Vector-Borne Disease Prediction

Aditya Yadav, Megha, Priyanshu Sehrawat, Vivaswan Nawani
Indraprastha Institute of Information Technology, Delhi
aditya21374@iiitd.ac.in, megha21337@iiitd.ac.in, priyanshu21407@iiitd.ac.in, vivaswan21217@iiitd.ac.in

November 23, 2023

1 Motivation

Vectors, often overlooked but biologically significant, serve as agents for pathogen transmission from host to host. These vectors are responsible for propagating vector-borne diseases, a category comprising 17% of all infectious diseases.
As per the World Health Organization's 2020 estimates, these diseases contribute to a substantial annual mortality rate of 700,000 individuals, presenting a formidable global health challenge. This challenge becomes even more acute in regions with limited healthcare infrastructure, where the absence of prompt diagnostic and predictive tools results in disproportionately high mortality.
We believe that the answer to solving this global problem could be found using machine learning and that is what prompted us to take on this topic.

2 Introduction

This research project aims to solve the pressing global issue of vector-borne diseases by leveraging Machine Learning techniques. Our primary objective is to conduct comprehensive experiments utilizing diverse machine-learning models to explore effective solutions. Specifically, we aim to construct a robust machine-learning model for the early detection of vector-borne diseases, utilizing a patient's symptomatology as input features to facilitate accurate prognosis. This innovative approach holds significant promise in reducing the impact of these diseases and reducing morbidity and mortality on a global scale.

We began working on this project, by exploring studies on vector borne diseases, to better understand the problem that we are dealing with and its gravity. This initial inquiry commenced with an in-depth analysis of information provided by organizations, such as the World Health Organization (WHO) and the Centers for Disease Control and prevention.

To further deepen our understanding on the work done in this field in the context of machine learning, we did an extensive literature survey, encompassing a rigorous review of research papers. Two studies of particular significance were identified: the seminal work of Mello Román et al.[2] and K. Indhumathi et al.[3]

Subsequently, data was sourced from a Kaggle dataset, marking the commencement of the practical part of our project. To ensure the quality of our data, a pre-processing phase was initiated. This phase started with data visualization. We used scatter plots, pie charts and bar graphs to better understand visualize our data. This was followed by feature engineering and dimensionality reduction using Principal Component Analysis (PCA). These critical steps were indispensable in transforming our raw data into a structured format suitable for model training, validation and testing and improve the quality of our data to improve our peformance.

After that, we started with model selection. For now, we selected with 3 models. Logistic Regression, Support Vector Classifier and Random Forest. We split the data into training dataset, validation dataset and testing dataset.

Each model underwent model fitting, hyperparameter tuning, and performance evaluation across key metrics, including Accuracy, Precision, and F1 score. The results were further analyzed through the use of confusion matrices, enabling a comprehensive analysis and meaningful conclusions to be drawn from our experiments.

As this report progresses, it will delve into the empirical findings and analyses resulting from our experimentation. These insights will provide an academic perspective on the potential of Machine Learning to address the challenges posed by vector-borne diseases. Ultimately, this paper represents a scholarly commitment to contribute to the mitigation of a pressing global health crisis through innovative research and empirically validated solutions in the domain of machine learning.

3 Literature Survey

Recent years have seen a rise in medical research that predominantly concentrates on the prediction and identification of variables that contribute to disease causation. The foundation for treatment recommendations in healthcare lies in the substantiation provided by these predictive findings, a principle underscored by Isinkaye et al.[1]in their 2015 study. Over recent years, the field has witnessed the emergence of an array of detection systems and decision support tools, designed to augment the diagnostic capabilities of medical professionals.

- In 2019, Mello Román et al.[2] conducted a study comparing two machine learning algorithms, Support Vector Machines (SVM) and Artificial Neural Networks (ANN), for disease diagnosis using data from the Paraguayan public healthcare system collected between 2012 and 2016. They found that with minimal data preprocessing, the Artificial Neural Network achieved impressive results, with 97 percent specificity, 96 percent sensitivity, and 96 percent reliability. Support Vector Machine (SVM) also performed well, with results exceeding 90percent for specificity, sensitivity, and accuracy.

  Furthermore, the researchers proposed a smart Healthcare Recommendation System (HRS) using a Deep Learning approach, combining Restricted Boltzmann Machines (RBM) and Convolutional Neural Networks (CNN). This approach emphasized the potential of big data analytics to enhance healthcare recommendation systems, opening new opportunities for the medical sector. In the context of a telehealth system, their approach demonstrated superior performance in terms of Mean Absolute Error (MAE) and Root Mean Square Error compared to existing strategies, indicating fewer errors in prediction.

- In 2021, K. Indhumathi et al.[3] conducted a survey on seasonal prediction and diagnosis linked to climate change using big data. They introduced the Gaussian Process Regression Conceptual Model and found a significant association between climate change and bacterial contamination. Rising temperatures have led to an increase in seasonal infections like dengue fever, diarrhea, salmonella, and giardia lamblia. However, the Gaussian Process Regression Method is more suitable for large datasets.

  To forecast dengue outbreaks in Selangor, Malaysia, the researchers explored various machine learning methodologies, including 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. Notably, the Support Vector Machine (linear kernel) achieved the highest forecasting accuracy, with a 70percent accuracy rate, 14percent sensitivity, 95percent specificity, and 56percent precision.

- In 2021, N. Reddy et al.[?] conducted research on machine learning (ML) methods for disease diagnosis, focusing on the identification of illnesses. They explored various ML algorithms, including Linear Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Trees (DT), and Naive Bayes (NB).

For the analysis of liver-related illnesses, the study employed methods such as Back Propagation, SVM, RF, NB, and KNN, with Naive Bayes demonstrating the highest precision at an impressive rate of 95.1 percent. While the study primarily focused on chronic diseases like liver and kidney illnesses, the methodologies and algorithms used in ML can be adapted to assess and verify the effectiveness of predictive models for seasonal illnesses.

4 Dataset

In this section, we provide an overview of the dataset used for our analysis and the preprocessing steps applied to prepare the data for modeling.

4.1 Dataset Description

The dataset used in this study consists of two main components: the training dataset and the test dataset. The training dataset (`train.csv`) contains a set of records, each representing a patient's medical information. It includes the following key attributes:

- **ID**: A unique identifier for each patient.
- **Symptoms**: A set of symptoms or features indicating the patient's medical condition. These symptoms are used as input features for our predictive model. There are a total of 64 binary features, each representing the presence or absence of a specific symptom.
- **Prognosis**: The target variable, representing the medical prognosis or outcome for each patient.
- **Dimensionality Challenge**: Our dataset's 64 features outnumber the available 707 data observations, prompting the need for dimensionality reduction techniques like PCA.

In our dataset, binary symptom features indicate the presence (1) or absence (0) of specific symptoms for each patient. Our objective is to develop a binary classification model to predict medical prognosis based on these symptom profiles

4.1.1 Initial Data Observations

Upon initial examination of our dataset, the following noteworthy observations emerged:

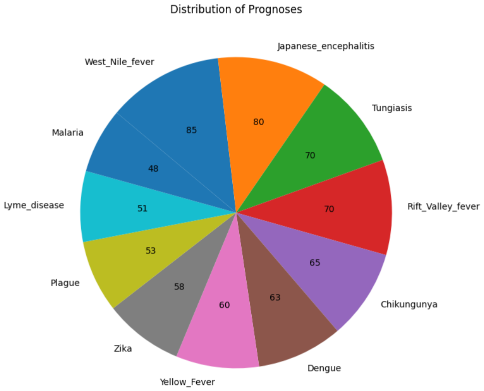

Figure 1: Class sample distribution

Data Completeness: The dataset is free of missing values, ensuring data integrity. **Clean Structure**: No duplicate records exist in the dataset, maintaining its cleanliness. **Prognosis Variety**: We have 11 distinct prognosis categories, reflecting the complexity of conditions we aim to predict. **Binary Symptom Encoding**: All features are one-hot encoded, representing binary symptom information. **Class Distribution**: While not severely imbalanced, there are variations in the frequency of prognosis categories.

In this section, we delve into visualizations that provide valuable insights into our dataset and its characteristics.

4.1.2 Correlation Matrix

To gain a deeper understanding of feature relationships within our dataset, we computed the correlation matrix for the binary symptom features. The correlation matrix reveals several intriguing observations:

- **Feature Collinearity**: The full correlation matrix shows a fair amount of feature collinearity, indicating the presence of relationships between certain symptoms. This suggests that some

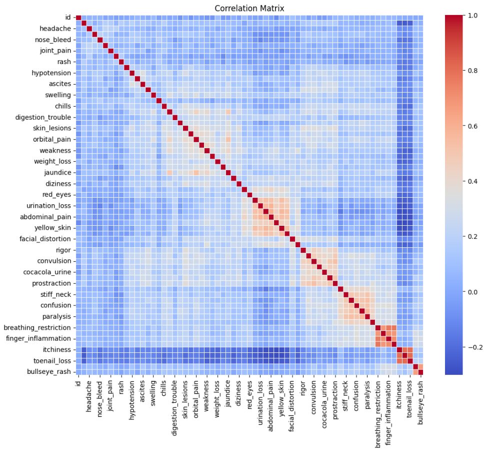

Figure 2: Correlation Matrix

symptoms tend to co-occur more frequently than others.

- **Strong Feature Correlations**: Notably, there are quite a few features that exhibit strong and statistically significant correlations with one or more other features. These relationships highlight the interdependencies between symptoms and underscore the potential presence of redundant or closely related features.

These insights from the correlation matrix reinforce the notion that dimensionality reduction techniques, such as Principal Component Analysis (PCA), could prove highly beneficial in addressing feature redundancy and optimizing our dataset for subsequent modeling.

4.2 Data Preprocessing

In this section, we detail the essential data preprocessing steps undertaken to prepare our dataset for subsequent modeling tasks.

4.2.1 Label Encoding

To facilitate the application of machine learning algorithms, we initiated the data preprocessing phase by encoding the categorical labels. Specifically, we encoded the ”Prognosis” column, which represents the medical prognosis or outcome for each patient, using a Label Encoder.

4.2.2 Dataset Splitting

With the labels successfully encoded, we proceeded to split the dataset into training and validation sets. The training dataset, comprising 80

4.2.3 Principal Component Analysis (PCA)

To tackle the dimensionality challenge posed by the dataset's 64 binary symptom features, we applied Principal Component Analysis (PCA). This technique reduces feature count while preserving vital information. We retained the top 15 principal components, capturing significant dataset variability.

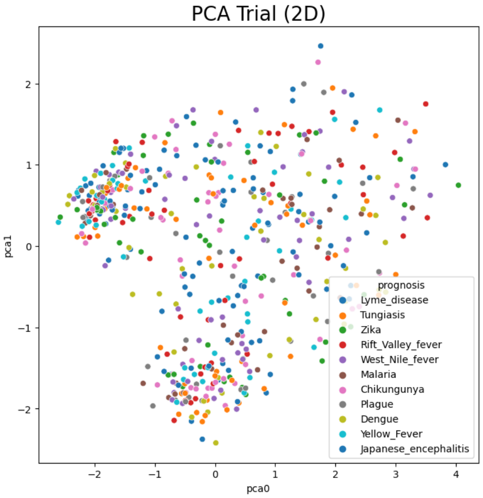

Figure 3: PCA Scatter Plot

4.2.4 Modified Training and Validation Sets

Following the selection of the top 15 principal components, we created modified training and validation sets that exclusively contained these components. This step streamlines our dataset, making it more manageable for subsequent modeling tasks while preserving the vital information necessary for accurate predictions.

4.2.5 Outlier Detection and Decision

We used Mahalanobis distance and Otsu thresholding to identify outliers. Due to our small dataset, many data points were classified as outliers. However, we opted to retain the entire dataset for modeling, avoiding information loss and model overfitting. This choice aligns with our goal of maximizing data utility for accurate predictions.

5 Methodology

Our group chose three models for the purpose of multi class classification. These models are Logistic Regression, Random Forest Classifier and SVC classifier. Initially we pre-processed our data for optimal train and test split. Further, we trained the models on the reduced dimensional train data set and performed evaluation metrics such as accuracy scores, precision scores and f1 scores on both train and test data sets. Initially we ran our models with default parameters, here logistic regression and Random Forest classifier showed high accuracy in training data set but relatively low accuracy on the test data set. It signaled over-fitting in the model. Hence, we tuned the hyper-parameters in the models. This reduced training accuracy and increased testing accuracy in the models. This addressed the issue of higher variance, bias and over-fitting of the model.

- ### 5.1 Logistic Regression

  Primarily logistic regression is used for binary class classification but we modified it using the multi-class parameter and set it to multinomial. In this methodology, the function assigns different probabilities to each class and the probabilities for each class sums up to 1. Here the predicted class is the one with highest probability. We used L2 regularization and controlled the strength of this regularization using the C parameter. A larger value for C would result in loose regularization whereas smaller values result in strong regularization. The best model fitting was found to be near C=1.

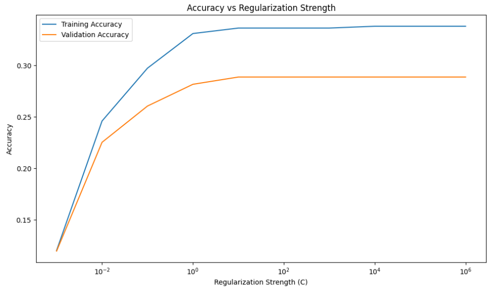

Figure 4: Regularization Strength vs Accuracy for Logistic regression

- ### 5.2 Random Forrest Classifier

  This classifier makes several decision trees upon the test data and predicts the most voted class label from these ensemble of decision trees. We used bootstrapping for making subsets of the data. By varying sample splits, we encountered lesser variance and improved overall model fitting.

- ### 5.3 SVC Classifier

  This classification produces an optimal decision boundary that separates the class labels from each other. Kernels help in projecting the data points in higher dimensions which helps when data has non-linearity. Most commonly used strategies in prediction multi class labels is one versus one strategy. Here the classifier two classes from the available classes and perform standard svm classification to produce a decision boundary. This is done for each pair of classes. Hence, the model is trained to distinguish between every pair of classes in order to predict the correct class label. We changed kernels and manually varied the parameters such as degrees, C, and gamma parameter in order to determine the best fit. RBF kernel proved to be the best fitting for the model as it showed improved accuracy as compared to other kernels and also reduced over-fitting.

- ### 5.4 XGB Classifier

XGB or Extreme Gradient Boosting classifier is a popular option for performing multi class classification tasks. It is based upon decision trees and improves on the random forest approach by using gradient boost framework. It is quite fast computationally. Given its advantages such as stronger predictions due to its ensemble nature and ability to recognize non linear pattern in data, it was expected for it to perform better than most of the other models.

- ### 5.5 Multi Layer Perceptrons

MLP is an artificial neural network that mimics the way our brain neurons operate. It is based on the feed forward mechanism and receives error signals in back propagation phase. It is primarily used for the data that can not be separated linearly. Hence, this makes it ideal for multi class classification.

- ### 5.6 Naive Bayes Classifier

This classifier uses the Bayes theorem to classify labels. It is a probabilistic classifier, which means that it assigns each class with its own probability. The class with highest probability is the predicted class. It assumes independence of features among themselves. This assumption severely affected the performance of the model as the features showed to be not independent from each other.

- ### 5.7 Voting Classifier

Voting classifier uses ensemble methodology for making predictions. It considers the predictions from all the base classifiers that it takes as input and finally predicts the most voted class as the predicted label. It is very useful for accounting for weaknesses in each separate classifier it takes as input. This helps in making more reliable predictions which may not be possible for a single classifier. We used Voting classifier on the top performing models as seen from the metrics. Top 3 models which were found to have provided accuracy scores of above 60 percent were given as input to the voting classifier. The Voting classifier performed better than all of the models' performances taken separately.

6 Result

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
Naive Bayes	66.56	1.86
Voting Classifier	95.58	1.2915

Table 1: Train Result (values are 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
Naive Bayes	57.75	2.01
Voting Classifier	67.4	1.904

Table 2: Test Result (values are in %)

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
Plague	78.57
Lyme Disease	29.41
Malaria	83.33 West Nile Fever 66.67

Table 3: Classwise Accuracy for Voting Classifier (values are in %)

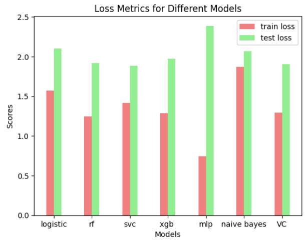

Figure 5: Loss metrics

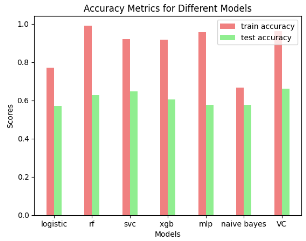

Figure 6: accuracy metrics

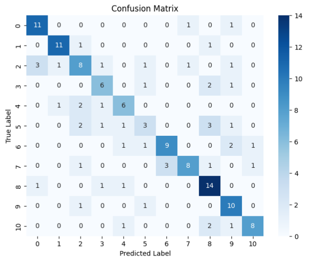

Figure 7: Confusion Matrix for Voting Classifier

Key Observations:

- Best Performance on testing data is given by Support Vector Classifier with 66.9% accuracy and 1.88 log loss
- Support Vector classifiers have high variance, indicating overfitting to some extent.
- XG Boost and Random Forrest are the next best models with the same accuracies of 60.56%, they also show high variance
- MLP classifier has the least variance , indiating that it has maintained its generalizability while having a test accuracy of 54.23%

7 Conclusion

We could see that the top performing models were random forest, XGB classifier and SVC classifier. We incorporated the best performing models further in the voting classifier. We used top k accuracy as the metric criteria for this selection process. Voting classifier used the most voted outcome from these models and hence, it improved the final prediction and outperformed all individual models.

References

[1] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, Recommendation systems: Principles, methods and evaluation, Egyptian Informatics Journal, Volume

16, Issue 3, 2015, Pages 261-273, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2015.06.005. Keywords: Collaborative filtering; Content-based filtering; Hybrid filtering technique; Recommendation systems; Evaluation

[2] Mello-Román JD, Mello-Román JC, Gómez-Guerrero S, García-Torres M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Computational and Mathematical Methods in Medicine. 2019 ;2019:7307803. DOI: 10.1155/2019/7307803. PMID: 31485259; PMCID: PMC6702853.

[3] Indhumathi K, Sathesh Kumar K. A review on prediction of seasonal diseases based on climate change using big data. Mater Today Proc. 2021;37:2648-2652. doi: 10.1016/j.matpr.2020.08.517. Epub 2020 Oct 2. PMID: 33024706; PMCID: PMC7530581.