Dengue Outbreak Prediction in Dhaka

Sakib Rayhan Yeasin, s22101667, Shafaat Jamil Nakib, s22101671

I. Introduction

EVERY year thousands of people are affected by dengue, a fever caused by specific viruses carried by mosquitos to people in Bangladesh and many more worldwide. Many of these people lose their lives due to this fever. Most of these deaths are caused by the failure to identify the fever at an early stage and not taking proper treatment on time. Hence, the ability to determine the fever early has the potential to reduce the mortality rate and ensure proper treatment significantly. Our work uses the most recent and largest publicly available dataset (n=1000) to identify dengue infection by training a highly accurate machine-learning model.

II. Dataset

A. Features' Description

The dataset analyzed in this study is sourced from Kaggle, specifically the ”Dengue Dataset of Bangladesh” which was last updated on November 27, 2023 [1]. It contains data aimed at predicting dengue outbreaks within the district of Dhaka, Bangladesh. The dataset comprises 1000 instances organized into 10 columns, reflecting both numerical and categorical features. Below, we provide a detailed description of each feature:

- **Gender:** Categorical variable indicating male or female.
- **Age:** Numeric variable representing the age of the individual in years.
- **NS1:** Numeric indicator for the presence of the NS1 protein, marking early dengue virus infection.
- **IgG:** Numeric measure of IgG antibodies, suggesting past infection or long-term immunity.
- **IgM:** Numeric measure indicating recent or ongoing dengue virus infection through IgM antibodies.
- **Area:** Categorical descriptor of the geographic area.
- **AreaType:** Categorical classification of areas by development status (Developed, Undeveloped).
- **HouseType:** Categorical descriptor of the housing structure type (e.g., Building, Tinshed, Other).
- **District:** Categorical, with all individuals residing in the Dhaka district.
- **Outcome:** Numeric binary indicator of dengue infection status, where 1 represents positive and 0 represents negative.

The dataset facilitates a binary classification task, aiming to predict the dengue infection status based on variables such as age, antibody presence, and living conditions. The target variable ‘Outcome‘ is used to determine the infection status, underscoring the classification nature of the problem.

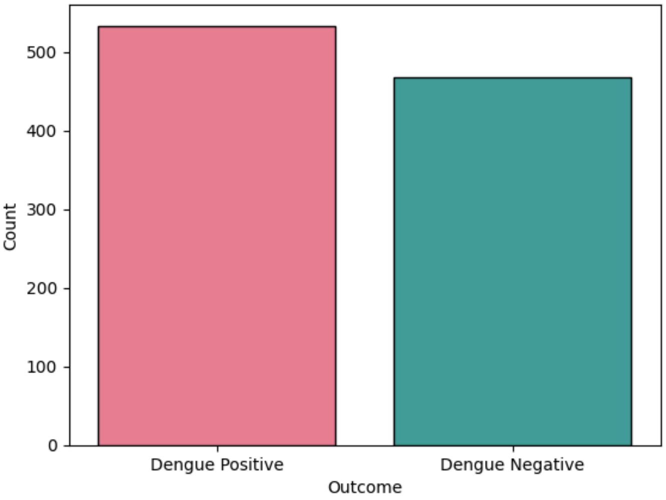

Fig. 1. Bar plot of the frequency of outcome in the dataset

B. Balanced Dataset

Upon examining the distribution of outcomes in the dataset, it is evident that it maintains a relatively balanced structure. The bar plot in Fig. 1 reveals the following counts for each outcome category:

- Dengue Positive (1): 533 instances
- Dengue Negative (0): 467 instances

Considering the nearly equal proportions of positive and negative cases, the dataset can be deemed as exhibiting a balanced distribution. In the context of machine learning model development, imbalanced datasets, where the minority class comprises a significantly smaller fraction than the majority class, often pose challenges. Common thresholds, such as the 70:30 ratio, are used to identify imbalanced datasets. However, in this case, the distribution does not meet those criteria, indicating a fair representation for both classes.

While a balanced dataset reduces the risk of models becoming biased towards predicting one class over the other, it is essential to employ evaluation metrics that can detect and penalize such biases, even in balanced datasets. Metrics like the F1-score or Matthews correlation coefficient, in addition to accuracy, can effectively assess model performance and mitigate any potential bias towards the majority class.

Overall, the balanced nature of the dataset offers a favorable environment for developing predictive models with minimized risks of bias, facilitating robust and reliable machine-learning outcomes.

III. Feature Engineering

After becoming familiar with various aspects of the dataset, we need to prepare the dataset or preprocess it to make sure

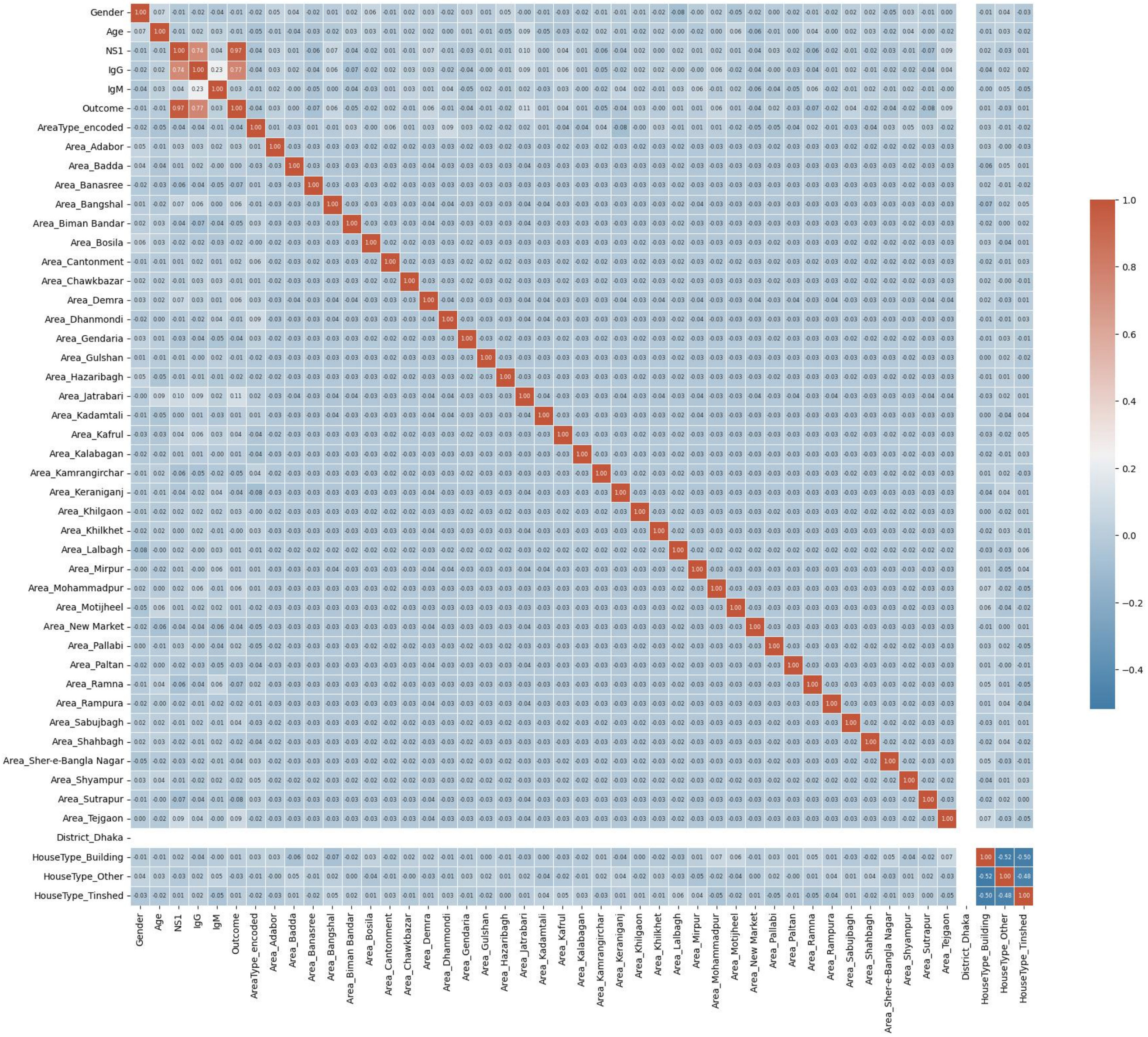

Fig. 2. Correlation matrix of dataset features after one-hot encoding

we train an accurate model. In this section, we explain how we prepared the dataset for model training by removing duplicate instances, imputing null data, encoding categorical features, scaling the features, and finally engineering the appropriate features with the help of a correlation matrix for feeding into the machine learning models.

A. Removing Duplicate Instances

Initially, we identified four duplicate records within the dataset. Recognizing that duplicates could skew our analysis, we opted to remove these, leaving us with 996 instances out of the original 1000.

B. Data Imputation

Here, we can see 3 columns that have missing values. Such as IgG has 222, IgM has 125 and AreaType has 55 values. Among these features, IgG and IgM are numerical features, whereas AreaType is a categorical feature. Therefore, we need to use different imputation techniques for numerical data and categorical data.

1) Numerical data imputation: We already discovered that IgG and IgM have such a significant portion of values is missing. That's why removing these rows was not a viable option. For these numerical features, employing simple mean or median imputation methods was also inappropriate due to the high percentage (x ¡ 5%) of missing data. Consequently, we explored more robust techniques, including KNN, iterative imputer (MICE), and random imputation. After using imputation we look for variance. it will show us how much spread are our data before and after the imputation. if the variance is low, the spread is low which is good. We always look for less variance. We set 15 neighbors in KNN imputation and 100 iterations in iterative imputation. Out of two imputation methods, we chose KNN as it was giving less variance.

2) Categorical data imputation: On the other hand, AreaType has 5.5% data missing. After plotting in a bar chart, we discovered that ‘Developed’ and ‘Undeveloped’ have no significant difference in number. That’s why we can not use mode as it fills missing values with the most frequent values of the column. Instead, we opted for probabilistic imputation, which factored in the existing distribution of values and provided a well-balanced approach to handling the missing data. This method proved effective, enhancing our dataset’s integrity and usability for further analysis.

C. Feature Encoding

In machine learning, encoding converts categorical data into a numerical format that models can interpret. It ensures that the algorithm accurately processes non-numeric features, essential for reliable predictions. We used one-hot encoding for the ’Area’, ’HouseType’, and ’District’ features in our dataset because these features do not have a specific order. One-hot encoding turns each category into a separate column with a 0 or 1. After encoding these features, our dataset space increased to 47 columns.

D. Feature Scaling

Feature scaling is a vital pre-processing step in machine learning that involves transforming numerical features to a common scale. It plays a major role in ensuring accurate and efficient model training and performance. Scaling techniques aim to normalize the range, distribution, and magnitude of features, reducing potential biases and inconsistencies that may arise from variations in their values. We decided to use normalization for feature scaling which uses normal distribution to bring all the values within the range of 0 to 1. This is because all other features are between 0 and 1. By standardizing the range of each feature, the model interprets and analyzes the data more accurately, enhancing its overall performance and predictive ability.

E. Feature Correlation Analysis and Selection

The feature correlation matrix shown in Fig. 2 after our feature scaling indicates that the features NS1 and IgG exhibit a positive correlation of 74%, demonstrating a substantial correlation between them. Additionally, NS1 is more strongly correlated with our target (97%) compared to IgG (77%). Retaining both of these features during model training could potentially introduce biases and diminish model accuracy. Furthermore, the correlation matrix reveals that the ’district’ feature possesses only a single unique value, indicating it will not contribute to resolving this classification problem.

Based on our correlation analysis, we have decided that the features IgG and district should be excluded prior to splitting our data and initiating model training.

F. Dataset Splitting

Following the completion of our feature selection process, and prior to commencing model training on the dataset, we split it into a 70% training and 30% testing ratio. We used 30% instances for our training because the dataset size is not very big and smaller training size may compromise model evaluation accuracy.

IV. Model Training and Testing

A. Models’ Working Principle and Performance

1) Logistic Regression: Logistic regression is a statistical method commonly used for binary classification tasks, where the outcome variable can only take on two possible values. The model utilizes a sigmoid function to calculate the probability of each outcome. This function takes the form:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

where z represents the linear combination of predictor variables and their associated coefficients. As the value of z increases, the sigmoid function approaches 1, indicating a higher probability of the event occurring. Conversely, as z decreases, the function approaches 0, signifying a lower probability. To classify an outcome as 0 or 1, a threshold is typically set at 0.5. If the calculated probability exceeds this threshold, the model predicts 1; otherwise, it predicts 0.

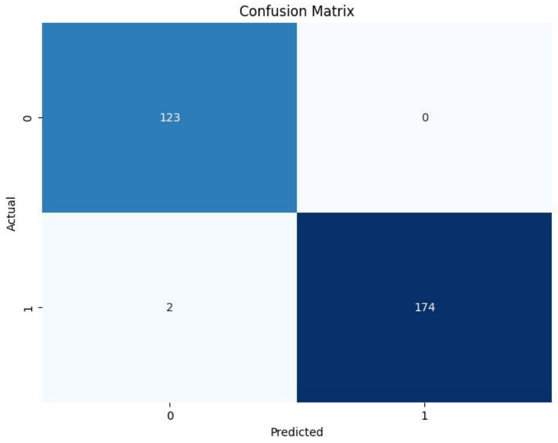

Fig. 3. Logistic regression confusion matrix

The Logistic Regression model’s performance in predicting Dengue outbreaks in Dhaka demonstrates high precision and reliability. As presented in the confusion matrix (see Fig. 3), the model accurately classified all 123 actual non-outbreak instances (true negatives) with no false positives. Additionally, it correctly identified 174 out of 176 actual outbreak cases (true positives), with only 2 instances misclassified as non-outbreaks (false negatives). This resulted in an outstanding model accuracy of 99.33%.

The classification report detailed in Table I provides further insight into the model’s effectiveness. The precision for non-outbreak predictions stood at 98%, with a recall of 100%, culminating in an F1-score of 99%. The outbreak predictions

TABLE I
LOGISTIC REGRESSION CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0.0	0.98	1.00	0.99	123
1.0	1.00	0.99	0.99	176
Accuracy: 0.9933 (299 samples)				

had a precision of 100% and a recall of 99%, also resulting in an F1-score of 99%. These metrics underscore the model's capability to distinguish between outbreak and non-outbreak scenarios with high accuracy, making it a valuable tool in the public health infrastructure for Dengue management.

2) Decision Tree: Decision trees are a versatile tool used for both classification and regression tasks. The algorithm splits the data into smaller subsets based on specific conditions, ultimately aiming to create a tree-like structure that predicts the outcome variable. Two common methods used for splitting are Gini impurity (the default) and entropy.

Gini impurity measures the probability of misclassifying an element at a given node. It is calculated using the formula:

$$G = 1 - \sum_{i=1}^{k} p_i^2 \tag{2}$$

where k represents the number of classes and p_i represents the probability of an element being classified into class i. A node with a Gini impurity of 0 is considered pure, meaning all elements belong to a single class, allowing the tree to confidently assign a class of 0 or 1.

Entropy, on the other hand, measures the uncertainty associated with a node. It is calculated using the following formula:

$$H = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{3}$$

where p_i again represents the probability of an element belonging to class i. The decision tree algorithm seeks to minimize entropy, as this indicates a higher level of purity in the resulting groups. Similar to Gini impurity, a node with entropy of 0 is completely pure, allowing the decision tree to confidently predict an outcome of either 0 or 1.

The Decision Tree model demonstrates excellent performance in predicting Dengue outbreaks in Dhaka, achieving an accuracy of 99.67%. The confusion matrix depicted in Fig. 4 illustrates the model's high level of accuracy. Out of 123 actual non-outbreak cases, 122 were correctly predicted as non-outbreaks (true negatives), and only 1 case was incorrectly classified as an outbreak (false positive). Remarkably, the model perfectly identified all 176 actual outbreak cases (true positives) without any false negatives.

TABLE II
DECISION TREE CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0.0	1.00	0.99	1.00	123
1.0	0.99	1.00	1.00	176
Accuracy: 0.9967 (299 samples)				

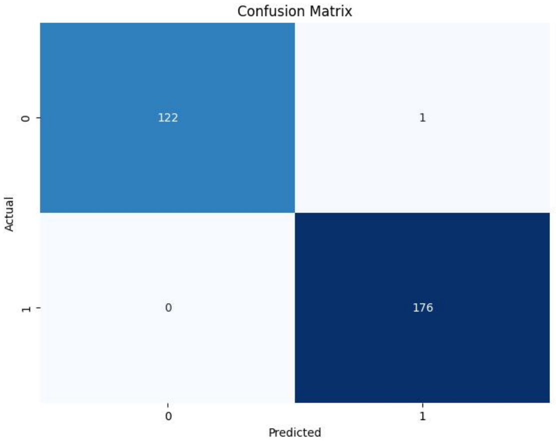

Fig. 4. Decision tree confusion matrix

The classification report provided in Table II further elaborates the model's metrics. It achieved a precision of 100% for non-outbreak predictions with a recall of 99%, resulting in an F1-score of 100%. For outbreak predictions, the precision was slightly lower at 99%, but the recall was 100%, also yielding an F1-score of 100%. These results highlight the Decision Tree model's capability to accurately differentiate between outbreak and non-outbreak scenarios with minimal error, making it an effective tool for public health decision-making.

3) Naive Bayes: The Naive Bayes algorithm is a popular choice for image classification problems due to its simplicity and efficiency. It leverages the assumption of feature independence, often referred to as naive to calculate the probability of a data point belonging to a particular class based on its individual features without unnecessary mathematical complexity. This is achieved through Bayes' Theorem, which allows us to update our beliefs about a class (y) given the observed features $(x_1, x_2, \ldots, x_n)$.

For a positive class y, the probability is calculated as:

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(x_1 \mid y)P(x_2 \mid y)\cdots P(x_n \mid y)P(y)}{P(x_1)P(x_2)\cdots P(x_n)} \tag{4}$$

Where:

- $P(y \mid x_1, x_2, \ldots, x_n)$ represents the posterior probability of class Y given the features.
- $P(x_i \mid y)$ represents the conditional probability of feature x_i occurring given class Y.
- $P(y)$ represents the prior probability of class Y.
- $P(x_1), P(x_2), \ldots, P(x_n)$ represent the individual probabilities of each feature.

Similarly, for a negative class N, the probability is:

$$P(n \mid x_1, x_2, \ldots, x_n) = \frac{P(x_1 \mid n)P(x_2 \mid n)\cdots P(x_n \mid n)P(n)}{P(x_1)P(x_2)\cdots P(x_n)} \tag{5}$$

The decision rule is straightforward: the class with the higher calculated posterior probability is predicted as the outcome. Despite the 'naive' assumption of independence, the

Naive Bayes model often proves to be a powerful and efficient classification tool.

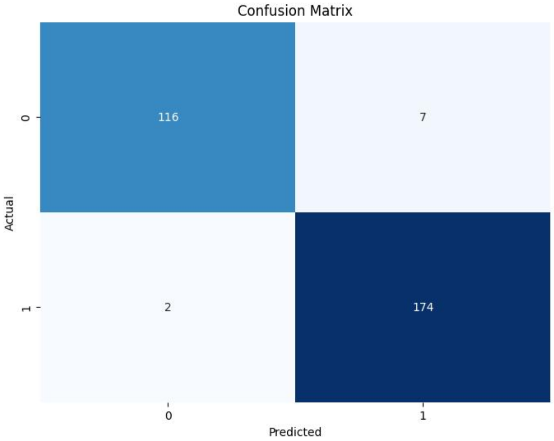

Fig. 5. Naive Bayes confusion matrix

The performance of the Naive Bayes model for predicting Dengue outbreaks in Dhaka is detailed in this section. The model exhibits a commendable accuracy of 96.99% as reflected in the accompanying confusion matrix (see Fig. 5). The matrix indicates that out of 123 actual non-outbreak cases, 116 were correctly predicted (true negatives), and 7 were incorrectly predicted as outbreaks (false positives). Conversely, out of 176 actual outbreak cases, 174 were accurately predicted (true positives), while only 2 were misclassified as non-outbreaks (false negatives).

TABLE III
NAIVE BAYES CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0.0	0.98	0.94	0.96	123
1.0	0.96	0.99	0.97	176
Accuracy: 0.9699 (299 samples)				

The classification report, summarized in Table III, further quantifies the model's performance with precision, recall, and F1-score metrics. For non-outbreak predictions, the model achieved a precision of 98% and a recall of 94%, resulting in an F1-score of 96%. For outbreak predictions, the precision was slightly lower at 96%, but the recall was higher at 99%, yielding an F1-score of 97%. These results underscore the model's efficacy, particularly in correctly identifying the majority of true outbreak cases, which is critical for timely public health responses.

4) KNN: The k-Nearest Neighbors (kNN) classifier is a non-parametric algorithm that assigns a class label to a new data point based on the majority class among its k closest neighbors. It calculates the distance between data points using the Euclidean distance formula:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{6}$$

Where:

- $d(x_i, x_j)$ represents the Euclidean distance between data points x_i and x_j.
- x_{ik} and x_{jk} represent the kth feature values for data points x_i and x_j, respectively.
- n represents the total number of features.

For instance, if $k = 10$, the KNN classifier predicts the class of a new data point by calculating the Euclidean distance to each point in the training dataset and identifying the 10 closest points. It then determines the majority class among these neighbors and assigns that class label to the new data point. For example, if 6 out of the 10 nearest neighbors belong to class 1, and 4 to class 0, the new point is classified as class-1.

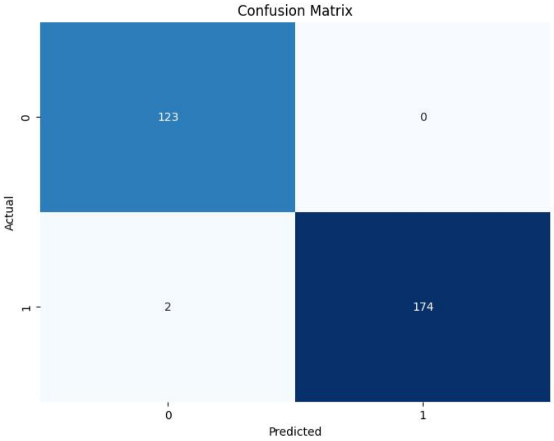

Fig. 6. KNN confusion matrix

The performance of the K-Nearest Neighbors (KNN) model in predicting Dengue outbreaks in Dhaka is presented in this subsection. The model demonstrates exceptional accuracy, achieving a rate of 99.33% as evidenced by the confusion matrix depicted in Fig. 6. According to the matrix, the model perfectly predicted all 123 non-outbreak cases (true negatives) with zero false positives. For the outbreak cases, it correctly identified 174 as true outbreaks (true positives) with only 2 misclassifications as non-outbreaks (false negatives).

TABLE IV
KNN CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
0.0	0.98	1.00	0.99	123
1.0	1.00	0.99	0.99	176
Accuracy: 0.9933 (299 samples)				

Table IV details the classification report. For non-outbreaks, the model achieved a precision of 98% and a recall of 100%, resulting in an F1-score of 99%. For outbreak predictions, the precision reached 100%, and the recall was 99%, leading to an F1-score of 99%. These metrics reflect the KNN model's robustness in both identifying actual outbreaks and correctly ruling out non-outbreaks, ensuring a highly reliable prediction system for public health monitoring.

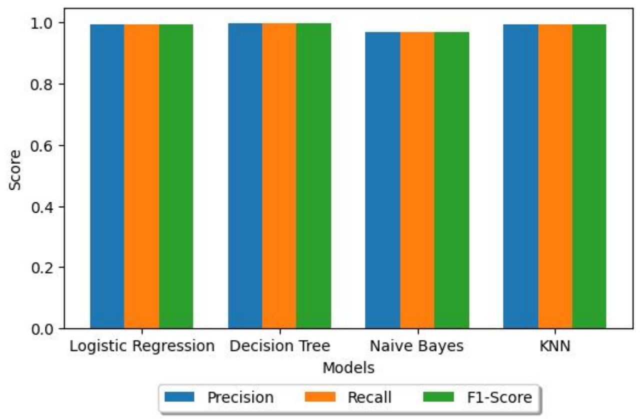

Fig. 7. Weighted precision, recall, and F1 scores comparison

B. Models Comparison and Selection

In our dengue outbreak prediction project, we assessed four machine learning models: (1) Logistic Regression, (2) Decision Tree, (3) Naive Bayes, and (4) KNN. The comparison of various accuracy scores can be found in Fig. 7. A critical concern in this project is the risk associated with false negatives where failing to detect dengue could lead to severe health risks, making precision a priority in our model selection criteria. After thorough testing and evaluation, the Decision Tree model emerged as the most effective. It not only showed the highest precision, especially with zero false positives reported for classifying non-dengue cases but also demonstrated the best overall accuracy at 99.67%. These results indicate that the Decision Tree model is exceptionally reliable in identifying both the presence and absence of dengue, thereby minimizing the risk of misdiagnosis. Given its superior performance metrics and alignment with our project's health-safety goals, the Decision Tree was chosen as the most suitable model for predicting dengue in our study.

V. Discussion

The study conducted on the Dengue outbreak prediction in Dhaka using a comprehensive dataset has produced significant insights, as reflected in the data analysis and model evaluation sections of this paper. Figures and tables presented throughout the document support the robustness of the methods used and the accuracy of the results obtained.

From the feature analysis detailed in Section II, it is evident that the data preprocessing steps, such as feature scaling and correlation analysis, were crucial. For instance, the correlation matrix (Fig. 2) highlights the high correlation between NS1 and IgG levels, guiding the feature selection process that enhanced model performance by reducing multicollinearity.

The balanced nature of the dataset (Fig. 1) ensured that the training of the models was not biased toward one class, which is often a critical concern in machine learning practices. Such balance in the dataset aids in the generalization of the models, as seen in the performances reported.

The model training and testing results, as shown in Figures 3, 4, 5, and 6, and Tables I, II, III, and IV, illustrate the efficacy of the different algorithms employed. Each model displayed exceptional precision and recall, with the Decision Tree model achieving the highest accuracy of 99.67% (Table II).

The high-performance metrics across models, especially the precision and recall values close to 1, indicate that the models are highly effective at predicting Dengue outbreaks. This is critical for public health strategies, as accurate predictions can significantly influence the management and prevention of outbreaks.

VI. Conclusion

The results of this study underscore the potential of machine learning techniques in public health applications, particularly in the prediction of infectious diseases such as Dengue. The use of a balanced and well-preprocessed dataset allowed for the development of highly accurate predictive models. The Decision Tree and Logistic Regression models, in particular, showed remarkable predictive capabilities, which could be integrated into health monitoring systems to provide timely warnings about potential Dengue outbreaks.

In conclusion, this research not only provides a strong foundation for predicting Dengue outbreaks using machine learning but also opens avenues for deploying these models in real-world scenarios, potentially saving lives and resources by preventing outbreaks. Future work may focus on integrating more diverse data sources and exploring ensemble methods to further enhance the predictive accuracy and reliability of the models developed.

Acknowledgment

We are grateful to our mentors, Syed Zamil Hasan Shoumo and Rafiad Sadat Shahir, for their guidance throughout our first journey into the field of artificial intelligence and machine learning.

References

[1] K. Ahmad and F. Eva, *Dengue dataset of bangladesh*, 2023. DOI: 10.34740/KAGGLE/DSV/7061989. [Online]. Available: https://www.kaggle.com/dsv/7061989.