Dengue Fever Prediction using Machine Learning: A Case Study of Tropical Cities

Yash Patel
School of Computer Science
University of Nottingham
Nottingham, UK, NG8 1BB
yash97828@gmail.com

Prajwal Gamare
School of Computer Science
University of Nottingham
Nottingham, UK, NG8 1BB
prajwalgamare28@gmail.com

Sanyog Chavhan
School of Computer Science
University of Nottingham
Nottingham, UK, NG8 1BB
sanyogchavhan2016@gmail.com

***Abstract*— The research utilises machine approaches to predict epidemics of dengue fever, a mosquito-transmitted viral illness, in San Juan and Iquitos cities, characterised by tropical and subtropical climates. With climate change affecting environmental conditions, precise prediction of dengue outbreaks is essential in these areas due to the significant public health threat posed by the disease. The predictive models developed examine historical data on climate factors, population demographics, and past dengue occurrences to estimate the total dengue cases for each city, year, and week. We applied data analysis to gain insights and data preprocessing to structure the data appropriately, enhancing the robustness of our models. Using Decision Trees, Random Forest, XGBoost, and Gradient Boosting, we optimised predictions through hyperparameter tuning with grid search CV and Optuna. We also utilised TPOT Regressor for automated machine learning and model selection. Notably, Gradient Boosting, fine-tuned with Optuna, outperformed other models, boosting proactive interventions for public health officials and policymakers. This research contributes to data-centric approaches for managing infectious diseases and highlights the importance of predictive modelling in addressing emerging health issues exacerbated by climate change.**

***Index Terms*—Random Forest, XGBoost, Gradient Boosting, Optuna, TPOT Regressor, Predictive modelling, Public health, Hyperparameter tuning, Data Preprocessing**

I. Introduction

Infectious diseases are notably becoming a concern for global public health particularly in tropical and subtropical regions. Among these diseases, dengue fever, transmitted by mosquitoes, stands out as a substantial threat, affecting millions annually. With climate change exacerbating environmental conditions, accurate prediction of dengue outbreaks is vital for effective control measures. Notably, the work of researchers such as Rachel Lowe et al. [5] emphasizes the importance of climate services in predicting and managing the evolution of dengue seasons in vulnerable regions like Machala, Ecuador.

This study employs advanced machine learning techniques to predict dengue fever outbreaks in San Juan, Puerto Rico, and Iquitos, Peru. These regions are susceptible to outbreaks due to their tropical climates. The research aims to develop robust predictive models based on historical data encompassing climate factors, demographics, and previous dengue occurrences.

Our methodology begins with extensive exploratory data analysis (EDA) to understand dataset characteristics, revealing a disparity in dengue cases between San Juan and Iquitos. Following EDA, data preprocessing involves handling missing values, eliminating redundant columns, and addressing outliers. Categorical variables undergo ordinal encoding, and feature scaling is executed using StandardScaler.

For predictive modelling, we employ supervised learning algorithms like Decision Trees Regressor, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor [1] [3] . Model performance is assessed through metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R^2), with hyperparameter tuning conducted via GridSearchCV, Optuna [2] and TPOT Regressor for optimising the model's performance.

This research aims to provide actionable insights for healthcare practitioners and policymakers, facilitating proactive interventions and resource allocation to mitigate the impact of dengue fever outbreaks amidst evolving climate dynamics. By advancing data-driven approaches, we contribute to managing infectious disease outbreaks and address emerging health challenges.

II. Literature Review

A. XGBoost: A Scalable Tree Boosting System

The machine learning and data mining sector was revolutionized by a scalable tree boosting system known as XGBoost proposed by Chen and Guestrin (2016) among other things. Researchers realized that there had been little previous work done on scale algorithms for tree boosting which led them to develop XGBoost. This review article points out the significance of XGBoost in dealing with scalability issues as well as its huge contribution towards advancing methods in machine learning. [1]

B. Optuna: A Next-generation Hyperparameter Optimisation Framework

Optuna, a next generation hyperparameter optimisation framework designed by Akiba et al. (2019), fills an essential gap in efficient optimisation techniques required for machine learning This is a breakthrough in this area that gives researchers an instrument to automate hyperparameter tuning processes. [2]

C. Greedy function approximation: A gradient boosting machine

Gradient boosting machine, a powerful algorithm for regression and classification tasks, was introduced by Friedman (2001). The concept of greedy function approximation serves as the basis for this milestone. [3]

D. Fast binary feature selection with conditional mutual information

To select informative features in machine learning tasks quickly, Fleuret (2004) came up with a method that selects binary features based on conditional mutual information. By identifying relevant features while reducing computational

overheads, this approach has greatly contributed to improving model performance. [4]

E. Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador

Lowe et al.'s (2018) climate services for health is a predictive model designed to forecast dengue fever outbreaks based on climate data. Their contribution underscores the need for integrating climate information into healthcare systems to adequately respond to infectious disease outbreaks. [5]

III. Data Science Project Lifecycle

Fig. 1. Project Workflow

IV. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is essential for understanding dataset structure and characteristics, guiding feature selection and model development. In this project focusing on predicting outbreaks of dengue fever, EDA enables the identification of missing values, trends, and outliers, thereby informing subsequent modelling decisions for accurate forecasting.

The dataset utilised for the dengue prediction project comprised 1456 instances with 24 features from the dengue_features_train.csv file, and additional information on city, year, weekofyear, and total cases from the dengue_labels_train.csv file. Initially, these datasets were merged into a single training dataset based on common identifiers such as city, year, and weekofyear, resulting in a dataset with dimensions 1456×25. A separate test dataset (dengue_features_test.csv) containing 416 instances was also employed.

The info () function was utilised to examine the dataset's structure and data types. Notably, the week_start_date feature required conversion to datetime format and was subsequently broken down into three separate features (start year, start_month, start date). Missing values were identified across various features, categorised into numerical and temporal variables. Visualisation techniques, including line plots and histograms, were employed to assess the impact of missing values on the target variable (total_cases). Additionally, categorical variables like 'city' were analysed for distribution and their relationship with the target variable.

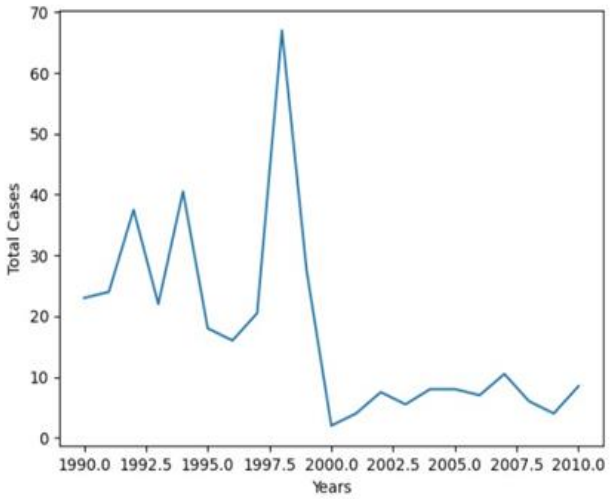

Fig. 2. Line graph showing the relationship between total cases and years.

The line graph in Figure 2 reveals a notable spike in dengue cases between 1997 and 1999, suggesting a significant increase in disease incidence during this period.

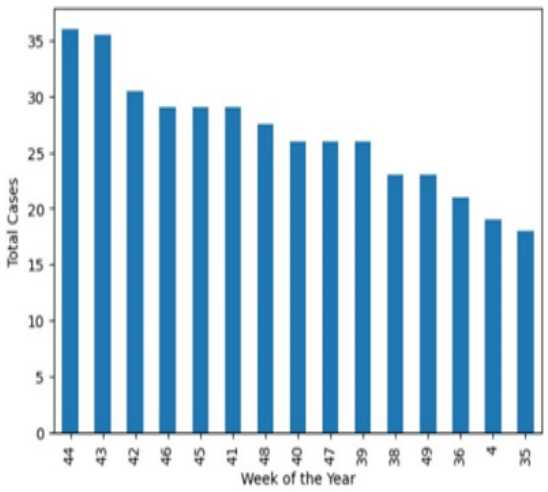

Fig. 3. Histogram showing the relationship between total cases and weeks.

Similarly, Figure 3 demonstrates a pronounced surge in total cases observed between weeks 35 and 49, indicating a possible seasonal pattern or environmental factor contributing to disease transmission.

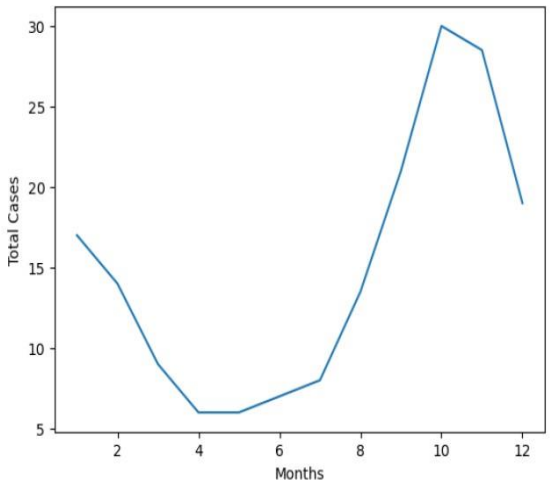

Fig. 4. Line graph showing the distribution of total cases across months.

The line graph analysis presented in Figure 4 unveils a distinct peak in dengue cases during months 8 to 10, highlighting a period of heightened disease activity.

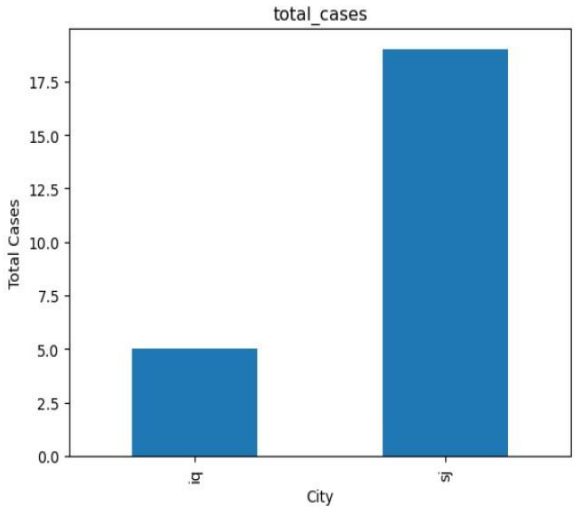

Fig. 5. Bar plot showing the distribution of total dengue cases across San Juan and Iquitos

Additionally, upon examining the bar plot of total cases vs city, it becomes evident that San Juan had significantly more cases than Iquitos, emphasising geographical disparities in disease burden.

V. Feature Engineering

Dengue fever is influenced by various factors, encompassing environmental conditions, demographics, and historical disease patterns. Through feature engineering techniques, including addressing missing data, managing outliers, transforming variables, and encoding categorical information, the dataset underwent refinement to capture these intricate connections and trends more effectively. This preprocessing step was fundamental in elevating the performance of machine learning models, ensuring they received meaningful and representative input data. Moreover, effective feature engineering contributed to the robustness and generalisability of the models, enabling them to provide accurate and reliable predictions of dengue fever outbreaks across different regions and time periods.

In the process of preparing the dataset for model training, several key steps were taken to enhance the quality and compatibility of the features:

A. Handling Missing Values

Missing values were imputed with the median values of their respective features to ensure robustness against outliers and prevent bias from mean imputation. This choice prevents extreme values from skewing the imputations, ensuring that the replaced values are more representative of the dataset's central tendency.

B. Removing Irrelevant Columns

Certain columns were identified as irrelevant for the modelling process and were subsequently dropped from the dataset. These included columns such as week_start_date and start year, which were not expected to contribute significantly to the prediction of dengue fever outbreaks.

C. Addressing Skewed Numerical Variables

To address skewness in numerical variables, a log-normal distribution transformation was applied, ideal for data with skewed distributions. By logarithmically transforming the data, this method compresses the long tail, reducing the influence of extreme values and yielding a more symmetric distribution. This enhances the suitability of variables for modelling and improves the interpretability and performance of statistical analyses.

D. Handling Outliers

Outliers can disrupt the underlying patterns and relationships within the data, resulting in skewed predictions from the model. Outliers in the numerical variables were detected and handled by establishing upper and lower boundaries based on the interquartile range (IQR) for each feature. Values exceeding the upper boundary were replaced with this boundary value, and values below the lower boundary were substituted similarly. This method ensured that extreme values did not skew subsequent analyses or model predictions, preserving the dataset's integrity and enhancing the reliability of insights derived from the data. For a visual representation of this outlier handling process, refer to Figure 6, which illustrates the impact of outlier removal on one of the numerical variables.

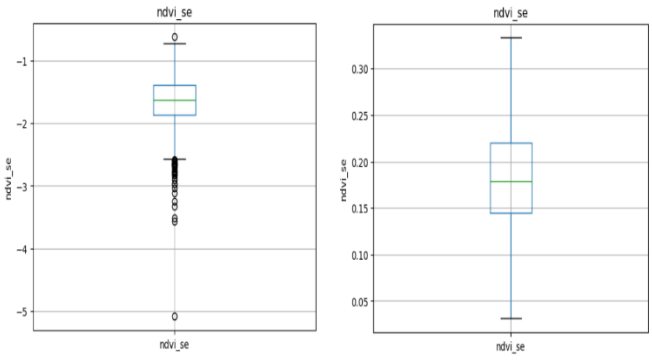

Fig. 6. Before and After handling outliers

E. Encoding Categorical Variables

Categorical variables, such as 'city', play a crucial role in predictive modelling but need to be converted into numerical format for algorithmic processing. In this study, 'San Juan' was mapped to 0 and ' Iquitos' to 1, simplifying city labels into numerical representations. This encoding improved the model's ability to process categorical data effectively, aiding in accurate predictions.

F. Transforming Temporal Variables

Temporal variables, which include features like 'year', were transformed separately to ensure they are compatible with machine learning algorithms. These variables often require special treatment due to their chronological nature and encoding them using ordinal encoding maintains their temporal order while providing numerical representations for model training.

G. Standardisation

Standardisation, a preprocessing technique, scales the numerical features to have a mean of 0 and a standard deviation of 1, making them comparable and thereby improving model performance. StandardScaler was employed to standardise the selected features in the dataset. The scaler was applied to features excluding the target variable "total cases" using the transform method, ensuring consistency in the scale of the features across the dataset. This process enhances the interpretability of model coefficients and aids algorithms in converging faster during training.
The standardisation formula is given by:

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

Where: z denotes the standardised value, x denotes the feature value, μ signifies the mean of the feature's values, and σ indicates the standard deviation of the feature's values.

VI. Feature selection

Choice of feature is important as it aids in recognising the most useful predictors for a response variable thus increasing model accuracy and interpretability through noise reduction and overfitting avoidance on the dataset. Mutual Information (MI) Regression analysis was employed for this task. MI regression measures the amount of information that flows between variables which are features with respect to an outcome or target variable.

Higher mutual information scores signify stronger relationships with the target variable, guiding us in selecting the top k features using the SelectKBest method [4]. With k set to 15, we aimed to streamline the dataset's dimensionality while retaining the most significant insights. These carefully chosen features were then utilised in subsequent modelling endeavours to develop accurate forecasts for dengue fever outbreaks.

VII. Model building

The section on model building delineates the steps involved in constructing predictive models for forecasting dengue fever outbreaks. In this research, we applied several machine learning models to predict dengue fever outbreaks, including Decision Trees, Random Forest Regression, Gradient Boosting Regression and XGBoost Regression.

A. Decision Tree Regression

Decision Tree Regression is a machine learning technique for predicting continuous numerical values. This is accomplished by separating the data into smaller subsets depending on the input attributes, with the goal of maximising the purity of the target variable. For each interior node, it chooses a feature and split point that will yield the purest child nodes as measured by criteria such as mean squared error (MSE). The division continues until stopping conditions are satisfied so that leaf nodes have estimated target values in them. This means that this algorithm inherently does feature selection too; it selects those informative features and splits which maximise purity over all subgroups. Here, purity refers to how much alike are the objects within a node with respect to their outcomes; thus, higher levels of homogeneity indicate stronger predictive power among different classes at any given level. Decision trees can serve as base models in ensemble methods like random forests or gradient boosting where they are designed to improve overall prediction accuracy.
The formula for MSE is:

$$\text{MSE} = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$$

where n signifies the total data points, yi stands for observed target value, and $y\hat{}i$ represents the predicted target value.

The results of the Decision Tree model on our dataset are outlined below:

- Mean Squared Error: 0.597
- Mean Absolute Error: 0.557
- R-squared: 0.629

B. Random Forest Regression

Random Forest Regression is a method in machine learning that enhances accuracy and prevents overfitting by combining predictions from multiple decision tree regressors. It uses a technique called bagging, where each decision tree is trained on a random subset of the training data with replacement. Additionally, it utilises only some features chosen randomly during the creation of each tree for splitting at any node. By so doing, this procedure ensures that trees are less correlated which reduces variance and raises the overall predictive power of the ensemble. Ultimately, the last calculation involves taking averages from all the decision tree predictions. Random Forest Regression takes advantage of individual trees' strengths while neutralising their weaknesses through this strategy of merging many weaker learners; thus, producing strong and highly accurate models for regression analysis which can capture complex nonlinearities in data especially well.

The results of the Random Forest Regression model on our dataset are outlined below:

- Mean Squared Error: 0.363
- Mean Absolute Error: 0.472
- R-squared: 0.774

C. Gradient Boosting Regression

Gradient Boosting Regressor is a strong approach that combines weak regression models, specifically decision trees, to provide accurate predictions. It does this by adding new models one by one, each focusing on the errors left behind by the previous ones. Each new model is trained on data instances that were poorly predicted by the existing ensemble, with higher weights assigned to those instances. The boosting approach allows the algorithm to iteratively learn and correct for errors made by the previous models, thereby capturing complex nonlinear patterns in the data. The final prediction is a weighted sum of all the individual model predictions in the ensemble. This boosting ensemble approach, where each new model tries to compensate for the weaknesses of the previous ensemble, leads to improved predictive performance over a single model. Gradient Boosting Regressor combines the strengths of decision trees and boosting, making it a powerful technique for regression tasks.

The results of the Gradient Boosting Regression model on our dataset are outlined below:

- Mean Squared Error: 0.454
- Mean Absolute Error: 0.521
- R-squared: 0.717

D. XGBoost Regression

XGBoost, also known as Extreme Gradient Boosting, is a highly effective version of the gradient boosting algorithm widely recognised for its superior predictive modelling performance. What makes XGBoost stand out is its efficient implementation and unique features. It tackles overfitting by using regularisation, which helps prevent overly complex models. Additionally, XGBoost employs a faster tree-building process called "greedy function approximation," which speeds up model training while maintaining accuracy. It also introduces "column subsampling," allowing for the selection of random subsets of features, adding further robustness to the model. These enhancements, along with its scalability and versatility, make XGBoost a popular choice for various machine learning tasks, especially in structured data scenarios.

Mathematically, the objective function for XGBoost Regression can be expressed as:

$$\text{Objective} \quad = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \Omega(f)$$

where n denotes the total number of samples, yi represents the observed target value, yˆi signifies the predicted target value, and Ω(f) stands for the regularisation term which penalizes the model's complexity.

The results of the XGBoost Regression model on our dataset are outlined below:

- Mean Squared Error: 0.364
- Mean Absolute Error: 0.462
- R-squared: 0.774

In summary, we implemented and evaluated multiple machine learning models for predicting dengue fever outbreaks. Each model demonstrated varying levels of performance, with XGBoost Regression achieving the highest R2 score of 0.774.

VIII. HYPERPARAMETER TUNNING

This section discusses the optimisation of model hyperparameters to improve performance and generalisation ability. Methods such as grid search or random search can be utilised to refine model parameters. Similarly, genetic algorithms like Optuna and TPOT Regressor are also effective in optimising model performance through automated search processes.

A. Grid Search Cross-Validation (Grid Search CV)

Grid Search CV is a technique to automatically tune the hyperparameters of a machine learning model. It involves training and evaluating the model across a wide grid of hyperparameter value combinations. Cross-validation is employed on the training data to evaluate the model's performance for each parameter combination. Cross-validation on the training data is used to estimate the model's performance for each combination. The hyperparameter values that result in the best performance metric are chosen as optimal. Grid Search CV eliminates manual tuning and guesswork in finding ideal hyperparameters. However, it can be computationally expensive for large datasets or search spaces.

In our project, Grid Search CV was particularly useful for fine-tuning the hyperparameters of our machine learning models, such as Random Forest Regression, Gradient Boosting Regression, and XGBoost Regression. Grid Search CV optimises regression model performance by exploring multiple hyperparameter combinations inside a given grid. It focuses on metrics like reducing mean squared error and mean absolute error, as well as increasing the R^2 score to enhance model accuracy and lower prediction errors.

The best performing model, Gradient Boosting Regression, optimised through Grid Search, achieved the following results on our dataset:

- Mean Squared Error: 0.324
- Mean Absolute Error: 0.449
- R-squared: 0.799

B. Optuna

Optuna is an automated hyperparameter optimisation framework that uses genetic algorithm approaches to systematically explore the hyperparameter space to find the most effective hyperparameter combinations. Unlike traditional methods like Grid Search, which perform an exhaustive search over predefined hyperparameter values, Optuna dynamically samples hyperparameters based on their performance, iteratively improving the model's performance.

In our project, Optuna played a pivotal role in fine-tuning the hyperparameters of our machine learning models, encompassing Gradient Boosting, Random Forest, and XGBoost Regression. This strategic application of Optuna ensured the optimisation of model parameters, particularly enhancing the performance of Gradient Regression. By intelligently sampling hyperparameters using Bayesian optimisation, Optuna effectively navigated the high-dimensional search space and identified promising configurations that maximised the model's performance

metrics, such as R^2 score, mean squared error, and mean absolute error.

The importance of Optuna in our project lies in its ability to efficiently explore the hyperparameter space and adaptively adjust the sampling strategy based on the observed performance of previous configurations. This enables Optuna to converge to the optimal set of hyperparameters quickly and effectively, even in complex and high-dimensional search spaces.

The best performing model, Gradient Boosting Regression, optimised through Optuna, achieved the following results on our dataset:

- Mean Squared Error: 0.297
- Mean Absolute Error: 0.426
- R-squared: 0.815

C. Tree-based Pipeline Optimization Tool(TPOT)

TPOT Regressor is an automated machine learning tool that employs genetic programming techniques to efficiently explore the model pipeline space and find the optimal sequence of preprocessing steps and machine learning models. Unlike traditional methods that require manual selection and tuning of algorithms, TPOT dynamically constructs and evaluates multiple pipeline configurations, iteratively improving the model's performance.

In our project, the TPOT Regressor was pivotal in automating the process of selecting models and tuning hyperparameters for regression tasks. By intelligently combining various preprocessing techniques, such as feature selection and feature scaling, with regression models like decision trees or linear models, TPOT effectively navigated the vast space of possible pipeline configurations and identified the optimal combination that maximised performance metrics like R^2 score, mean squared error, and mean absolute error.

The importance of TPOT Regressor in our project lies in its ability to efficiently explore the pipeline space and adaptively adjust the search strategy based on the observed performance of previous configurations. This enables TPOT to converge to the optimal pipeline quickly and effectively, even in complex and high-dimensional datasets, saving significant time and effort compared to manual experimentation.

The results of the TPOT Regression model on our dataset are outlined below:

- Mean Squared Error: 0.387
- Mean Absolute Error: 0.483
- R-squared: 0.759

IX. Results

In this research, we examined the performance of various machine learning models on our dataset, focusing on three key evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 Score. These metrics provide insights into the models' predictive accuracy, precision, and goodness of fit, respectively. Across all experiments, we meticulously tracked these metrics to comprehensively evaluate the efficacy of different models and hyperparameter tuning methods.

The baseline results revealed promising performance across all models, with XGBoost achieving the lowest MSE of 0.364 and the highest R2 Score of 0.774. However, upon applying hyperparameter tuning techniques, significant improvements were observed.

When employing Grid Search CV, Gradient Boosting exhibited remarkable performance with an MSE of 0.324 and an R2 Score of 0.799, showcasing the effectiveness of this method in optimising model parameters. Optuna, another hyperparameter tuning technique, further refined the models' performance. Particularly noteworthy was the performance of Gradient Boosting Regressor with Optuna, achieving an MSE of 0.297 and an R2 Score of 0.815, surpassing the results obtained through other tuning methods.

Additionally, while TPOT didn't achieve the lowest MSE or the highest R2 Score, it provided a balanced improvement across all metrics, with an MSE of 0.387, an MAE of 0.483, and an R2 Score of 0.759. TPOT's strength lies in its ability to dynamically construct and evaluate multiple pipeline configurations, making it a valuable tool for comprehensive model exploration.

Overall, our findings highlight the importance of hyperparameter tuning in improving the predictive performance of machine learning models, with Optuna emerging as particularly effective strategy when combined with Gradient Boosting in our experimental setup.

Here is the comparison table of the model performances and hyperparameter tuning methods:

Model	**Hyperparameter Tuning Method**	**MSE**	**MAE**	**R2 Score**
Decision Trees	None	0.597	0.557	0.629
Random Forest	None	0.363	0.472	0.774
Gradient Boosting	None	0.454	0.521	0.717
XGBoost	None	0.364	0.462	0.774
Random Forest	Grid Search CV	0.567	0.607	0.647
Gradient Boosting	Grid Search CV	0.324	0.449	0.799
XGBoost	Grid Search CV	0.339	0.458	0.789
Random Forest	Optuna	0.359	0.470	0.777
Gradient Boosting	Optuna	0.297	0.426	0.815
XGBoost	Optuna	0.315	0.444	0.804
TPOT	AutoML	0.387	0.483	0.759

X. Discussion

The outcomes of this investigation illustrate the remarkable potential of machine learning techniques, particularly ensemble methods like XGBoost and Gradient Boosting, in accurately forecasting dengue fever outbreaks.

By leveraging historical data on environmental factors, population demographics, and previous disease occurrences, these models exhibited superior predictive capabilities compared to traditional approaches.

Notably, the iterative nature of ensemble methods, which combine multiple weak learners and focus on residual errors, enabled them to capture complex nonlinear patterns underlying dengue transmission dynamics effectively. The synergy between Gradient Boosting and the Optuna hyperparameter optimisation framework further accentuated the model's performance, achieving the lowest mean squared error (0.297) and the highest R-squared score (0.815) on the dataset.

The success of Gradient Boosting can be attributed to its powerful capabilities in addressing the complexities of modelling, including regularisation techniques that mitigate overfitting and the model's ability to sequentially enhance the accuracy of predictions. The algorithm's focus on refining weak learners through the optimization of decision trees enables it to capture intricate, nonlinear patterns within the data. In conjunction with Optuna's efficient Bayesian optimization approach, Gradient Boosting demonstrated its adaptability and proficiency in navigating high-dimensional hyperparameter spaces, facilitating rapid convergence towards optimal configurations. This combination resulted in outstanding model performance and robust generalization across the dataset.

While Grid Search CV and TPOT also demonstrated their effectiveness in optimising model performance, Optuna emerged as the most efficient and effective hyperparameter tuning technique in this study. Its ability to dynamically sample hyperparameters based on observed performance and adaptively adjust the sampling strategy enabled rapid convergence to optimal configurations, even in complex search spaces.

XI. Conclusion

This research study has effectively demonstrated the potential of machine learning techniques, particularly ensemble methods like XGBoost and Gradient Boosting, to accurately forecast dengue fever outbreaks. Through the integration of preprocessing steps, such as data cleaning, feature engineering, and feature scaling, the models were prepared to effectively leverage environmental factors, demographic information, and historical disease patterns. This comprehensive preprocessing approach ensures that the developed models can provide timely alerts to public health authorities, facilitating proactive interventions and resource allocation.

Moreover, the integration of advanced hyperparameter tuning techniques, such as Optuna, further optimised model performance, emphasising the importance of rigorous parameter exploration in machine learning applications. The collaborative use of Optuna and Gradient Boosting emerged as the most effective approach, achieving superior predictive accuracy and generalisation capabilities.

By contributing to data-driven approaches for managing infectious diseases, this study underscores the significance of not only predictive modelling but also robust preprocessing techniques in addressing emerging health challenges exacerbated by climate change and environmental factors. The findings pave the way for developing comprehensive early warning systems and tailored interventions, ultimately contributing to improved public health outcomes and resource allocation strategies.

XII. Recommendations for Future Research

While this research has yielded promising results, several avenues for future work exist:

A. Incorporation of Additional Factors:

Exploring a wider range of features, such as air pollution levels, urbanisation patterns, and health infrastructure, could potentially enhance the models' predictive capabilities.

B. Extension to Other Geographic Regions:

Applying the developed models to diverse geographic locations with varying climatic conditions and dengue prevalence would assess their generalisability and facilitate tailored interventions.

C. Integration with Early Warning Systems:

Collaborating with public health authorities and policymakers to integrate the predictive models into comprehensive early warning systems would facilitate timely interventions and resource allocation.

D. Expansion to Other Infectious Diseases:

Adapting the methodological framework to forecast outbreaks of other climate-sensitive infectious diseases, such as malaria or Zika, could broaden the impact of this research on global health initiatives.

Exploring these future paths can help researchers improve and expand the use of machine learning for public health issues, creating a more proactive and data-driven approach to managing infectious diseases amid environmental changes.

XIII. References

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[2] e. a. Takuya Akiba, "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[3] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, pp. 1189-1232, 2001.

[4] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1531-1555, 2004.

[5] e. a. Rachel Lowe, "Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador," *The Lancet Planetary Health*, vol. 2, no. 12, pp. 508-517, 2018.