, there is a critical need to develop predictive models to analyze their impact. This research aims to develop an ensemble model to predict the effects of heat waves in Odisha, explicitly focusing on the correlation between meteorological factors and the number of deaths due to heat waves. Recent studies have highlighted Odisha's vulnerability to the risk of natural disasters caused by heat waves. The state has experienced increased heat wave intensity and frequency, adversely impacting human health. Understanding the changing patterns of climatic variability and the direct impact on human health is crucial for adequate precautions. This research employs an ensemble-based approach to predict heat wave events in Odisha. The model utilizes historical data on air pressure, wind speed, temperature, humidity, and other relevant meteorological factors. By analyzing the data from past heat wave events and corresponding mortality rates, the model aims to predict the potential impact of future heat waves on human health. We collected the previous year's data on the number of deaths on a particular date and analysed the Meteorological factors of the particular day. Also, we are collecting daily data on temperature, humidity, AQI, pressure, and wind speed to predict the weather conditions of a particular day. The ensemble approach in this research leverages advanced algorithms to analyze the complex relationships between meteorological variables and heat wave fatalities. We use an ensemble model with multiple linear regression, ridge regression, and support vector regressor at level zero and fully connected neural networks at level one. We are getting promising results regarding the R-squared score and Root Mean Squared Error (RMSE). Preliminary results show the high accuracy of the model predicting heat wave deaths in Odisha. Factors like air pressure, temperature, and humidity play significant roles.

Keywords:

Heat waves; Meteorological factors; Heatwave fatalities; R-squared score; RMSE;

1 Introduction

A heatwave is a period of unusually high temperatures, lasting for days or even weeks. [Attia et al., 2021] defines it's like a massive hot blanket covering an area, making everything feel excessively warm. This phenomenon happens when a high-pressure system traps warm air in a region, preventing cooler air from coming in. Heat waves, driven by global warming, are becoming more frequent and severe, significantly impacting human health by increasing risks of dehydration, heatstroke, and cardiorespiratory issues as research by [Robinson, 2001] and [Xu et al., 2016]. They are defined variably, complicating assessments and necessitating localized heat health warnings. According to [Guo et al., 2017] Children and the elderly are particularly vulnerable, with risks linked to renal and respiratory diseases. The urban heat island effect exacerbates these impacts, leading to stressed power grids and water shortages. Effective mitigation and adaptation strategies are essential for minimizing these health threats and ensuring climate resilience. Developing consistent definitions and protective measures is crucial for future climate adaptation in India.

Recent research highlights the increasing threat of heatwaves due to climate change, particularly in India. Studies by [Maharana et al., 2024] and [Akhtar, 2024] project a significant rise in the frequency and intensity of heatwaves, especially under high-emission scenarios. Vulnerable regions, such as northern and central India, are expected to face severe heat-related discomfort and health risks. Agricultural workers, in particular, are at heightened risk of heat-related illnesses, underscoring the need for improved awareness and preventive measures. Some researcher's [Boyaj et al., 2023] study examines the increasing frequency and impact of heat waves in Bhubaneswar, India. Using the Weather Research and Forecasting model, various radiation and urban canopy schemes were tested for predicting heat waves. The Community Atmospheric Model and Single-Layer Urban Canopy Model showed the best performance, predicting events two days in advance with reduced errors. Research by [Mishra et al., 2022] develops high-resolution climate data 4x4 km for April to June (2001–2016) using the Weather Research and Forecasting model, analyzing heat waves and their impact in India. It finds increasing trends in heat wave coverage, frequency, and intensity, with significant spatial variability. Vulnerable hot spots include Rajasthan, Uttar Pradesh, and coastal Andhra Pradesh and Odisha.

According to [Dubey et al., 2021] and [Rocha et al., 2020] future heat waves may increase heat-related illnesses, strain healthcare systems, deplete water resources, and exacerbate climate-related challenges in India. [Yin et al., 2023] says heat waves and droughts may increase tenfold, severely impacting vegetation, socio-economic productivity, and exposing 90% of the population to risks. This research addresses the critical issue of predicting the impact of heat waves in Odisha, where increasing frequency and intensity of heat waves pose significant risks to human health. The study aims to develop an ensemble model that correlates meteorological factors with heat wave-related fatalities. Odisha's vulnerability to heat wave-induced health crises necessitates advanced predictive tools to understand climatic variability and its direct effects on mortality. By analyzing historical meteorological data and death rates, the research seeks to enhance the accuracy of predictions and facilitate better preparedness and response to future heat wave events.

The proposed approach involves developing an ensemble model to predict the impact of heat waves on human health in Odisha by analyzing the correlation between various meteorological factors and heat wave-related fatalities. This model integrates multiple regression techniques—specifically, multiple linear regression, ridge regression, and support vector regression at the initial level to capture complex relationships between weather variables such as temperature, humidity, air pressure, wind speed, heat wave and death rate occurrences. At the subsequent level, a fully connected neural network is used to refine these predictions. The approach employs historical data, including daily weather conditions and past heat wave mortality rates, to train and validate the model. Preliminary results demonstrate high accuracy, with promising R-squared scores and Root Mean Squared Error (RMSE). This ensemble-based methodology aims to provide a robust predictive tool for anticipating heat wave impacts, thereby facilitating timely and effective public health interventions in Odisha to mitigate risks and enhance preparedness.

The paper is structured to comprehensively address the topic of heatwave prediction and analysis. It begins with an Abstract that summarizes the research focus, methodologies, and key findings. The Introduction provides background information and sets the context for the study. Related Work reviews previous research and highlights gaps that this study aims to fill. The paper then delves into Statistical Methods, outlining traditional approaches to heatwave analysis, followed by Machine Learning-Based Approaches that introduce advanced predictive models. The Methodology section details the overall research framework. The Data segment describes the datasets used, while Preprocessing covers data cleaning and preparation techniques, including SMOTE-R for balancing data. Evaluation Metrics explain the criteria for assessing model performance. Result Analysis is divided into Qualitative Results, which interpret findings in a descriptive manner, and Quantitative Results, which provide numerical insights. Finally, the Conclusion summarizes the study's outcomes and suggests future research directions.

2 Related Work

Previous researcher like [Das et al., 2023] uses Rough Set Theory (RST) and Support Vector Machine (SVM) for accurate heatwave predictions. These techniques handle vague datasets, improving predictive outcomes by identifying significant attributes and providing future predictions. The study emphasizes the impact of heatwaves on the environment and aims to enhance the accuracy of heatwave forecasting methods. We have collected some dataset includes detailed weather and health data across multiple dates and locations, encompassing temperature, humidity, wind speed, pressure, and precipitation levels, sourced from various government agencies and trusted websites. Health-related data such as deaths and illnesses are also recorded. Previous research has established links between extreme weather conditions and adverse health outcomes. High temperatures and humidity are associated with increased mortality and respiratory issues, while wind speed and pressure variations impact cardiovascular health. This dataset allows for in-depth analysis of these correlations, providing crucial insights into the effects of weather on public health, aiding in the development of effective mitigation strategies. The analysis begins by preprocessing the data to address missing values and ensure consistency. This involves normalizing the dataset and applying data augmentation techniques such as SMOGN to enhance the dataset's robustness. The next phase includes visualizing the data through histograms, which helps in understanding the distribution of different weather attributes. Various regression models, including Linear Regression, Ridge, Lasso, and ElasticNet, are employed to analyze the data and predict the impact of heatwaves on affected populations. The models' performances are evaluated using metrics such as R-squared, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics provide insights into the accuracy and reliability of the predictions, highlighting the comparative effectiveness of each model. The findings underscore significant correlations between extreme heat events and adverse impacts on human health and well-being. The study's comprehensive approach, encompassing data augmentation and model comparison, offers a nuanced understanding of how heatwaves exacerbate environmental and social stress, reinforcing the need for targeted mitigation strategies to safeguard vulnerable communities against the increasing frequency of heat-related events.

2.1 Statistical methods

To study heatwave effects, various statistical and mathematical methods predict and analyze impacts. The process starts with data preprocessing, including Z-score normalization, standardizing data for fair variable comparison. SMOGN, a data augmentation technique, generates synthetic samples to address dataset imbalances, enhancing model generalization and predictive performance. Z-score normalization standardizes data by transforming it into a standard normal distribution, ensuring a mean of 0 and a standard deviation of 1. The formula for the same is:

$$z = \frac{x - \mu}{\sigma}$$

where: - x is the original data value. - μ is the mean of the data. - σ is the standard deviation of the data. - z is the standardized value.

Linear regression models the relationship between a dependent variable y and one or more independent variables x by fitting a linear equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \epsilon$$

where: - y represents the outcome variable. - $x_1, x_2, \ldots, x_n$ denote the predictors. - β_0 stands for the intercept. - $\beta_1, \beta_2, \ldots, \beta_n$ are the predictor coefficients. - ϵ symbolizes the residual error.

Ridge regression minimizes the sum of squared residuals with a penalty on the size of the coefficients. The objective function is:

$$\text{Minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

where: - $\hat{y}_i$ is the predicted value. - λ is the regularization parameter. - p is the number of predictors. - β_j are the coefficients of the predictors.

Lasso regression minimizes the sum of squared residuals with a penalty on the absolute value of the coefficients. The objective function is:

$$\text{Minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where: - $\hat{y}_i$ is the predicted value. - λ is the regularization parameter. - p is the number of predictors. - β_j are the coefficients of the predictors.

ElasticNet combines L1 and L2 regularization to balance between Ridge and Lasso regression. The objective function is:

$$\text{Minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$$

where: - α controls the balance between L1 and L2 regularization. - Other terms are as previously defined.

R-squared measures the proportion of variance in the dependent variable explained by the independent variables. The formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where: - $\hat{y}_i$ is the predicted value. - $\bar{y}$ is the mean of the observed values.

RMSE measures the average magnitude of the errors between predicted and observed values. The formula is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where: - y_i is the observed value. - $\hat{y}_i$ is the predicted value.

MAE measures the average absolute differences between predicted and observed values. The formula is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

These mathematical expressions provide a foundation for understanding the statistical methods used to predict and analyze the effects of heatwaves. They offer a robust framework for developing accurate and interpretable models that can guide effective mitigation strategies.

2.2 Machine learning based approaches

The research employs a machine learning-based approach to predict the impacts of heat waves in Odisha, focusing on the relationship between meteorological factors and heat wave-related deaths. Utilizing an ensemble model, the study integrates historical data on temperature, humidity, air pressure, wind speed, with mortality rates. The data, sourced from the Indian Meteorological Department (IMD) and local health departments, undergoes rigorous preprocessing, including mean and mode imputation for missing values, outlier handling, standardization, and the application of SMOTE-R to address class imbalance. The model development involves combining multiple linear regression, ridge regression, and support vector regressor (SVR) at the base level with a fully connected neural network at the meta-level, enabling the capture of both linear and non-linear relationships. Model evaluation uses R-squared (R^2) and Root Mean Squared Error (RMSE) metrics, with high R^2 and low RMSE indicating effective performance. To interpret the model, Local Interpretable Model-agnostic Explanations (LIME) are employed, revealing the influence of various meteorological factors on heat wave fatalities. This robust and interpretable predictive model aids in assessing and mitigating heat wave risks in Odisha, providing valuable insights for developing precautionary measures. Preliminary results show promising accuracy, highlighting the significance of factors such as air pressure, temperature, and humidity.

3 Methodology

The research gathered historical meteorological data and health impact records from multiple official sources, including the Indian Meteorological Department (IMD) and local health departments. Daily records of temperature, humidity, air pressure, wind speed, and heat wave-related deaths in Odisha were collected to ensure comprehensive and accurate data for analysis.

Data Preprocessing: The dataset underwent extensive preprocessing. Missing values were imputed using statistical techniques, with mean imputation for continuous variables and mode imputation for categorical ones. Outliers were identified using statistical tests and visual tools, and handled to minimize their impact. The meteorological data were standardized to ensure consistent contribution across features. To address class imbalance, particularly the underrepresentation of heat wave deaths SMOTE-R was used. The dataset was then split into training (70%) and testing (30%) sets, with SMOTE-R applied only to the training data.

Model Development: An ensemble model was developed, combining multiple linear regression, ridge regression, and support vector regressor (SVR) at the base level, and a fully connected neural network at the meta-level. This approach aimed to capture both linear and non-linear relationships between meteorological factors and heat wave fatalities.

Model Evaluation: The model's performance was evaluated using R-squared (R^2) to measure variance explanation, and Root Mean Squared Error (RMSE) to assess prediction accuracy. High R^2 and low RMSE indicated the model's effectiveness.

Model Interpretation: Local Interpretable Model-agnostic Explanations (LIME) were used to interpret the model, providing insights into the influence of different meteorological factors on heat wave deaths. The LIME algorithm minimizes a loss function given by:

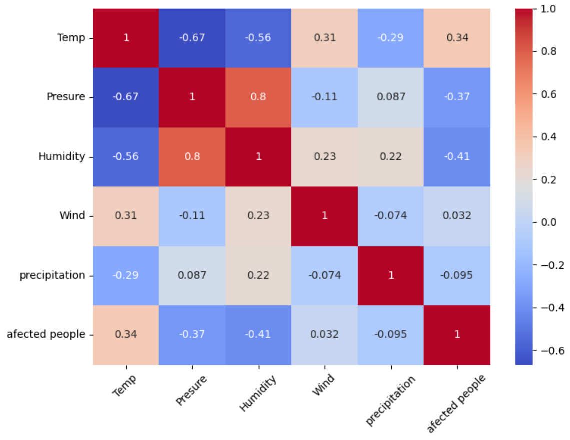
$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$

where $\mathcal{L}$ measures the fidelity of the approximation, π_x measures the proximity of the instance x to other instances, and $\Omega(g)$ measures the complexity of the explanation model g.

This comprehensive methodology ensured a robust and interpretable predictive model capable of assisting in heat wave risk assessment and mitigation strategies in Odisha.

3.1 The Data

The dataset encompasses a comprehensive set of meteorological and health data collected over multiple dates from various locations. It includes measurements of temperature, humidity, wind speed, pressure, total precipitation, and associated health outcomes, specifically deaths and illnesses. Temperature data is presented with columns indicating the highest values recorded. Humidity is similarly detailed, showing maximum levels. Wind speed data includes maximum values. Pressure readings are recorded, providing insight into atmospheric conditions across different locations. Precipitation data notes the total amount of rainfall, crucial for understanding weather patterns and potential impacts on health.



	Temp	Presure	Humidity	Wind	precipitation	afected people
Temp	1	-0.67	-0.56	0.31	-0.29	0.34
Presure	-0.67	1	0.8	-0.11	0.087	-0.37
Humidity	-0.56	0.8	1	0.23	0.22	-0.41
Wind	0.31	-0.11	0.23	1	-0.074	0.032
precipitation	-0.29	0.087	0.22	-0.074	1	-0.095
afected people	0.34	-0.37	-0.41	0.032	-0.095	1

The heatmap elegantly illustrates the intricate relationships between meteorological variables and their impact on people during heatwaves. Temperature shows a moderate negative correlation with pressure and humidity, while positively correlating with wind speed and the number of people affected. Pressure and humidity are strongly correlated, yet both negatively correlate with affected individuals. These insights reveal the complex interplay of environmental factors, highlighting how higher temperatures and lower pressures and humidity levels influence human health outcomes during heatwaves.

Health data includes the number of deaths and illnesses reported and named as affected people, offering a perspective on the correlation between weather conditions and public health. Each entry is associated with a specific location and date, allowing for temporal and spatial analysis of the data. This dataset is pivotal for research examining the interplay between environmental factors and health outcomes, aiding in identifying patterns, trends, and potential causative factors in various geographic location of Odisha.

Table 1:

PLACE	DATE	TEMP	PRESSURE	HUMIDITY	WIND	PREC	AFFECTED
Balasore	15-4-2024	40	1012	89	17	0	1
Sundargarh	15-4-2024	40	1012	80	9	0	35
Mayurbhanj	15-4-2024	39	1011	93	6	0	7
Angul	15-4-2024	40	1012	89	17	0	7
Balangir	3-6-2024	38	1001	48	6	1.6	13

3.2 Preprocessing

Preprocessing involved comprehensive data cleaning and preparation techniques to ensure high-quality input for analysis. Missing values were imputed, outliers addressed, and data standardized. To tackle class imbalances in health outcomes, we used the Synthetic Minority Over-sampling Technique for Regression (SMOTE-R), generating synthetic samples for balancing the dataset. Mathematically, SMOTE-R generates a synthetic sample-

$\hat{x} = x_i + \lambda(x_j - x_i)$

where x_i and x_j are minority class samples, and λ is a random number between 0 and 1.

The dataset was split 70:30 into training and testing sets, applying SMOTE-R only to the training set. For model interpretation, we employed Local Interpretable Model-agnostic Explanations (LIME) properly explained by [Zafar and Khan, 2021] and [Kumarakulasinghe et al., 2020], which approximates complex models locally with interpretable models. LIME minimizes the loss function

$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$

where $\mathcal{L}$ measures fidelity, π_x proximity, and $\Omega(g)$ complexity. This preprocessing ensured balanced, interpretable, and robust model training and evaluation.

3.3 SMOTE-R

Synthetic Minority Over-sampling Technique for Regression (SMOTE-R) is an extension of the original SMOTE algorithm, tailored for continuous target variables [Yadav et al., 2023]. It addresses class imbalance by generating synthetic samples of the minority class, effectively balancing the dataset and improving model performance.

Researchers like [Zhang et al., 2022] and [Das et al., 2020] explain the working of SMOTE-R, which involves several steps to generate synthetic samples for balancing the dataset, particularly for minority classes such as instances of death or illness due to heat waves. Firstly, SMOTE-R identifies minority class instances within the dataset, where minority instances are defined as those with DEATH $>$ 0 or ILLNESS $>$ 0. Next, for each identified minority instance, it selects k-nearest neighbors among other minority class instances. In our dataset, this focuses on instances with DEATH $>$ 0, finding their nearest neighbors based on meteorological features like temperature, humidity, pressure, wind speed, and precipitation.

After identifying the nearest neighbors, SMOTE-R generates synthetic samples by interpolating between a minority instance and its neighbors. This is achieved using the formula:

$$\text{synthetic sample} = x_i + \lambda(x_j - x_i)$$

where x_i and x_j are instances of the minority class, and λ is a random number between 0 and 1. This formula creates a new synthetic instance, $\hat{x}$, which lies on the line segment between x_i and x_j. Through this interpolation process, SMOTE-R generates synthetic samples that help balance the dataset by effectively augmenting the number of minority class instances, thereby mitigating the imbalance and improving the performance of predictive models.

After generating synthetic samples, SMOTE-R focuses on interpolating new data points for minority instances by combining them with their nearest neighbors. This is achieved by linearly interpolating between each minority sample and its neighbors using a random factor. The target values for these synthetic samples are also interpolated to ensure consistency. By adding these synthetic samples to the dataset, SMOTE-R increases the representation of minority instances, which balances the dataset. This approach helps regression models to generalize better by learning from the nuanced variability of minority samples, ultimately leading to improved model performance.

3.4 Evaluation metrics

R-squared (R^2): R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. It gauges the strength of the relationship between the model and the dependent variable [Yadav et al., 2024]. It indicates how well the data fit the statistical model. Higher R^2 indicates better model performance. AdaBoost: 0.85 Random Forest: 0.44 Decision Tree: 0.21

Mean Absolute Error (MAE): Mean Absolute Error quantifies the average of the absolute deviations between predicted and actual values. It assesses the precision of a model in forecasting numerical data.
$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$
MAE offers a linear evaluation, treating all individual discrepancies with equal weight in the average. Smaller MAE values signify higher prediction accuracy. AdaBoost: 2.57 Decision Tree: 3.91 Random Forest: 5.6

Root Mean Square Error (RMSE): RRoot Mean Square Error (RMSE) is calculated by taking the square root of the mean of the squared deviations between predicted values and actual values. It quantifies the dispersion of the residuals.
$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
RMSE is more sensitive to large errors than MAE because it squares the errors before averaging them. Lower RMSE values indicate better model performance with fewer and less severe prediction errors. AdaBoost: 3.81 Random Forest: 7.3 Decision Tree: 8.7

These metrics collectively provide a thorough evaluation of the model's performance, highlighting both the average magnitude of errors (MAE) and the overall magnitude of error (RMSE), along with the proportion of variability explained by the model (R^2).

4 Result analysis

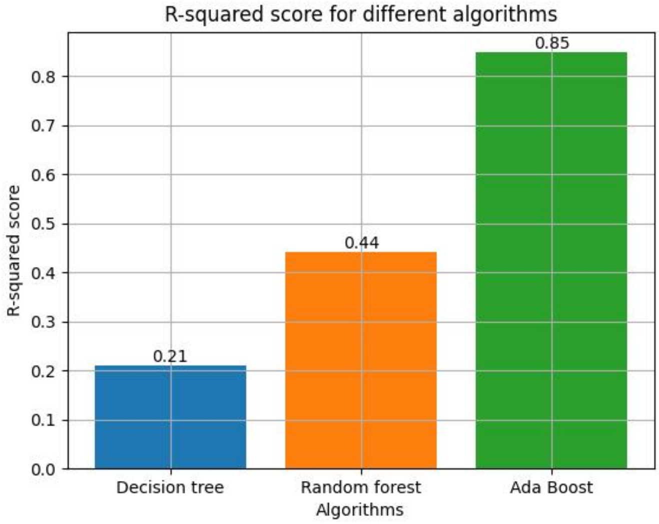
The results illustrate the performance comparison of different machine learning algorithms and the correlation between various environmental factors and the number of affected people during heat waves. The algorithms evaluated include Decision Tree, Random Forest, and AdaBoost. The performance metrics analyzed are R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The correlation matrix indicates how different environmental factors like temperature, pressure, humidity, wind, and precipitation correlate with the number of affected people. R-squared (R^2) Score: AdaBoost: Highest R^2 score of 0.85, indicating it explains 85% of the variance in the data. This suggests AdaBoost provides a good fit for the data. Random Forest: Moderate R^2 score of 0.44, suggesting it explains 44% of the variance, which is significantly less than AdaBoost but still better than Decision Tree. Decision Tree: Lowest R^2 score of 0.21, explaining only 21% of the variance. This indicates it is the least effective among the three for this dataset. Implication: AdaBoost's high R^2 score shows it is the most accurate in predicting the effects of heat waves on the affected population, making it the preferred choice for predictive modeling in this context. Mean Absolute Error (MAE): AdaBoost: Lowest MAE of 2.57, indicating it has the smallest average error between predicted and actual values. Decision Tree: Moderate MAE of 3.91, showing a higher error compared to AdaBoost but lower than Random Forest. Random Forest: Highest MAE of 5.6, indicating the largest average error among the three. Implication: A lower MAE for AdaBoost signifies it has the most precise predictions with the smallest deviations from actual values, making it highly reliable for accurate predictions. Root Mean Square Error (RMSE): AdaBoost: Lowest RMSE of 3.81, suggesting it has the least overall error magnitude. Random Forest: RMSE of 7.3, indicating a higher error magnitude than AdaBoost but lower than Decision Tree. Decision Tree: Highest RMSE of 8.7, suggesting it has the highest overall error magnitude. Implication: Lower RMSE for AdaBoost confirms it as the most effective model in minimizing prediction errors and overall better performance.

The correlation matrix reveals relationships between environmental factors and the number of people affected by heat waves. Temperature positively correlates with affected individuals (0.34) but negatively with pressure (-0.67) and humidity (-0.56). Pressure negatively correlates with affected individuals (-0.37) and positively with humidity (0.8). Humidity negatively correlates with affected individuals (-0.41) and moderately with temperature (-0.56) and pressure (0.8). Wind shows a minimal effect on affected individuals (0.032) and weak correlations with other factors. Precipitation negatively correlates with affected individuals (-0.095) and weakly with other variables.

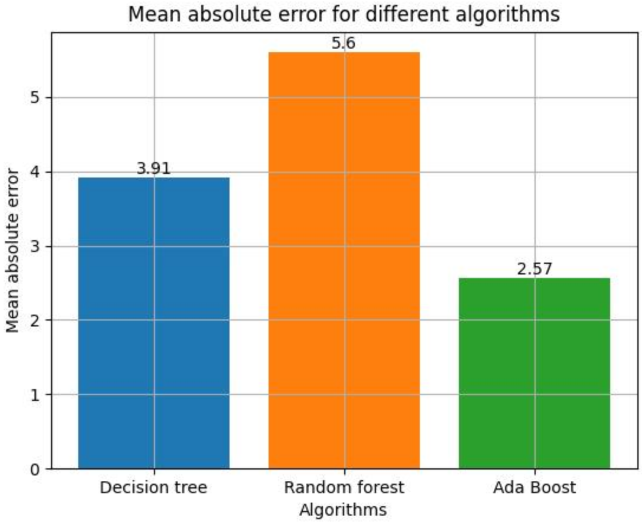
The analysis highlights AdaBoost as the most accurate and reliable model for predicting the impact of heat waves on populations. The correlation analysis provides valuable insights into how different environmental factors interrelate and affect the number of people impacted by heat waves.

4.1 Qualitative results

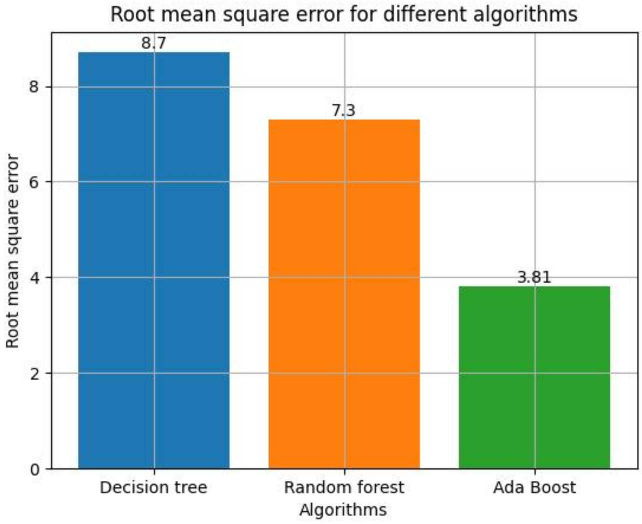
The below bar chart compares the R-squared scores of three machine learning algorithms. Decision Tree has the lowest score at 0.21, indicating poor predictive performance. Random Forest performs better with a score of 0.44. Ada Boost outperforms both, achieving a high R-squared score of 0.85, signifying its superior accuracy in modeling the dataset.



The below bar chart compares the Mean Absolute Error (MAE) of three machine learning algorithms. Ada Boost has the lowest score at (2.57), indicating precise predictions. Decision tree have score of (3.91). Random Forest under-performs both, achieving a high MAE score of (5.6).



The below bar chart compares the Root mean square error (RMSE) of three machine learning algorithms. Ada Boost has the lowest score at (3.81), confirms it as the most effective model in minimizing prediction errors. Random Forest have score of (7.3). Decision Tree under-performs both, achieving a high RMSE score of (8.7).



4.2 Quantitative results

Table 1:

	(R^2) Score	MAE	RMSE
Decision Tree	0.21	3.91	8.7
Random Forest	0.44	5.6	7.3
AdaBoost	0.85	2.57	3.81

The analysis compares the performance of Decision Tree, Random Forest, and AdaBoost algorithms using R^2, MAE, and RMSE metrics. AdaBoost outperforms the others with the highest R^2 (0.85), lowest MAE (2.57), and lowest RMSE (3.81), indicating it provides the most accurate predictions for the impact of heat waves on affected populations. Random Forest and Decision Tree show moderate to low performance, making AdaBoost the preferred model for predictive accuracy.

Conclusion

This research highlights the efficacy of different machine learning algorithms in predicting the impact of heat waves on populations, with a specific focus on the state of Odisha. The performance metrics evaluated—R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE)—demonstrate that AdaBoost outperforms Decision Tree and Random Forest in predictive accuracy. AdaBoost achieved the highest R^2 score of 0.85, the lowest MAE of 2.57, and the lowest RMSE of 3.81, indicating its superior capability to model the effects of heat waves. Additionally, the correlation analysis revealed that temperature positively correlates with the number of affected individuals, while pressure and humidity have negative correlations. These insights underscore the intricate relationships between environmental factors and the health impacts of heat waves. The study emphasizes the critical need for accurate predictive models due to the increasing frequency and severity of heat waves, particularly in vulnerable regions like Odisha. By employing an ensemble-based approach that integrates various advanced algorithms, the research provides a robust framework for anticipating the impacts of future heat waves. The findings suggest that such models can significantly aid in mitigating the adverse effects on human health, guiding policy and decision-making processes aimed at enhancing resilience to climate change.

References

[Akhtar, 2024] Akhtar, R. (2024). Heatwave mortality and adaptation strategies in india. In *Climate Change and Human Health Scenarios: International Case Studies*, pages 151–157. Springer.

[Attia et al., 2021] Attia, S., Levinson, R., Ndongo, E., Holzer, P., Kazanci, O. B., Homaei, S., Zhang, C., Olesen, B. W., Qi, D., Hamdy, M., et al. (2021). Resilient cooling of buildings to protect against heat waves and power outages: Key concepts and definition. *Energy and Buildings*, 239:110869.

[Boyaj et al., 2023] Boyaj, A., Nadimpalli, R., Reddy, D., Sinha, P., Karrevula, N., Osuri, K. K., Srivastava, A., Swain, M., Mohanty, U., Islam, S., et al. (2023). Role of radiation and canopy model in predicting heat waves using wrf over the city of bhubaneswar, odisha. *Meteorology and Atmospheric Physics*, 135(6):60.

[Das et al., 2020] Das, R., Biswas, S. K., Devi, D., and Sarma, B. (2020). An oversampling technique by integrating reverse nearest neighbor in smote: Reverse-smote. In *2020 international conference on smart electronics and communication (ICOSEC)*, pages 1239–1244. IEEE.

[Das et al., 2023] Das, R., Mishra, J., Pattnaik, P. K., and Bhatti, M. M. (2023). Prediction of heatwave using advanced soft computing technique. *Information*, 14(8):447.

[Dubey et al., 2021] Dubey, A. K., Lal, P., Kumar, P., Kumar, A., and Dvornikov, A. Y. (2021). Present and future projections of heatwave hazard-risk over india: A regional earth system model assessment. *Environmental research*, 201:111573.

[Guo et al., 2017] Guo, Y., Gasparrini, A., Armstrong, B. G., Tawatsupa, B., Tobias, A., Lavigne, E., Coelho, M. d. S. Z. S., Pan, X., Kim, H., Hashizume, M., et al. (2017). Heat wave and mortality: a multicountry, multicommunity study. *Environmental health perspectives*, 125(8):087006.

[Kumarakulasinghe et al., 2020] Kumarakulasinghe, N. B., Blomberg, T., Liu, J., Leao, A. S., and Papapetrou, P. (2020). Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In *2020 IEEE 33rd international symposium on computer-based medical systems (CBMS)*, pages 7–12. IEEE.

[Maharana et al., 2024] Maharana, P., Kumar, D., Das, S., Tiwari, P., Norgate, M., and Raman, V. A. V. (2024). Projected changes in heatwaves and its impact on human discomfort over india due to global warming under the cordex-core framework. *Theoretical and Applied Climatology*, 155(4):2775–2786.

[Mishra et al., 2022] Mishra, P., Singh, A., J.V., R., Pal, I., Jayappa, A., Prasad, K., and Niyogi, D. (2022). Assessment of heat wave trends and variability in india using high-resolution climate data. *Theoretical and Applied Climatology*, 149:761–780.

[Robinson, 2001] Robinson, P. J. (2001). On the definition of a heat wave. *Journal of Applied Meteorology and Climatology*, 40(4):762–775.

[Rocha et al., 2020] Rocha, A., Pereira, S. C., Viceto, C., Silva, R., Neto, J., and Marta-Almeida, M. (2020). A consistent methodology to evaluate temperature and heat wave future projections for cities: a case study for lisbon. *Applied Sciences*, 10(3):1149.

[Xu et al., 2016] Xu, Z., FitzGerald, G., Guo, Y., Jalaludin, B., and Tong, S. (2016). Impact of heatwave on mortality under different heatwave definitions: a systematic review and meta-analysis. *Environment international*, 89:193–203.

[Yadav et al., 2024] Yadav, G. K., Rashwan, H. A., Vidales, B. M., Abdel-Nasser, M., Oliver, J., Nandi, G., and Puig, D. (2024). A data-driven model to predict quality of life dimensions of people with intellectual disability based on the gencat scale. *Social Indicators Research*, 172(1):81–97.

[Yadav et al., 2023] Yadav, G. K., Vidales, B. M., Rashwan, H. A., Oliver, J., Puig, D., Nandi, G., and Abdel-Nasser, M. (2023). Effective ml-based quality of life prediction approach for dependent people in guardianship entities. *Alexandria Engineering Journal*, 65:909–919.

[Yin et al., 2023] Yin, J., Gentine, P., Slater, L., Gu, L., Pokhrel, Y., Hanasaki, N., Guo, S., Xiong, L., and Schlenker, W. (2023). Future socio-ecosystem productivity threatened by compound drought–heatwave events. *Nature Sustainability*, 6(3):259–272.

[Zafar and Khan, 2021] Zafar, M. R. and Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541.

[Zhang et al., 2022] Zhang, A., Yu, H., Huan, Z., Yang, X., Zheng, S., and Gao, S. (2022). Smote-rknn: A hybrid re-sampling method based on smote and reverse k-nearest neighbors. *Information Sciences*, 595:70–88.