

# Vector Borne Disease Prediction

---

By - Aditya Yadav, Megha,  
Priyanshu Sehrawat, and Vivaswan Nawani

Date: 23rd Novembler,  
2023



INDRAPRASTHA INSTITUTE *of*  
INFORMATION TECHNOLOGY  
DELHI



## Why have we chosen this problem?

- According to WHO, vector borne diseases account for 17% of all infectious diseases and cause the death of 700,000 people annually.
- Now, most cases of vector borne diseases are treatable with modern medicine, however the main problem lies with the lack of early diagnosis.
- This problem becomes even more acute in underdeveloped regions of the world which lack severely in healthcare infrastructure.
- We believe that this problem could be solved to a great extent with Machine Learning and this is what made us choose this topic.

# Literature Review



designed by  freepik

# Study 1 : Machine Learning in Disease Diagnosis



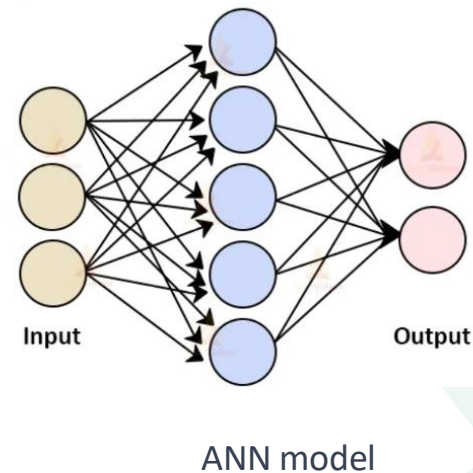
## Introduction and Background :

As dengue is a growing public health concern with approx. 50 million infections annually, the early diagnosis of it is very crucial. Machine learning can aid in early detection with improved accuracy, which will reduce diagnostic time and costs.

The study compared two machine learning algorithms, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) for predicting dengue.

## Research Method and Key Findings :

A dataset of 4,332 dengue cases in Paraguay (2012-2016) with patients information.



1. ANN : Multilayer Perceptron (MLP) and Radial Basis Function (RBF) networks.
2. SVM : Linear ,Gaussian, and Polynomial kernels.

## EVALUATION :

Accuracy, sensitivity and specificity were evaluated using confusion matrices on test datasets generated from randomized partitions of the dataset.

## Results and Conclusion :

ANN-MLP:96% accuracy, 96% sensitivity, and 97% specificity on average with low variation.

SVM Polynomial: Above 90% for all three indicators with acceptable variations.

This concludes that both have proven effective in diagnosing dengue with minimal data preprocessing.

# Study 2: Climate Change and Disease Prediction

---



## Introduction and Background :

Study focuses on seasonal prediction and diagnosis related to climate change using big data. The extreme weather condition creates various vector borne diseases among humans.

## Research Method :

Gaussian Process Regression Conceptual Model used. They found a significant association between climate change and bacterial contamination.

## Classification Models

They used CART, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Naive Bayes (NB).



---

They incorporated climate-related variables such as temperature, wind velocity, moisture, and wetness into their models.

## Results

The Support Vector Machine (linear kernel) achieved the highest forecasting accuracy, with a 70percent accuracy rate, 14percent sensitivity, 95percent specificity, and 56percent precision.



# Dataset Description

---



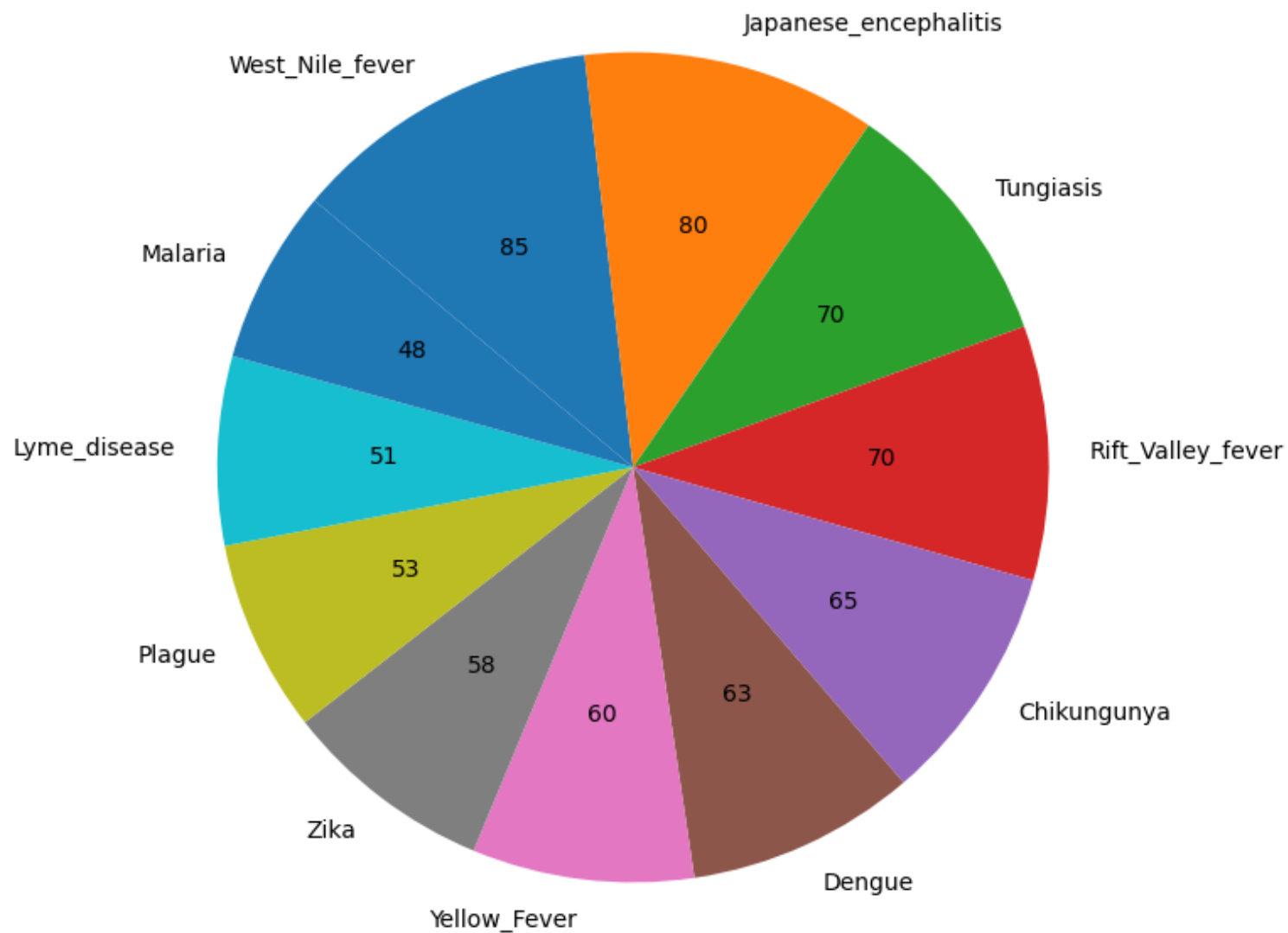
- **ID:** Unique identifier for each patient.
- **Symptoms:** A set of 64 binary features representing presence/absence of specific symptoms.
- **Prognosis:** Target variable representing medical prognosis.

## Initial Data Observations

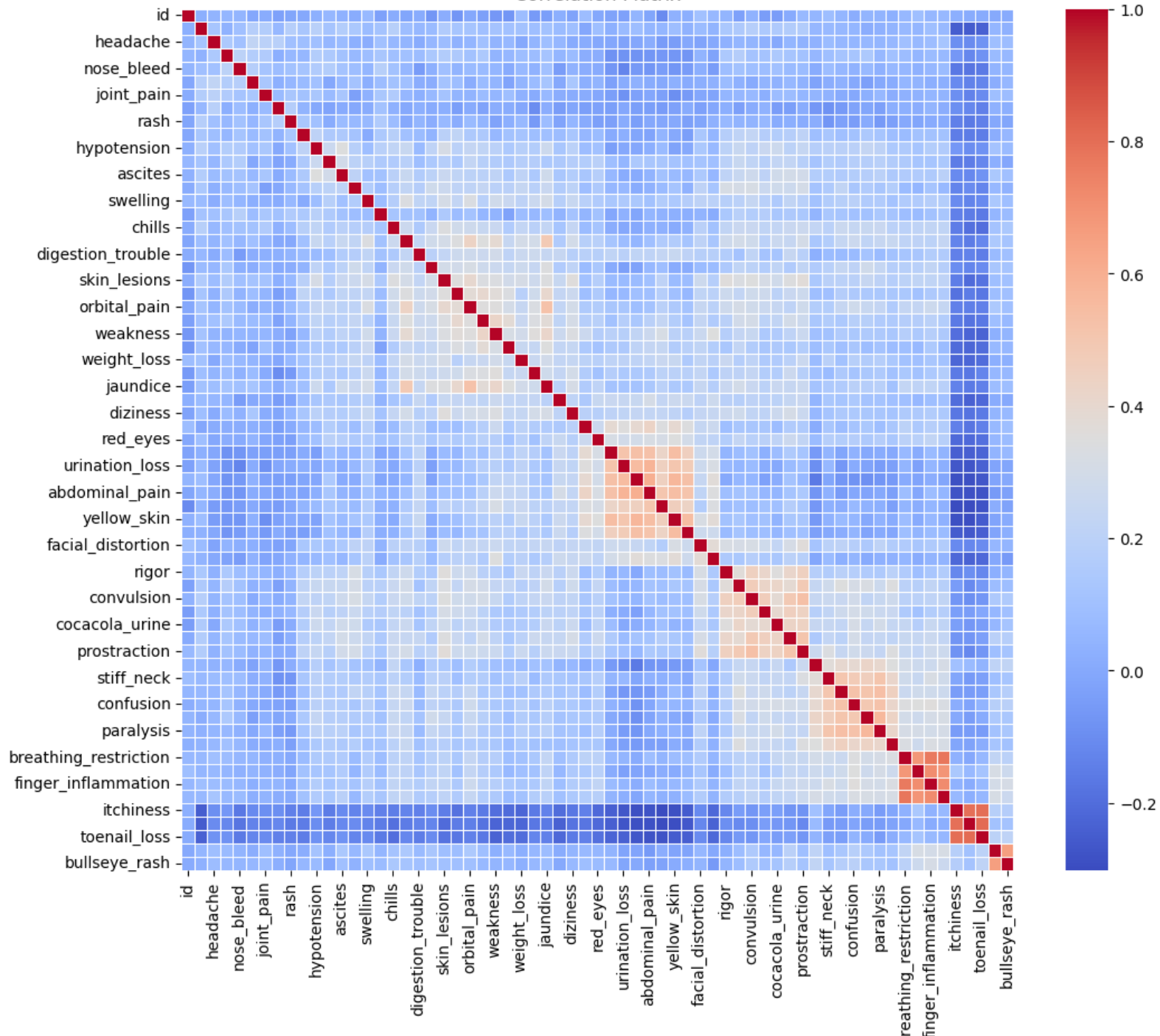
- **Data Completeness:** No missing values, ensuring data integrity.
- **Clean Structure:** No duplicate records.
- **Prognosis Variety:** 11 distinct prognosis categories.
- **Binary Symptom Encoding:** One-hot encoding for binary symptom information.
- **Class Distribution:** Variations in prognosis category frequencies.
- **Dimensionality Challenge:** 64 features vs 707 data observations, prompting dimensionality reduction like PCA



## Distribution of Prognoses

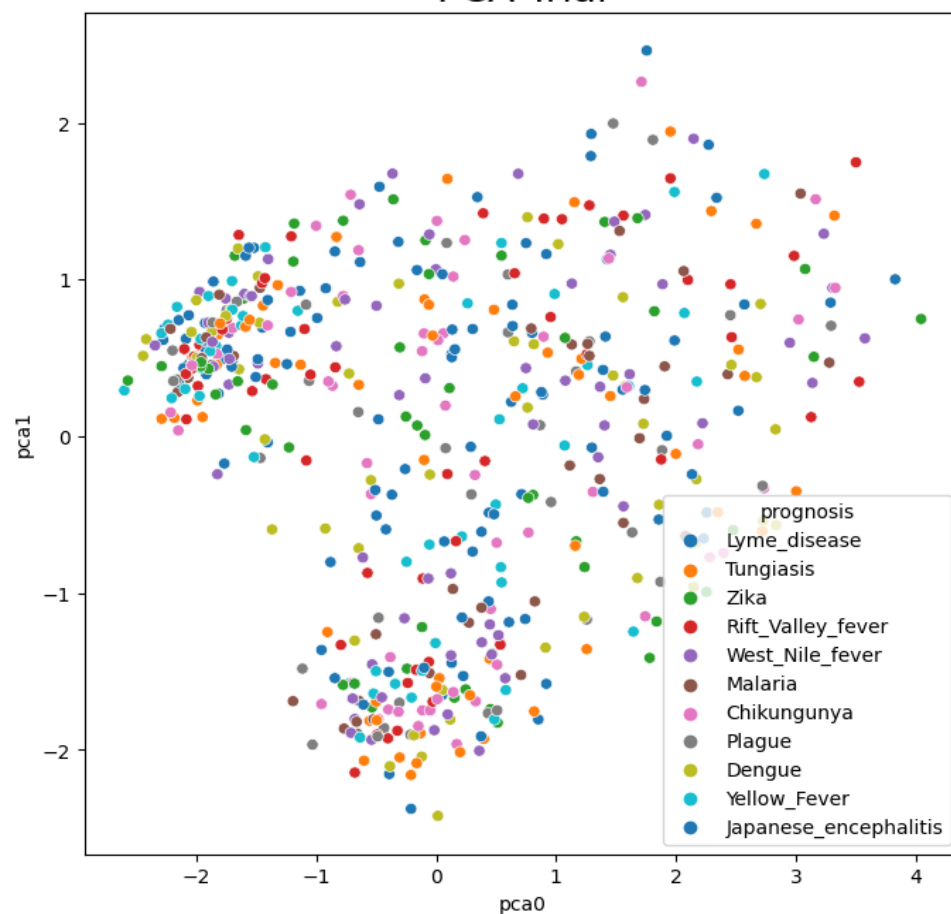


Correlation Matrix

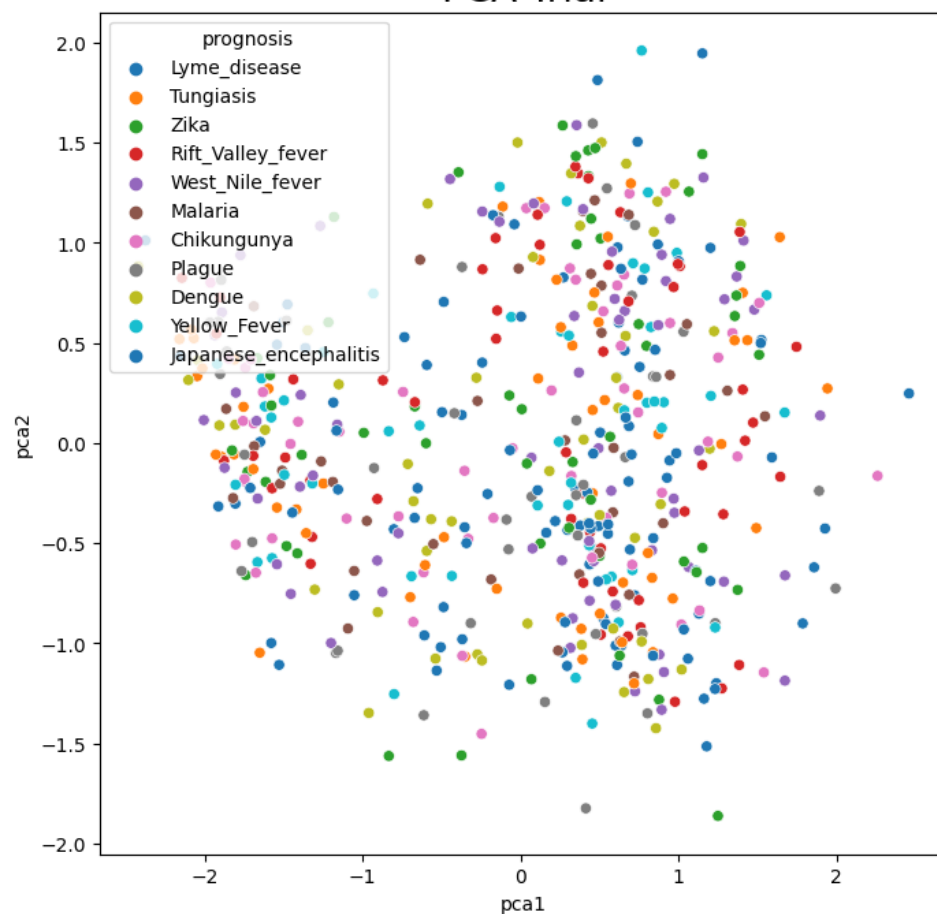


- After the initial Principal Components, the variance explained by each component becomes stagnant.
- Initial observation of the dimensionality reduction need is validated.

PCA Trial



PCA Trial



## 1. Outlier Detection

- **Methods Used:** Mahalanobis distance and Otsu thresholding.
- **Result:** Many data points classified as outliers due to a small dataset.

## 2. Decision to Retain Entire Dataset

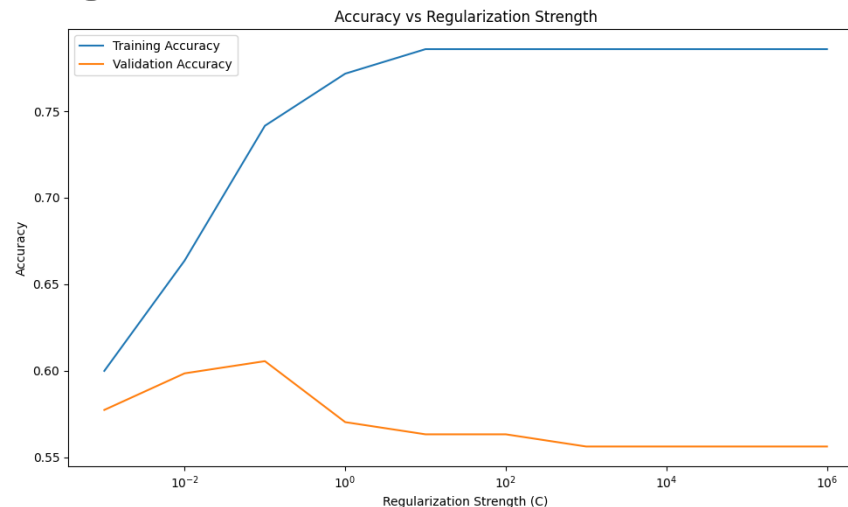
- **Reasoning:** To avoid information loss and model overfitting.
- **Alignment with Goal:** Maximizing data utility for accurate predictions.

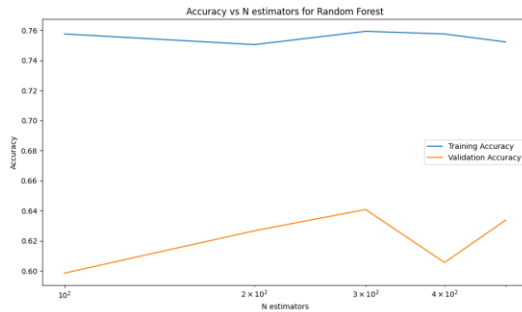


# Methodology

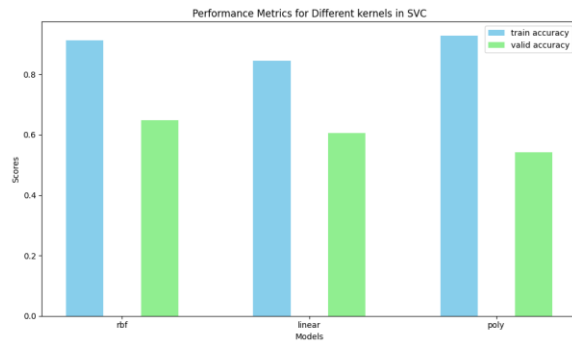


- Models used for the purpose of multi-class classification were Logistic Regression, Random Forest Classifier, SVC Classifier, XG Booster, MLP Classifier, Naive Bayes, Voting Classifier
- Reduced the dimensionality to the optimum point in order to get desirable results.
- In, Logistic Regression we used L2 regularization and fine-tuned C-parameter which controls the strength of this regularization. L2 was preferred over L1 as it showed much better evaluation.
- Later on, we performed grid search on all models, to fine tune their hyperparameters.






- Using ensemble methodology such as Random Forest Classifier helped attaining higher predictive accuracy as it uses multiple decision trees to predict the class label. We fine-tuned parameters and allowed bootstrapping for addressing overfitting.



- SVC can prove to be helpful in detecting complex boundaries among the classes. With the use of kernel tricks data non-linearity can be addressed that other two models might not have detected. Out of different kernels RBF proved to be the best as it is able to project data to infinite dimensions.

- 
- Used gradient boosting techniques provided in XGB classifier. This improved upon the previously used random forest. It is also quite fast computationally.
  - MLPs are a type of artificial neural network that can perform excellent on non-linear data. It can help in extracting features that are relevant for classification, making it helpful in learning hierarchical structure of features with classes.
- 
- A decorative graphic in the bottom right corner of the slide, consisting of several overlapping, light teal-colored rectangular blocks arranged in a staggered, geometric pattern.

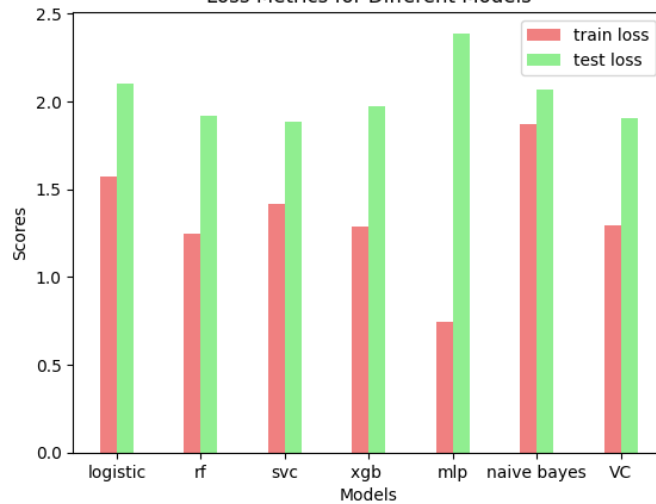
- Naïve Bayes classifier that is primarily based upon Bayes theorem helps in determining classes in probabilistic approach. It makes assumption of independent feature set. It assigned each class with its probability. Class with highest probability is the shown as the predicted class. Naïve classifier helps for higher dimensionality data set.
- Finally, we used ensemble methodology for making more reliable predictions. We selected top 3 performing models and used them as inputs for the Voting Classifier. This classifier uses the base models and their predictions in order to report the most voted class prediction. This methodology is expected to improve the predictions as compared to each individual models.



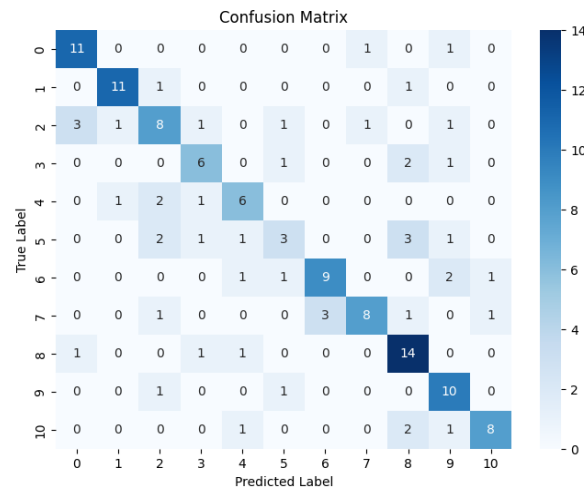
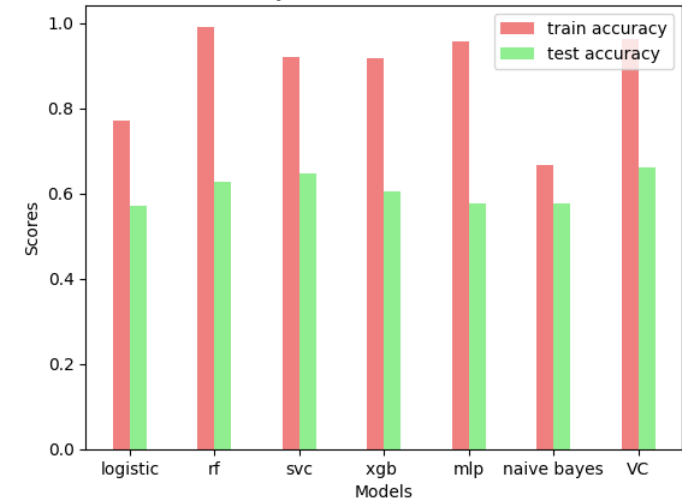
# Results



Loss Metrics for Different Models



Accuracy Metrics for Different Models



# 1) Train Accuracy(in %)

---



Model Type	Accuracy	Log loss
Logistic Regression	77.17	1.57
Random Forest	99.29	1.24
Support Vector Classifier	91.68	1.39
XGB	91.68	1.29
MLP	54.23	2.55
Naïve Bayes	66.56	1.86
Voting Classifier	95.58	1.2915

## 2) Test Result (in %)



Model Type	Accuracy	Log loss
Logistic Regression	57.04	2.10
Random Forest	60.56	1.93
Support Vector Classifier	66.9	1.88
XGB	60.56	1.97
MLP	54.23	2.55
Naïve Bayes	57.75	2.01
Voting Classifier	64.08	1.904

### 3) Class wise Accuracy for voting Classifier



Disease Name	Accuracy
Japanese Encephalitis	84.62
Tungiasis	84.62
Rift Valley Fever	37.5
Chikungunya	60
Dengue	40
Yellow Fever	36.36
Zika	71.43
West Nile Fever	66.61
Plague	78.57
Lyme Disease	29.41
Malaria	83.33

- We get the best model performance, when we select the top 35 features out of the given 45 features using PCA.
- SVM shows the lowest bias amongst all 7 models.
- Except MLP, all other models have high variance, indicating overfitting.
- All models have accuracies lying in between 55% - 65%
- The model's performance seems quite good, considering the accuracy for random selection is around 9% and we are getting accuracy of around 67% for Voting classifiers and almost between 55% - 65% for the other rest.
- High top 3 accuracies but relatively less favorable log loss result signaled that although models are predicting correct within 3 ranks of probabilities but aren't predicting with reliable probabilities for those 3 ranks.

# Conclusion

---



- Our project on "Vector-Borne Disease Prediction" represents a significant step forward in leveraging machine learning techniques to address the global challenge of vector-borne diseases.
- We embarked on this research motivated by the alarming statistics provided by the World Health Organization, highlighting the substantial mortality rates associated with these diseases.
- Our top-performing models, namely Support Vector Classifier, Random Forest demonstrated promising results in terms of accuracy. However, we acknowledged the challenge of model overfitting, especially in the case of Support Vector Classifier, and addressed it through hyperparameter tuning.
- The integration of these top-performing models into a Voting Classifier further enhanced predictive accuracy, surpassing the individual models. The Voting Classifier emerged as a robust solution, showcasing improved performance on both training and testing datasets.
- While the journey through this project revealed the complexities of predicting vector-borne diseases, the results underscore the potential of machine learning in contributing to the mitigation of this global health crisis.

➤ Yes, we have been on track with our dates and have completed this project within according to our timeline.

➤ **Our current timeline is as follows:**

- I. Data Pre-processing: 1st Sep, 2023 – 14th Sep, 2023
- II. Feature Selection and Extraction: 15th Sep, 2023 – 30th Sep, 2023
- III. Model Selection and Training: 1st Oct, 2023 – 20th Oct, 2023
- IV. Model Testing and Evaluation: 15th Oct, 2023 – 30th Oct, 2023
- V. Analysis and Documentation: 1st Nov, 2023 – 15th Nov, 2023

# Individual Task

---

- Aditya Yadav: Data pre-processing, Feature extraction, Model analysis
- Megha: Literature review ,Data Preprocessing
- Priyanshu Sehrawat : Model selection, model training and testing
- Vivaswan Nawani: Model Training, Model Testing

