Analysis On Dengue Severity Using Machine Learning Approach

by

Sanjana Sayeed
17301189
Iktisad Rashid
16231004
Muktadir Rabbi Sotej
16101113

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
January 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sanjana Sayeed

Sanjana Sayeed
17301189

Muktadir Rabbi Sotej
16101113

Iktisad Rashid

Iktisad Rashid
16241003

The thesis/project titled "Analysis On Dengue Severity Using Machine Learning Approach" submitted by

1. Sanjana Sayeed (17301189)
2. Iktisad Rashid(16231004)
3. Muktadir Rabbi Sotej (16101113)

Of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 08, 2021.

Examining Committee:

Supervisor:
(Member)

Rasif Ajwad

Rasif Ajwad
Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

faisal

Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:

Prof. Mahbub Majumdar
Chairperson
Dept. of Computer Science & Engineering
Brac Universitv

Mahbubul Alam Majumdar, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Dengue is a viral disease that spreads in tropical and subtropical regions and is especially prevalent in South-East Asia. To some certain extent, this mosquito-borne disease triggers nationwide epidemics, which results in large number of fatalities. In our study, we mainly worked with two data sets from two countries: Bangladesh and Vietnam. For the Vietnamese data set we have used supervised learning and implemented different prediction models like Decision Tree Classifier, Random Forest, Gradient Boosting, Ada Boosting, XG-Boosting Classifier Model and have taken the best fitted one (that being XG-Boosting Classifier) to predict the severity amongst the dengue infected patients. After predicting severity we analyzed the data set further to identify what might be the possible cause leading towards the DSS or the DHF for the clinical data. In parallel, for the Bangladeshi data set we applied the unsupervised learning technique, Hierarchical Clustering, to find the different clusters of various vital components of the patients according to their blood report. We then analyzed the clusters further to find the severity among the patients, which led them to DSS or DHF.

Keywords: dengue, DSS, DHF, supervised, unsupervised, hierarchical clustering, xg-boosting,clinical data

Acknowledgement

We would like to thank both of our Supervisor Rasif Ajwad, Lecturer of Brac University and Co-Supervisor Faisal Bin Ashraf, Lecturer of Brac University. Sir Rasif Ajwad and Faisal Bin Ashraf have been like a guardian to us and an inspirational figure to aspire toward our goal on our research from the very beginning. We will sincerely like to thank our Co-Supervisor Faisal Bin Ashraf Sir to mentor and guide us throughout the entire process of the research. Without the contribution of both Rasif Ajwad Sir and Faisal Ashraf sir we could not have completed the research. We would also like to render our appreciation to Dr. Md Arifur Rahman and Dr. Subrata Kumar Mondal for their active support throughout the whole data collection journey. Finally we express our gratitude to the Computer Science And Engineering Department, BRAC University for giving us the opportunity to conduct the research.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ada	Adaptive
DHF	Dengue Haemorrhagic Fever
DSS	Dengue Shock Syndrome
FN	False Negative
FP	False Positive
HCT	Hematocrit
KNN	K-Nearest Neighbors
MICE	Multivariate Imputation By Chained Equations
PICU	Pediatric Intensive Care Unit
SHAP	SHapley Additive exPlanations
TF	True Positive
TN	True Negative
WBC	White Blood Cell
XG	eXtreme Gradient

Chapter 1

Introduction

The events of dengue cases have grown tremendously around the world over the last few decades[49][45]. Dengue fever is an acute febrile viral disease spread via the bite of Aedes mosquitoes carrying any of the four viral serotypes of dengue. Recent research suggests that there are 390 million dengue infections every year and picture dengue transmission is omnipresent throughout the tropics, with the highest risk in the American and Asian regions [16] [47]. The build up of dengue cases throughout the Southeast Asia Region are periodic and the situation of an epidemic in various countries varies greatly from time to time and from place to place [42]. Bangladesh as a South Asian country has made tremendous progress in the management of communicable diseases. But however , the country still has problems managing dengue cases, sometimes scaling to epidemic levels. A special problem is vasculopathy marked by endothelial dysfunction and plasma leakage that occurs several days well after disease, often throughout the time of defervescence; this condition appears to be much more severe in children and can be significant enough to induce hypovolaemic shock, i.e. dengue shock syndrome(DSS) [10] [24] [12]. Results of analysis of 40,476 cases of Bangladesh occurring during 2000–2017 indicated that 49.73% of the dengue cases occurred during the monsoon season (May–August) and 49.22% during the post-monsoon season (September–December) [16].

1.1 Motivation

Dengue virus outbreak places about 20 million people at risk each year in tropical and subtropical countries[37]. The continuum of disease symptoms varies from asymptomatic to moderate inflammation to various degrees of thrombocytopenia and vein leakage characteristic of dengue hemorrhagic fever (DHF) to extreme shock syndrome and multi-organ failure[37]. The motivation behind our analysis on severity of dengue or dengue shock syndrome(DSS) is the recent epidemic events in Bangladesh. There was an enormous outbreak of Dengue recently in Bangladesh which caused remarkable casualties with observation of severe cases more than ever as well as the impact of this disease on this region throughout the decade.

1.2 Major Contribition

In our following research we tried machine learning approach to our data sets which is a function of artificial intelligence (AI) that provides the ability to automatically learn and develop experience without being specifically programmed from the given data sets. Throughout applying several Prediction model like- Decision Tree Classifier,Random Forest,Gradient Boosting, Ada Boosting, XG-Boosting Classier Model we came to a conclusion that XG-Boosting Classifier model works the best in predicting shock(DSS) among patients in Vietnam data sets. Later this specific model was used for further analysis of severity among dengue patients.On the other hand, for Bangladesh data sets we did Hierarchical Clustering to find clusters among the given data and analyze the severity of dengue or dengue shock syndrome(DSS)/Dengue Hemorrhagic Fever(DHF) among the clusters.

1.3 Thesis Orientation

The following sections of this paper has been ordered in the way described below. Chapter 2 is Literature Review which is a description of the related works and various methods and techniques similar to our proposed methodology. Chapter 3 consists of the background study and methodologies of our thesis which includes the working principles, description of dengue fever classifications, dengue shock syndrome, serologies and their differences as well as other related terms. Chapter 4 describes the types of algorithms that were used in our research. Chapter 5 comprises of the Vietnam data set description along with their data constructions, feature selections and proposed models.Chapter 6 consists of the analysis and discussion of Vietnam data set. Chapter 7 consists of the Bangladesh data set description along with their data constructions, feature selections and proposed models. Chapter 8 consists of the analysis and discussion of Bangladesh data set.The entire paper is concluded and summarized with explaining the similarities between features of the two mentioned subtropical countries in Chapter 9.

Chapter 2

Literature Review

In this research paper[4] Chen, C.-C., Chang, H.-C.to estimate the outbreak of dengue they used approximate entropy algorithm and pattern recognition technique Here by using an approximate entropy algorithm they calculated ApEn they formed a sliding window which was used for moving along with each data point throughout the time series. This complexity calculator ApEn was for the measurement of local irregularities. Their data set was designed for weekly cases. As for the pattern recognition the study was divided with two patterns, pattern ‘A’ consisted of one screen and pattern ‘B’ consisted of two screens combined by an “OR” operator. Growing patterns had either established a positive or negative prediction. They measured both methods ’sensitivity and specificity, and reported the time(in weeks) differences between each true positive prediction and their subsequent outbreak. To calculate the ApEn to apply both methods MATLAB was used for analyzing the data.

In this particular paper[42], Ramadona AL, Lazuardi L, Hii YL, Holmner Å, Kusnanto H, Rocklöv has used statistical analysis of meteorological variables relating to dengue time series and the predictiveness of self-inheritance within the dengue time series itself to predict the dengue outbreak based on disease inspection and related data. The study consists of a collection of predictive models and compares predictive ability with the number of diseases to identify and forecast the temporal trends of dengue and dengue outbreaks. In this study the initial analysis of meteorological variables and their connection with disease trends were developed over time using all monthly lag with a temperature ranging from 0–3 , rainfall and modified moisture in the model and took all the values into consideration. The reason this method was chosen to disconnect in the results the different lag contributions factors of the similar meteorological component.

The authors of paper [1], prepared a report that can differentiate dengue from other febrile diseases and predict the critical health condition in adults. They passed clinical and haematological parameters in a C.45 classifier which can recognize dengue patients in comparison with the non dengue patients in the first 72 hours of the illness.A 25% of standard technique named pruning confidence was applied to delete branches where the algorithm was 25% or more optimistic in order to avoid making different branches which would not be suitable for generalisation. This prevented the data from overfitting. In their study they used a parameter called ‘minimal cases’ which represents a stop criterion for further data partitioning at different nodes in the decision. When the decision tree reached a level less than or equal to the minimum threshold value of ‘minimal cases’ then the decision making process of the tree was terminated. This prevented the tree from being subdivided into excessively narrow nodes which have very small amounts of data to support. They used ‘K’ fold validation technique for choosing the appropriate value for missing cases. The diagnostic algorithm reached an accuracy of 84.7%. Furthermore, in different disease prevalence they predicted that their algorithm can be used differently to yield positive and negative predictive values that are clinically useful. Besides this, they used platelet counts, crosshold threshold value of a real time RT-PCR for dengue viral RNA and presence of pre-existing anti dengue IgG antibodies in sequential order which had a specificity and sensitivity of 80.2% and 78.2% respectively.

In this research paper [6] Jiaqi Liu, James Ma,Jiaojiao Wang, Daniel Dajun Zeng, Hongbin Song, Ligui Wang, Zhidong, Cao applied a big data technique on electronic data record to investigate sex and age specific detection levels of comorbid conditions of hypertension and find their relationship to reveal the risk of hypertension of patients. Moreover, they used the network analysis method to find the concurrent relationship of the comorbidities of the diseases. They considered that there is a co-occurrences relation between the comorbidity pair if two comorbidities occurred in the same electronic medical report. The network nodes reflected comorbidities, and the edge in the network showed the concurrences of the two comorbidities with whom the edge is connecting. The number of co-occurrences between the comorbidities was determined by the weight of edges of the network. Whenever there contains more than one comorbidities in the electronic report the relationship between each comorbidities in that report would increase by one. After investigating all the medical reports related to hypertension they held elevated frequencies among the peak 20 comorbidities. The high frequency relationships represent the weight of more than 1% of the total number of hypertension-related records.

In the comprehensive study [15] they built a telecommunication and computer based technology system to monitor patients with comorbid conditions. In this research they developed a monitoring system for patients suffering from chronic disease (WPW) and acute disease (AF). In their experiment , ECG features of a particular patient are partially collected by a device which contains a group of sensors and the data was entered manually by the health staff. The ECG Features are dynamic and transferred to the hospital server via wireless network and internet. With this arriving data the decision support unit in the data server will make initial decisions on patient condition and treatment plans.By referring to these initial decisions and treatment plans the doctor will make necessary diagnosis and give necessary feedback. In addition to that , they proposed a method to evaluate the performance of their monitoring system by linking system accuracy and detects the conflicts with the distribution of the ECG feature.

Paper [30] findings show the comorbid conditions existed among the ICU patients and also compared the comorbidity across different demographic group.Moreover they also observed different comorbid conditions that co-exist among the patients with hospital acquired pressure injury (HAPI). They collected 4 years de-identified patient data from an academic institution in central Ohio,United States.The data contained demographic and clinical information from 12652 patients.For analysing the comorbid conditions of the ICU patients they collected data containing patients of different age ,gender,race and different health issues that exists between the patients.They summarized the comorbid conditions of the ICU patients by using descriptive statistics. Additionally for comparing comorbid conditions among different demographic groups they used the chi-square test and ANOVA. Next they filtered the conditions with HAPIs and explored the comorbid conditions in the specific group of patients.Finally they applied the univariate analysis to determine the comorbid association.

Paper [27] showed a model for multiple comorbidities of diabetes patients by the usage of Bayesian functional networks of latent variables. In their study they showed that the comorbidities related to diabetes can be complicated and hard to predict, particularly when loosely coupled.They mainly focused on two comorbidities for their research. One is the disorder of the lipid metabolism and another one is the non alcoholic chronic liver disease.They used the dynamic Bayesian networks from the bootstrapped data using the REVEAL algorithm. They included a latent variable that connects to all the data points and parameterized using the Expectation Maximization algorithm. It is being observed that the latent process captures the dynamic process of the comorbidities and also shows the interaction with clinical variables. Through their study they expected that different manifestations of diabetes can be understood.

In this research [28] Ms. Sanjana Das and Ms. Abha Thakral tried to conduct a research and used R predictive analysis to foretell dengue and malaria disease. By applying it on the data using R technique to evaluate time-series generic X-Y plotting and linear regression analysis, they used time series analysis. One of their key goals of time series analysis was to predict forthcoming values of the series. Furthermore for the analysis of their data, for X-Y plotting the generic function has been used. The different lines will reflect the different years in which cases occurred during the period from 2010 to 2015. A pattern has been observed during this time period to reach their conclusion.

In this comprehensive study [10] the authors have taken data analysis approach with monitoring the platelet and hematocrit count in blood from children who are laboratory confirmed dengue cases to predict the dengue shock syndrome(DSS). They also have taken the data of Vietnamese children aged from 5-15 years admitted to the Hospital with clinically suspected dengue case between 2001 and 2009. All the data sets were comprised of laboratory confirmed dengue case within 1-4 days of illness. For both univariate and multivariate analyses, logistic regression was the dominant statistical model for their research. Using graphs and separate regression models fitted for each day of illness, the prognostic value of daily haematocrit and platelet levels was assessed.

Chapter 3

Background And Methodology

3.1 Dengue Outbreak Throughout the Region

The events of dengue cases have grown tremendously around the tropical and subtropical island mostly in the South-East Asia region over the last few decades[49][46]. Recent research suggests that there are 390 million dengue infections every year and picture dengue transmission is omnipresent throughout the tropics, with the highest risk in the American and Asian regions[16][48]. The build up of dengue cases throughout the Southeast Asia Region are periodic and the situation of an epidemic in various countries varies greatly from time to time and from place to place[42]. Bangladesh as a South Asian country has made tremendous progress in the management of communicable diseases. But however , the country still has problems managing dengue cases, sometimes scaling to epidemic levels. Results of analysis of 40,476 cases of Bangladesh occurring during 2000–2017 indicated that 49.73% of the dengue cases occurred during the monsoon season (May–August) and 49.22% during the post-monsoon season (September–December)[42].

3.2 Dengue

Dengue is a virus carried by female Aedes mosquitoes carrying any of the four viral serotypes of dengue (DENV-1, DENV-2, DENV-3, DENV-4).

3.2.1 Dengue Fever(DF)

Dengue fever is a tropical disease transmitted by Aedes mosquito bites caused by the dengue virus. Symptoms normally start to show within three to fourteen days after infection. This can include high fever, fatigue, vomiting, muscle and joint pain, and a distinctive skin rash. Recovery usually takes between two and seven days.

3.2.2 Dengue Haemorrhagic Fever(DHF)

Dengue hemorrhagic fever may occur if a person is bitten by a mosquito or subjected to dengue-infected blood. Infected mosquitoes are by far the most common sources of infection. There are four distinct kinds of dengue viruses. Once someone has been infected with one of the viruses, he or she will develop immunity to the virus for the rest of your life. However, this immunity will not protect the person from other viruses. It is possible to be infected with all four different types of the dengue virus in someone's lifetime. Constant exposure to dengue virus can increase the likelihood that an individual will develop dengue hemorrhagic fever.

3.2.3 Dengue ShockSyndrome(DSS)

Shock syndrome is a severe complication of dengue infection and is linked to increased mortality rates. Extreme dengue develops as a result of secondary infection with another virus serotype. Increased vascular permeability, coupled with myocardial impairment and dehydration, leads to shock production, resulting in multi-organ failure.The onset of shock in dengue can be dramatic, and its progression relentless. The onset of shock in dengue can be drastic, and its persistence unceasing [29].

3.3 Dengue Serotype

Dengue illnesses are associated with four strongly linked viruses, DENV-1, DENV-2, DENV-3 and DENV-4. These four pathogens are called serotypes because they have distinct associations with antibodies in human blood serum.

3.4 Serological Differentiation of dengue virus

Glycoprotein NS1 is widely used in both serotypes and can actually identify whether the patient is in the primary or secondary stages of dengue. Additionally, IgM and IgG examinations can be useful for both primary and secondary diagnosis.

3.5 Dengue Detection Test

The most widely used diagnostic tool remains the identification of dengue-specific IgM and IgG-class antibodies. Seroconversion generally occurs 3 to 7 days after exposure and thus acute and convalescent sera monitoring can be appropriate for diagnosis.

3.5.1 RT-PCR

Real-time RT–PCR is a nuclear-derived tool for identifying the involvement of unique genetic material in any pathogen, even a virus.

3.5.2 NS1

NS1 generally refers to a sensitivity test. It does not confirm that a person has dengue or not but it will confirm the presence of a pathogen.

3.5.3 IgG/IgM

IgG/IgM conducts a specificity test. This is basically the antibody test which confirms the presence of dengue virus.

3.5.4 Complete Blood Count(CBC)

CBC consists of the following components [23] :

- **White blood cell count (WBC or Leukocyte count)**
- **White blood cell count (WBC differential count)**
- **White blood cell count (Red blood cell count (RBC or erythrocyte count))**
- **White blood cell count (Hematocrit (Hct))**
- **Hemoglobin (Hbg)**
- **Mean corpuscular volume (MCV)**
- **Mean corpuscular hemoglobin (MCH)**
- **Mean corpuscular hemoglobin concentration (MCHC)**
- **Red cell distribution width (RDW)**
- **Platelet count**
- **Mean platelet volume (MPV)**

3.6 Methodology

3.6.1 Vietnam datasets

We collected the the datasets from the supplementary section of Paper [10] and predicted the severity among dengue infected patients on the basis of the clinical data.The patients were from the age 5 to 15 years and were admitted to a hospital of Vietnam in the city of Ho Chi Minh between the year 2001 to 2009. The datasets required preprocessing, Where we used KNN Imputation to account for missing values. Then the pre-processed data set was fitted in different predictive model to achieve the best fitted which can accurately predict severity of dengue infected patients. The best fit model is further analyse the relevant features which affected probability to develop DSS/DHF. For the Vietnam dataset we choose supervised learning because the data was well structured. There was a labeled column output for every patient entry which was used as a target column to train the model.

The working flowchart for Vietnam Data set study is shown in Figure 3.1:

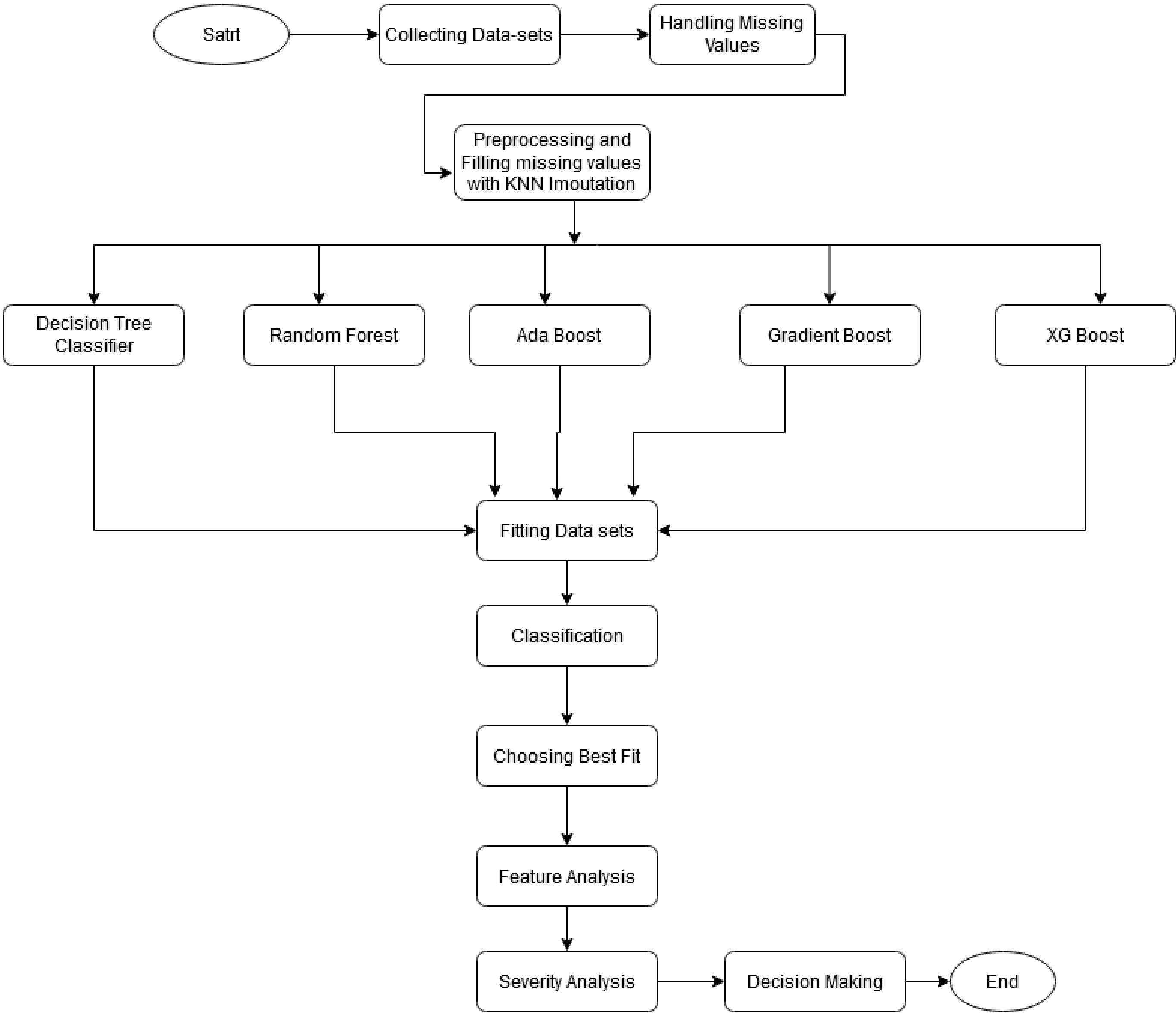


Figure 3.1: Proposed Plan Flowchart For Vietnam Data-sets

3.6.2 Bangladesh datasets

From Bangladesh we collected data from 'DR MR Khan Shishu Hospital' and 'Central Hospital'. 'DR MR Khan Shishu Hospital' data contained information of 69 pediatric patients with the age ranging from 8 month to 15 years. On the other hand the 'Central Hospital' datasets contains data of almost 100 patients comprising of both adults and children.The two data sets were merged together to make a single dataset and the rows of the newly created dataset was shuffled for a robust study. The dataset contained a lot of missing values which were handled with the interpolation technique. Then the dataset was passed through an unsupervised learning technique called Hierarchical Clustering to form clusters among dengue infected patients on the basis of different components of the complete blood count report. The clusters that were formed were further analysed to observe severity among patients. We chose unsupervised learning for Bangladesh dataset because there was no clear indication of labeled output column which could potentially say which patients are at risk of DSS. So we needed an algorithm to label and specify which patients could be at risk factor.

The working Flow diagram for Bangladeshi dataset is shown in Figure 3.2:

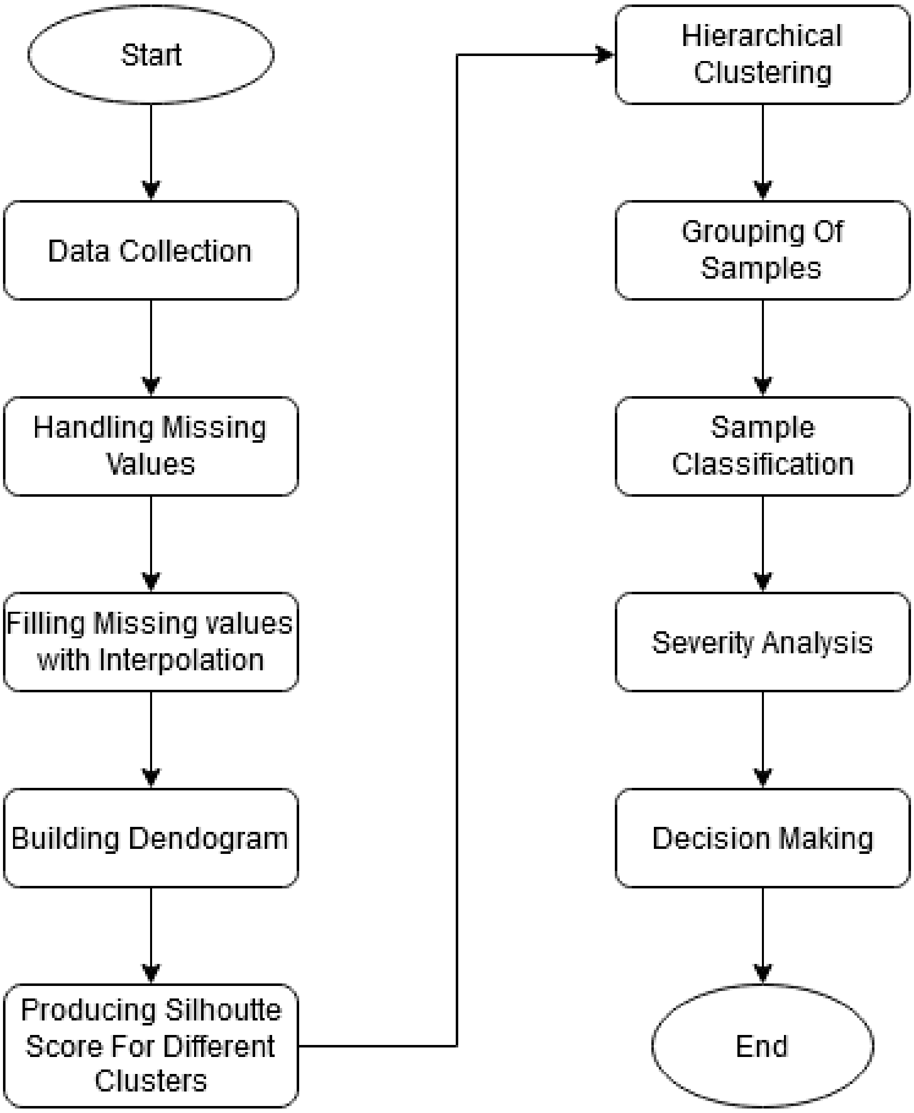


Figure 3.2: Working Flowchart For Bangladesh Datasets

Chapter 4

Machine Learning

4.1 What is Machine Learning?

Machine learning is an application and a subset of artificial intelligence (AI) that transforms systems, to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that feed on data to make better predictions over time.

4.2 Types Of Machine Learning

There are basically four types of Machine learning approach-Supervised Learning,Unsupervised Learning,Semi-supervised and lastly Reinforcement Learning.

- **Supervised Learning:** Supervised learning is by far the most common subbranch of machine learning, where the program has labeled input data and anticipated output results. The algorithm in the method is explicitly told what to look for, so the model is conditioned until it can identify the underlying patterns and associations, allowing it to generate as reliable a result as possible, when presented with newly presented data. Supervised learning is strong at grouping and regression questions, such as deciding which type of news item belongs to or forecasting the amount of revenue for a given future date. The aim of supervised learning is to make use of the data for precise calculations.

- **Unsupervised Learning:** In comparison to supervised learning, there is an unsupervised learning approach that seeks to make sense of the data itself. There are no external metrics or instructions for unsupervised learning; the algorithm only wants to grasp the data and to identify correlations or similarities. The most popular unsupervised learning approach is cluster analysis, which is used for exploratory data analysis to identify underlying correlations or data clustering. Clusters are represented using a similarity measure described by metrics such as Euclidean or probabilistic space.

- **Semi-Supervised Learning:** The study of how computers and natural systems, such as humans, learn in the presence of both labeled and unlabeled data is concerned with semi-supervised learning. Traditionally, learning has been studied either in the unsupervised paradigm (e.g., clustering, outlier detection) where all the data are unlabeled, or in the supervised paradigm (e.g., classification, regression) where all the data are labeled. The purpose of semi-supervised learning is to understand how it can change the learning behavior by combining labeled and unlabeled data, and develop algorithms that take advantage of such a hybrid. In machine learning and data mining, semi-supervised learning is of high relevance because it can use unlabeled data that is readily available to improve supervised learning tasks when labeled data is scarce or expensive[35].

- **Reinforcement Learning:** Reinforcement Learning :- Reinforcement learning is Machine Learning technique, which takes suitable action to maximize reward in a particular situation i.e to find the best possible behavior or path. Reinforcement learning differs from supervised learning in not needing labelled input/output pairs to be present, and in not needing sub-optimal actions to be explicitly corrected. Instead the focus is on finding a balance between exploration of uncharted territory and exploitation of current knowledge[31].

4.3 Algorithms Implemented for the Research

We basically used Classification,Clustering and Boosting Algorithm for predicting the severity of dengue.The brief description about the implementation of those algorithms are as follows:

Decision Tree:

The reason for considering a Decision Tree involves creating a testing model that can be used to forecast the goal variable's class or value by learning simple rules of decision derived from prior training data. We track the branch referring to that value on the basis of comparison, and leap to the next node. The decision to make strategic splits influences significantly on the accuracy of the decision tree. Conditions of decision for grouping and regression trees are different. Decision tree uses mixed algorithms to determine whether to divide a node in two or more child nodes. The decision tree divides the nodes into necessary variables that are needed to be picked and then separates resulting in the majority of homogeneous child nodes. The selection of the algorithm is focused on the goal variables. The Figure 4.1 will explain the working flow of Decision Tree Algorithm:

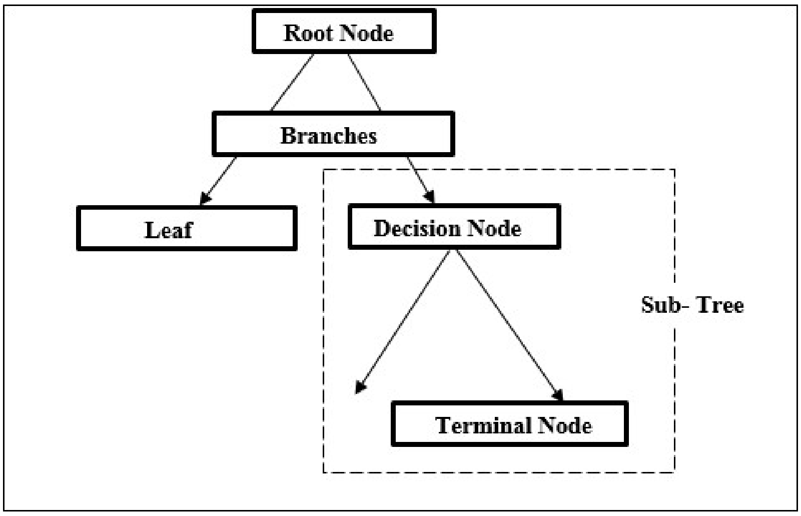


Figure 4.1: Decision Tree Working Principle [40].

In Figure 4.2 accuracy for decision tree model will both criteria Entropy and Gini with depth from 1 to 20 is shown:

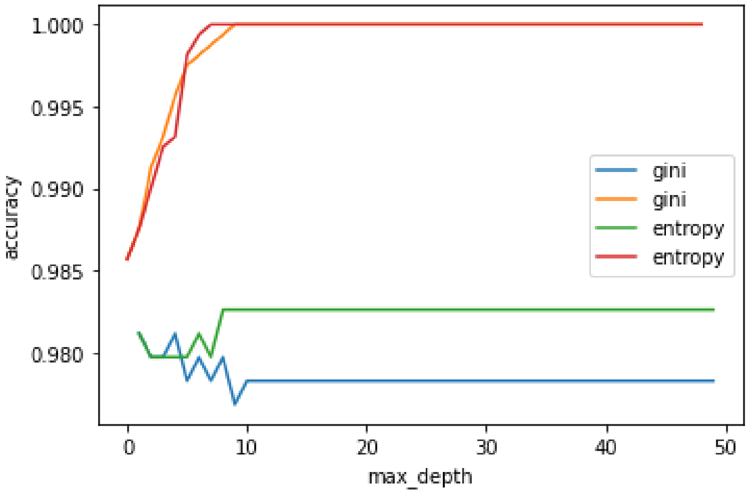


Figure 4.2: Testing and Training Accuracy for different hyperparameter for Decision Tree Classifier Model.

Random Forest:

Random Forest is an algorithm for supervised learning which is used for classification as well as regression.The main purpose of using Random Forest Algorithm is, it creates multiple Decision Trees on the given data samples, then gets a prediction from each individual sample by picking the best solution via voting. In Random Forest the process of finding nodes and splitting the feature nodes is completely random. Firstly the algorithm will start with the random sample selection from the datasets. Secondly it will merge the two decision trees and try getting prediction results from each of the decision trees.The beauty of this Algorithm is that a significant number of fairly uncorrelated models(trees) acting as a committee outweigh all of the actual constituent models therefore giving a more stable and accurate prediction as the final outcome. Lastly there is less chance of error in this method because the algorithm reduces the over-fitting of the decision tree by averaging the result. The following Figure 4.3 will explain the working flow of Random Forest Algorithm:

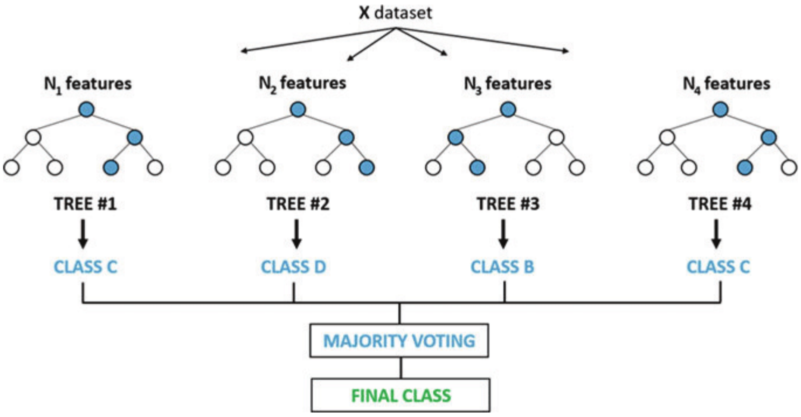


Figure 4.3: Random Forest Working Principle [11].

In Figure 4.4 accuracy for decision tree model will both criteria Entropy and Gini with depth from 1 to 20 is shown:

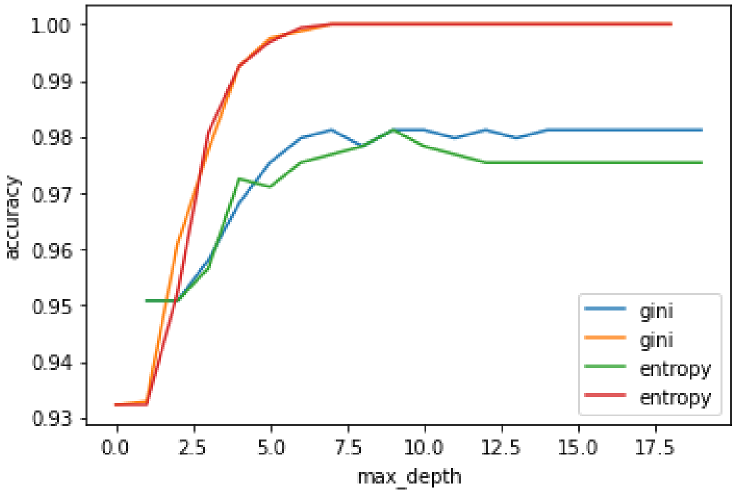


Figure 4.4: Testing and Training Accuracy for different hyperparameter for Random Forest Model.

Gradient Boost:

In Gradient Boost , base learners are generated sequentially in such a way that the present base learner is always more effective than the previous one. The overall model improves sequentially with each iteration. Only difference with other boosting techniques is that the weights for the mis-classified outputs are not incremented. Instead optimization of the loss function of the previous base learner is utilized.The Figure 4.5 shows the working principle of gradient boost model.

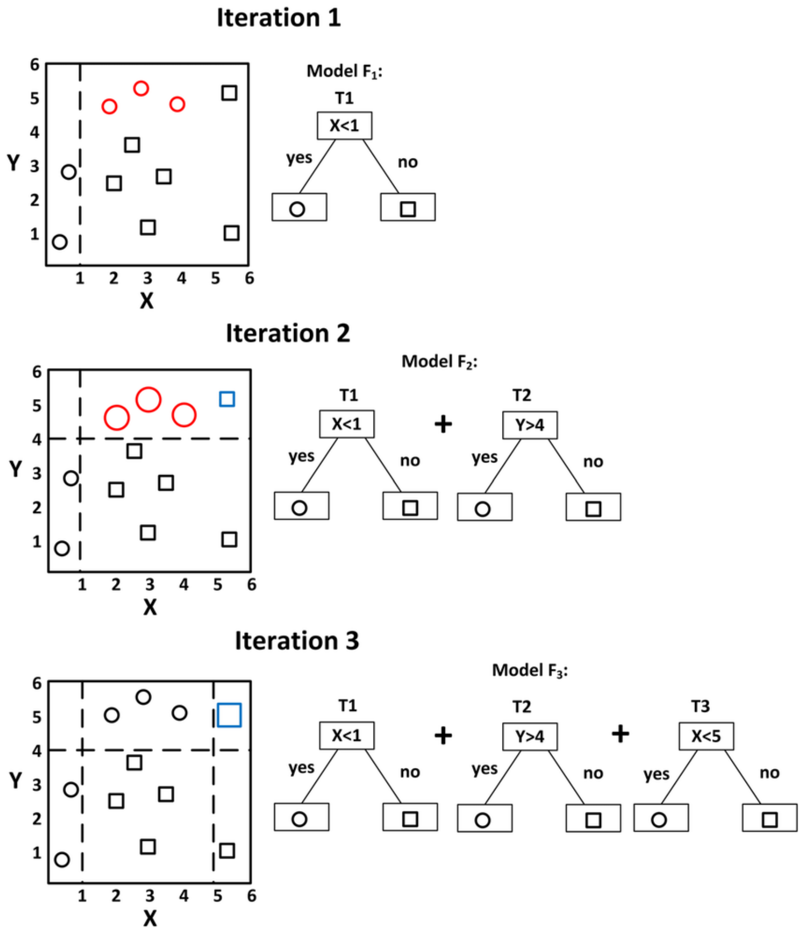


Figure 4.5: Gradient Boosting Working Principle [41].

Adaptive(Ada) Boost:

Adaptive (Ada) boost is a boosting algorithm which takes multiple weak learners and combines them to make a strong prediction rule. In ada boosting, the algorithm reads the data and assigns equal weight to each sample observation. After that it will analyze data and try to draw a decision stump. Progressing to the next stage, any false predictions are assigned to the next base learner with a higher weight on the incorrect predictions and then it will repeat the step of classifying data based on decision stumps.

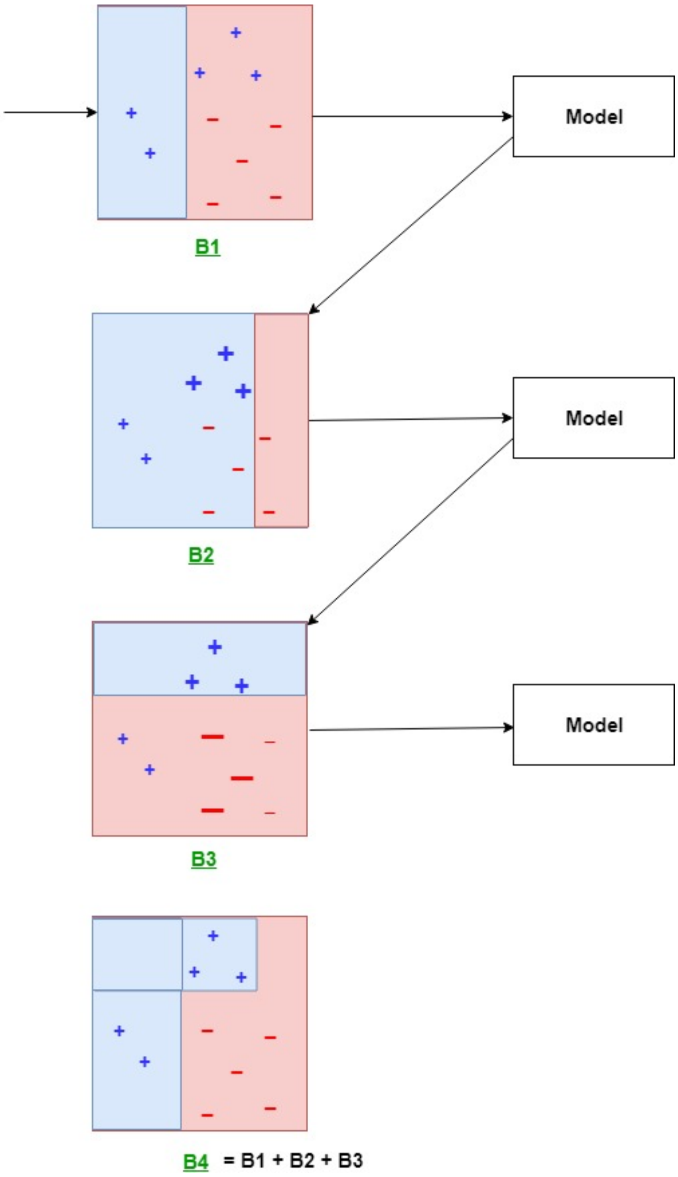


Figure 4.6: Adaptive Boosting Working Principle [14].

XG Boost:

XG Boost is an advanced version of Gradient Boosting that is designed to focus on computational speed and model efficiency. Here the parallelization process takes place , where parallel decision trees are created. Cache optimization is done in order to make the best use of hardware and software resources. Out of core computing is performed to analyze large and varied datasets and finally it supports distributed computing methods to evaluate large and complex models.The following Figure 4.7 shows the working principle for XGBoosting.

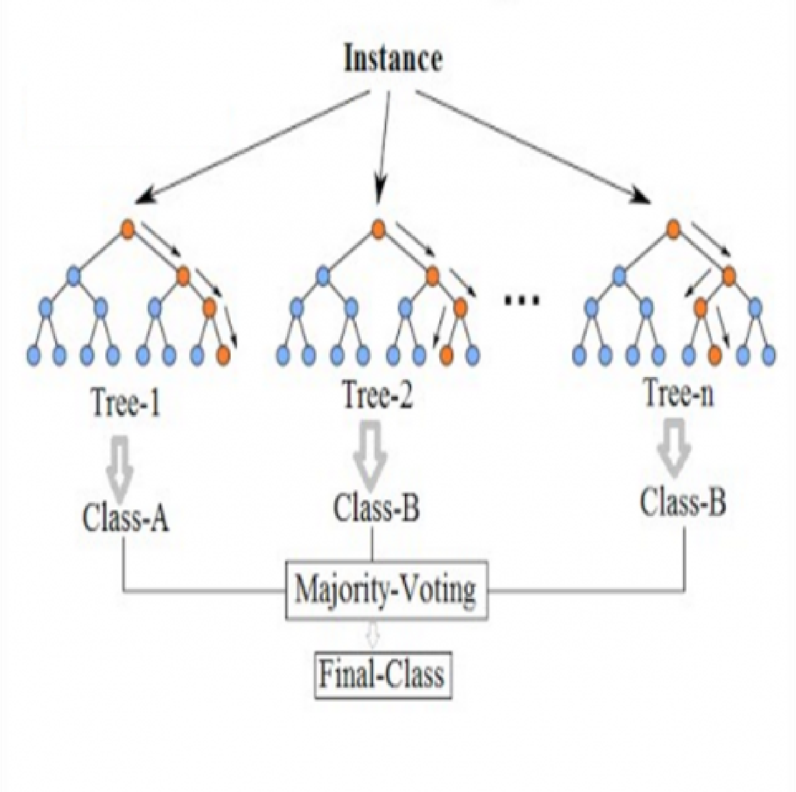


Figure 4.7: XG Boosting Working Principle [17].

K Means Clustering:

The KNN algorithm assumes similar items that occur very close to each other. It is considered to be an supervised learning algorithm which relies on labeled input to enter in order to learn a function, if new unlabeled data are provided, produces an acceptable output. In KNN an object is categorized by majority votes of its neighbour, assigning the object to the most common class of its nearest k neighbors. 'K' in this context means clustering e.g. if the value of k=1,it means the object is allocated to the nearest single neighbor's class.
The working principle of the KNN algorithm is pretty simple.Firstly, the value of k chosen in any single case in the data .Then the distance from the query example and the existing example from the data is measured by the Euclidean Distance.Other than that Manhattan Distance or Minkowski Distance can also be used to calculate the distance. Next the distance and the index of the example is added to the ordered collection and this collection is sorted by sizes, from the shortest to the highest. Then the preliminary k entries from the list of sorted collections is determined to predict the position where the query example will fit.

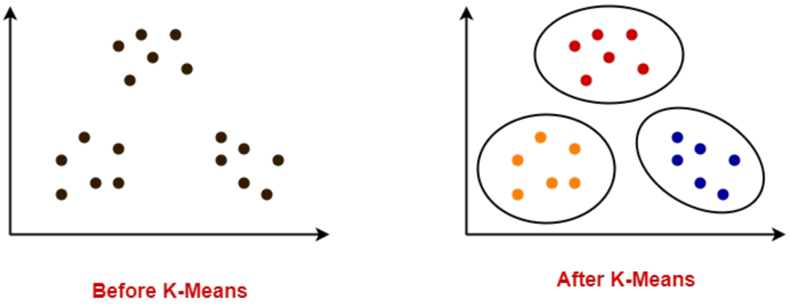


Figure 4.8: KNN Working Principle [25].

Hierarchical Clustering:

Hierarchical Clustering is often associated with heat maps where the columns represent different samples and the rows represent measurements from different targeted data. Red marking in the map represents a high expression of the targeted data. Blue markings in the map represent low expression of the targeted data. The clustering orders rows and columns based on similarity which makes it easy to see correlations in the given data. Hierarchical clustering is often accompanied by a dendogram, which indicates both similarity and the order that the clusters were formed. In Figure 4.9 it is shown how clusters are formed in hierarchical clustering

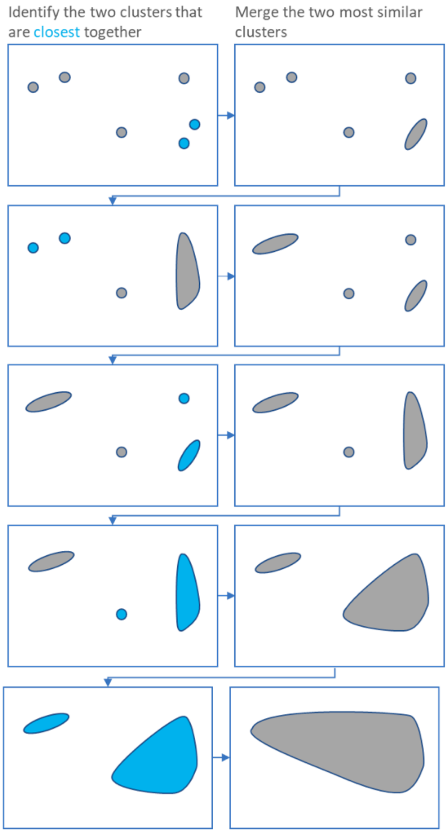

Identify the two clusters that are closest together
Merge the two most similar clusters

Figure 4.9: Hierarchical Clustering Working Principle [26].

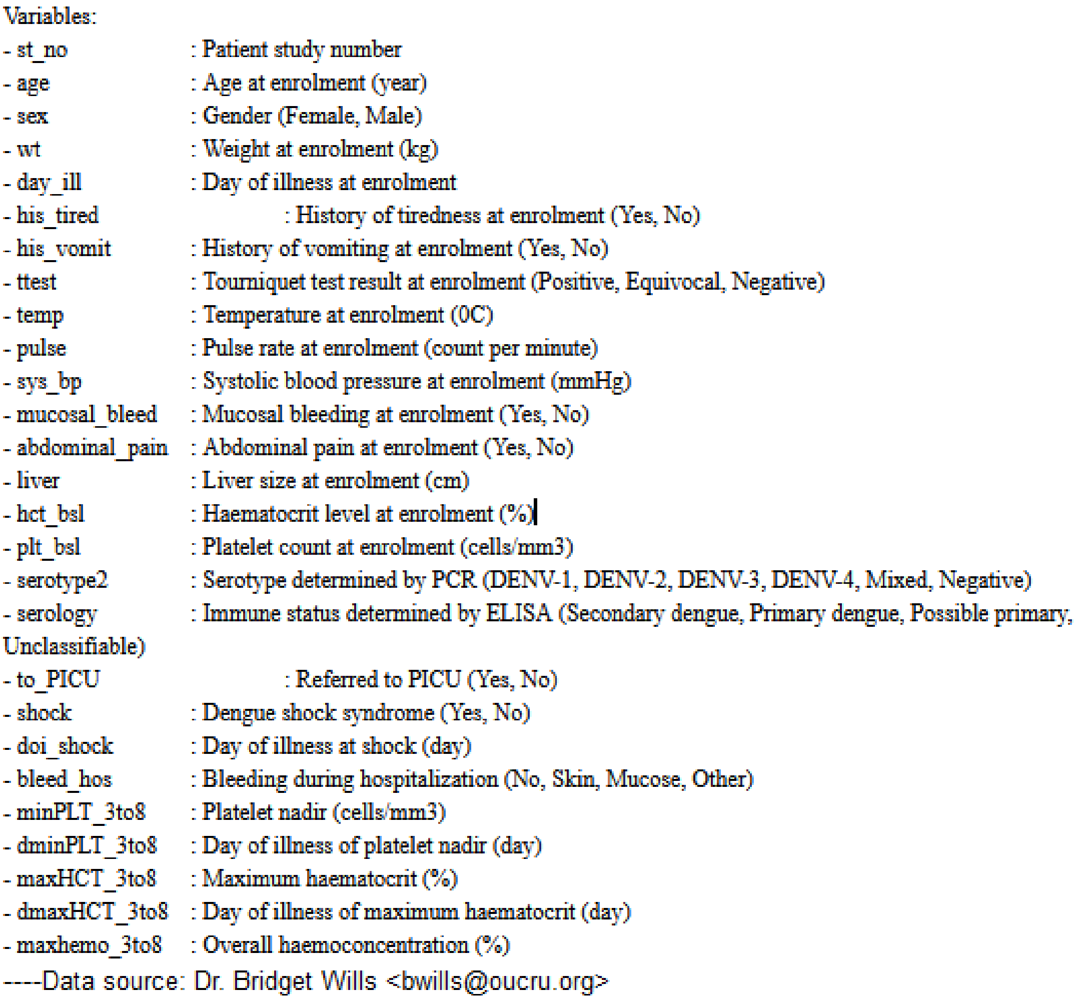
Chapter 5

Vietnam Data-set Analysis

5.1 Data Description

In this research,Vietnam dataset was collected from the journal ‘The value of daily platelet counts for predicting dengue shock syndrome: Results from a prospective observational study of 2301 Vietnamese children with dengue’ was used [10]. This dataset contain clinical observable data of 2301 children suffering with dengue in the Md Cohort. The patients were admitted to a hospital of Vietnam in the city of Ho Chi Minh between the year 2001 to 2009 [10]. Among 2301 patients 143 patients i.e almost 6.21% progressed to Dengue Shock Syndrome(DSS) and the rest 2158 patients i.e 93.79% patients did not reach DSS but they were suffering from normal dengue disease. All the patients from the data set ranged between the age 5 to 15 years old [10]. The data set contain different information like age, gender, weight temperature, pulse rate at the day of enrollment of the patients in the hospital. After observing the dataset it was seen that most of the patients were confirmed dengue between 1 to 4 days of the enrollment [10]. The dataset also contained information regarding the platelets count (cell/mm3) and hematocrit concentration(%) of the patients on the day they were admitted to the hospital. In addition to that the serology, serotype, tourniquet test results of patients suffering from dengue were also noted down. Some symptoms of dengue like abdominal pain, tiredness, vomit, mucosal bleeding were also taken into consideration for determining the severity and listed down on the dataset at the day of enrollment of the patient. For further analysis the minimum platelets count and maximum hemoconcentration of patients between day 3 to day 8 were recorded and the information were stored in the dataset.

The different features for Vietnam dataset are as follows:



Variables:

- st_no	: Patient study number
- age	: Age at enrolment (year)
- sex	: Gender (Female, Male)
- wt	: Weight at enrolment (kg)
- day_ill	: Day of illness at enrolment
- his_tired	: History of tiredness at enrolment (Yes, No)
- his_vomit	: History of vomiting at enrolment (Yes, No)
- ttest	: Tourniquet test result at enrolment (Positive, Equivocal, Negative)
- temp	: Temperature at enrolment (0C)
- pulse	: Pulse rate at enrolment (count per minute)
- sys_bp	: Systolic blood pressure at enrolment (mmHg)
- mucosal_bleed	: Mucosal bleeding at enrolment (Yes, No)
- abdominal_pain	: Abdominal pain at enrolment (Yes, No)
- liver	: Liver size at enrolment (cm)
- hct_bsl	: Haematocrit level at enrolment (%)
- plt_bsl	: Platelet count at enrolment (cells/mm3)
- serotype2	: Serotype determined by PCR (DENV-1, DENV-2, DENV-3, DENV-4, Mixed, Negative)
- serology	: Immune status determined by ELISA (Secondary dengue, Primary dengue, Possible primary, Unclassifiable)
- to_PICU	: Referred to PICU (Yes, No)
- shock	: Dengue shock syndrome (Yes, No)
- doi_shock	: Day of illness at shock (day)
- bleed_hos	: Bleeding during hospitalization (No, Skin, Mucose, Other)
- minPLT_3to8	: Platelet nadir (cells/mm3)
- dminPLT_3to8	: Day of illness of platelet nadir (day)
- maxHCT_3to8	: Maximum haematocrit (%)
- dmaxHCT_3to8	: Day of illness of maximum haematocrit (day)
- maxhemo_3to8	: Overall haemoconcentration (%)

----Data source: Dr. Bridget Wills <bwills@oucru.org>

Figure 5.1: Features of the Vietnam dataset [10]

An overview of features from the Vietnam dataset for patients suffering from dengue i.e patients both for shock and non-shock are briefly described below in the bar diagram.

5.1.1 Gender

In Figure 5.2, '0(blue)' refers to female patients and '1(orange)' refers to male patients. It is seen that among 93.79 % non shock patients, the percentage of male patients were more. Similarly it is seen that among 6.21% of patients progressing in shock, 4.2% of the patients are male.

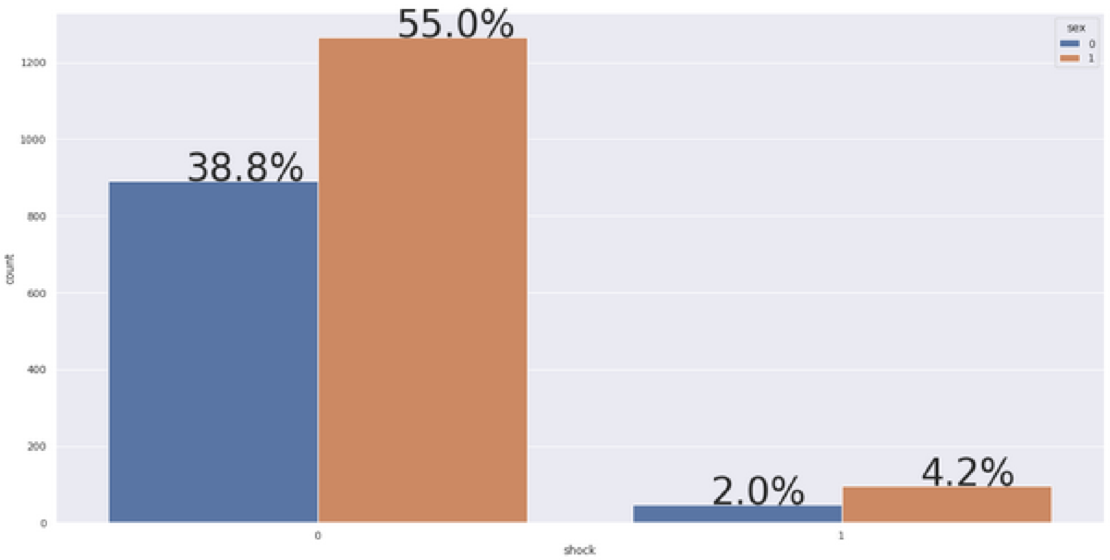


Figure 5.2: Bar Diagram showing the count of Males and Females for both shock and non-shock condition

5.1.2 History Of Tiredness At Enrolment:

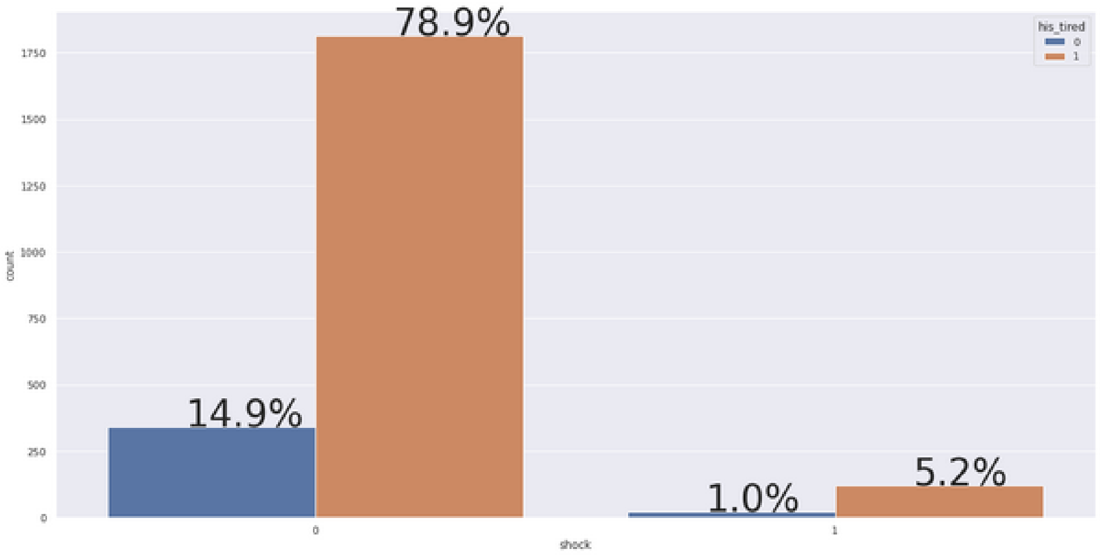


Figure 5.3: Bar Diagram showing the count of patients showing tiredness at the day of enrollment for both shock and non-shock condition

In the Figure 5.3, '0(blue)' refers to patients not showing tiredness and '1(orange)' refers to patients showing tiredness at the day of enrollment in the hospital. Among 93.79% patients with no shock i.e normal dengue patients, 78.9% patients showed tiredness. Similarly, among 6.21% patients with shock 5.2% showed the sign of tiredness. So tiredness was visible among patients irrespective of going into shock or not. It is also understood most of the patients in our dataset tend to show the symptoms of tiredness on the day of enrollment.

5.1.3 History Of Vomit At Enrollment:

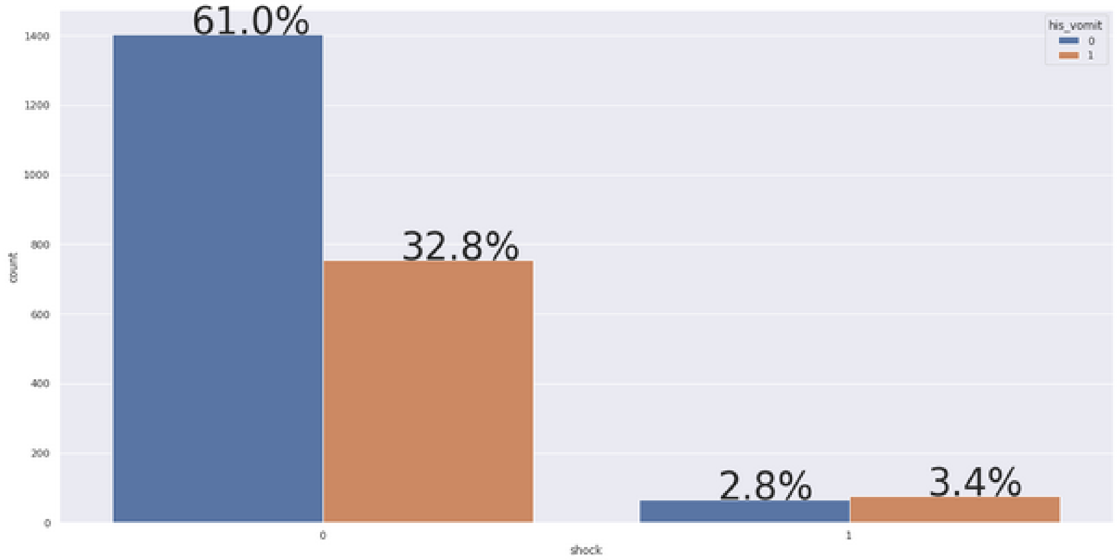


Figure 5.4: Bar Diagram showing the count of patients vomiting at the day of enrollment for both shock and non-shock condition

In the Figure 5.4, ’0(blue)’ refers to patients not vomiting and ’1(orange)’ refers to patients with vomiting tendencies at the day of enrollment. From our dataset it is seen that the percentage of patients not going to shock have less tendency to vomit. But higher percentages of patients progressing to shock showed higher tendencies of vomiting.

5.1.4 Abdominal Pain At Enrolment:

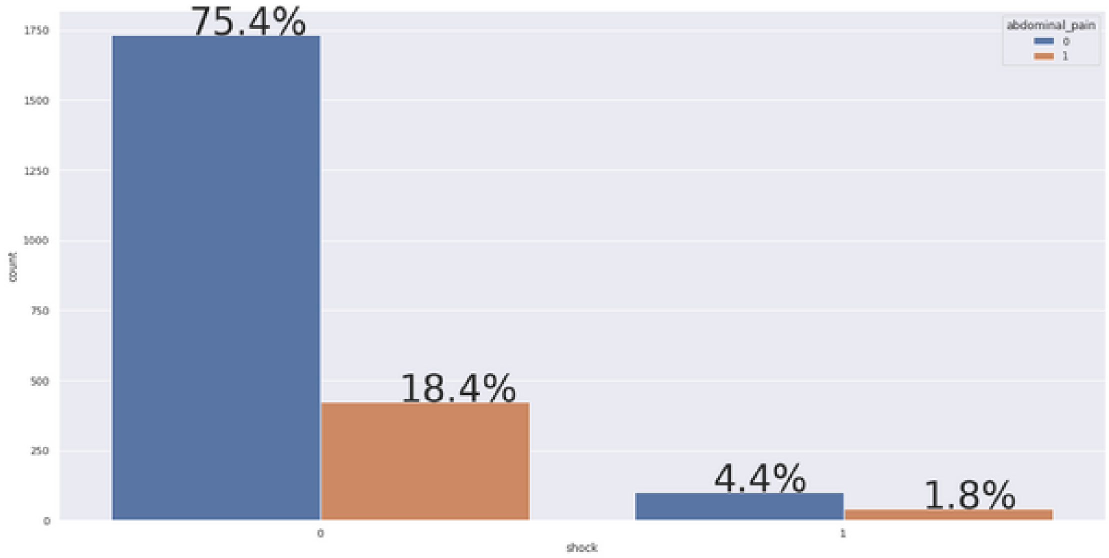


Figure 5.5: Bar Diagram showing the count of patients having abdominal pain at the day of enrollment for both shock and non-shock condition

In the Figure 5.5, '0(blue)' color refers to patients not having abdominal pain and '1(orange)' refers to patients with abdominal pain at the day of enrollment. Both patients with shock and no shock did not experience abdominal pain. So considering our dataset patients with dengue has a low chance of abdominal pain.

5.1.5 Mucosal Bleed At Enrolment:

In the Figure 5.6, '0(blue)' refers to the patients showing no mucosal bleed and '1(orange)' refers to the patients showing mucosal bleed at the day of enrollment. Both patients with shock (87.4%) and no shock (5.7%) have less percentage of showing no mucosal bleed at the day of enrollment.

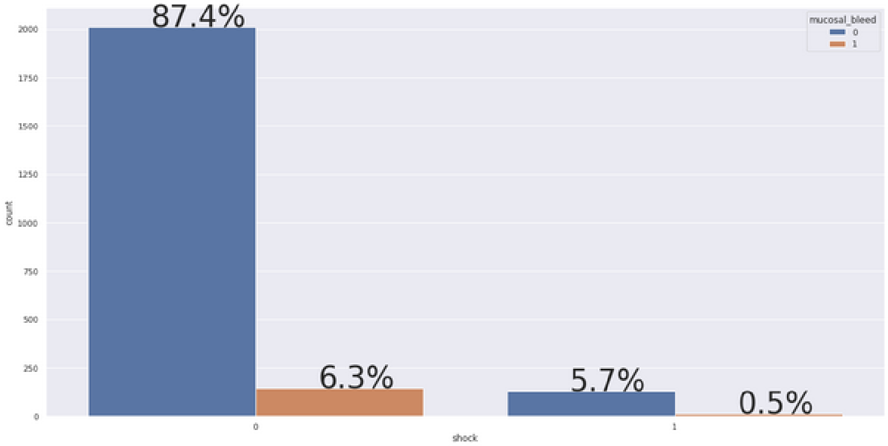


Figure 5.6: Bar Diagram showing the count of patients' mucosal bleed at the day of enrollment for both shock and non-shock condition

5.1.6 Bleeding During Hospitalization:

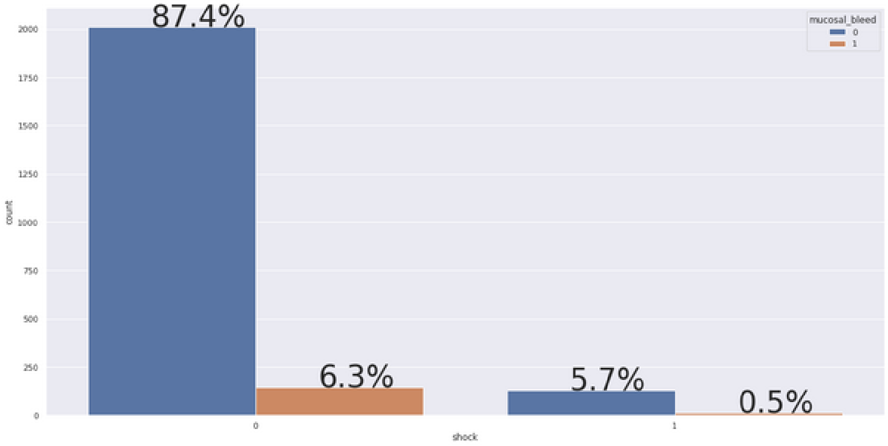


Figure 5.7: Bar Diagram showing the count of patients showing bleeding symptoms at the day of enrollment for both shock and non-shock condition.

In the Figure 5.7, 0 refers to mucosal bleed, 1 refers to no bleeding at all, 2 refers to bleeding through other organs except skin or mucus and lastly 3 refers to bleeding through skin. From our dataset patients with no shock shows mucosal bleeding more during the day of enrolment to the hospital. But Patients with shock shows bleeding through only skin at the day of enrolment.

5.1.7 Tourniquet Test At Enrolment:

The Tourniquet Test is part of the current WHO Dengue case definition. The tourniquet test indicator is a marker of capillary fragility which can be used as a triage technique to separate acute gastroenteritis patients from dengue patients. The Tourniquet test gives positive outcomes where there is more than one petechiae per square inch [33]. Petechiae is formed when tiny blood vessels called capillaries break open. When these blood vessels rupture, the blood would disperse to the tissues. In the case of IgM and NS1 ELISA kits, the outcome was negative if less than 9,equivocal between 9 and 11 and positive if greater than 11. If both the dengue and the JE, IgM results were positive, the JE result will be separated by the Dengue result [3].

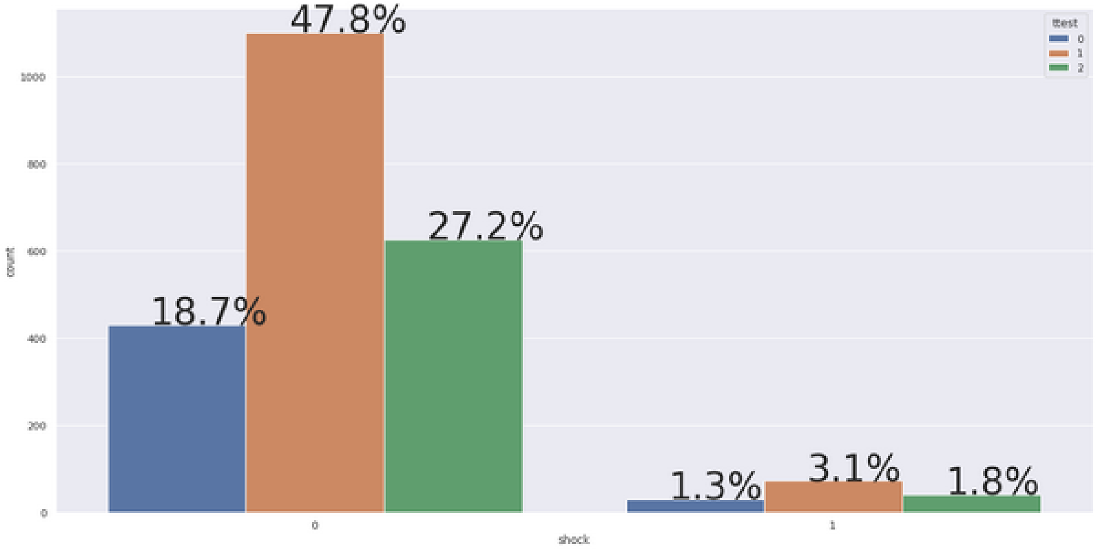


Figure 5.8: Bar Diagram showing the percentage count of patients showing tourniquet test results at the day of enrollment for both shock and non-shock condition.

In our data from Figure 5.8, 0(blue) color refers to Equivocal, 1(orange) refers to negative and 2(green) refers to the positive results for the tourniquet test. If the percentages from the above bar diagram is observed then it it seen that greater percentages of patients going in shock(3.1%) and not going in shock(47.8%) tested negative on the day of enrolment.

5.1.8 Serology:

A glycoprotein NS1 is commonly prevalent in every serotype and can help to detect whether the patient is in primary or secondary stages of dengue. Moreover IgM and

IgG tests can be helpful for primary and secondary detection.

In the Figure 5.9, 0(blue) color refers to patients who are probably suffering from dengue virus, 1(orange) refers to Primary Dengue, 2(green) refers to Secondary Dengue and 4(red) refers to patients whose serotype was unclassified.

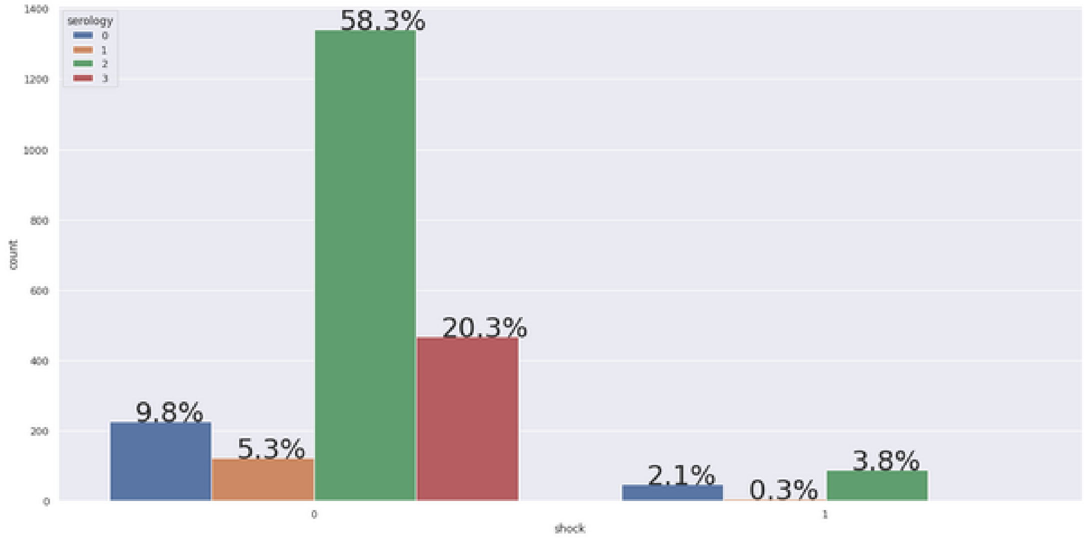


Figure 5.9: Bar Diagram showing the percentage count of patients suffering from different serology of dengue virus at the day of enrollment for both shock and non-shock condition.

If we look carefully at the bar diagram 5.9, it is seen that more percentages of patients with both shock and no shock have developed Secondary dengue which refers that these patients had suffered from dengue in the past. Patients with secondary dengue mostly show thrombocytopenia, bleeding and abdominal pain.

5.2 Missing Data Imputation

In our dataset there contains 2301 rows but there were some missing values. So the dataset require a bit of preprocessing for handing the missing values. For a particular dataset identifying the missing values and transforming them into a numerical form is known as missing data imputation. We applied MICE Algorithm and KNN Imputer for missing value imputation. It was observed that our dataset fitted more efficiently using the KNN imputer model and thus KNN was finally used to handle the missing values.

The KNN Imputer model is generally a regressive model and for predicting missing values. Input variables are required to be numerical. But in our datasets among 24 columns, 11 of the columns contain categorical values. So those specific categorical columns were converted to numerical values using Label Encoder. Label Encoder is a library of Scikit-learn which is used to convert the categorical values into numerical values. When all the values of each columns were converted into numbers, KNN Imputer was used to fill up the missing values using the K-Nearest algorithm. This imputation method works by searching the whole dataset to find the similar instances in order to fill up the missing data. The K-Nearest Neighbours (KNN) identifies the neighboring points in the dataset by calculating the distance. Then the missing values can be measured using completed values of neighboring observations [21]. The distance is usually measured by Eucledian's distance formula. In case of missing values, the Euclidean distance is determined by ignoring the missing values and scaling the weight of the missing coordinates[21].

$$dxy = sqrt(weight * squareddistancefrompresentcoordinates) \tag{5.1}$$

where, weight=Total Number Of Coordinates / Number Of Present Coordinates[21].

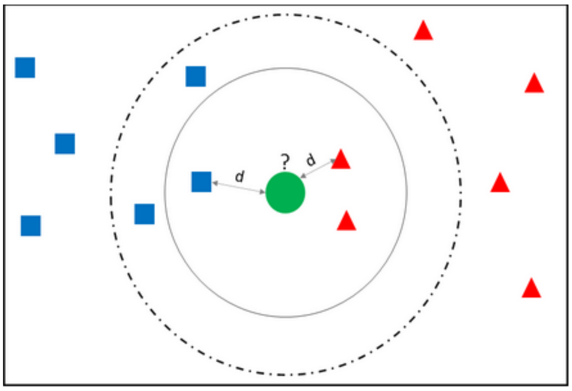


Figure 5.10: Working principle of KNN Imputation for determining missing values.[36]

Thus for determining the missing values we used KNN Imputer model. The missing values were filled by determining the Euclidean' distance of the closest neighbouring points.

5.3 Proposed Model

The primary aim of our research was to analyse the severity of dengue patients and to analyse the dataset to find the predominant reason for progressing to DSS. We applied Decision Tree Algorithm, Random Forest, Gradient Boosting, Ada Boosting and XG Boosting on our dataset and these models were trained and tested to predict patients who will either progress to shock or will not proceed to shock.

Classification models like Decision Tree and Random Forest were fitted on our dataset applying both the criteria 'Gini' and 'Entropy' with maximum depth ranging from 1 to 20. The data was fitted for each depth with both the criterion and tested to find whether the model can predict shock accurately. Furthermore Boosting algorithm such as Ada Boosting, XG-Boosting and Gradient Boosting was used. The hyper-parameters such as learning rate(0.05, 0.075, 0.1, 0.25, 0.5, 0.75) and maximum depth of the following algorithms were tuned to achieve maximum optimisation. Classification report containing different metrics like sensitivity, specificity, misclassification, precision, f1_Score, PPV, NPV were further analysed to see which model is fitting the best with the data sets to predict severity among dengue infected patients.

Each of the terms related to classification report are briefly described below:

- **Sensitivity**: It is the measure of proportion of actual positive values that are predicted positive [9].

$$sensitivity = TP/(TP + FN) \tag{5.2}$$

- **Specificity**: It is the measure of proportion of actual negative values that are predicted positive [9].

$$specificity = TN/(TN + FP) \tag{5.3}$$

- **Precision**: It refers to the proportion of actual positive identifications that are actually correct [32].

$$precision = TP/(TP + FP) \tag{5.4}$$

- **Recall**: It refers to the portion of positive values that were calculated correctly [32].

$$recall = TP/(TP + FN) \tag{5.5}$$

- **f1_Score**: It is the harmonic mean of precision and recall [15].

$$f1\_Score = 2 * ((Precision * Recall)/(Precision + Recall)) \tag{5.6}$$

- **PPV**: It stands for the Positive Predictive Values. It refers to the probability that the person with a positive result is likely to have the disease or have a higher chance to develop the disease [39].

$$PPV = NumberofTP/(NumberofTP + NumberofFP) \tag{5.7}$$

- **NPV**: It stands for Negative Predictive Values. It refers to the probability that if a patient is tested negative he/she will genuinely not have that specific disease [22].

$$NPV = NumberofTN/(NumberofTN + NumberofFN) \tag{5.8}$$

5.4 Feature Engineering

After fitting our datasets with the XG-Boost Classifier model the most important features are determined. We have used a summary plot here. For all the features and all samples in the selected range, this type of plot aggregates SHAP values. SHAP values are sorted then, so the most important feature is the first one seen. Furthermore, we provide details about how each feature affects the performance of the model. The important features were selected so that it can help us during our analysis for finding the severity among patients whether they will reach to DSS or not.

The Figure 5.11 plot uses the SHAP values obtained from our XG-Boost Classifier model which combines important features with feature effects. The horizontal axis contains the SHAP values for our predicted output i.e shock. The positive values along the right side of the horizontal axis refer to shock positive(1) and negative values on left side refer to shock negative(0). Along the vertical axis the position is determined by the features from our dataset according to the most important features being on the top and least important features at the bottom. Threshold color Red defines high, deep blue defines medium and blue is low value.

- **to_PICU**: When to_PICU high i.e the patients have reached to pediatric unit then the shock syndrome is positive(1). When to_PICU is low, shock value is negative(0).

- **minPLT_3to8**: When minimum platelets count of 3 to 8 days from enrollment is low or blue value then shock value is positive(1), when minimum platelets count is mid to high range(deep blue to red) shock value is negative(0)

- **maxhemo_3to8**: When maximum hemoconcentration value of 3 to 8 days from enrollment is high or red then shock is positive(1), and maximum hemoconcentration value is mid to low or deep blue or light blue color then shock value is negative(0)

- **serology**: When serology is high the shock is negative(0) and when it's low shock is positive(1)

- **plt_bsl**: When platelets count at the day of enrollment is high the shock is negative(0) and when it's low shock is positive(1)

- **pulse** : When the pulse rate of patients in our data sets were mid value they did not reach to shock i.e shock negative(0), but pulse rate with low or high value is shock positive(1).

- **serotype_2**: High to mid value is shock negative(0), low value is shock positive(1).

- **his_vomit**: When the patients did not show any sign of vomit on the day of enrollment then it appears patients are shock negative(0), high i.e show the tendency if vomit appear to be shock positive(1).

- **bleed_hos** : When the patients at the day of enrollment did not show any symptoms of bleeding have less probability for not reaching in shock,but patients with bleeding symptoms appear to shock.

- **maxHCT_3to8**: When maximum hematocrit count of of patients for 3 to 8 days from enrollment is high or red then shock is positive(1), and maximum hematocrit value is mid to low or deep blue or light blue color then shock value is negative(0)

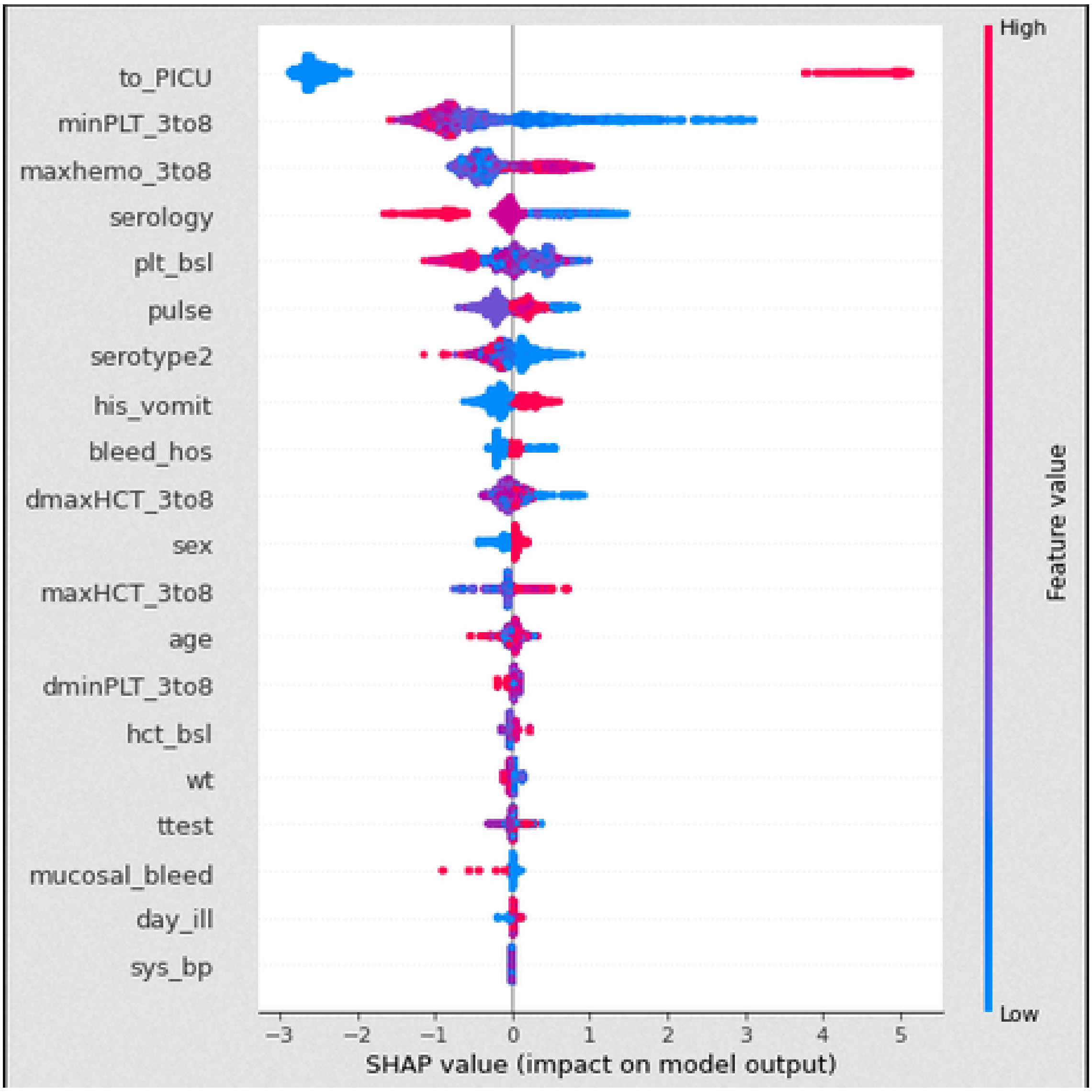


Figure 5.11: Summary plot of different features of the data sets using SHAP values.

The important features of data sets are shown as well with the help of bar diagram where the top most feature has the highest importance and bottom one has the lowest importance. After finding the percentages of important features it is seen that to_PICU column is the most dominating feature as its percentage is the highest. From this we can say that patients who have reached to shock were sent to PICU or patients admitting in the PICU gradually reached to shock.

The bar diagram showing important features are shown in Figure 5.12:

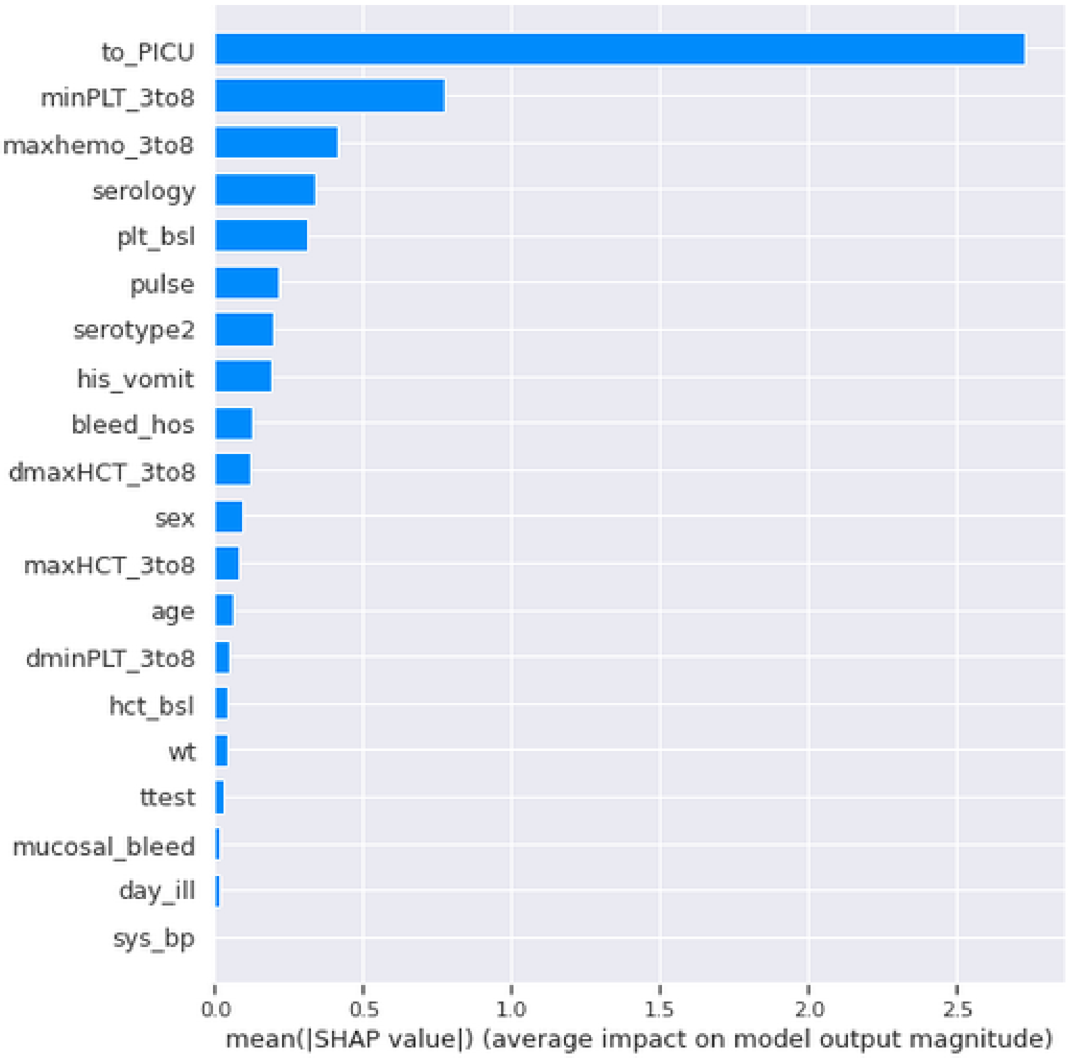


Figure 5.12: Bar diagram showing importance of different features for Vietnam dataset

The correlation of different features for Vietnam dataset 5.13:

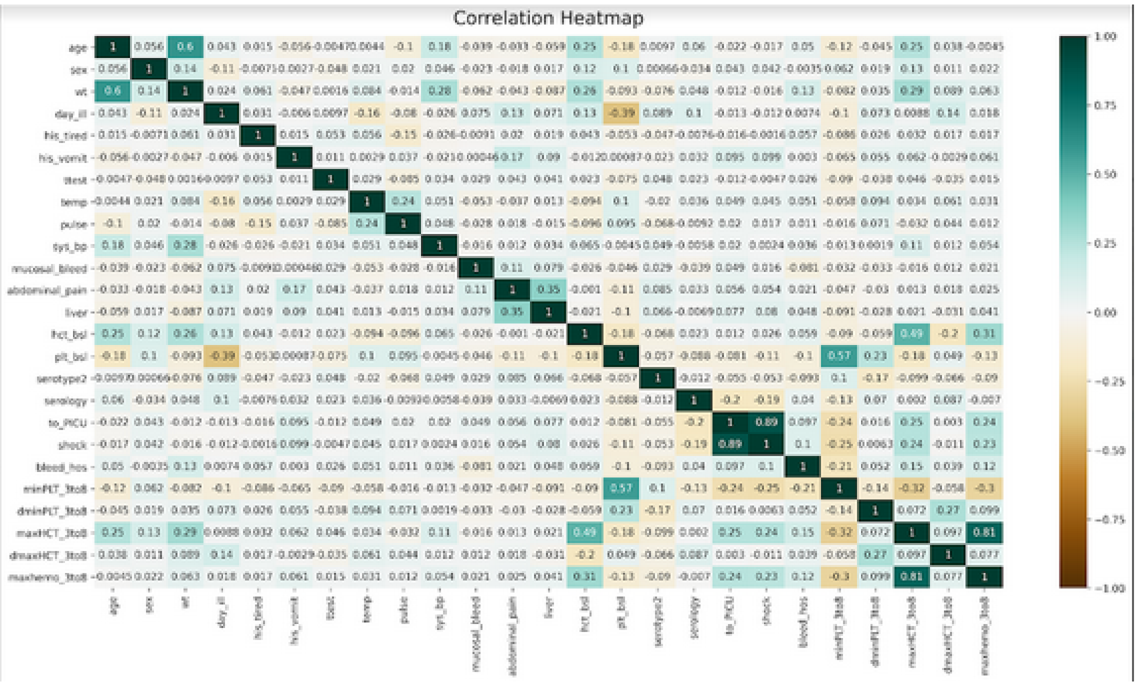


Figure 5.13: The correlation of different features for Vietnam dataset

A heatmap is used to show the correlation among the selected features from the data sets. Heatmap helps in visualization of the correlation matrix as the highly concentrated values are given one color and the less concentrated value is given another color. The correlated values of the two interrelated features ranges between -1 to +1. The lowest value -1 means that the features are weakly related and +1 means that the two features have a strong relationship.

Chapter 6

Results And Discussion On Vietnam Datasets

Among 2301 patients from the Vietnam dataset 2158 patients did not reach to shock and 143 patients reached to shock. Considering the Vietnamese dataset 93.79% of patients did not reach shock and 6.21% reached to shock. As to_PICU is the most important feature so SHAP dependence plot is drawn between to_PICU and other important features from our datasets. SHAP dependence plot illustrates the marginal effect on the expected result between one or two variables. It indicates whether the relationship is linear, monotonic, or more complex between the target and the variable [19].

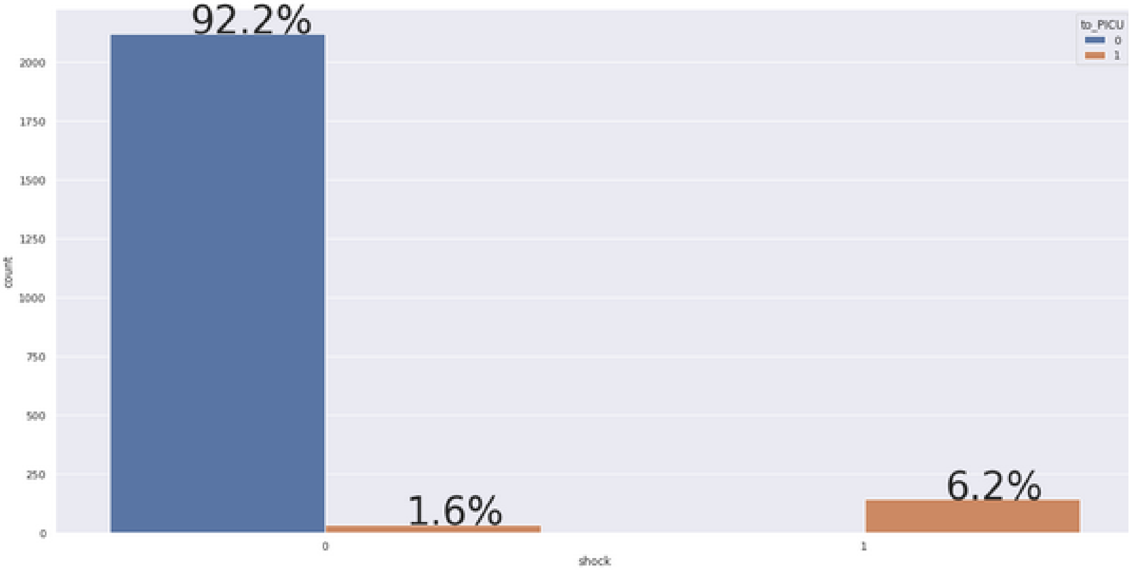


Figure 6.1: Bar diagram showing percentages of patients in pediatric unit and also percentages of patients not sent to pediatric unit

'0(BLUE)' refers to patients not sent in the PICU and '1(ORANGE)' refers to patients who were send to_PICU. It is seen among 6.21% percent with shock, all the patients with severe dengue were sent to the Pediatric Intensive Care Unit(PICU).

Table 6.1: CLASSIFICATION REPORT AFTER APPLYING DIFFERENT MACHINE LEARNING APPROACH ON THE DATA-SET

Name Of the Algo	**Training Accuracy**	**Test Accuracy**	**Sensitivity**	**Specificity**	**Misclassification**	**Precision**	**f1_Score**	**PPV**	**NPV**
AdaBoosting Learning Rate=1	0.998	0.981	0.94	0.98	0.02	0.98	0.96	0.74	1
XG Boosting Learning Rate=1 Depth=9	1	0.986	0.94	0.99	0.01	0.98	0.96	0.8	1
Random Forest Criterion=Gini Depth=8	1	$0\dot{9}78$	$0\dot{8}2$	$0\dot{9}9$	$0\dot{0}2$	$0\dot{9}9$	$0\dot{9}$	$0\dot{7}6$	$0\dot{9}9$
Random Forest Criterion= Entropy Depth=10	1	0.979	0.94	0.98	0.02	0.98	0.96	0.73	1
Decision Tree Classifier Criterion= Entropy Depth=8	1	0.982	0.94	0.98	0.02	0.98	0.96	0.76	1

Classification Report is shown in Table 6.1
We ran the predictive algorithm as shown in Table 6.1 on all the variants of our dataset several times. The whole datasets were divided into two portions - a)Training Sets and b)Testing Sets. We randomly split the datasets and each time 70% of the total datasets were selected and fitted to the model for training and 30% data was selected for testing whether the model can accurately predict the shock in dengue patients or not. After fitting the datasets with the mentioned predictive algorithms randomly rows from the testing datasets were selected to see the predictive output. These outputs were matched with the original data to check the correctness of the model. It was found that XG-Boosting Classifier model could give the highest accurate result. Apart from that other metrics like sensitivity ,specificity, PPV, NPV, precision, f1_Score were measured to check the best model for our datasets. From the Table 6.1 it is seen XG-Boosting gave the highest testing accuracy i.e 98.6%. On top of that, the model gave 94% specificity which means that the model could predict 94% patients merged in shock and missed 6% of the patients with shock. Furthermore, the model gave 99% specificity which refers that the model could identify perfectly 99% of the patients who did not reach to the shock level. As we were dealing with medical data so we took two more metrics into consideration the PPV value and NPV value. The PPV result for XG Boosting model was 80% which means that if dengue patients with shock get tested, there is 80% chance that the patient will actually reach to DSS. Similarly the value of NPV is 100% which refers that if a

patient did not reach to shock then our model can identify perfectly those patients with no shock. The XG-boosting model gave only 1% misclassification which was less compared to the other models we also have used. Lastly the training and testing accuracy were compared to find out whether the model was over-fitting the datasets or not and it was found not to overfit.

For fitting our datasets we set the following conditions to the model as it gave the highest accuracy with those specific parameters. The parameters are objective = binary logistic, colsample_bytree = 0.3, learning rate = 1, max_depth=9, alpha = 10, n_estimators = 10. Though the XG-Boost Classifier model still gave the highest accuracy, we further analysed to see whether the model was actually able to predict the output properly. Thus for retrieving the performance of the model on the evaluation datasets we plotted the log loss of the XG-Boost Classifier model to get a clear insight about the model being working properly. We provide an array of test sets and training sets to the eval_matrix argument which will provide us with an insight of how well the model is performing while training the datasets. The performance for each evaluation sets for the model can be seen by eval_result() function. The size of the epochs are taken equal to the length of the evaluation sets. The Figure6.2 is the log loss which indicates the model behavior on the train and test datasets over the training epochs. In real sense the log loss for training sets will be always less than the testing sets. But when the generalization gap is small between the training and testing log loss curve then it can be understood that the model has a good fit.

The function for log loss is as follows:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i)\cdot log(1-p(y_i))$$

Binary Cross-Entropy / Log Loss

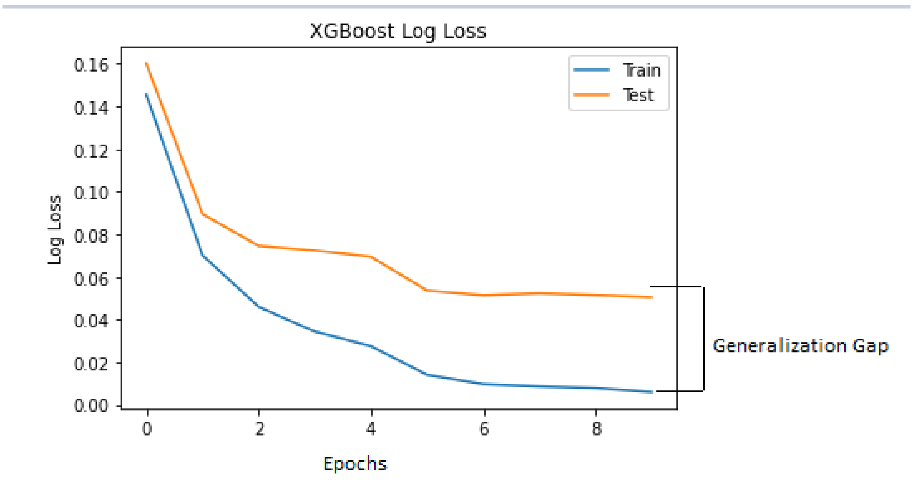


Figure 6.2: XGBoost Log loss curve.

Table 6.2: CONFUSION MATRIX FOR XGBOOSTING CLASSIFIER

TP	32
TN	649
FP	8
FN	2

Classification Error Curve was also plotted for both the train and test to visualize the misclassification among data points. It was seen in the last epoch the number of misclassified samples for train was almost tends to 0 and for test it is approximately 0.015 or 1.5%.

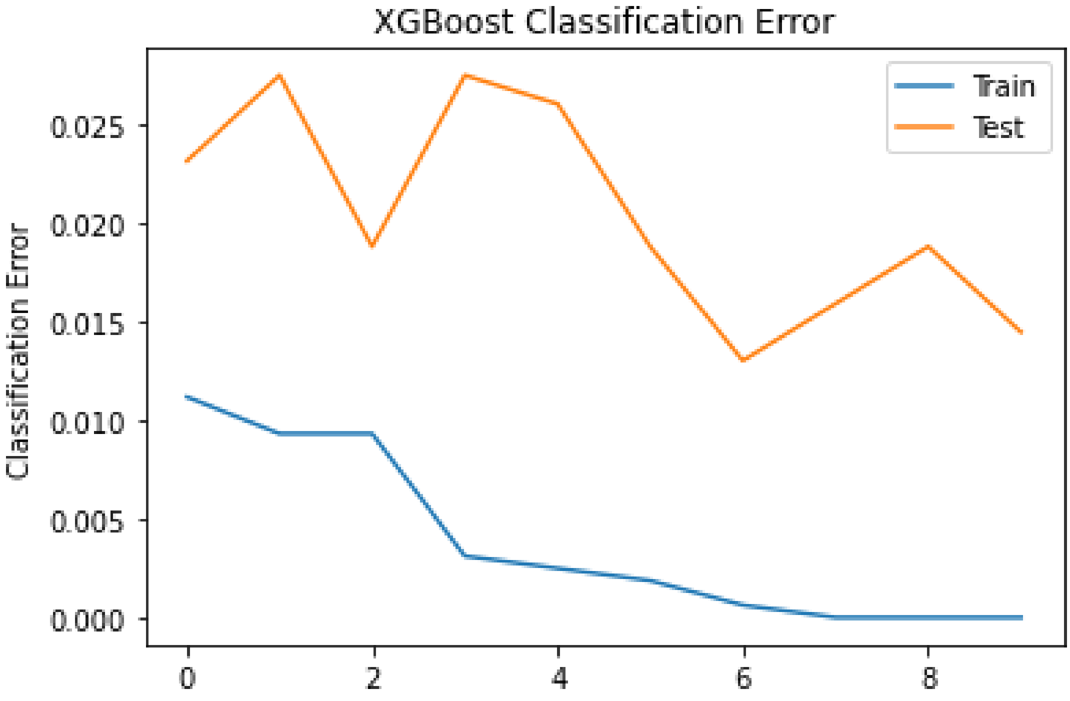


Figure 6.3: XGBoost Classification Error curve.

We also plotted the ROC and AUC curve to further assess the performance of our model. ROC curve contains the total proportion of accurately classified observations [8]. AUC curve summarizes the performance and gives a metric which lies between 0 to 1.If the value tends to 1 then it is a good classier model.From the Figure 6.4 we see for our XG-Boost model it is 0.993.

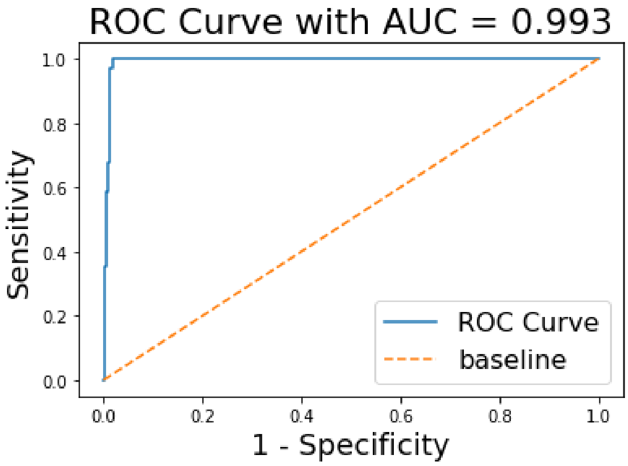


Figure 6.4: ROC and AUC Curve.

The features that are being used in our proposed model are-'age', 'sex', 'wt', 'day_ill', 'his_tired', 'his_vomit', 'ttest', 'temp', 'pulse', 'sys_bp', 'mucosal_bleed', 'abdominal_pain', 'liver', 'hct_bsl', 'plt_bsl', 'serotype2', 'serology', 'to_PICU', 'shock', 'bleed_hos', 'minPLT_3to8', 'dminPLT_3to8', 'maxHCT_3to8', 'dmaxHCT_3to8', 'maxhemo_3to8'. Among all these features to_PICU, 'maxHCT_3to8', 'minPLT_3to8', 'maxhemo_3to8', 'serotype2' were considered the most important features and the analysis was done by taking all these aforementioned features into consideration .

- **TO_PICU AND MAXHEMO_3TO8 AND MAXHCT_3TO8** :

  Haemoconcentration basically refers to the decrease in the plasma volume which indicates the simultaneous increase in the concentration of the red blood cells. Moreover, an increase of haemoconcentration may lead to the increase in the blood viscosity.Severe dengue fever may lead to the Dengue Haemorrhagic Fever(DHF) and Dengue Shock Syndrome (DSS) which is predominantly characterized by Plasma Leakage and manifestation of uplifted hematocrit concentration which later may leads to Hypoproteinemia.

  The hematocrit concentration(%) of children at the day of enrollment is noted down and stored in the'hct_bsl' column of our datasets. In addition, between the third and eighth days, a daily hematocrit concentration count of children suffering from dengue is analyzed and the maximum amount is noted and stored in our datasets in the column 'maxHCT'3to8'. Similarly, overall hemoconcentration(%) of the patients are recorded and listed to the column named 'maxhemo_3to8'in our datasets. And as mentioned earlier 'to_PICU' column contains the records of the patients who were sent to PICU.

  If we consider Figure 6.5 (**Figure:A**) horizontal axis, 'maxHCT_3to8' is plotted which indicates the actual value from the dataset, and the vertical axis indicates value that has an impact on the prediction ie. on the severity of the dengue fever whether the patient will go in shock or not. The SHAP value above 0 indicates the prediction result is positive i.e the patients has gone to shock and the value below 0 indicates the patient still has not reached the shock level or developed Dengue Shock Syndrome(DSS). If we look clearly the circled portion in Figure A then it is seen patients having hematocrit concentration count above 45% are most likely to go in shock and develop DSS. Moreover most of the patients from those circled region were sent to PICU as well as red dots represents those patients who have been admitted to PICU.

If Figure 6.5 (**Figure:B**) is closely observed then a scatter plot between 'maxhemo_3to8' and 'to_PICU' is plotted. Like the above mentioned graph the vertical axis indicates that impact of maximum hemoconcentration on determining whether the patient will be in shock or not. If the circled portion region is detected carefully the it is observed that patient having hemoconcentration greater than 20% has the most probability of going in shock in spite of he has been sent to PICU or not. This increase in hemoconcentration may lead to plasma leakage and the patient may suffer from Dengue Haemorrhagic Fever (DHF) or in worse condition can develop Dengue Shock Syndrome(DSS).

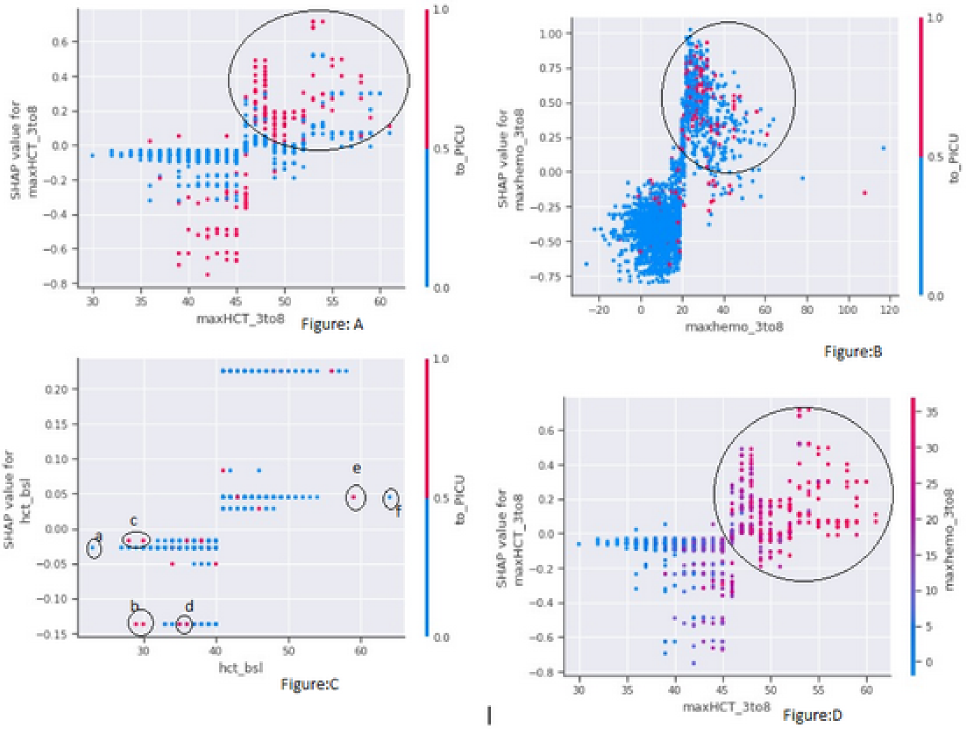


Figure 6.5: SHAP dependence plot of 'maxHCT_3To8', 'maxhemo_3To8', 'hct_bsl','to_PICU'

Figure 6.5 (**Figure:C**) contains a SHAP dependence scatter plot between 'hct_bsl' and 'to_PICU'. Hematocrit(HCT) testing is used to measure the degree of leakage of plasma. Usually hematocrit concentration is monitored regularly to observe the hemorrhagic manifestation associated with hemoconcentration. HCT concentration plays a crucial part in determining whether the patient is proceeding towards shock or is just having a normal dengue fever. Thus for proper prediction of DHF, HCT is monitored every 24 hrs and for DSS HCT level is monitored every 3 to 4 hours(2). Though Figure C is not giving a very clear idea but close monitoring of hematocrit level at the day of enrollment some vital decisions regarding patients health conditions can be taken. If we see the circled region marked ‘a’, then the patient having hematocrit level less than 20% have less tendency to go to shock. But above 20% in marked region ‘b’,’c’,’d’any patient between 25% to 45% either sent to_PICU or not has little scope of proceeding towards shock. On the other hand if circled region marked ‘e’ and ‘f’ are considered then it is seen if HCT level is above 45% then for both scenario whether patients sent to sent to_PICU or not will reach towards Dengue Shock Syndrome.

Figure 6.5 (**Figure:D**) shows a relation between the maximum hematocrit concentration for 3 to 8 days and hemoconcentration. The circle portion indicates if hematocrit concentration is more than 45% and the overall hemoconcentration is above 20% then the patients is going to suffer from a hazardous health issues which may heads to either DHF or DSS.

- **TO_PICU AND MINPLT_3TO8 AND PLT_BSL** :

  The platelets count(cells/mm3) of children at the day of enrollment is noted down and stored in the‘plt_bsl’ column of our datasets and, between the third and eighth days, daily platelets count of children suffering from dengue is analyzed and the minimum amount is noted and stored in our datasets in the column ’minPLT_’3to8’.

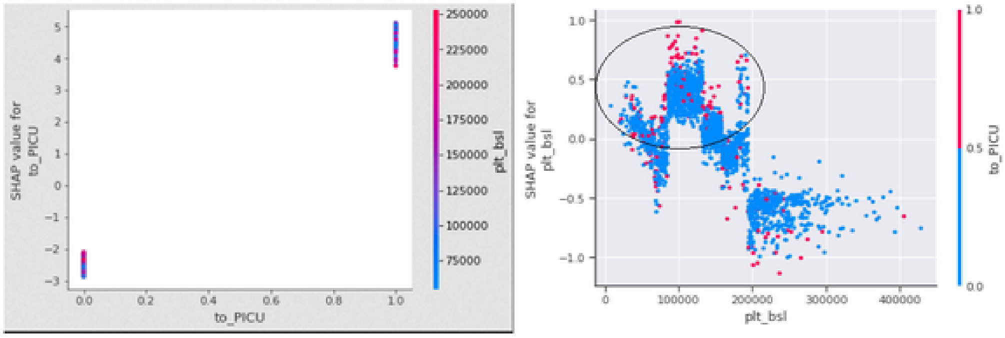


FIG:A

FIG:B

Figure 6.6: SHAP dependence plot of’plt_bsl’and’to_PICU’

If the Figure 6.6 for **Figure A** and **Figure B** is observed then in the first figure along the horizontal axis ‘to_PICU’ is plotted and in the second one in the horizontal axis ‘plt_bsl’ is plotted and on the vertical axis of both the figure SHAP values are given which will tell us how the specific feature on X axis affects the prediction column.

From **Figure A**(Figure 6.6) it is slightly understood that when the platelets are low i.e. in the blue region then the patient may enter into shock. But when we closely look into**Figure B**(Figure 6.6), the plot is scattered too much and it is hard to analyze the graph to decide which patients will be in shock and which patient will be not in shock. Moreover, if we see the circled region in the second picture at top left, in that region from 25000(cell/mm3) to nearly 200000(cells/mm3) the dots are having the positive SHAP values which indicates the patient has a chance of going to dengue shock if he/she belong from that group. But by seeing the two graph and examining the platelets counts at the enrollment day we couldn't come into any conclusion. That's why in the next we have taken the minimum platelets count between 3 to 8 days into consideration for finding the severity among dengue patients.This will give a idea of which patients will develop DSS, DHF.

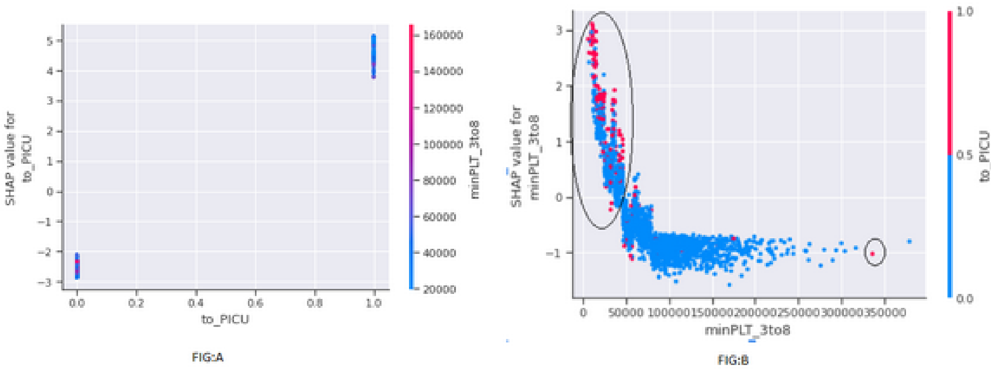


Figure 6.7: SHAP dependence plot of'minPLT_3To8'and'to_PICU'

In the Figure 6.7,(**Figure:A**) horizontal axis 'to_PICU' is plotted which indicates the actual value from the dataset, and the vertical axis indicates value that has an impact on the prediction. The fact this slopes upward says that more the value of 'to_PICU' ie. 1 (positive), higher the chance for the model to predict that the patient will go into shock.

Moreover in the above Figure 6.7,(**Figure:B**) horizontal axis 'minPLT_3to8' is plotted which indicates the actual value from the dataset, and the vertical axis indicates value that has an impact on the prediction. The fact this slopes upward says that more the value of 'to_PICU' i.e 1 (positive), higher the chance for the model to predict that the patient will go into shock. As there is a spread of dots so another feature 'to_PICU' is combined to find the possibility of patients going to shock.

Thus by combining both the above scatter plot we can predict the patient going in shock. Now if we see the SHAP summary plot in (**Figure:B**)(Figure 6.7) then when minPLT_3to8(Platelets nadir (cells/mm3) have low values then there is a higher possibility that the patient will go in shock. But on the other hand when the value of minPLT_3to8 ranges from mid value to higher value then there is less tendency for the patient to go in shock.Moreover most of the children from our datasets who were approaching towards DSS(Dengue Shock Syndrome) or most likely were in vulnerable condition were sent to Pediatric Intensive Care Unit(PICU). Now if we look into the above dependence plot between to_PICU and minPLT_3to8 then the red dots represents patients who are admitted to PICU and the blue dots represents the patients who are not sent into PICU. So patients who are admitted to PICU having platelets count from 15000-50000(cells/mm3) have positive SHAP value which indicates that those patients are most likely to have Dengue Shock Syndrome(DSS). Again if we look at the right bottom corner though the patient has been sent to PICU but as the platelet count is nearly 300000 cells/mm3 and has a negative SHAP value so it is unlikely for him to go to shock.

- **MINPLT_3TO8 AND MAXHEMO_3TO8**:
  A SHAP dependence plot was plotted between minPLT_3TO8 and maxhemo_3To8 to show the relation between minimum paltelets count and maximum haemoconcentration of patients from our datasets.

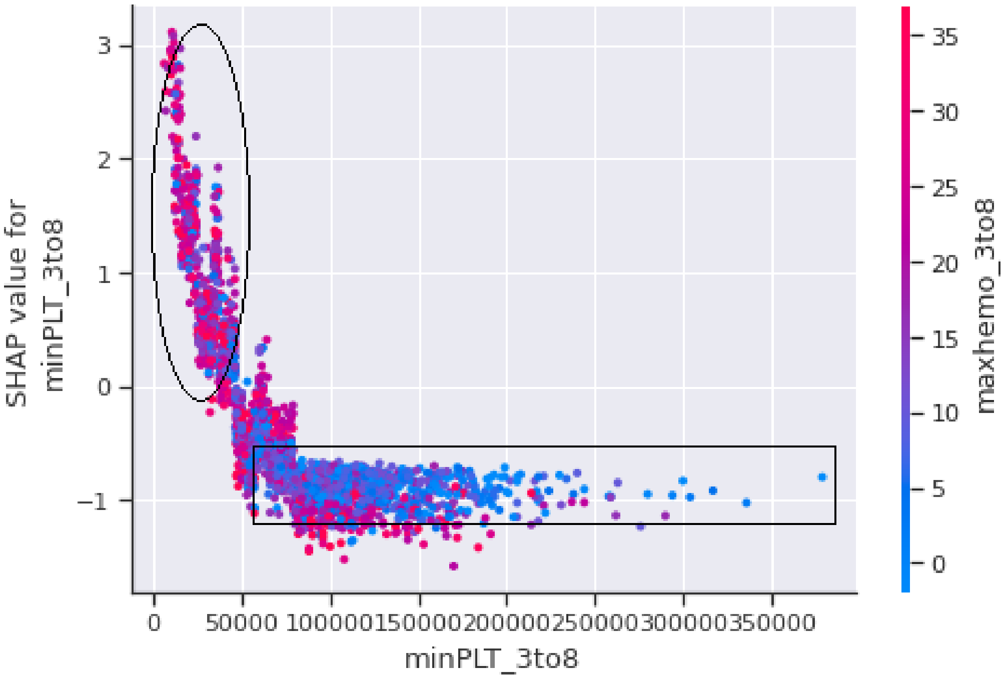


Figure 6.8: SHAP dependence plot of 'minPLT_3To8' and 'maxhemo_3TO8'

Now if the Figure 6.8 is considered then the circle region suggests that the patients is those region has a greater probability in going to shock.It is clearly seen that patients with minimum platelets count ranging between 20000(cells/mm3) and 50000(cells/mm3) and having haemoconcentration higher than 20% leads to shock. On the other hand, minimum platelets count above 50000(cells/mm3) and hemoconcentration less than 20% suggests that the patient is less likely to develop Dengue Shock Syndrome(DSS).

- **SEROTYPE2**:

  Serotype basically refers to the recognizable variation of bacteria or virus or immune cells of different individuals within a species.At present in Subtropical and Tropical region of Asia and America dengue fever is becoming a very challenging issue because of the rapid spread of dengue and its adverse effect on the economy and health care services of those regions[21].Primarily,four types of dengue virus(DENV) are prevalent among human beings. Infection with DENV can cause symptoms of varying severity [3]. It has been assumed that the four dengue serotypes are distinct in terms of symptoms incidence and clinical symptoms.For our datasets the category of the serotype is determined by PCR.

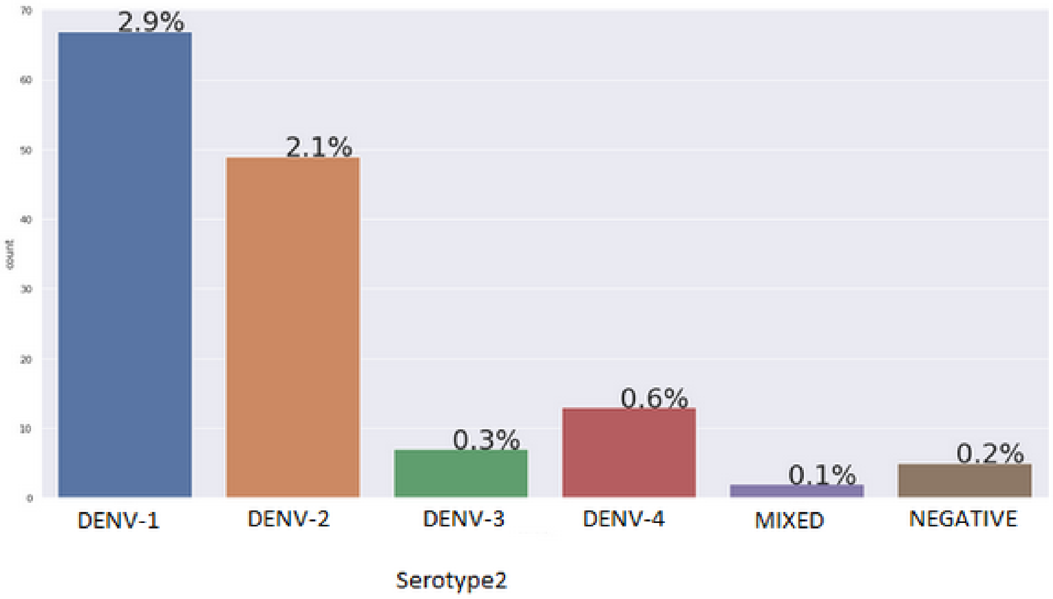


Figure 6.9: Percentage of Serotype2 of patients going into shock

As mentioned earlier in our datasets there were 2301 patients with dengue who have not reached into shock and 143 patients who were into shock. So considering the whole datasets 6.21% patients have reached into shock and have the tendency to develop Dengue Shock Syndrome. If we look into the Figure 6.9 we can observe the percentages of patients from our datasets who has suffered from different serotypes. Among 143(6.21%) patients with shock in our datasets, there were 67(2.9%) patients with DENV-1, 49(2.1%) with DENV-2, 7(0.3%) with DENV-3, 13(0.6%) with DENV-4, 2(0.1%) with mixed serotype and 5(0.2%) with no serotype were inspected from our Vietnam datasets. A bar diagram is also plotted above on the basis of our Vietnam datasets which show patients with DENV1 and DENV2 are most likely to associate with shock where the patients has high tendency to develop with Dengue Shock Syndrome(DSS).

- **SEROTYPE2,MINPLT_3TO8,MAXHCT_3TO8 AND MAXHEMO_3TO8**: In the Figure 6.10,

  - 0=DENV1
  - 1=DENV2
  - 2=DENV3
  - 3=DENV4
  - 4=MIXED
  - 5=NEGATIVE

  In the Figure 6.10 'serotype2' is plotted along horizontal axis is plotted which indicates the actual value from the dataset, and the vertical axis indicates value that has an impact on the prediction i.e on the severity of the dengue fever whether the patient will go in shock or not. As mentioned earlier the SHAP value above 0 indicates the prediction result is positive i.e the patients has gone to shock and the value below 0 indicates the patient still has not reached the shock level or developed Dengue Shock Syndrome(DSS).

  If we look clearly the circled portion of Figure 6.10 it is seen patients with DENV-1 are most likely to go in shock and develop DSS. Taking into consideration, the minimum platelets, maximum Hematocrit and hemoconcentration of patients monitored for 3 to 8 the mean value for all these three features with respect to 'serotype' was calculated. After calculation it was found patients falling in the category of DENV-1 serotype and having mean value of 30942(cells/mm3) platelets, 48% hematocrit and 25% hemoconcentration are very much likely to develop DSS. Though Figure 6.9 and Figure 6.10 suggests that patients with DENV-1 have higher chances to associate with shock but the bar diagram of the Figure 6.9 also suggests that patients with DENV-2 can also be highly probable of going into shock. So the mean values of platelets, hematocrit and hemoconcentration is also calculated DENV2. And the calculation gives a mean value of 25603(cells/mm3) platelets, 49% hematocrit and 28% hemoconcentration of patients having DENV-2 associated with shock.

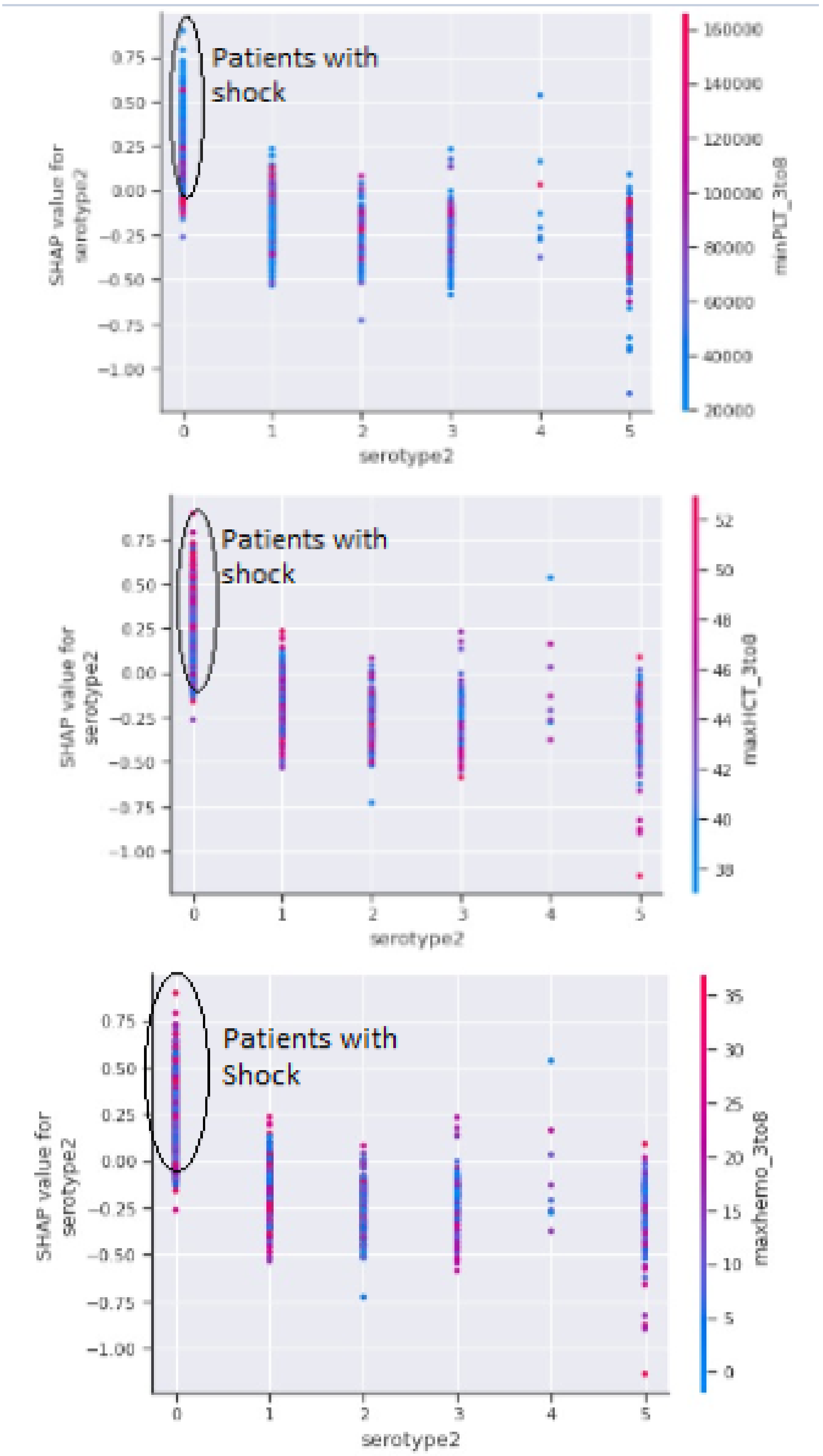


Figure 6.10: SHAP Dependence Plot Of Serotype2 And Minplatelets,MacHct, Max-Hemo.

Furthermore, KDE graph was plotted to visualize the probability of minimum platelets and maximum hematocrit count of the patients from day 3 to day 8 of the enrollment. A KDE plot is basically used to see an clear distribution of observation in a dataset [43]. It helped us to observe on which day from the day of enrollment to the hospital the patients have the highest probability of gaining dengue severity. This density graph was plotted fully on the basis of our Vietnam dataset to get an insight about how patients where proceeding towards severity(shock).

Density of minimum Platelets count of patients with shock between 3 to 8 days of Enrollment due to suffering from dengue:

In the Figure 6.11 KDE plot the probability density of minimum platelets of patients with shock are plotted between Day 3 from enrollment day at hospital to Day8. As mentioned by WHO platelets count less than 100000(cells/mm3) refers to as thrombocytopenia which leads the patient to develop Dengue Haemorrhagic Fever (DHF) or in severe case will lead to Dengue Shock Syndrome(DSS) [9]. If we look into the above density curve ,patients having platelets less than 50000(cells/mm3) at Day 6 counted from day of enrollment of patients have the highest density probability. This aforementioned patients have gone into shock and have the higher tendency of developing DSS. Similarly the density probability of patients going in shock with minimum platelets for Day7(purple),Day5(green),Day4(teal),Day8(red) are shown in the above figure with the help of KDE plot.

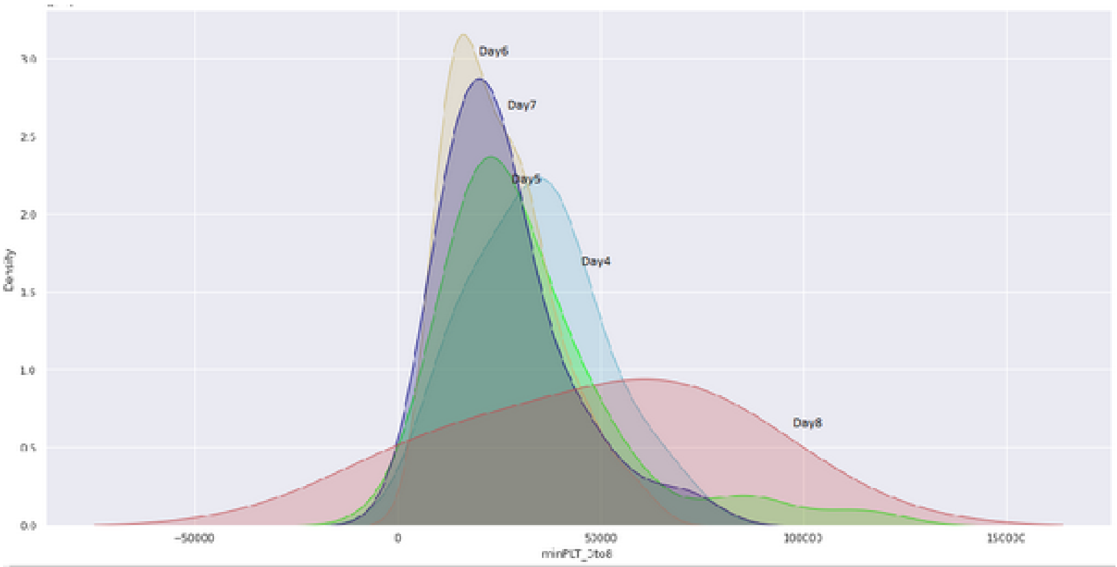


Figure 6.11: KDE plot plot of minimum platelets count of patients with shock with respect to day from enrollment in the hospital.

Density of Maximum Hematocrit Concentration count of patients with shock between 3 to 8 days of Enrollment due to suffering from dengue:

In the Figure 6.12 plot the probability density of maximum hematocrit(%) of patients with shock are plotted between Day 3 from enrollment day at hospital to Day

8. Here it is seen that patients with more than hematocrit concentration greater than 45% at day 8 have the highest density probability. According to WHO hematocrit more than 20% refers that the patient is leading to severity [9]. The density probability of patients with maximum Hematocrit for Day 5(green), Day 6(yellow), Day 4(teal), Day 7(blue) and Day 8(black) are shown in the above figure with the help of KDE plot.

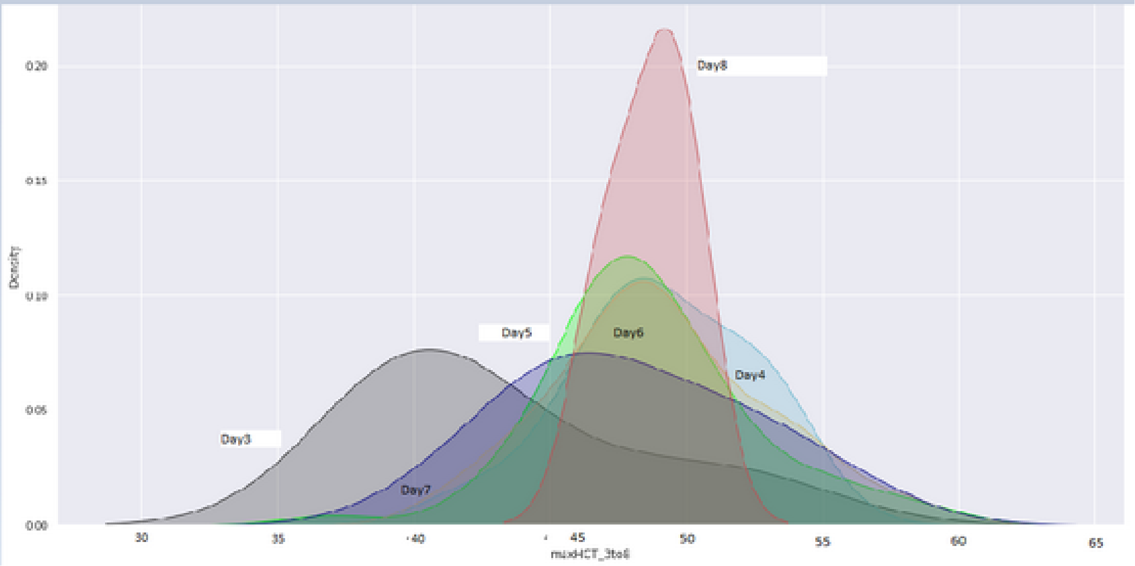


Figure 6.12: Kernel density estimation plot of maximum hematocrit (%) of patients with shock and day from enrollment to the hospital.

As we have observed the SHAP values of different graphs aforementioned which defines negative value means shock negative and positive value means shock positive, the most significant relations we have seen are from to_PICU, MAXHEMO_3to8 and MINPLT_3to8 graphs. We can observe that patients with minimum platelets count ranging between 20000(cells/mm3) and 50000(cells/mm3) and having haemoconcentration higher than 20% leads to shock syndrome(DSS) which leads them to pediatric intensive care units (PICU). Similarly, observing the serotype bar diagram relative to our datasets of all patents in shock we can see that most of the patents who developed shock syndrome (DSS) had formed mostly DENV-1 and a slight amount had formed DENV-2 dengue virus. Monitoring 3 to 8 days of enrollment for platelet counts day 6(20000-25000)cells/mm3 has shown the higher density or probability of going towards severity. Also for max hematocrit for day 8 we saw value greater than 45% leads toward dengue severity.

Chapter 7

Bangladesh Data-set And Analysis

7.1 Datasets Collection

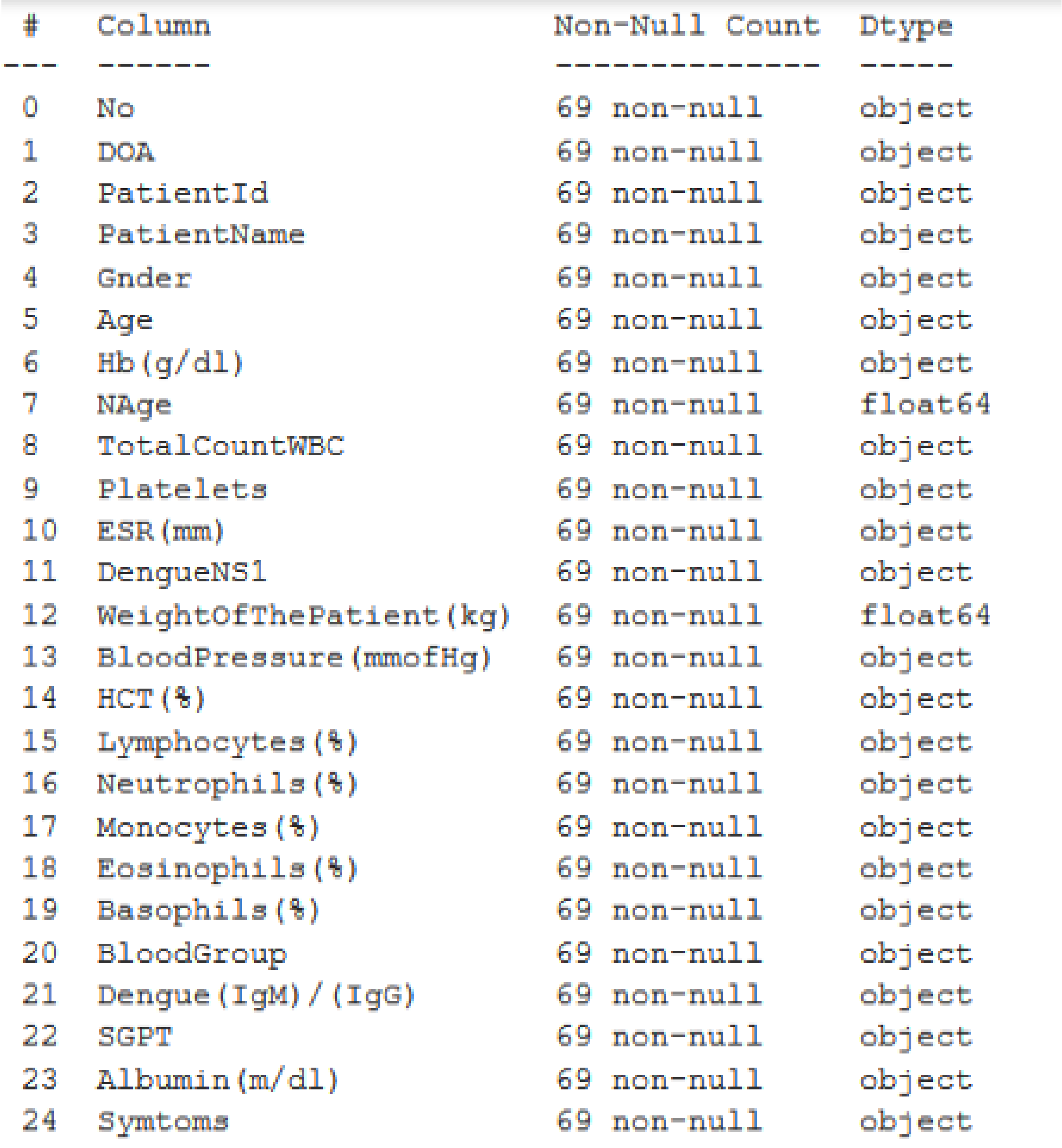
Data collection phase has been the toughest part of our journey, we have searched most of the renowned public hospitals specifically Dhaka Medical College (DMC), Salimullah Medical and private hospitals Birdem, Health And Hope, Bangladesh University of Health Science (BUHS). We also reached out to the top institution of research ICDDR,B for dengue patient data but they refused to give it to us due to the patient consent law prevailing in their institution. However in the last minute of our search we could come in contact with Dr Md Arifur Rahman and Dr Subrata Kumar Mondol who were interested in our research and therefore thought of helping us to gather data. Working with their aid in the initial phase, we again met with some failures as we learned that most hospitals in our country do not store data of the patients rigorously. Whatever tests and results are carried, the documents are given away with the patients. In the later phase, with constant support from Dr Md Arifur Rahman, we could get in touch with "DR MR KHAN SHISHU HOSPITAL" and "CENTRAL HOSPITAL" who were willing to help by providing us with data but pledged to keep the patients identity to remain anonymous. Finally from continual assists from these highly qualified doctors we could start working on something.

From the accumulated data, we followed machine learning approach which is a multidisciplinary field that explains the past and predicts the future by means of data analysis. Furthermore we have analyzed the datasets, preprocessing it and pushing it through machine learning algorithms in hope to achieve the best results.

7.2 Datasets Description

We collected 69 data of patients from the Shishu hospital which mainly contains data of children aged between 8 month to 15 years. The data sets contain all the components of blood like WBC, Platelets count, Lymphocytes, Monocytes etc that a normal blood test report contains and symptoms that was visible at the initial stage when the disease was detected. In addition to that it has columns containing the NSI, IgM and IgG Test results of those children. Similarly the dataset of Central Hospital contains a Hematology report and Laboratory report of Dengue Test of around 100 patients of different ages. In our study we aggregated dengue patient information from two hospitals and intend to develop a model by analysing the phenotypic characteristics of the patient by merging the file in one datasets and selecting the similar features among the datatsets. It will be a challenging thing for us to conduct the severity research and reach a satisfying accuracy level due to the limited size of the datasets we got from the hospitals.

The Figure 7.1 shows the attributes of the datasets we gather from the “DR MR KHAN SHISHU HOSPITAL” which is as follows:



```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   No                         69 non-null     object
 1   DOA                        69 non-null     object
 2   PatientId                  69 non-null     object
 3   PatientName                69 non-null     object
 4   Gnder                      69 non-null     object
 5   Age                        69 non-null     object
 6   Hb(g/dl)                   69 non-null     object
 7   NAge                       69 non-null     float64
 8   TotalCountWBC              69 non-null     object
 9   Platelets                  69 non-null     object
 10  ESR(mm)                    69 non-null     object
 11  DengueNS1                  69 non-null     object
 12  WeightOfThePatient(kg)     69 non-null     float64
 13  BloodPressure(mmofHg)      69 non-null     object
 14  HCT(%)                     69 non-null     object
 15  Lymphocytes(%)             69 non-null     object
 16  Neutrophils(%)             69 non-null     object
 17  Monocytes(%)               69 non-null     object
 18  Eosinophils(%)             69 non-null     object
 19  Basophils(%)               69 non-null     object
 20  BloodGroup                 69 non-null     object
 21  Dengue(IgM)/(IgG)          69 non-null     object
 22  SGPT                       69 non-null     object
 23  Albumin(m/dl)              69 non-null     object
 24  Symtoms                    69 non-null     object
```

Figure 7.1: Showing Attributes of the Shishu Hospital Dataset

Table 7.1: ATTRIBUTES OF DATASET COLLECTED FROM CENTRAL HOSPITAL

Name Of the Attributes	Unit	Name Of the Attributes	Unit
Date Of Arrival	dd/mm/yy	Monocytes	%
Patient Id	-	Basophils	%
Patient Name	-	HCT	%
Gender	0(M)/1(F)	MCV	fl
Age	year	MCH	pg
Hemoglobin	g/dl	MCHC	g/dl
WBC Count	/cmm	RBC Count	million/cm
Platelets	K/L	Dengue NS1	Positive/Negative
Neutrophils	%	Dengue IgG	Positive/Negative
Lymphocytes	%	Dengue IgM	Positive/Negative
Eosinophils	%	-	-

As we could not collect enough data from both the hospitals so we decided to merge the two datasets and randomly shuffled them .So a total of 169 data we had for analysing the severity among patients. Among 169 patients all of them were dengue patients. So we selected the common features among the datasets and also made sure that all the features from the datasets are having the same unit. For instance the unit for Platelet in CENTRAL HOSPITAL was 'K/L' which was converted to 'cells/mm3' to make it equivalent with the Platelets count obtained from the SHISHU HOSPITAL. Finally the following features were selected:

- Sex
- Age(yr)
- Hb(g/dl)
- HCT(%)
- Platelets(cells/mm3)
- WBC(/cmm)
- Lymphocytes(%)
- Neutrophils(%)
- Monocytes(%)

7.3 Missing Values Imputation

As the amount of data for Bangladesh data-set was very small so we did not want to drop the rows with missing values. So the dataset require a bit of preprocessing for handing the missing values. We used interpolation to fill up the missing values.Interpolation is a mathematical analysis that adjusts a function to our dataset and using that function the missing value is deduced [38]. The formula of interpolation which was used to fill up the missing values in our dataset are as follows:

$$y = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)} \tag{7.1}$$

where,
y = linear interpolation value
x = independent variable
x_1, y_1 = values of the function at one point
x_2, y_2 = values of the function at another point

Interpolation is a very simple technique which estimates the unknown value by observing the distinct dataset with known data points [13]. Among all other missing values imputations approach interpolation worked more efficiently than other techniques for our small Bangladesh dataset.

7.4 Feature Engineering

Data preprocessing is challenging as it requires the right attributes that need to be used in order to do a relevant analysis. In our dataset, we have some important parameters such as age, IgG and IgM, blood pressure values, ESR values which could have enriched our dataset in optimizing the final feature sets. IgG and IgM is the specificity test for past and present dengue infection and ESR values is an indicator for any illness present in the patient. But most of the values were missing from the above mentioned features. Since most medical values cannot be predicted without proper tests we concluded to filter it out except the age of patients. There were two missing age parameters in pediatric patients data provided by Dr MR Khan Shishu(Children) Hospital and most of the age data was in years with months e.g 11 year 4 months. We initially needed to round the ages in years only e.g if a pediatric patient was 6 months old then we rounded it to 0.5 years old. Since we had complete data on weights column, we estimated those two missing values of the age column by using modified APLS equation: weight = 3*(age)+7 which is valid for weights from age one to puberty [18].

In the second phase, we made sure both hospital data collected had common fields and are in the same units.

In the third phase we used python libraries such as seaborn, pandas, numpy, matplotlib to do some univariant and bi-variant analysis to find interrelation between different features. This helps us to separate the dependent and independent variables and assist us to reduce the initial data sets from raw data to more manageable groups.

In the final phase we made progress by using Pearson's Correlation to find the connected entities. Then finally we passed the relevant attributes through an entropy algorithm along with the decision tree[5][2] to identify more relevant attributes and filter out the rest of the attributes.

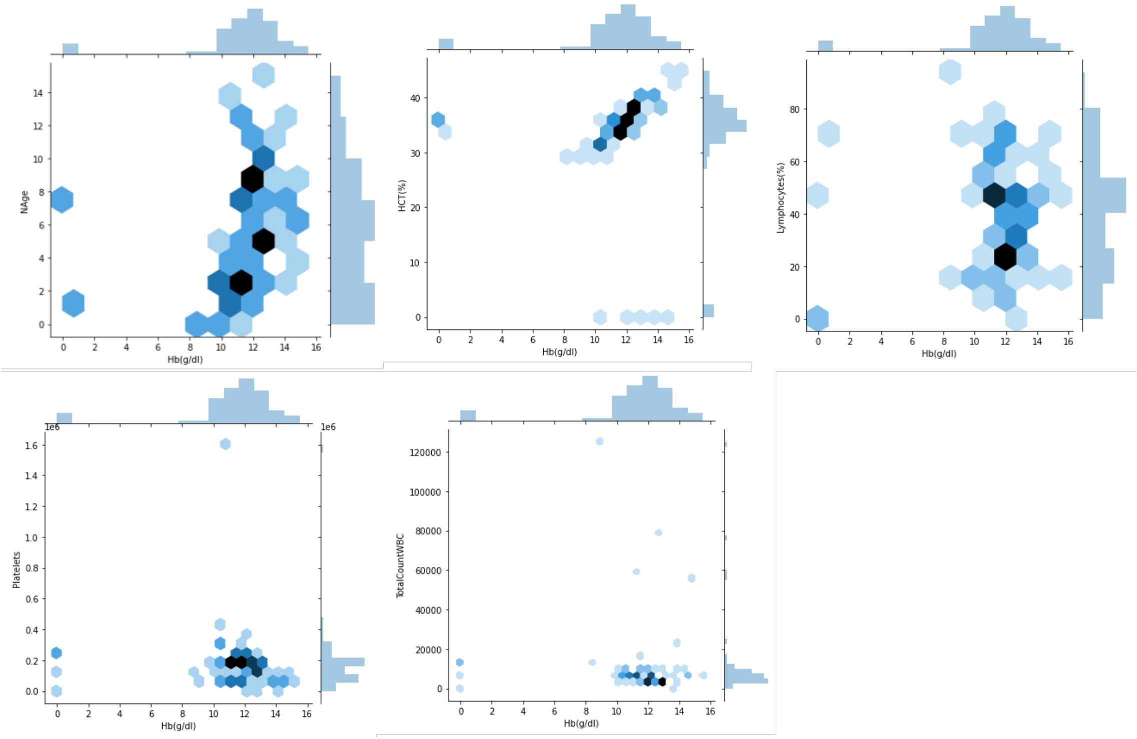


Figure 7.2: Bi-variant Relation between different features of the Bnagladeshi dataset

Figure 7.2 shows bi-variant analysis between Haemoglobin and other features like Platelets, HCT, WBC and Lymphocytes. The plot is used to calculate two events occurring at the same point in time. It creates a regression line between two events and generates a probability. The darker blue region implies higher concentration.

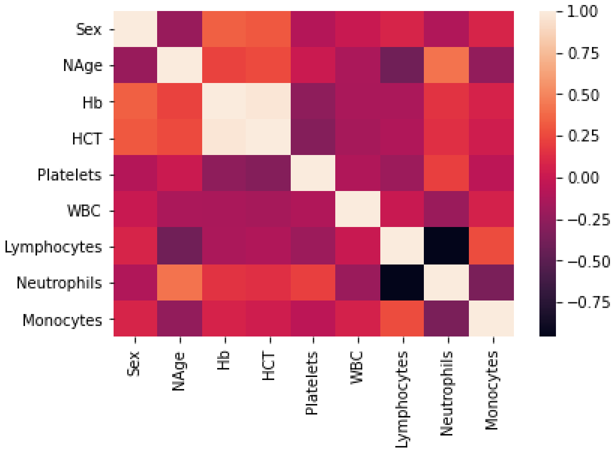


Figure 7.3: The correlation of different features for Bangladeshi dataset

The correlation heat map was produced shown in Figure 7.3 to see the co relation among different features for proper understanding the effects of each features.

After analysing the bivalent plot we could not come to a certain conclusion regarding the most important features in the dataset. So furthermore, we applied ExtraTreeClassifier model to find out the important features. Extra Trees Classifier

(Extremely Randomized Trees Classifier) is a type of ensemble learning technique that combines the outcomes of many de-correlated decision trees obtained in a forest to generate the outcome of classification [20]. To select important features the model builds forests for each of the features. Thus,during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini importance and the top k features is selected.

In the figure 7.4 it is seen that Platelets has the highest score so it is more relevant or important feature towards finding the severity among patients.Apart from Platelets, HCT, Lymphocytes, Neutrophils and WBC are also slightly important.

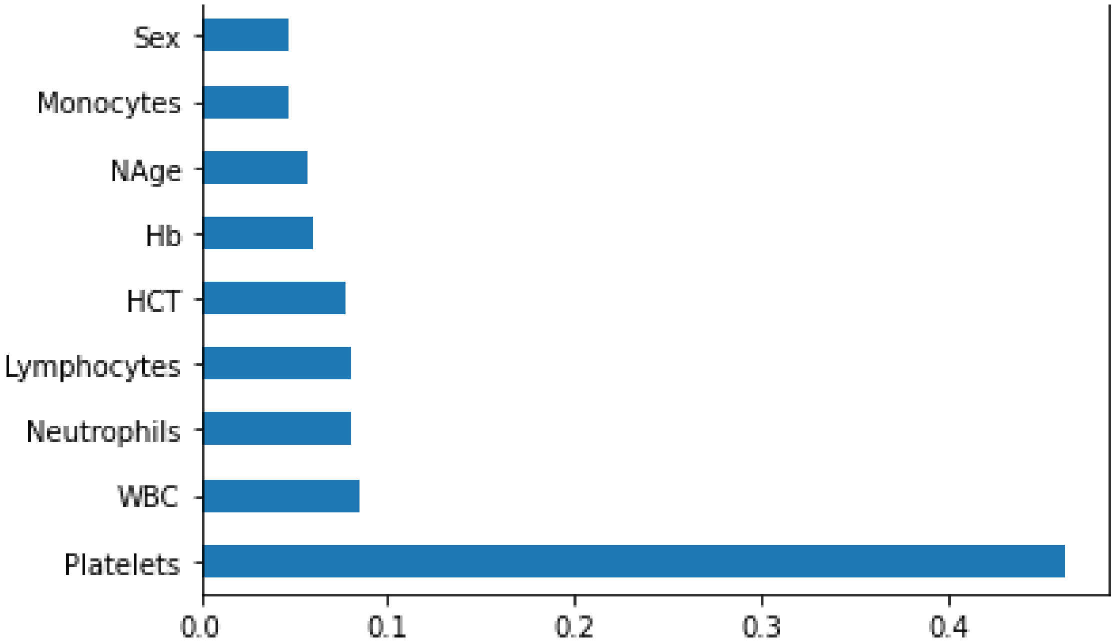


Figure 7.4: Bar Diagram showing Important Features of Bangladeshi Dataset

7.5 Proposed Model

For the Bangladeshi dataset it was very hard to reach any decision regarding the severity of dengue whether the patient has developed DHF or DSS due to insufficient data. Most of the hospitals in Bangladesh do not have proper database to store information regarding patients. So the data we collected from the two hospitals did not contain any strong column which can be provided to any supervised classification algorithm to predict severity among patients. So we used an unsupervised model called hierarchical clustering to cluster unlabelled data points. Later the guidelines of WHO was followed to analyse the severity among the clusters.

Firstly we considered a dendogram to find the arrangements of clusters from our dataset. But it was difficult to decide the arrangements of clusters from the dendogram.The first reason for this could be the unavailability of features or the length of the dataset being too short. Figure 7.5 shows the dendogram for our dataset. So later we used Silhoutte Score to determine the best cluster arrangements for Bangladesh dataset.

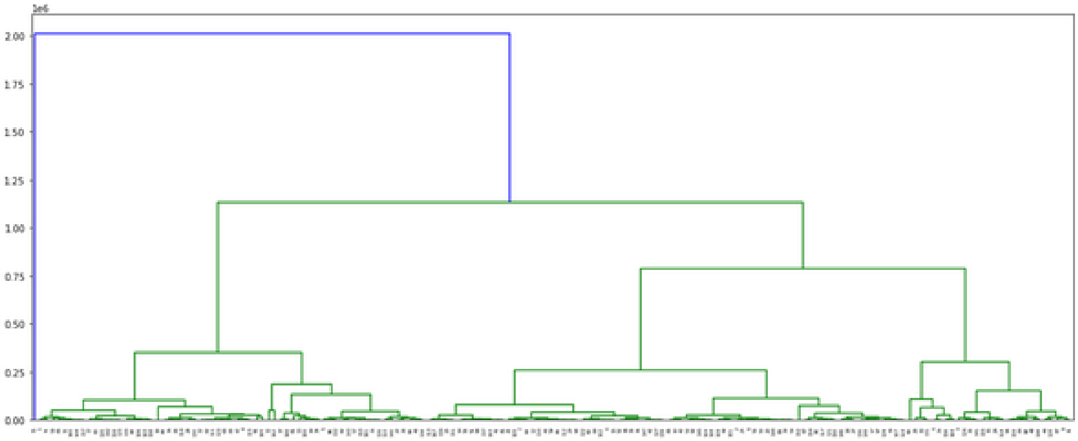


Figure 7.5: Dendogram of Bangladeshi Dataset

Silhouette refers to the method of analysis and accuracy validation within data clusters. The methodology offers a straight forward graphical representation of how well each object has been categorized. The meaning of the silhouette is an indicator of how close an object is to its own cluster (cohesion) relative to other clusters (separation). The silhouette varies from -1 to +1, where a high value means that the object is well matched to its own cluster and badly matched to the surrounding clusters. If most objects have a high value, a clustering setup is acceptable. If many points have a low or negative value, there could be too many or too few clusters in the cluster configuration. The silhouette can be measured using distance metric, such as the Euclidean distance or the Manhattan distance [44].

And if the Figure 7.6 for Silhouette score bar diagram is observed, then for clusters = 2 the silhouette score is close to 1 considering our datasets. So we do hierarchical agglomerative clustering by providing cluster number = 2.

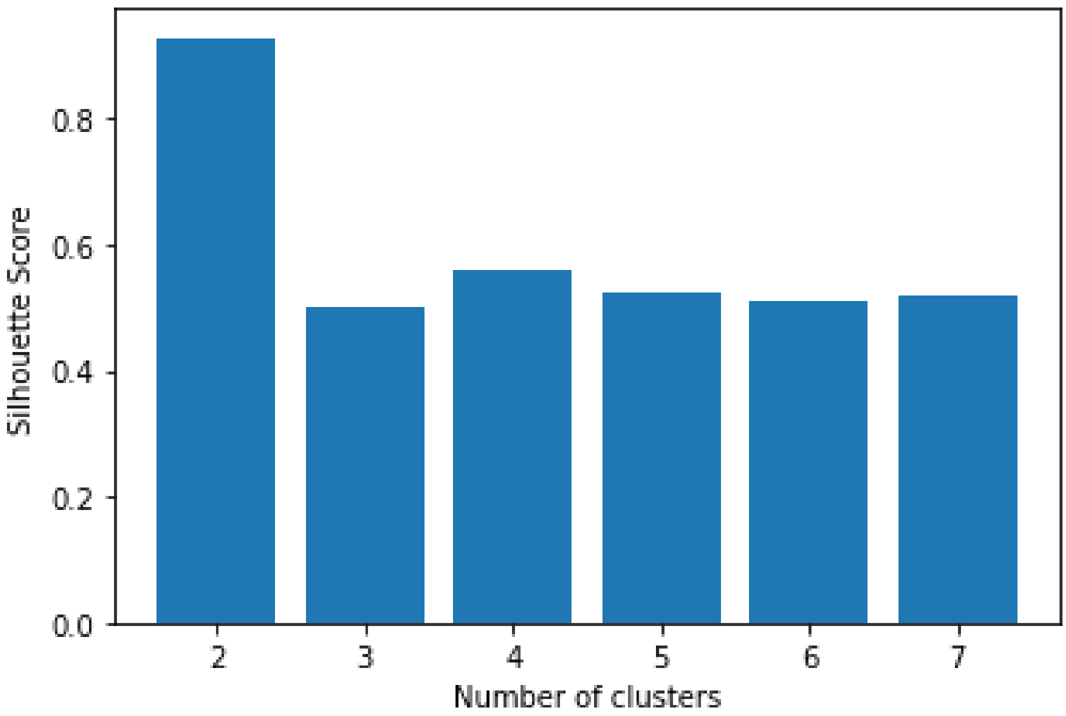


Figure 7.6: Silhoutte Score for different clusters arrangement for Bangladesh dataset

Chapter 8

Results And Discussion On Bangladesh Datasets

Our datasets was fitted with the Agglomerative Hierarchical Clustering to determine different clusters of patients from the datasets. After fitting the datasets with the model it starts its processing by finding all the dissimilarities between all the data points from the data sets. Then, two objects that minimize a given agglomeration criterion when clustered are taken under the same group, creating a class consisting of these two objects. In the next step the dissimilarity the aforementioned class and other objects from the datasets are calculated with the agglomerative criterion and the classes or group of objects whose clustering minimize the agglomerative criterion fall under the same cluster as they show similarities among the features. Clustering will keep continuing by computing distance between every data points and for measuring this distance we used the Euclidean Distance Technique.

Two clusters could be formed after applying hierarchical clustering. Mean and standard deviation values of all the features for both the clusters are as follows after fitting the datasets with hierarchical clustering.

Table 8.1: MEAN AND STANDARD DEVIATION OF DIFFERENT FEATURES FOR **CLUSTER 0**

-	**Platelets (cells/mm3)**	**HCT(%)**	**Lymphocytes (%)**	**Monocytes (%)**	**Neutrophils (%)**	**WBC**	**Hb (g/dl)**
Mean	221085	37.94	28.50	4	65	6918	12
Std Deviation	63918	4.77	16.26	1.99	18.16	5812	1.61

Table 8.2: MEAN AND STANDARD DEVIATION OF DIFFERENT FEATURES FOR **CLUSTER 1**

-	**Platelets (cells/mm3)**	**HCT(%)**	**Lymphocytes (%)**	**Monocytes (%)**	**Neutrophils (%)**	**WBC**	**Hb (g/dl)**
Mean	93714	40.63	40	4	52	10521	12
Std Deviation	3596	5.96	18.57	1.95	19.82	18762	2.02

Figure 8.1 shows two cluster and that is shown through the Scatter plot. As the whole graph is very bigger in size so it has been divided into three parts for our convenience. Moreover the two clusters that we obtained are shown in Figure 8.1 **Blue** color dots represent **cluster 1** and **red color** dots represent **cluster 0**.

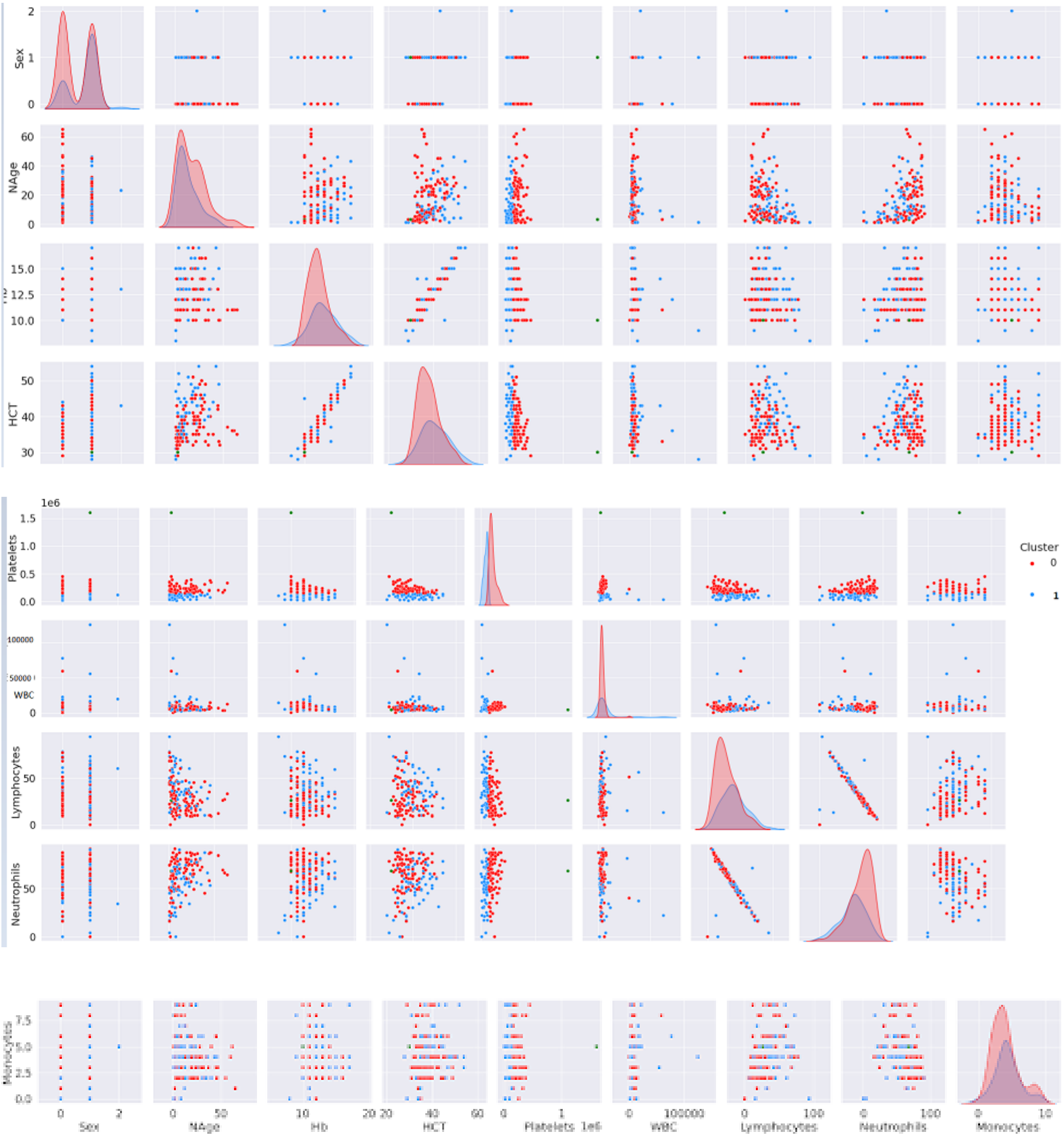


Figure 8.1: Scatter plot Showing two clusters.

Later the two clusters are analysed further to see if patients from any of the clusters proceed to severity or not.The WHO guideline was taken into consideration to detect patients with DHF or DSS. According to the guideline given by WHO patients with platelets less than 100000(cells/mm3) may lead to thromocytopenia and hemoconcentration with the hemetocrit level more than 20% will lead to plasma leakage [34]. These two reasons are the evidence that the patient is proceeding towards DHF/DSS which is more severe than the normal dengue.

Pair plot of two clusters were plotted individually to get an insight about the patients showing severity. The severity decision was taken following the guideline of WHO.

- **CLUSTER 0**

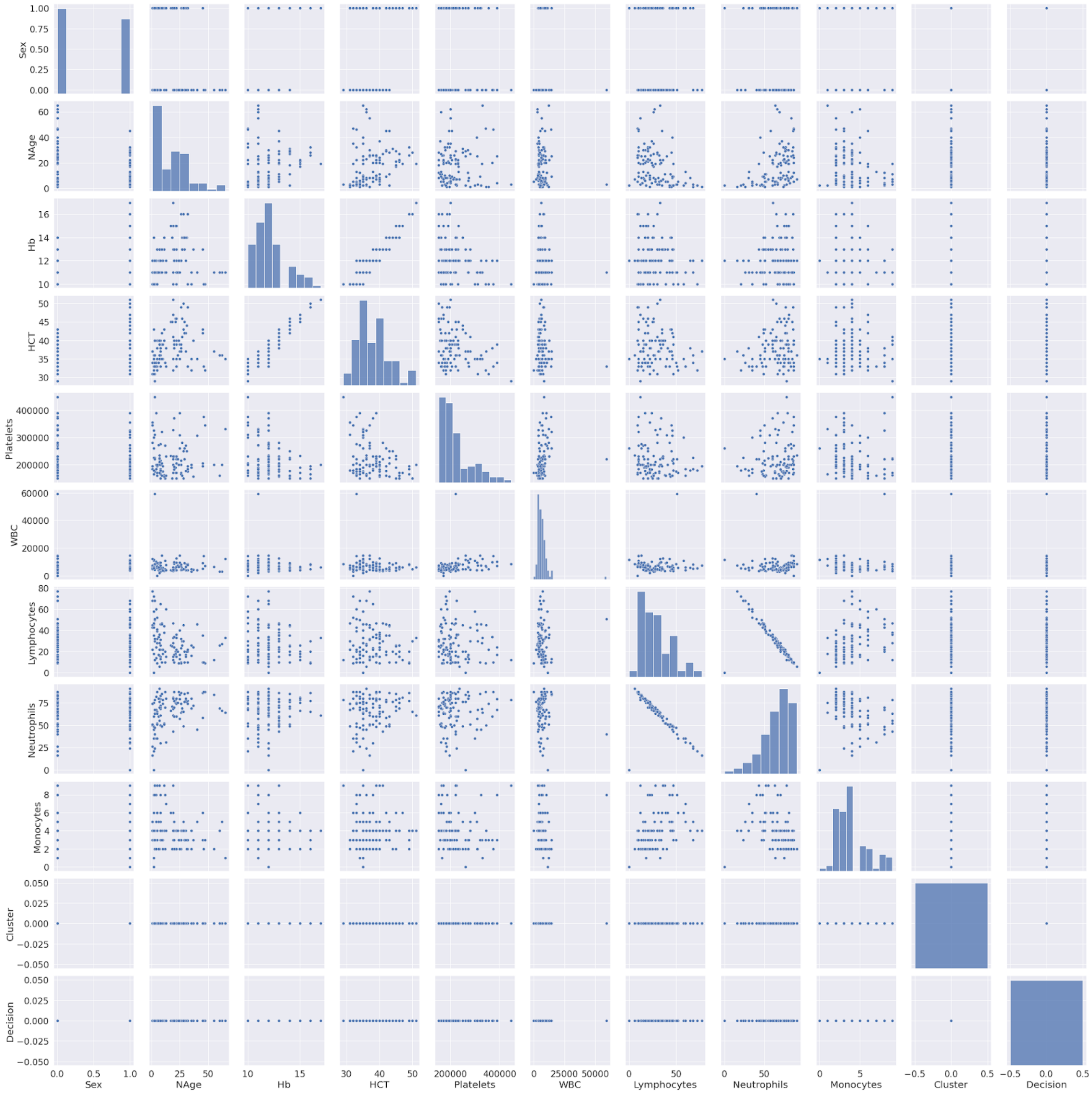


Figure 8.2: Pair plot for cluster 0.

Here observing Figure 8.2 patients with cluster 0 have less probability of suffering from DHF or DSS as Platelets count of patients of this group starts from 200000(cells/mm3). As their platelets count is more than 100000(cells/mm3) they have less chance of having thrombocytopenia.

- **CLUSTER 1**

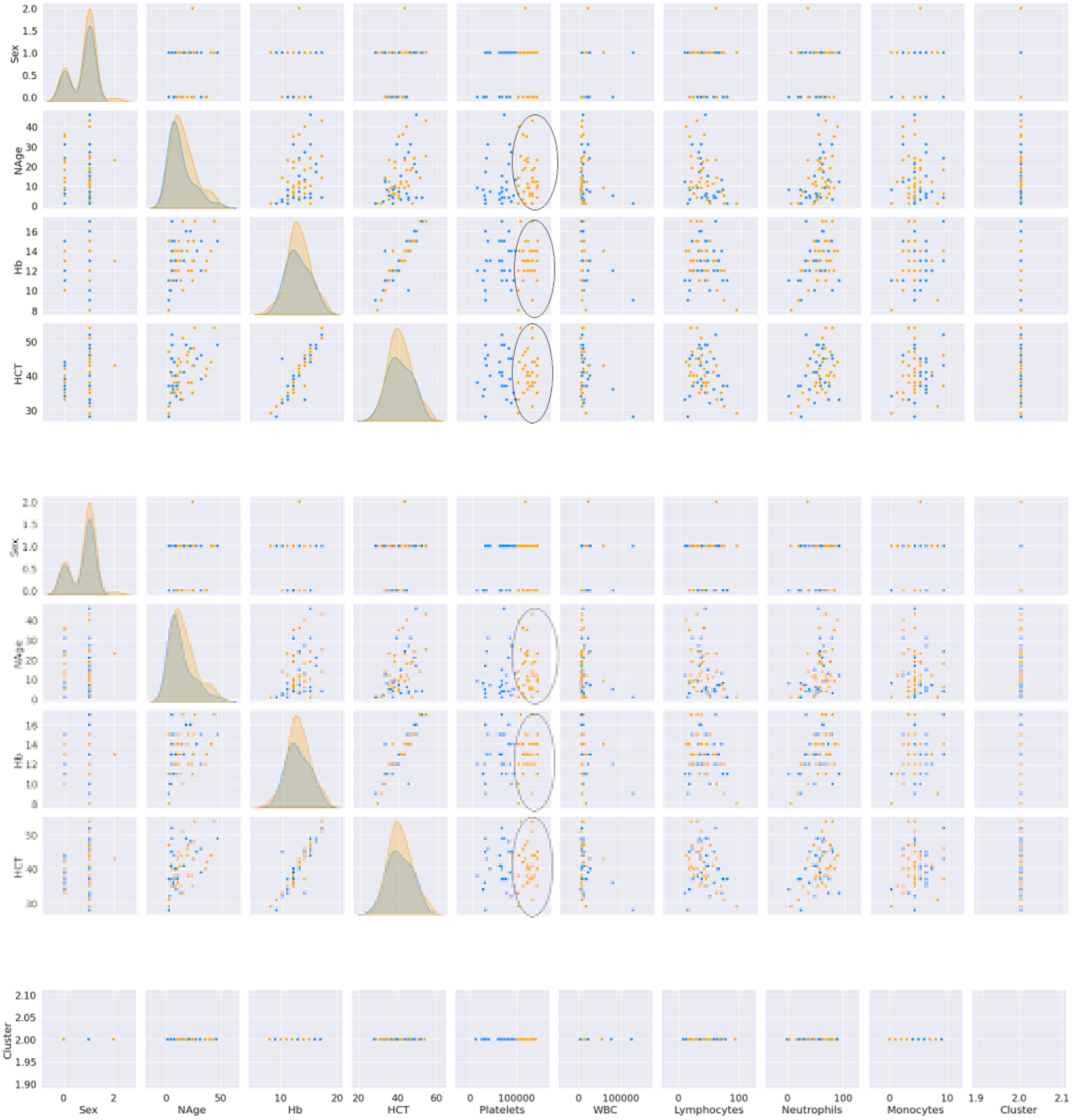


Figure 8.3: Pair plot for cluster 1.

Here after observing Figure 8.3 for cluster 1 unlike cluster 0 there is a possibility of patients suffering from severe dengue. If we further observe the above pair plot then two separated groups can be observed one with blue color dots and another with yellow color dots. If we observe the yellow color dots bounded with circles then those patients platelets count is greater than 100000(cells/mm3). On the contrary the blue color scatter dots beside the circle region resembles patients with platelets count less than 100000(cells/mm3). So the patients in the blue color dotted region have higher tendency to develop thrombocytopenia which is a significant criteria for developing DHF. On the other hand patients in the yellow dotted region are less likely to develop devere dengue rather they are suffering from normal DF.

Thus, after analysing both the clusters, among total 169 patients 106 patients belong to Cluster 0 and the rest 63 belong to Cluster 1.

According to world health organization Normal dengue DF is characterized by a platelet count less than 150000(cells/mm3) and rising hematocrit (5-10)% with no plasma leakage and leukopenia(WBC count(less than 5000cells.mm3)). In case of DHF or in more serious case DSS, it is characterized by thrombocytopenia(less than 100000 cells/mm3) and hematocrit concentration greater than 20% [7].

In case of Cluster 0 patients all the patients did not develop a severe dengue rather they were suffering from normal DF.
In case of Cluster 1 some patients develop normal dengue fever and some patients have the probability to reach DHF. It was seen that 44% of people of Cluster 1 has a very high chance of reaching towards severity and as a result develop DHF. On the other hand, 55% patients from cluster 1 have normal DF only.

Moreover if we observe then patients in Cluster 0 have a mean neutrophils count of 65% and patients with Cluster 1 have a mean neutrophils count of 52%. Neutrophils are the part of white blood cells which helps the body to fight against any foreign body or any sorts of infection and helps the injured tissues to heal faster.But as patients in Cluster1 are more likely to develop neutropenia which is characterized by low levels of neutrophils which increases the higher risk of getting infected by different types of infections [7]. In addition to that the mean HCT percentage in Cluster1 patients is slightly higher than the Cluster 0 patients.Though Cluster 1 have higher risk of DHF or DSS but the mean percentage of lymphocytes for Cluster 1 is more than Cluster 0 patients.The mean percentage of lymphocytes count for Cluster 0 is 28.5% and Cluster 1 is 40%. If the lymphocytes count reduces more in Cluster 0 patients then there is a possibility of developing lymphocytopenia. Furthermore the mean percentage count for Hb and Monocytes is same for both the clusters.

Chapter 9

Conclusion And Future Work

9.1 Future Work

Our plans for future include adding more data in our Bangladeshi data set since finding adequate data from different hospitals were quite challenging due to mismanagement of data safe storage as well as some hospital's reluctance towards providing more data due to their policy of patient's consent. Initially our plan for this project was to work with Bangladeshi data set mostly to analyse and predict the severity out from the important features extracted but due to extreme difficulty in finding adequate data and lack of required attributes and information from patient reports we could not manage to produce a perfect outcome. Hence, we took Vietnamese data set as a reference to find the similarities in order to analyse our own Bangladesh data set. Additionally, the effect of COVID-19 lockdown has also slowed us down in our collection of data for our research. But we are still determined to collect as much data as we can to enrich our data set of Bangladesh to analyse further and conduct a much bigger research on much bigger data set for maximum efficiency.

9.2 Conclusion

In our study our goal was to analyse and predict severity among dengue infected patients using different machine learning approaches. We observed that for two subtropical countries Vietnam and Bangladesh there is a strong correlation between DSS or DHF with patients platelet count and HCT concentration. While, for the Vietnam data sets we implemented different classification and boosting algorithm like-Decision Tree Classifier, Random Forest, Gradient Boosting, Ada Boosting, XG-Boosting Classier Model to predict severity among dengue infected patients. We managed to produce almost perfect accuracy of 98.6% using XG-Boost Classifier model after fitting our data sets containing 2301 patients from Vietnam in our mentioned model. Our Vietnam data sets required a bit of pre-processing for handling missing data and KNN Imputation was used to fill up the missing values. For the analysis of the same data set we identified the important features and used the SHAP value to observe and find the higher probability of a patient developing DSS/DHF. On the other hand, for Bangladeshi data set we collected data from both "Shishu Hospital" and "Central Hospital" and applied Hierarchical Clustering to find the various clusters of different vital elements of the patients blood report. Then we further analyzed the clusters to find the seriousness of the patients that lead them to DSS or DHF. While applying the Hierarchical Clustering we observed two clusters among all the patients. After further analysis we found that in ”cluster 0” the probability of leading to severity is quite less than ”cluster 1”. Lastly, we hope that our work helps to analyze more research studies in related field.

References

[1] L. Tanner, M. Schreiber, J. G. H. Low, A. Ong, T. Tolfvenstam, Y. L. Lai, L. C. Ng, Y. S. Leo, L. Thi Puong, S. G. Vasudevan, and et al., *Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness*, Mar. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18335069.

[2] ——, *Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness*, Mar. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18335069.

[3] M. Mayxay, R. Phetsouvanh, C. E. Moore, V. Chansamouth, M. Vongsouvath, S. Sisouphone, P. Vongphachanh, T. Thaojaikong, S. Thongpaseuth, S. Phongmany, and et al., *Predictive diagnostic value of the tourniquet test for the diagnosis of dengue infection in adults*, Oct. 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3156.2010.02641.x/abstract.

[4] C.-C. Chen and H.-C. Chang, *Predicting dengue outbreaks using approximate entropy algorithm and pattern recognition*, Apr. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0163445313000686.

[5] ——, *Predicting dengue outbreaks using approximate entropy algorithm and pattern recognition*, Apr. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0163445313000686.

[6] J. Liu, J. Ma, J. Wang, D. D. Zeng, H. Song, L. Wang, and Z. Cao, *Comorbidity analysis according to sex and age in hypertension patients in china*, Mar. 2016. [Online]. Available: https://arizona.pure.elsevier.com/en/publications/comorbidity-analysis-according-to-sex-and-age-in-hypertension-pat.

[7] A. Giorgi, *Neutropenia: Definition and patient education*, Mar. 2017. [Online]. Available: https://www.healthline.com/health/neutropenia.

[8] Kassambara, Visitor, Jg, Kassambara, and Mahmoud, *Evaluation of classification model accuracy: Essentials*, Mar. 2018. [Online]. Available: http://www.sthda.com/english/articles/36-classification-methods-essentials/143-evaluation-of-classification-model-accuracy-essentials/.

[9] A. Kumar, *Ml metrics: Sensitivity vs. specificity - dzone ai*, Sep. 2018. [Online]. Available: https://dzone.com/articles/ml-metrics-sensitivity-vs-specificity-difference.

[10] P. K. Lam, T. V. Ngoc, T. T. T. Thuy, N. T. H. Van, T. T. N. Thuy, D. T. H. Tam, N. M. Dung, N. T. H. Tien, N. T. T. Kieu, C. Simmons, and et al., *The value of daily platelet counts for predicting dengue shock syndrome: Results from a prospective observational study of 2301 vietnamese children with dengue*, Nov. 2018. [Online]. Available: https://research.monash.edu/en/publications/the-value-of-daily-platelet-counts-for-predicting-dengue-shock-sy.

[11] A. R, *Applying random forest (classification) - machine learning algorithm from scratch with real...* Jul. 2018. [Online]. Available: https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57.

[12] W.-K. Wang and D. J. Gubler, Apr. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5925702/.

[13] W. Badr, *6 different ways to compensate for missing data (data imputation with examples)*, Jan. 2019. [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779.

[14] *Boosting in machine learning: Boosting and adaboost*, May 2019. [Online]. Available: https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/.

[15] *F-score*, May 2019. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score.

[16] P. Mutsuddy, S. Tahmina Jhora, A. K. M. Shamsuzzaman, S. M. G. Kaisar, and M. N. A. Khan, *Dengue situation in bangladesh: An epidemiological shift in terms of morbidity and mortality*, Mar. 2019. [Online]. Available: https://www.hindawi.com/journals/cjidmm/2019/3516284/.

[17] M. Pathak, *Using xgboost in python*, Jan. 2019. [Online]. Available: https://computational-communication.com/xgboost/.

[18] Y. Zhu, L. M. Hernandez, Y. Dong, J. H. Himes, L. E. Caulfield, J. M. Kerver, L. Arab, P. Voss, S. Hirschfeld, M. R. Forman, and et al., *Weight estimation among multi-racial/ethnic infants and children aged 0-5·9 years in the usa: Simple tools for a critical measure*, Jan. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6312489/.

[19] Dansbecker, *Advanced uses of shap values*, Oct. 2020. [Online]. Available: https://www.kaggle.com/dansbecker/advanced-uses-of-shap-values.

[20] *Ml: Extra tree classifier for feature selection*, Jul. 2020. [Online]. Available: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/.

[21] K. is a data science professional, a trainer with more than 7 years of experience working in the consulting domain, and around 3 years of teaching experience., *Knnimputer: Way to impute missing values*, Jul. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/.

[22] E. Sharp·Statistics, *Statistics: Sensitivity, specificity, ppv and npv*, Jul. 2020. [Online]. Available: https://geekymedics.com/sensitivity-specificity-ppv-and-npv/.

[23] M. Siamak N. Nabili, Nov. 2020. [Online]. Available: https://www.emedicinehealth.com/complete_blood_count_cbc/article_em.htm.

[24] C. P. Simmons, A. A. the Oxford University Clinical Research Unit, W. T. M. O. P. (C.P.S., L. R. Baden, F. P. Polack, M. C. Castells, E. J. Phillips, *et al.*, *Dengue: Nejm*, Dec. 2020. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra1110265.

[25] A. Singhal, *Akshay singhal*, Jan. 2020. [Online]. Available: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/.

[26] *What is hierarchical clustering?* Dec. 2020. [Online]. Available: https://www.displayr.com/what-is-hierarchical-clustering/.

[27] [Online]. Available: https://twitter.com/hashtag/Antibioticresistance.

[28] [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7813712.

[29] [Online]. Available: http://www.onlinejets.org/article.asp?issn=0974-2700;year=2011;volume=4;issue=1;spage=120;epage=127;aulast=Rajapakse.

[30] *[solved] what are the similarities and differences between 'features' of reinforcement learning and 'features' of deep neural network?: Course hero.* [Online]. Available: https://www.coursehero.com/tutors-problems/Artificial-Intelligence/27794730-What-are-the-similarities-and-differences-between-features-of-reinfo/.

[31] *[solved] what are the similarities and differences between 'features' of reinforcement learning and 'features' of deep neural network?: Course hero.* [Online]. Available: https://www.coursehero.com/tutors-problems/Artificial-Intelligence/27794730-What-are-the-similarities-and-differences-between-features-of-reinfo/.

[32] *Classification: Precision and recall — machine learning crash course.* [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall.

[33] *Clinical assessment.* [Online]. Available: https://www.cdc.gov/dengue/training/cme/ccm/page73112.html.

[34] *Dengue shock syndrome.* [Online]. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/dengue-shock-syndrome.

[35] *Introduction to semi-supervised learning.* [Online]. Available: https://www.morganclaypool.com/doi/abs/10.2200/S00196ED1V01Y200906AIM006.

[36] *Most popular distance metrics used in knn and when to use them.* [Online]. Available: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html.

[37] E. PMC. [Online]. Available: http://europepmc.org/articles/PMC3097561.

[38] L. Portella, *What to do when data is missing?* [Online]. Available: https://leportella.com/missing-data/.

[39] *Ppv.* [Online]. Available: https://medical-dictionary.thefreedictionary.com/PPV.

[40] *Python machine learning tutorial.* [Online]. Available: https://www.python-course.eu/Decision_Trees.php.

[41] Z. Z. G. Y. G. F. R. J. M. S.-A. R. S. A. K.-V. R. R. A. P. B. C. M. E. J. T. F. M. C. L. C. E. C. A. R; *Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning.* [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30006563.

[42] A. L. Ramadona, L. Lazuardi, Y. L. Hii, Å. Holmner, H. Kusnanto, and J. Rocklöv, *Prediction of dengue outbreaks based on disease surveillance and meteorological data.* [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152688.

[43] *Seaborn.kdeplot¶.* [Online]. Available: https://seaborn.pydata.org/generated/seaborn.kdeplot.html.

[44] *Selecting the number of clusters with silhouette analysis on kmeans clustering¶.* [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html.

[45] B. S. P. O. J. A. C. J. J. A. O. M. D. T. G. C. T. J. SI; *The global distribution and burden of dengue.* [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23563266/.

[46] ——, *The global distribution and burden of dengue.* [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23563266/.

[47] *The lancet.* [Online]. Available: https://www.sciencedirect.com/journal/the-lancet/vol/394/issue/10215.

[48] *The lancet.* [Online]. Available: https://www.sciencedirect.com/journal/the-lancet/vol/394/issue/10215.

[49] *World health statistics.* [Online]. Available: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/world-health-statistics.