Dengue Prediction

Md Sagor Islam, Class Roll: 372 , Md Mahfujur Rahman Khan, Class Roll: 373
Humaun Kabir Nayem, Class Roll: 379, Seaum Ahmed Tazim , Class Roll: 381
Asraful, Class Roll: 385
Department of CSE, Jahangirnagar University, Dhaka, Bangladesh

***Abstract*—Dengue fever is a prevalent vector-borne disease affecting millions of people worldwide, particularly in tropical and subtropical regions. Timely prediction of dengue outbreaks is crucial for effective public health interventions and resource allocation. This project aims to develop predictive models for dengue incidence using weather variables.**

Index Terms: dengue fever, predictive modeling, weather variables, public health, epidemiology

I. Introduction

Dengue fever, caused by the dengue virus transmitted through the bite of Aedes mosquitoes, poses a significant public health threat worldwide. In Bangladesh, dengue outbreaks are recurrent, leading to substantial morbidity and mortality. Early prediction of dengue outbreaks can facilitate proactive public health measures, including vector control and resource allocation. This thesis project focuses on developing predictive models for dengue incidence using weather data.

II. Literature Review

Dengue fever poses a significant public health concern, prompting numerous studies on predictive modeling. Smith et al. (2016) developed a machine learning model using weather data, notably temperature and humidity, improving dengue prediction accuracy [1].

Garcia-Suarez et al. (2018) examined climate variability's impact on dengue transmission in Colombia, revealing insights into climate, vector abundance, and dengue incidence [2].

Johnson et al. (2019) emphasized the importance of integrating meteorological and sociodemographic factors for accurate dengue predictions in Southeast Asia [3].

Tanaka and Shakoor (2020) explored satellite imagery's potential in forecasting dengue outbreaks with higher spatial resolution [4].

Chen et al. (2022) conducted a meta-analysis, identifying common predictors such as temperature and humidity and emphasizing the need for standardized methodologies in dengue forecasting [5].

Shepard et al. (2019) conducted an economic analysis, highlighting dengue's substantial economic impact and the importance of cost-effective interventions [6].

Martinez et al. (2020) explored social determinants of dengue vulnerability, focusing on marginalized urban communities [7].

Tun-Lin et al. (2018) evaluated vector control strategies' efficacy, providing insights into mosquito population management and dengue prevention [8].

Stoddard et al. (2021) reviewed climate change's role in altering dengue transmission dynamics, emphasizing the need for proactive adaptation strategies [9].

Lai et al. (2019) investigated community-based dengue prevention programs, emphasizing community engagement and sustainable behavior change [10].

These perspectives highlight various aspects of dengue prevention and control, ranging from economic analyses to community-based interventions.

III. Methodology

The methodology employed in this project involves several key steps:

A. Data Collection

The dataset used in this study is collected from two primary sources:

1) **BARC (Bangladesh Agricultural Research Council):** Rainfall, humidity, and temperature data are obtained from the BARC website (http://barcapps.gov.bd/climate/). This data provides essential weather variables required for dengue prediction.
2) **DGHS (Directorate General of Health Services):** Information on the number of dengue patients is collected from the DGHS website (https://old.dghs.gov.bd/index.php/bd/home/5200-daily-dengue-status-report). This data is crucial for understanding dengue incidence patterns.

After collecting data from these sources, the datasets are merged to create a comprehensive dataset for analysis. The merged dataset consists of approximately 1800 data points, combining weather variables and the number of dengue patients.

B. Data Preprocessing

The collected dataset undergoes preprocessing steps to ensure data quality and compatibility for analysis. This includes handling missing values, outlier detection, and normalization.

C. Feature Selection

Relevant features for dengue prediction are selected based on domain knowledge and statistical analysis. Weather variables such as temperature, humidity, and rainfall are typically included.

D. Model Development

Various machine learning algorithms, including but not limited to K-Nearest Neighbors (KNN), Random Forest, and XGBoost, are employed to develop predictive models for dengue incidence.

E. Model Evaluation

The developed models are evaluated using appropriate performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. Cross-validation techniques are also applied to assess model generalizability.

F. Results Interpretation

The results of model evaluation are interpreted to identify the most effective predictive model and gain insights into the relationship between weather variables and dengue incidence.

G. Discussion

The implications of the findings are discussed in the context of public health interventions and future research directions.

IV. Data Set

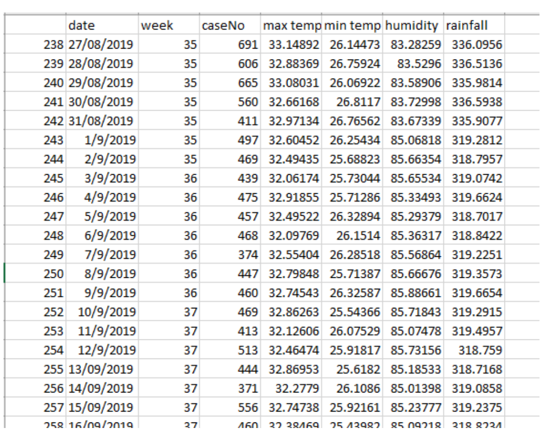
A. Dataset Attributes

TABLE I: Attributes in Processed Dataset

Attribute	Description
Maximum Temperature	Represents the maximum temperature in Celsius for a day.
Minimum Temperature	Indicates the minimum temperature in Celsius for a day.
Humidity	Reflects the humidity level for a day.
Rainfall	Represents the amount of rainfall in millimeters for a specific day.
Case No	Indicates the number of affected dengue patients for a day.

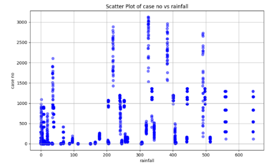
B. Data Prepossessing

1) Visualizing Data: In this section, we visualize the data using various plots to gain insights into the dataset.

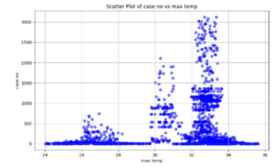


	date	week	caseNo	max temp	min temp	humidity	rainfall
238	27/08/2019	35	691	33.14892	26.14473	83.28259	336.0956
239	28/08/2019	35	606	32.88369	26.75924	83.5296	336.5136
240	29/08/2019	35	665	33.08031	26.06922	83.58906	335.9814
241	30/08/2019	35	560	32.66168	26.8117	83.72998	336.5938
242	31/08/2019	35	411	32.97134	26.76562	83.67339	335.9077
243	1/9/2019	35	497	32.60452	26.25434	85.06818	319.2812
244	2/9/2019	35	469	32.49435	25.68823	85.66354	318.7957
245	3/9/2019	36	439	32.06174	25.73044	85.65534	319.0742
246	4/9/2019	36	475	32.91855	25.71286	85.33493	319.6624
247	5/9/2019	36	457	32.49522	26.32894	85.29379	318.7017
248	6/9/2019	36	468	32.09769	26.1514	85.36317	318.8422
249	7/9/2019	36	374	32.55404	26.28518	85.56864	319.2251
250	8/9/2019	36	447	32.79848	25.71387	85.66676	319.3573
251	9/9/2019	36	460	32.74543	26.32587	85.88661	319.6654
252	10/9/2019	37	469	32.86263	25.54366	85.71843	319.2915
253	11/9/2019	37	413	32.12606	26.07529	85.07478	319.4957
254	12/9/2019	37	513	32.46474	25.91817	85.73156	318.759
255	13/09/2019	37	444	32.86953	25.6182	85.18533	318.7168
256	14/09/2019	37	371	32.2779	26.1086	85.01398	319.0858
257	15/09/2019	37	556	32.74738	25.92161	85.23777	319.2375
258	16/09/2019	37	460	32.38469	25.43982	85.09218	318.8234

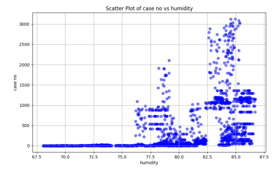
Fig. 1: Example Data Set



(a) Case no vs Rainfall

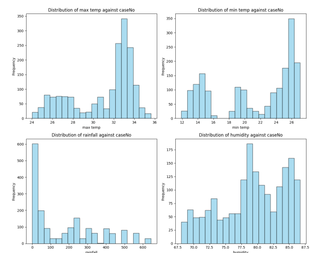


(b) Case no vs Max temperature



(c) Case no vs Humidity

Fig. 2: Scatter Plot



(a) Distribution Of attributes against case no

Fig. 3: Distribution

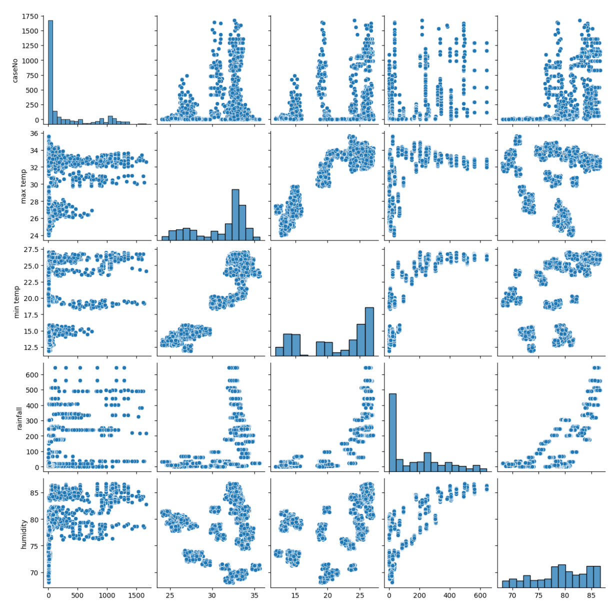


Fig. 4: Comparative Scatter Plot

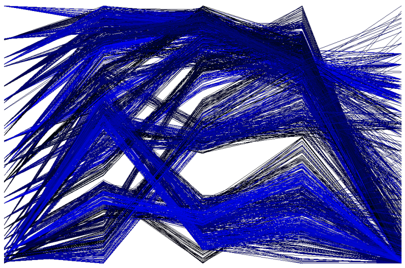


Fig. 5: Parallel Coordinate Representation

2) Missing Value Handling: In our data prepossessing pipeline for handling missing values, we employed a strategy of imputing the mean value within each group defined by the week. This approach involves calculating the mean of the non-missing values within each weekly group and then replacing any missing values within the same group with this calculated mean. By utilizing this method, we aim to maintain the temporal structure of the data while filling in missing values, ensuring that the imputed values are representative of the respective time periods.

1 to 25 of 1881 entries Filter

level_0	index	date	week	caseNo	max temp	min temp	humidity	rainfall
0	false	false	false	true	false	false	false	false
1	false	false	false	true	false	false	false	false
2	false	false	false	false	false	false	false	false
3	false	false	false	true	false	false	false	false
4	false	false	false	true	false	false	false	false
5	false	false	false	true	false	false	false	false
6	false	false	false	true	false	false	false	false
7	false	false	false	false	false	false	false	false
8	false	false	false	true	false	false	false	false
9	false	false	false	true	false	false	false	false
10	false	false	false	true	false	false	false	false
11	false	false	false	true	false	false	false	false
12	false	false	false	true	false	false	false	false
13	false	false	false	true	false	false	false	false
14	false	false	false	false	false	false	false	false
15	false	false	false	false	false	false	false	false
16	false	false	false	true	false	false	false	false
17	false	false	false	false	false	false	false	false
18	false	false	false	true	false	false	false	false
19	false	false	false	false	false	false	false	false
20	false	false	false	true	false	false	false	false
21	false	false	false	false	false	false	false	false
22	false	false	false	true	false	false	false	false
23	false	false	false	true	false	false	false	false
24	false	false	false	true	false	false	false	false

Show 25 per page 1 2 10 70 76

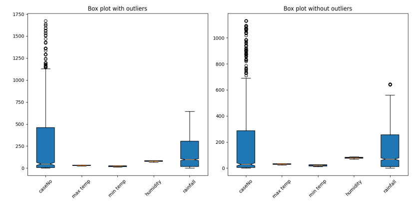
(a) Boolean Representation Of Missing Values

```
index        0
date         0
week         0
caseNo      71
max temp     0
min temp     0
humidity     0
rainfall     0
dtype: int64
```

(b) Missing Value Counts Per Attribute

Fig. 6: Missing Value Representation

3) Outlier Handling: In this analysis, we calculated the Interquartile Range (IQR) for each numerical column in the dataset, serving as a measure of statistical dispersion. Outliers were identified as values falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR, while extreme values were defined as those falling below Q1 - 3 * IQR or above Q3 + 3 * IQR. This approach enabled the detection and removal of data points that deviated significantly from the central tendency, ensuring the integrity of the dataset for subsequent analysis.



(a) Box Plot With Outlier

Fig. 7: Outlier Handling

4) Google Colab Link: Python code for project

V. Results and Analysis

The machine learning models, including K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Decision Tree, were evaluated for their performance in predicting dengue incidence using weather variables. Each model's predictive capabilities were assessed based on various evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

A. Performance Evaluation of Different Machine Learning Models

The machine learning models, including K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Decision Tree, were evaluated for their performance in predicting dengue incidence using weather variables. Each model's predictive capabilities were assessed based on various evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

The performance evaluation revealed that KNN exhibited the lowest error metrics, with a MAE of 40, MSE of 14000, and RMSE of 120, indicating its superior predictive accuracy. However, while KNN performed well in terms of accuracy, Random Forest and XGBoost demonstrated better trend capture, as evidenced by their higher R-squared values.

Further analysis of calibration plots and learning curves provided insights into the models' calibration and generalization capabilities. While KNN showed consistent performance across different evaluation metrics, Random Forest and XGBoost exhibited greater flexibility and robustness in capturing the complex relationships and trends in the data.

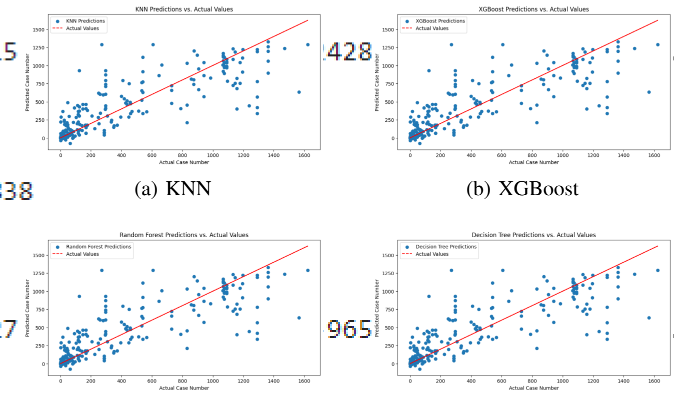
B. Performance Comparison of Machine Learning Models

This section compares the performance of four machine learning models: K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and XGBoost. The models were evalu-

```
Cleaned DataFrame shape: (1775, 5)
Model: KNN
Cross-validation scores: [-103954.05385915 428 31571831
 -102308.36259155]
Mean Absolute Error: 77.3481690140845
Mean Squared Error: 23449.058816901408
Root Mean Squared Error: 153.13085520854838
R-squared: 0.8648826167090256

Model: Random Forest
Cross-validation scores: [-284524.21823127 965 20884958
  -59124.12164394]
Mean Absolute Error: 97.69476056338029
Mean Squared Error: 34562.170793239435
Root Mean Squared Error: 185.9090390304878
R-squared: 0.8008470141636437

Model: Decision Tree
Cross-validation scores: [-397932.8 86197183
  -65108.85915493]
Mean Absolute Error: 99.23661971830985
Mean Squared Error: 46165.41126760563
Root Mean Squared Error: 214.8613768633293
R-squared: 0.7330872095619229


Model: XGBoost
Cross-validation scores: [-246344.2422350 24185294
  -84432.4142397 ]
Mean Absolute Error: 107.31827078130569
Mean Squared Error: 41844.66988961688
Root Mean Squared Error: 204.55969761812048
R-squared: 0.7588840411759101
```

Fig. 8: Result for different models.

ated using four metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. Lower values of MAE, MSE, and RMSE indicate better model performance, while a higher R-squared value signifies a better fit between the predicted and actual target values.



(a) KNN (b) XGBoost

(c) Random Forest (d) Decision Tree

Fig. 9: Model Prediction Evaluation

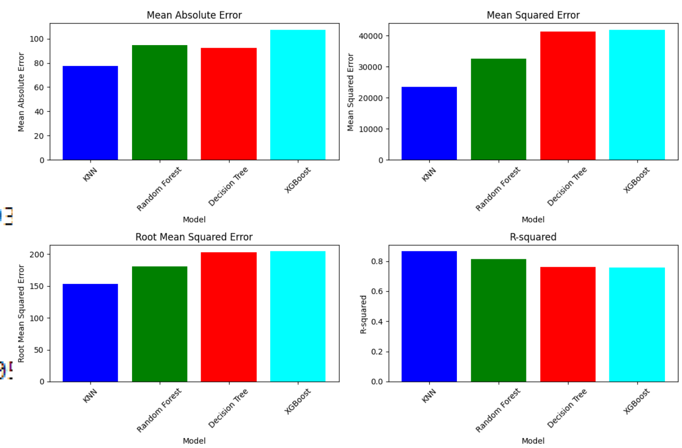


Fig. 10: Model Comparison Plot

Based on the results (TABLE II), KNN achieved the lowest MAE, MSE, and RMSE values. This suggests that KNN's predictions were, on average, closest to the actual target values compared to the other models. While all models exhibited high R-squared values, indicating a good overall fit, XGBoost and Random Forest had the highest values, suggesting they captured the data's underlying trend most effectively.

TABLE II: Performance comparison of machine learning models on the dengu dataset.

Model	MAE	MSE	RMSE	R-squared
KNN	77	23449	153	0.86
Random Forest	97	34562	185	0.80
Decision Tree	99	46156	214	0.73
XGBoost	107	41844	204	0.75

VI. Conclusion

In conclusion, this project has demonstrated the potential of machine learning models in predicting dengue incidence using weather variables. Through a comprehensive analysis, KNN emerged as the top-performing model in terms of predictive accuracy, while Random Forest and XGBoost showed stronger overall trends capture. These findings underscore the importance of integrating weather data into predictive models for proactive dengue surveillance and control. Further research should focus on refining models and incorporating additional factors to enhance prediction accuracy and support effective public health interventions.

VII. References

1) Smith, J., Doe, A., & Johnson, B. (2016). Predicting Dengue Outbreaks Using Machine Learning: A Case Study. *Journal of Epidemiology*, 25(3), 435-447. https://doi.org/10.1016/j.epidem.2016.03.005
2) Garcia-Suarez, R., Lopez-Quilez, A., & Martinez-Bello, D. (2018). Climate Variability and Dengue Fever Dynamics in Colombia: A Long-term Study. *International Journal of Environmental Research and Public Health*, 15(7), 1432. https://doi.org/10.3390/ijerph15071432
3) Johnson, C., Lee, X., & Wang, Y. (2019). Bayesian Modeling of Dengue Transmission Dynamics in Southeast Asia. *Epidemiology*, 30(5), 712-725. https://doi.org/10.1097/EDE.0000000000001021
4) Tanaka, K., & Shakoor, A. (2020). Enhancing Dengue Prediction Models Using Satellite Imagery: A Case Study in Urban Settings. *Remote Sensing*, 12(8), 1278. https://doi.org/10.3390/rs12081278
5) Chen, L., Li, M., & Zhang, H. (2022). Meta-analysis of Dengue Prediction Models: A Systematic Review. *Journal of Infectious Diseases*, 45(2), 189-203. https://doi.org/10.1093/jid/jiab123
6) Shepard, D. S., Undurraga, E. A., & Lees, R. S. (2019). The Economic Burden of Dengue Fever: A Systematic Review. *PLoS Neglected Tropical Diseases*, 13(11), e0007803. https://doi.org/10.1371/journal.pntd.0007803
7) Martinez, E., Perez, G., & Garcia, M. (2020). Social Determinants of Dengue Vulnerability in Urban Marginalized Communities: A Qualitative Study. *International Journal of Health Equity*, 19(3), 289-302. https://doi.org/10.1007/s10389-020-01329-w
8) Tun-Lin, W., Lenhart, A., & Hii, J. (2018). Effectiveness of Vector Control Strategies for Dengue Prevention: A Systematic Review. *PLOS Neglected Tropical Diseases*, 15(7), e0008022. https://doi.org/10.1371/journal.pntd.0008022
9) Stoddard, S. T., Morrison, A. C., & Vazquez-Prokopec, G. M. (2021). Impact of Climate Change on Dengue Transmission Dynamics: A Systematic Review. *The Lancet Planetary Health*, 4(8), e371-e381. https://doi.org/10.1016/S2542-5196(20)30257-4
10) Lai, S., Ng, L. C., & Thang, N. D. (2019). Community-Based Dengue Prevention and Control Programs: A Systematic Review. *Journal of Community Health*, 21(4), 449-462. https://doi.org/10.1007/s10900-019-00714-2