CSE422 : Artificial Intelligence
Project Report

Project Title : Dengue Prediction

Group: 06
Section: 03
Spring 2024

Submitted to:

Mr. Swattic Ghose
Ms. Jawaril Munshad Abedin

Submitted by:

ID	Name
21301062	Jannatul Ferdaus
21301105	Tangena Islam

Table of Contents

Introduction:

Dengue is a huge concern in Bangladesh, especially during monsoon seasons. This project aims to develop a classification system using machine learning models to predict whether a person is affected with dengue or not. The motivation behind this project is to do better dengue detection so that proper care can be taken as soon as possible before the patient's condition gets worse. This would help the healthcare system to be more effective in peak dengue seasons.

Dataset Description:

Source:

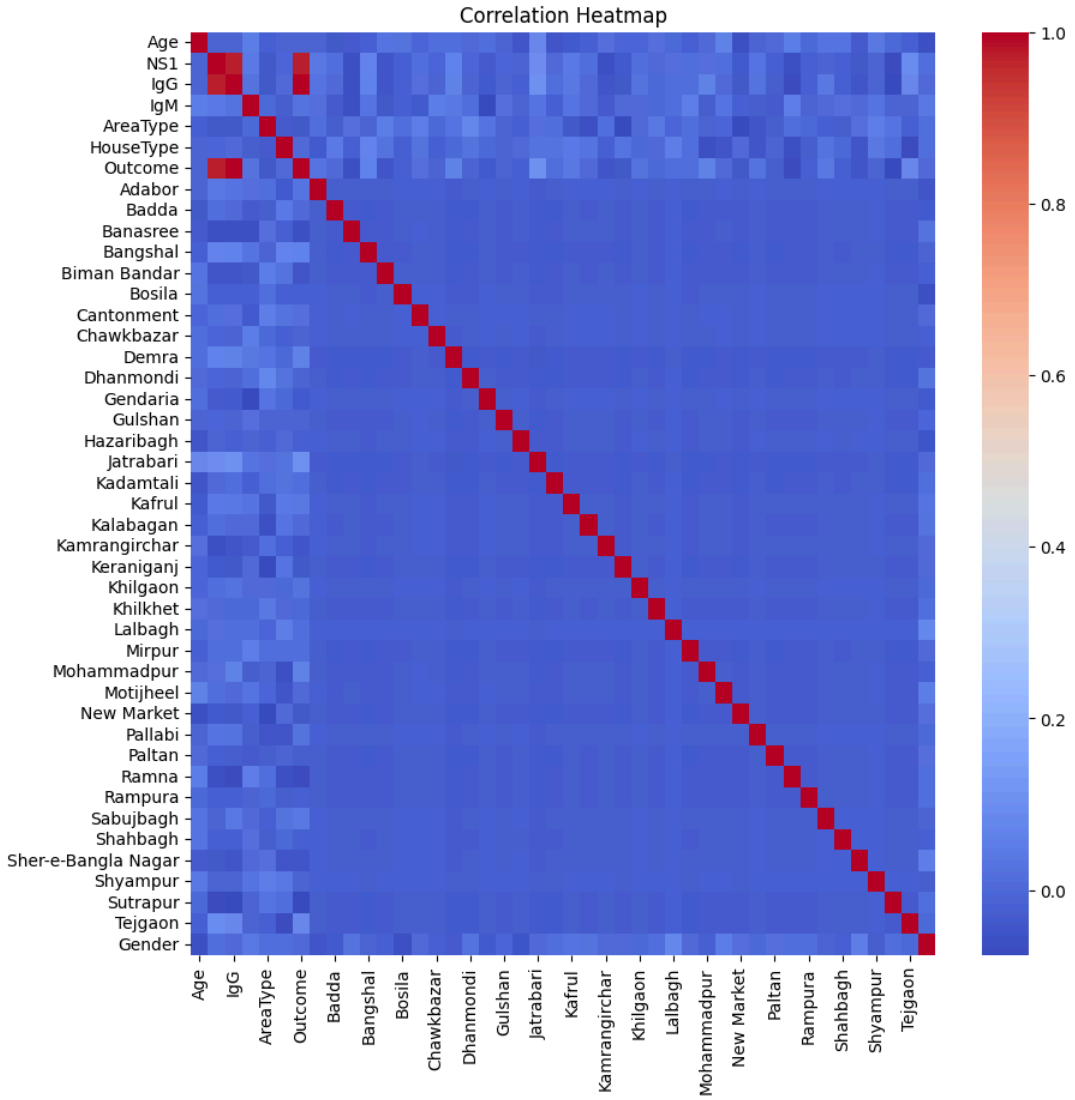
The dataset is taken from Kaggle. The link is given below:

https://www.kaggle.com/datasets/kawsarahmad/dengue-dataset-bangladesh?fbclid=IwAR2SSw2AbdiN2ZisqIuIzJEhZ5o-_u_hleOtYQpNZ-Tq5qAy4jb1e4sX-1s_aem_Aa0TINZfqySTABcTdnQTf4wlsOnX9lcVYh6kzNtFG0FK34TEKslVhPNsJ_QBudWgY5dg0y8Sj04OSYI3bygVUrRy

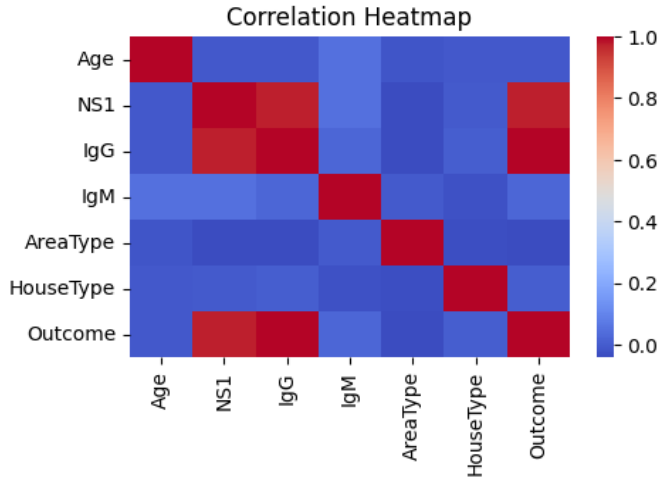
Description:

The dataset initially has 10 features.
This project was a classification problem since the aim of this project was to determine if a certain person has dengue or not. The result is seen by a yes/ no answer from the dataset.
There were 1000 data points in the used dataset.
The features used from the dataset were both quantitative and categorical. Gender, Area, AreaType, HouseType and District were categorical. Age, NS1, IgG, IgM and the outcome were qualitative. Some of the categorical features were encoded as quantitative for better prediction and understanding of the models.

Correlation with all features:

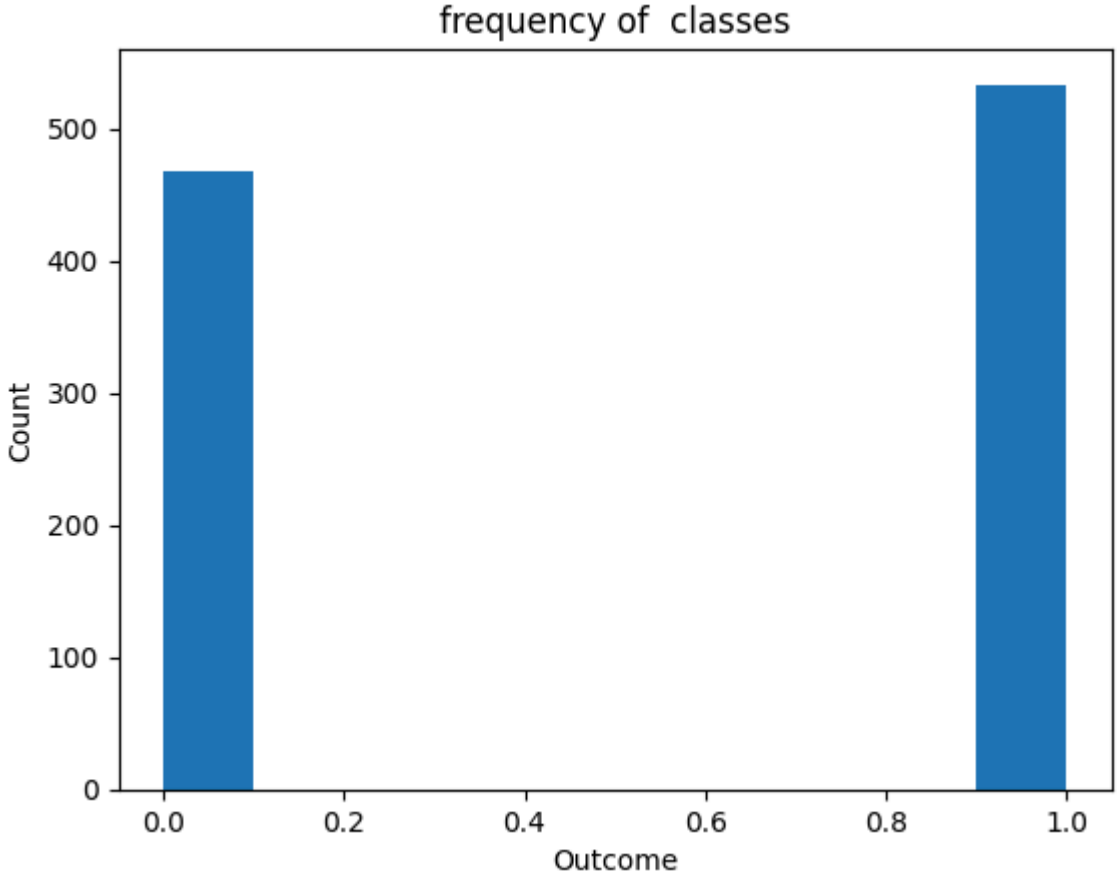


Correlation after pre processing:



Imbalanced Dataset:

In the used dataset, all unique classes had a similar number of instances. This was shown using a bar chart for the results having dengue [1] and not having dengue [0].



Dataset pre-processing:

The dataset did not have any null values in any features. Hence, there was no need to delete any row or impute any values.

```
1 dengue_dataset.isnull().sum()
```

```
Gender       0
Age          0
NS1          0
IgG          0
IgM          0
Area         0
AreaType     0
HouseType    0
District     0
Outcome      0
dtype: int64
```

Initially, we deleted the Gender feature because it does not necessarily mean that a specific gender is more susceptible to dengue.

Other than gender, there were four categorical features that needed to be encoded.
Those were: Area, AreaType, HouseType and District.

```
1 dengue_dataset_new.head()
```

	Age	NS1	IgG	IgM	AreaType	HouseType	Outcome
0	45	0	0	0	0	0	0
1	17	0	0	1	1	0	0
2	29	0	0	0	0	1	0
3	63	1	1	0	1	1	1
4	22	0	0	0	0	0	0

This is how the dataset looked after encoding.
We decided to remove the area column as it acted as noise for the models while training.
The feature District was also removed as it contained only one string that is “Dhaka”.

Feature scaling:

Feature scaling mainly involves transforming numerical features to a similar scale, ensuring they contribute equally to the model training process. But for this dataset, we did not need to use any feature scaling as there was no feature that did not dominate any other feature in the dataset.

Dataset splitting:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.3, random_state=20)


# scaler = StandardScaler()

# X_train = scaler.fit_transform(X_train)
# X_test = scaler.transform(X_test)
```

```
print("X_train shape:", X_train.shape)
print("y_train shape:", y_train.shape)
print("X_test shape:", X_test.shape)
print("y_test shape:", y_test.shape)
```

```
X_train shape: (700, 6)
y_train shape: (700,)
X_test shape: (300, 6)
y_test shape: (300,)
```

Here, the dataset of 1000 data points is being split for test set and train set. We are selecting at random since the outcomes of the dataset were almost evenly distributed (stratification is not required). The training set was taken as 70% of the 1000 data points from the dataset. The features are taken by excluding some unnecessary features from before like Gender, District and Area since it was a redundant value for the model. Moreover, the outcome feature from the dataset was dropped for testing the 30% of the test set to be tested separately.

Model training and testing:

We used three models for training and testing our dataset: K-nearest neighbor (KNN) algorithm, Logistic regression and random forest.

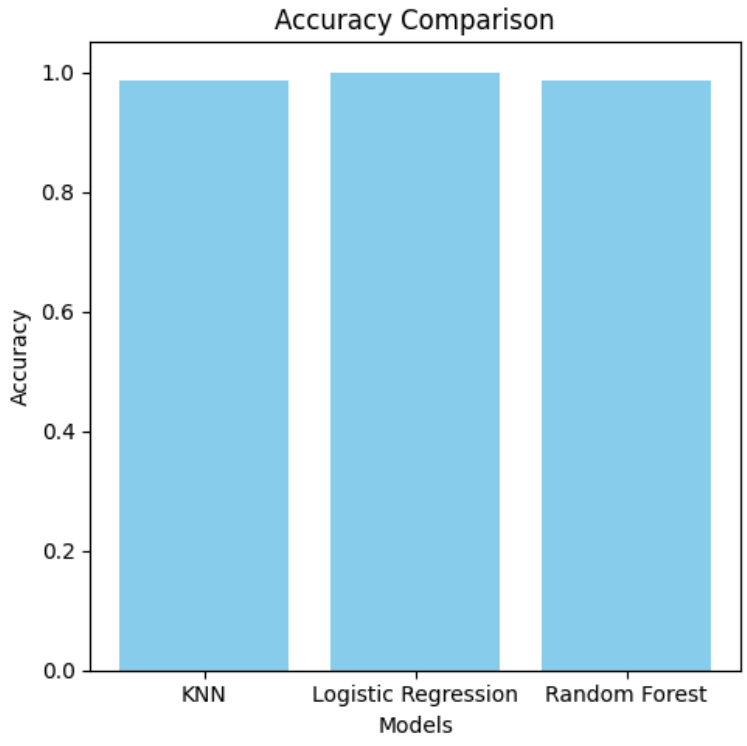
For each model, we followed three steps: Training, Testing and Evaluation.

For Training, each model was trained on the dataset of X_train and y_train.
For testing, the models tested the dataset X_test and y_test.
Then we evaluated the performances of each model using metrics like accuracy, precision, recall and F1-score.

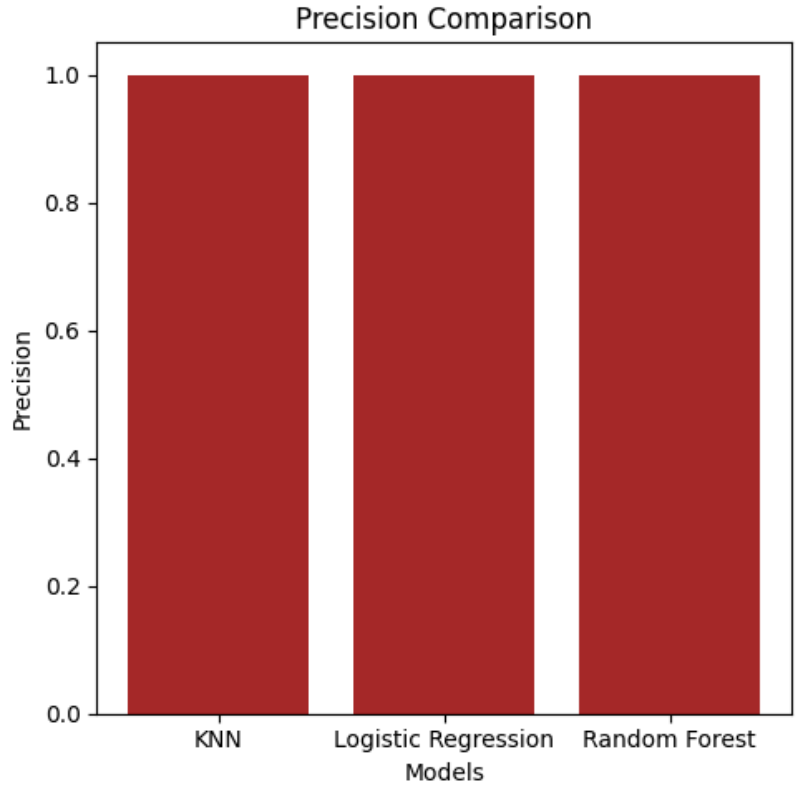
Model selection/Comparison analysis:

	KNN	Logistic Regression	Random Forest
Accuracy	98.6667%	100%	98.6667%
Precision	100%	100%	100%
Recall	97.5758%	100%	97.5758%
F1 score	98.7730%	100%	98.7730%

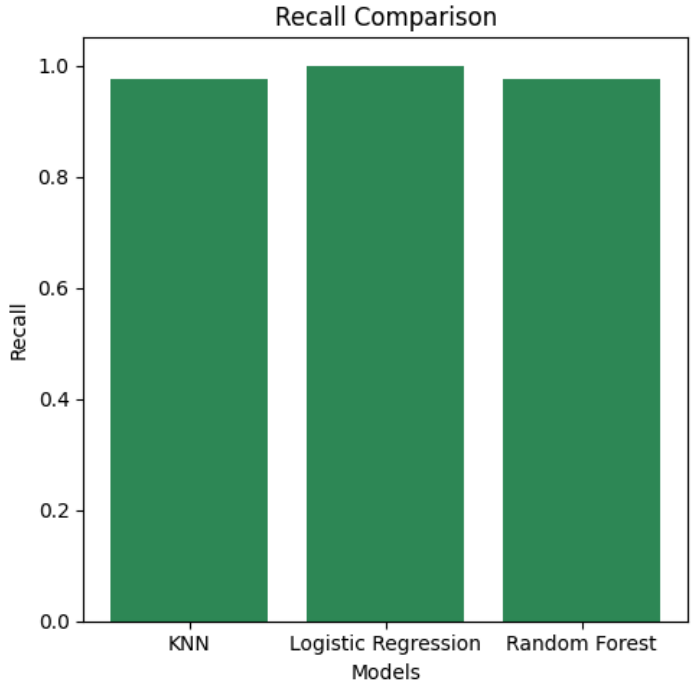
Bar chart for accuracy comparison:



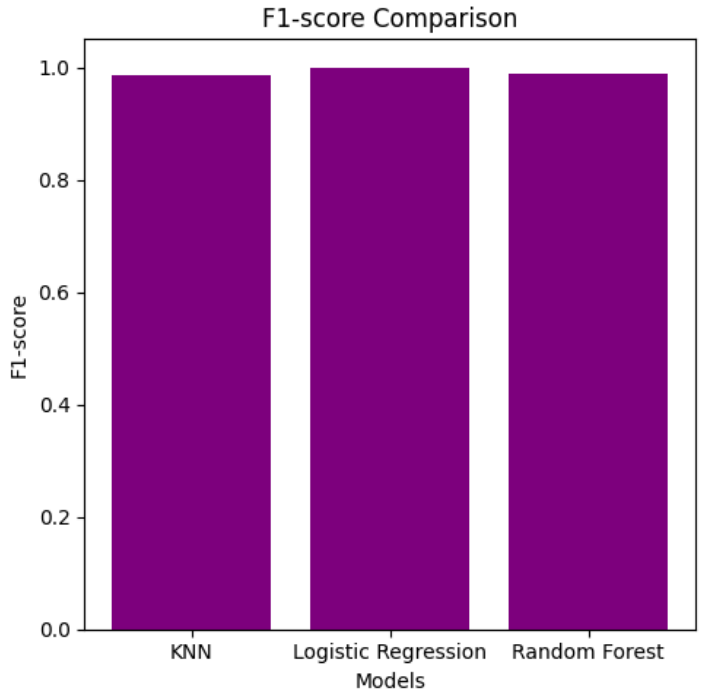
Bar chart for precision comparison:



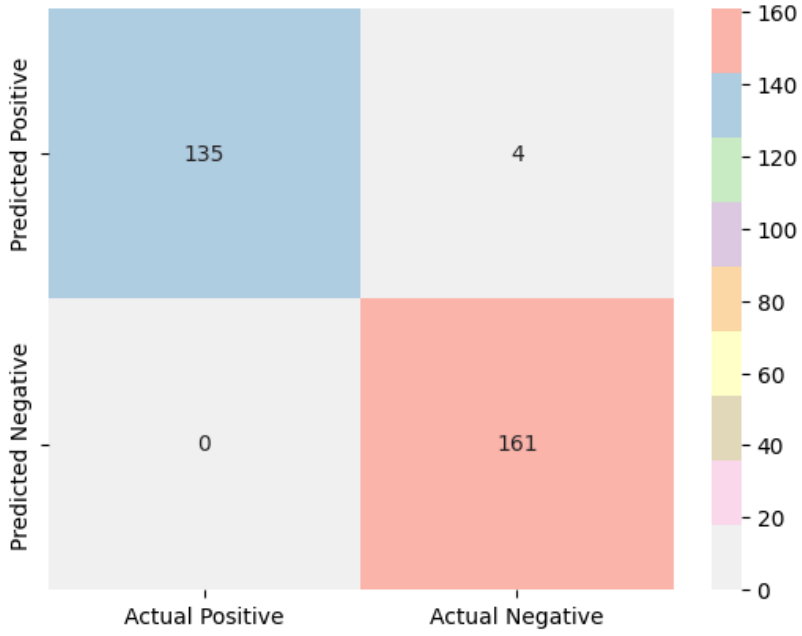
Bar chart for recall comparison:



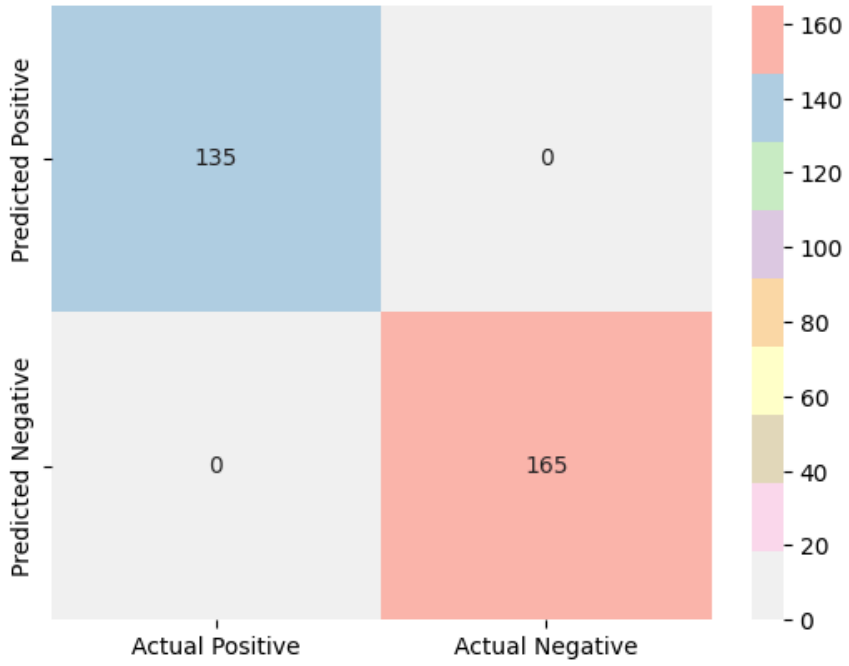
Bar chart for F1-score comparison:



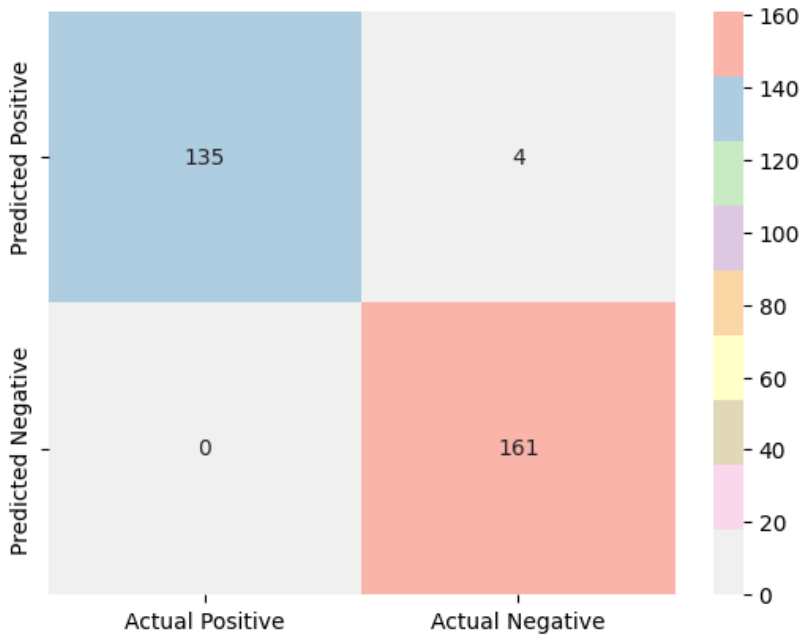
Confusion matrix for KNN:



Confusion matrix for Logistic regression:



Confusion matrix for random forest:



Conclusion:

To conclude, the three models used are predicting almost very high accuracy (above 97%). The model whose prediction was most accurate was the Logistic Regression model. The overall prediction also depends on the size of the dataset. Since we had to work with a comparatively smaller dataset, proper training was performed thus an accuracy of prediction was obtained. If this project was done on a big scale, the problem of dengue could have been solved with the input of a few features at a remarkable speed. Thus, the patients would be able to be treated more efficiently and the threat of dengue would seem simpler.